G.711.1 ANNEX D AND G.722 ANNEX B – NEW ITU-T SUPERWIDEBAND CODECS

Lei Miao[1], Zexin Liu[1], Chen Hu[1], Václav Eksler[2], Stéphane Ragot[3], Claude Lamblin[3], Balazs Kovesi[3], Jongmo Sung[4], Masahiro Fukui[5], Shigeaki Sasaki[5], Yusuke Hiwasaki[5]

[1]Huawei Technologies, China, [2]VoiceAge Corp., Canada, [3]France Telecom Orange, France,
[4]Electronics and Telecommunications Research Institutes, Korea, [5]NTT Cyber Space Laboratories, Japan

ABSTRACT

This paper presents high quality monaural superwideband extensions to G.711.1 and G.722, recently standardized as Recommendations ITU-T G.711.1 Annex D and G.722 Annex B. The superwideband (50-14000 Hz) functionality is achieved using embedded scalable structure that adds extension layers on top of the wideband core codecs. The bit rates are extended to 96/112/128 and 64/80/96 kbit/s for G.711.1 and G.722, respectively. The main technologies include lower and higher band (0-4 kHz and 4-8 kHz) enhancements, 8-14 kHz bandwidth extension and transform coding based on algebraic vector quantization. The codecs' performance is illustrated with listening test results extracted from formal ITU-T Characterization tests.

Index Terms— speech coding, audio coding, G.711.1 Annex D, G.722 Annex B, superwideband

1. INTRODUCTION

Wideband (WB, 50-7000 Hz) voice has gained momentum in fixed network services. Recommendation ITU-T G.722 standardized in 1988 [1] is the mandatory wideband codec specified for NG-DECT terminals, and it has been deployed for several years in fixed voice services. Furthermore, Recommendation ITU-T G.711.1 [2] was standardized in 2008 to provide low-delay, low-complexity, high-quality wideband speech and audio addressing the transcoding with legacy G.711 narrowband (300-3400 Hz) terminals.

Superwideband (SWB, 50-14000 Hz) is the next step for enriching audio quality to bring improved naturalness for voice signals and FM-broadcasting quality for non-voice signals. In 2008, ITU-T launched the G.711.1/G.722 SWB extension effort; the outcome consists of Recommendations ITU-T G.711.1 Annex D (G.711.1D) [3] and G.722 Annex B (G.722B) [4] approved in November 2010. These codecs are backward compatible with G.711/G.711.1 and G.722, respectively. Besides achieving speech quality at near transparency, they also provide very high music quality.

The paper presents an overview on ITU-T G.711.1D and G.722B, and is organized as follows. In Section 2, a general description of the SWB extensions is provided. In Sections 3 and 4, the encoder and decoder functionalities are described. Finally, the performance is discussed in Section 5.

2. CODEC MAIN FEATURES

The SWB extensions were developed jointly for both G.711.1 and G.722 core codecs, such that extension layers are computed using common algorithms. The input signal is processed in frames of 5 ms. By default, the encoder input and decoder output are sampled at 32 kHz. The actual operating bit rates depend on the selected core bit rate and enhancement layers.

The G.711.1D comprises two 16-kbit/s extension layers on top of G.711.1. Adding those layers corresponds to two bit rates at 96 and 112 kbit/s or at 112 and 128 kbit/s depending on the chosen G.711.1 core bit rate (80 kbit/s or 96 kbit/s).

Similarly, the SWB extension for G.722 core at 64 kbit/s comprises two 16-kbit/s extension layers corresponding to two bit rates at 80 and 96 kbit/s. Finally, the SWB extension for G.722 core at 56 kbit/s comprises one extension layer of 8 kbit/s for a total SWB bit rate of 64 kbit/s. This 8 kbit/s layer is also a part in a scalable way of the G.722 SWB extension layers at 80 kbit/s and 96 kbit/s.

3. ENCODER OVERVIEW

The encoder block diagram is shown in Fig. 1. A pre-processing high-pass filter is applied to the 32-kHz-sampled input signal to remove 0-50 Hz components. The pre-processed signal is divided into two 16-kHz-sampled wideband and super higher-band (SHB, 8-16 kHz) signals, using a 32-tap quadrature mirror filterbank (QMF). The wideband signal is divided into two 8-kHz-sampled lower band (LB, 0-4 kHz) and higher band (HB, 4-8 kHz) signals with the codec specific QMF. The LB and HB signals are coded with G.711.1 or an enhanced but bitstream interoperable version of G.722. The two codec dependent enhancement sub-layers (EL0 and EL1) further enhance the HB signal coding.

The SHB signal in the modified discrete cosine transform (MDCT) domain is coded by the SHB encoder into three SWB sub-layers (SWBLx, x = 0, 1, 2) common to

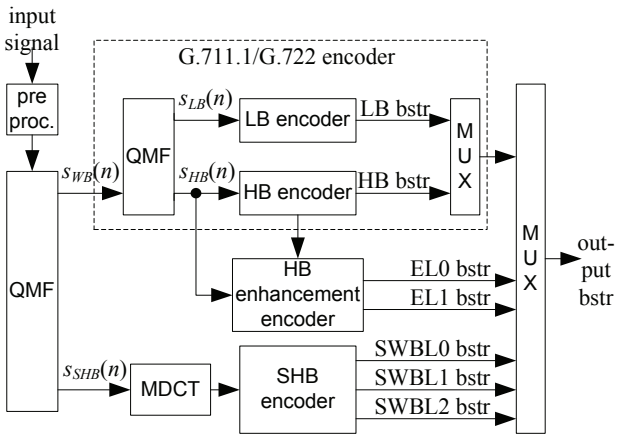

Fig. 1. G.711.1D/G.722B high-level encoder block diagram (*bstr* stands for bitstream).

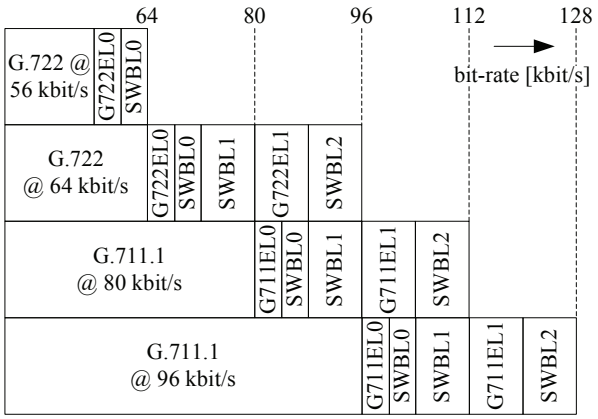

Fig. 2. Layered structure of G.711.1D/G.722B bitstreams.

both extensions. The SHB frequency coefficients are quantized using the bandwidth extension (BWE) and the algebraic vector quantization (AVQ) [5]. The summary of sub-layers as parts of two extension layers is shown in Fig. 2.

3.1. SWBL0 encoding

The SHB sub-layer SWBL0 is coded with a multi-mode BWE algorithm that switches the time and frequency resolution of the signal envelope depending on the signal classification.

Only the first 64 (out of 80) coefficients of SHB signal in MDCT domain, $S_{SHB}(k)$, $k = 0,..,63$, corresponding to 8.0-14.4 kHz are encoded. The last 16 MDCT coefficients corresponding to 14.4-16 kHz frequency range are discarded.

The input SHB time-domain signal $s_{SHB}(n)$ and the corresponding SHB spectrum $S_{SHB}(k)$ are used to classify the input signal. If the classifier [3] detects the previous or current frame as a transient frame, the transient class encoding is performed. In this case, EL0 is not coded and all 40 bits (16 kbit/s) are allocated to SWBL0. In case of non-transient frames, 21 bits (8.4 kbit/s) are allocated to the SWBL0 coding and 19 bits (7.6 kbit/s) are allocated to the EL0 coding. The non-transient frames are further classified as harmonic, normal or noise according to the spectral fluctuation. The classification information is coded with two bits. Furthermore, the global gain representing the SHB energy is scalar quantized with 5 bits.

The 64 MDCT coefficients $S_{SHB}(k)$ are split into four sub-bands for transient or eight sub-bands for non-transient frames, respectively. The spectral envelope is then computed as a set of root mean square (RMS) values per sub-bands and vector quantized. In case of transient frames, the time envelope is computed using the signal $s_{SHB}(n)$ as a set of RMS values per 20-sample sub-frames, and then vector quantized in addition to spectral envelopes. For non-transient frames, the time envelope is not computed.

3.2. SWBL1 and SWBL2 encoding

The SWBL1 and SWBL2 sub-layers code the fine structure of the normalized SHB MDCT coefficients $S_{SHB}(k)$, k=0,…,63, using the AVQ in eight sub-bands. Because of the limited bit budget, SWBL1 and SWBL2 (40 bits each per frame) cannot fully code spectrum in all sub-bands using the AVQ. Hereafter the sub-bands with zero bit allocation are called *zero sub-bands*.

The encoder for SWBL1 and SWBL2 sub-layers classifies the SHB signal into three SHB coding modes. The SHB modes 0 and 2 are chosen for spectrally sparse inputs. The difference between modes 0 and 2 is in the bit-allocation. The SHB mode 1 is selected for denser spectrum, aimed to reduce perceptual degradation caused by spectral holes when quantizing the dense spectrum by the AVQ designed for the sparse spectrum.

If the SHB mode does not equal 2, one bit in SWBL1 indicates whether the SHB mode is 0 or 1. Three SWBL1 bits are also used to refine the global gain coded in SWBL0. The remaining SWBL1 bits are allocated to the AVQ. All SWBL2 bits are allocated to the AVQ. The actual bit consumption of the AVQ parameters varies from frame to frame. The eventual bits unused by the AVQ are used to refine the zero sub-bands.

In SHB modes 0 and 2, the eight sub-bands are ranked in decreasing perceptual importance. More perceptually important sub-bands are coded in SWBL1 and then subsequently in SWBL2. The ranking order is derived from the spectral envelope obtained and quantized by the SWBL0 encoding. Zero sub-bands where the spectral envelope is not quantized sufficiently close to the original spectrum are detected and their energies are adjusted at the decoder. On a condition that SWBL1 and SWBL2 are transmitted and there are at least four AVQ unused bits in at least one of SWBL1 and SWBL2, spectrum coefficients in one or two zero sub-bands are searched using a maximum correlation search with the AVQ coded spectrum. One lag is coded in SWBL1 and another lag in SWBL2 (if bits are available).

Finally, if both SWBL1 and SWBL2 are transmitted and there are still some AVQ unused bits, the magnitudes of quantized spectrum in each sub-band are coded and multiplied using gradient adjustment factors.

In the SHB mode 1, to represent dense spectrum, only perpetually important coefficients are encoded and the others are generated using the spectral envelope at the decoder. The coefficients with amplitude greater than the spectral envelope are considered important and their differences to RMS are coded using the AVQ. Then zero sub-band filling is performed and zero coefficients in the AVQ quantized sub-bands are replaced by a coefficient with magnitude computed from the decoded spectral envelope.

3.3. G.711.1 enhancements

For non-transient frames, the G711EL0 enhancement sub-layer is coded with 19 bits (and none for transient frames). It encodes the 6.4-8.0 kHz residual noise in the MDCT domain (a difference between the HB input spectrum and the G.711.1 locally decoded spectrum) in two 8-coefficient sub-bands. The sub-band with greater energy is coded by the AVQ (usually the second one corresponding to the 7.2-8.0 kHz). To avoid the spectral gap between 6.4-7.2 kHz, either a vector quantization or a tilt coding is adaptively selected to represent the respective coefficients. The decision to use either scheme depends on the number of AVQ unused bits.

The subsequent G711EL1 coding is computed for all frames. It consists of two stages: a pulse coding of the MDCT residual coefficients using their respective amplitude and polarity. This is followed by a magnitude adjustment of the decoded MDCT coefficients from the G.711.1 core and G711EL0 by multiplication factors. The bit budget for multiplication factors is allocated dynamically.

3.4. G722 enhancements

For non-transient frames, the first enhancement sub-layer, G722EL0, is coded with 19 bits and enhances the 19 most perceptually important HB time samples (out of 40) using scalable scalar quantization. The second enhancement sub-layer, G722EL1 has a bit-budget of 40 bits to represent the whole frame (40 samples). It further refines the G.722 HB using an additional scalar quantization enhancement sub-layer (1 bit/sample). In addition to the scalable quantization in the G.722 HB, the LB part of legacy G.722 encoder is enhanced using noise feedback loop to perceptually shape the ADPCM coding noise. The noise feedback loop not only keeps the bit-interoperability with legacy G.722 bitstream, it gives a better subjective quality when decoded with legacy G.722.

4. DECODER OVERVIEW

In each 5-ms frame, the decoder can receive any of the supported bit rates, from 64 kbit/s up to 128 kbit/s for G.711.1D and from 64 kbit/s to 96 kbit/s or from 56 kbit/s to 64 kbit/s for G.722B respectively. The scalability in bit rates may result in bandwidth switching between adjacent frames. To mitigate the audible artifacts, the decoder smoothes out the spectral coefficients across frames with different bandwidth. In the case of an erased frame, in addition to the WB frame erasure correction (FEC) algorithm, the codecs support SHB FEC algorithm used to recover the SHB signal.

First, the EL0 and EL1 sub-layers are decoded if received and they enhance the HB spectrum. In G.722B, an MDCT domain post-processor is also applied to enhance the HB signal in non-transient frames. Then the SWBL0 is decoded. When SWBL1 and SWBL2 sub-layers are not received, the output SHB spectrum is derived from the wideband spectrum with adjusted spectral and time envelopes. The SHB signal class decoded in SWBL0 is further used to determine the SHB mode in SWBL1 and SWBL2 decoding. In the SHB mode 1, the decoded spectrum coefficients of each sub-band are obtained by adding the offset to the AVQ output. In case that the AVQ output has no amplitude, the polarity of the coefficient is randomly chosen. After the AVQ decoding and filling of some zero sub-bands, the decoded SHB spectrum can still contain one or more zero sub-bands. Those zero sub-bands are dealt in accordance to the SHB mode. In the SHB modes 0 and 2, the spectrum in zero sub-bands is replaced by the SWBL0 output spectrum. In the SHB mode 1, the spectrum in zero sub-bands is approximated by the spectral envelope and the sign extracted from the SWBL0 output spectrum. These zero sub-bands and the zero MDCT coefficients in non-zero sub-bands are subject to the BWE/AVQ adaptation. It is performed to reduce the perceptual noise and improve the subjective quality especially when the SWBL2 is not received. In case of the SHB mode 1, a spectrum post-processor is applied to smooth the decoded SHB spectrum.

The SHB MDCT coefficients are transformed to the time domain using the inverse MDCT overlap-add operation. The SHB synthesis is spectrally folded and together with the WB synthesis is filtered through the synthesis QMF to form the 32-kHz sampled output synthesis.

5. ITU-T PERFORMANCE EVALUATION

5.1. Subjective Quality

The G.711.1D (with core bit rate of 80 kbit/s) and G.722B superwideband codecs were formally evaluated in ITU-T Characterization tests in June 2010. The triple stimulus/hidden reference/double blind method ('Ref', 'A', 'B') with a five grade impairment scale, compliant with ITU-R BS.1116-1 [6], was used in the testing. Both codecs were evaluated in 3 sets of experiments (clean speech, noisy reverberant speech, music and mixed content) to check 45 requirement and 40 objective conditions. The G.722.1 Annex C codec was used as a reference in all experiments.

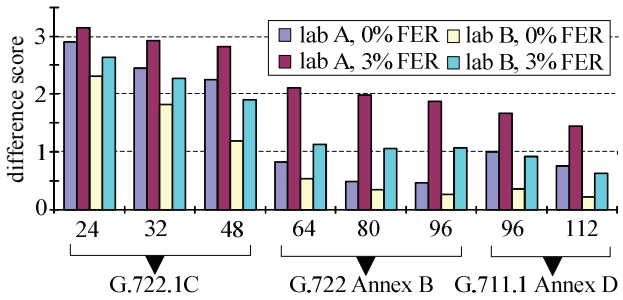

Fig. 3. Quality in clean speech (-26dBov) with clean channel and noisy channel with 3% frame error rate (FER).

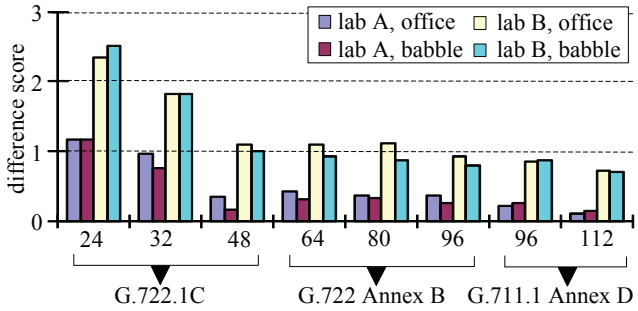

Fig. 4. Quality in noisy reverberant speech (office, babble).

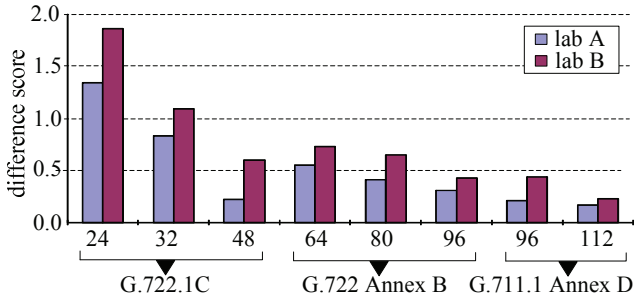

Fig. 5. Quality in music and mixed content.

Each experiment was run by two listening laboratories in different languages. Selected test results extracted from the G.711.1D and G.722B test report [7] are summarized in Figs. 3 to 5 by means of the difference scores (comparing the mean difference of the coded outputs to the originals). The numbers below the graph denote the codec bit rates in kbit/s. For G.711.1D, μ-law is used except for babble noise in Fig. 4 and music in Fig. 5 where A-law is used. At 95% confidence interval, both codecs passed all requirements in all experiments in both labs. For clean speech, all objectives were passed in both labs. For noisy reverberant speech and music and mixed content, all objectives but one passed in at least one lab.

5.2. Complexity and delay

The observed worst-case complexity and storage requirements in 16-bit words (encoder plus decoder with frame erasure rate of 10.7%) are based on the ITU-T Software Tool Library STL2009 [8] and are detailed in Table 1. The algorithmic delays of G.711.1D and G.722B are 12.8125 ms and 12.3125 ms, respectively. Those delay figures are among the lowest in the existing SWB codecs.

Table 1. Complexity and memory of G.711.1D/G.722B.

Codec		**G.711.1D, A-law**		**G.722B**	
Core codec bit rate		**80 kbit/s**	**96 kbit/s**	**56 kbit/s**	**64 kbit/s**
Complexity (WMOPS)	Encoder	12.139	12.417	7.926	10.935
	Decoder	8.741	9.081	10.613	11.826
	Overall	20.880	21.498	18.539	22.761
Dynamic RAM and Static RAM [kwords]		6.080		4.634	
Data ROM [kwords]		4.386		2.973	
Program ROM [basic ops]		5944		4905	

6. CONCLUSION

This paper presents the main features, coding architecture and performance of the recently standardized codecs G.711.1D and G.722B. These new codecs provide low-delay, state-of-the-art performance in superwideband speech and audio coding and are backward compatible with G.711.1/G.711 and G.722.

7. ACKNOWLEDGEMENTS

The authors would like to thank Simão Campos, Herve Taddei, Wu Wenhai, Xu Jianfeng, Lang Yue, Kimitaka Tsutsumi, Milan Jelínek, Vladimír Malenovský, and Tommy Vaillancourt for their contributions to the G.711.1D and G.722B work.

8. REFERENCES

[1] X. Maitre, "7 kHz audio coding within 64 kbit/s," *IEEE Select. Areas. Com.*, vol. 6, no. 2, pp. 283-298, Feb. 1988.
[2] 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.
[3] Rec. ITU-T G.711.1 Annex D (pre-published), Nov. 2010.
[4] Rec. ITU-T G.722 Annex B (pre-published), Nov. 2010.
[5] S. Ragot, B. Bessette, and R. Lefebvre, "Low-Complexity Multi-Rate Lattice Vector Quantization with Application to Wideband TCX Speech Coding at 32 kbit/s," In *Proc. IEEE ICASSP*, Montreal, QC, Canada, vol. 1, pp. 501-504, May 2004.
[6] Rec. ITU-R BS.1116-1, "Methods for the subjective assessment of small impairments in audio systems including multichannel sound systems," 1997.
[7] ITU-T WP3/16 TD289, "LS on G.722-SWB and G.711.1-SWB extension and G.718 post-processing," ITU-T Q7/SG12 rapporteurs, Geneva, Switzerland, July 2010.
[8] Rec. ITU-T G.191, "Software tools for speech and audio coding standardization," March 2010.