RE-ENGINEERING ITU-T G.722: LOW DELAY AND COMPLEXITY SUPERWIDEBAND CODING AT 64 KBIT/S WITH G.722 BITSTREAM WATERMARKING

Balazs Kovesi[1], Stéphane Ragot[1], Claude Lamblin[1], Lei Miao[2], Zexin Liu[2], Chen Hu[2]

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

ABSTRACT

This paper presents the lowest bitrate mode (64 kbit/s) of the new superwideband (SWB, 50-14000 Hz) coder, recently standardized as ITU-T G.722 Annex B. This mode provides a superwideband extension of G.722 at 56 kbit/s with one 8 kbit/s enhancement layer divided in two sub-layers. The resulting bitstream is compatible with ITU-T G.722 at 64 kbit/s and can be viewed as watermarking G.722 least significant bits (LSBs). The novel technologies in this mode include G.722 enhancements (noise feedback coding, scalable quantization in G.722 higher band), as well as multimode bandwidth extension (BWE). Selected ITU-T characterization test results and additional informal test results show that the 64 kbit/s mode of G.722 Annex B gives high SWB quality with low delay and complexity.

Index Terms— Speech coding, audio coding, G.722B, superwideband, bandwidth extension, watermarking

1. INTRODUCTION

Wideband voice (sampled at 16 kHz) – also known as High-Definition (HD) voice – offers a proven improvement in sound quality compared to traditional voice calls. HD voice deployments started few years ago in fixed voice over IP (VoIP) network services. In particular, New Generation DECT, which was later named CAT-iq (Cordless Advanced Technology - Internet and Quality), has been successfully introduced in the market, and specifies two mandatory codecs in portable terminations (handsets) and fixed terminations (base stations) [1,2]: one narrowband codec, ITU-T G.726 at 32 kbit/s, for backward compatibility with DECT, and one wideband codec, ITU-T G.722 at 64 kbit/s [3,4] for HD voice. To transport the 64 kbit/s bitrate of G.722, the bitrate per channel at the air interface is doubled from 32 kbit/s in traditional DECT to 64 kbit/s.

For future conversational services, superwideband (SWB, 50-14000 Hz, sampled at 32 kHz) audio is the next step in (mono) audio quality improvement. In 2008 ITU-T launched the G.722-SWB standardization to develop an embedded scalable SWB extension of G.722. This effort resulted in the recently standardized ITU-T G.722 Annex B (G.722B) coder [5] operating at 64, 80 and 96 kbit/s. The lowest bit rate mode (64 kbit/s) has been designed as an extension of G.722 at 56 kbit/s with an 8 kbit/s enhancement layer, to provide an overall bit rate fitting in existing 64 kbit/s CAT-iq transport channels.

A general description of G.722B can be found in [5,6]. This paper presents only a subset (or profile) of G.722B, corresponding to the lowest bit rate mode (64 kbit/s). This paper is organized as follows. In Sections 2 and 3, the encoder and decoder parts are described focusing on the novel technologies brought in G.722B at 64 kbit/s. The codec performance is discussed in Section 4, before concluding in Section 5.

2. G.722B ENCODER AT 64 KBIT/S

The input signal sampled at 32 kHz is processed in frames of 5 ms each. The encoder block diagram is shown in Fig. 1. A 1st order IIR 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 (WB, 0-8 kHz) and super higher-band (SHB, 8-16 kHz) signals $s_{WB}(n)$ and $s_{SHB}(n)$, using a 32-tap quadrature mirror filterbank (QMF). The WB signal is encoded by an enhanced 56 kbit/s G.722 core with 280 bits per frame. Additionally, an enhancement layer with 40 bits per frame is coded, comprising a wideband enhancement sub-layer (EL0) and a SHB sub-layer (SWBL0).

2.1. SWBL0 encoding

The SHB signal $s_{SHB}(n)$ is transformed in the frequency domain by the well known modified discrete cosine transform (MDCT) using a 10 ms sinusoidal window, and coded with a multi-mode BWE algorithm using adaptive envelope coding. 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 of $S_{SHB}(k)$, $k = 64,..79$, corresponding to 14.4-16 kHz are discarded.

The input SHB time domain signal $s_{SHB}(n)$ and the SHB spectrum $S_{SHB}(k)$ are used to classify the input signal. If the transient detector detects the previous or current frame as a

transient frame, the transient class encoding is performed. In this case, the 40 bits are allocated to the SWBL0 encoder

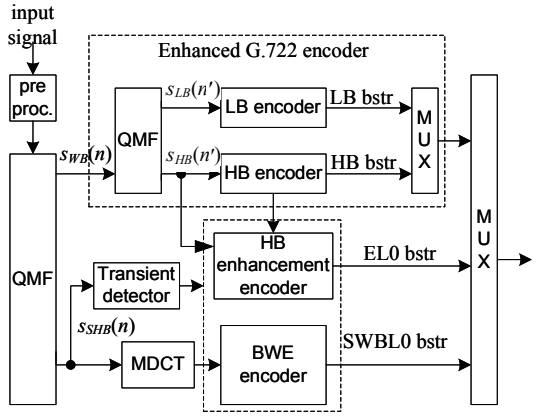


Fig. 1. 64 kbit/s G.722B high-level encoder block diagram ("bstr" stands for bitstream).

(i.e. 0 bit is allocated to the EL0 encoder). In case of non-transient frames, a bit-budget of 21 bits is used in the SWBL0 coding (i.e. 19 bits for the EL0 encoder).

The transient detection uses parameters computed from time envelopes of three consecutive frames - the current frame and the previous two frames. The time envelope is computed from the spectrally-folded SHB signal $s_{SHB}^{fold(m-m0)}(n)$, where $m0 = 0,...,2$ is the frame offset and m the current frame index, and it is defined as the log-RMS of blocks of 20 samples (the time frame is split in four blocks):

$$t_{rms}^{(m-m0)}(j) = \frac{1}{2}\log_2\left(\frac{1}{20}\sum_{n=0}^{19} s_{SHB}^{fold(m-m0)}(20j+n)^2\right), j = 0,...,3,$$

Three parameters are calculated for the transient detection [5]: the total RMS of time envelopes in three consecutive frames t_{rms_total}, the maximum difference between 12 consecutive time envelopes $d_{tenv\_\max}$ and the maximum difference between the time envelopes and the total RMS of the time envelopes $d_{tenv\_total\_\max}$. The current frame is classified as a transient frame when $t_{rms\_total} > 8$, $d_{tenv\_\max} > 2.4$ and $d_{tenv\_total\_\max} > 3.3$.

The non-transient frames are further classified as harmonic, normal or noise based on the SHB spectrum fluctuation evaluated by the spectral sharpness, $\kappa(j)$, j = 0,..,9, defined as the ratio between peak magnitude and average magnitude in "sharpness bands" of 6 coefficients [5].

The class information is coded on 2 bits. A global gain g_{glob} is calculated and encoded using 5-bit uniform scalar quantizer in the range [0,31]:

$$g_{glob} = round\left(\frac{1}{2}\log_2\left[\frac{1}{64}\left(\sum_{k=0}^{63}(S_{SHB}(k))^2\right)\right]\right).$$

The 64 MDCT coefficients $S_{SHB}(k)$ are split in 4 sub-bands of 16 coefficients in transient frames, and in 8 sub-bands of 8 coefficients otherwise. Then, the normalized sub-band spectral envelopes are quantized. In case of transient frames, the 4 time envelopes $t_{rms}^{(m)}(j), j = 0,...,3$ are also encoded.

2.2. Improved 56 kbit/s G722 core and EL0 encoding

At 56 kbit/s the legacy G.722 encoder splits the bit rate into 5 bits/sample for the lower band (LB, 0-4 kHz) and 2 bits/sample for the high band (HB, 4-8 kHz). In G.722B the G.722 LB encoder was modified to perceptually shape the ADPCM coding noise by noise feedback using a 4th order noise shaping filter derived from a linear prediction (LP) filter [5].

When high frequencies (>8kHz) are added on top of the G.722 synthesis, the quantization noise of the G.722 HB signal becomes more audible, therefore enhancements to G.722 HB coding were required to obtain good SWB quality. For non-transient frames only, the 2 bits/sample G.722 HB coding is refined by an enhancement sub-layer EL0 with 19 bits. The 19 most perceptually important HB samples are selected by an adaptive algorithm based on the core decoded HB samples [5]. For these 19 selected samples, a new embedded scalar quantizer (SQ) of 3 bits/sample extends the G.722 HB 2-bit SQ. Note that this quantization is performed by analysis by synthesis in a perceptually weighted domain.

2.3. Bitstream format as a form of watermarking

The 40 bits of two sub-layers (EL0 and SWBL0) replace in G.722 legacy bistream at 64 kbit/s the 40 least significant bits (LSBs). These LSBs are the LSB of the 40 coded LB samples (which are coded on 6 bits). Note that these LSBs are dropped if G.722 operates at 56 kbit/s. Furthermore, as explained in section 4.2, the G.722B SWB bitstream at 64 kbit/s can be decoded by a G.722 WB legacy decoder at 64 kbit/s without annoying degradation. So the scheme can be seen as watermarking embedding SWB information in WB bitstream, similarly to the scheme described in [13] to embed WB information in NB bitstream.

3. G.722B DECODER AT 64 KBIT/S

The decoder output is sampled at 32 kHz. First, G.722 decoding at 56 kbit/s is performed. Then, the 2-bit class information is decoded. In non-transient frames, the G.722 decoded HB signal is enhanced using EL0 sub-layer, which is further improved by an MDCT domain post-processor [5]. Then, the bandwidth is extended to SWB by SWBL0 decoding.

3.1. EL0 decoding and MDCT postprocessing

In non-transient frames only, the 19 most perceptually important samples are determined using the same adaptive algorithm as in the encoder. For these samples, the ADPCM decoder in G.722 HB uses an embedded 3 bits/sample scalar inverse quantizer to double the resolution. Then, an MDCT

domain HB post-processor is applied to improve the quality of the WB decoded signal in 4.4 – 8 kHz frequency range. An MDCT is performed on the decoded wideband signal. Two magnitude parameters are defined, called local masking magnitudes, $M_0(k)$, and local masked magnitudes, $M_1(k)$,

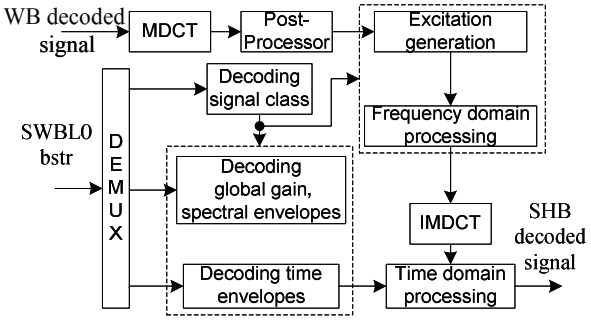


Fig. 2. BWE decoder diagram for SWBL0 sub-layer.

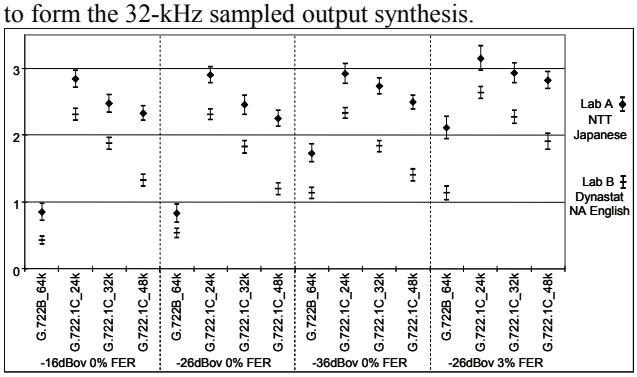
respectively. The local masking magnitudes are viewed as the "perceptual loudness" and local masked magnitudes as the estimated local "perceptual error floor". They are estimated by taking a weighted sum of the spectral magnitude neighbors. The ratio $M_0(k)/M_1(k)$ can reflect the local relative perceptual quality at bin *k*. In function of this ratio some spectra with enough quality are amplified with gain factors slightly larger than one whereas some poor quality spectra are attenuated or reduced below an estimated masking threshold [5].

3.2. SWBL0 decoding

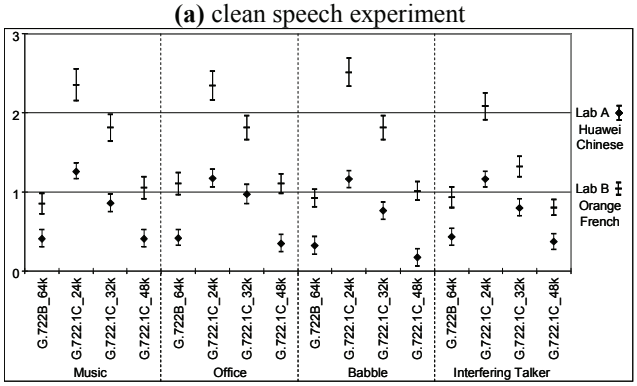
The BWE decoder is illustrated in Fig. 2. Depending on the SHB signal class, spectral envelopes and time envelopes are adaptively decoded: 4 spectral envelopes and 4 time envelopes for transient frames, only 8 spectral envelopes for non transient frames (no time envelope). The decoded spectral envelopes are smoothed and adjusted according to the SHB signal class.

Frequency excitations are also generated according to the SHB signal class. If the current frame is noise and the previous frame is not harmonic, a random noise is generated for the BWE excitation. Otherwise, the wideband MDCT coefficients after the higher band post-processor are used. The excitation signal is normalized to remove the envelope information. The adjusted spectral envelope is applied to the normalized excitation signal. Then, the SHB MDCT coefficients are transformed to the time domain using inverse MDCT overlap-add operation. Afterwards, the time envelope is applied if the current frame or the previous frame is classified as transient. When only the previous frame is transient, the current frame time envelope is obtained by attenuating the last time envelope in the previous frame.

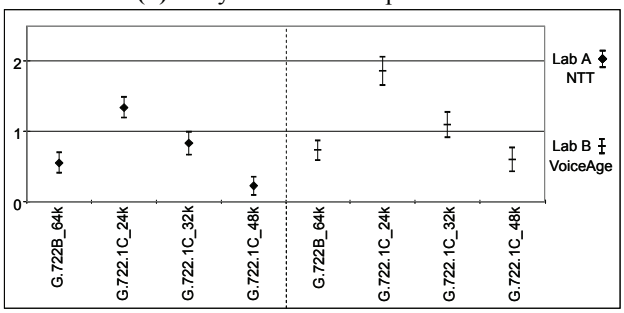
Finally, the SHB synthesis is spectrally folded and combined with the WB synthesis through the synthesis QMF to form the 32-kHz sampled output synthesis.



(a) clean speech experiment



(b) noisy reverberant experiment



(c) music and mixed content experiment

Fig. 3. G.722B@64kbit/s quality comparison.

3.3. Frame Erasure Concealment (FEC)

In case of erased frames, a WB FEC algorithm, derived from the G.722 App. IV [7], reconstructs the WB signal. The SHB signal is recovered by attenuating the inverse MDCT signal from the last good frame with overlap-add [5].

4. PERFORMANCE EVALUATION

4.1. ITU-T quality results for G.722B at 64 kbit/s

G.722B was formally evaluated in ITU-T Characterization tests in June 2010. The triple stimulus/hidden reference/double blind method ('Ref', 'A', 'B'), compliant with ITU-R BS.1116-1 [8] was used with a five grade impairment scale. The 64 kbit/s mode was evaluated

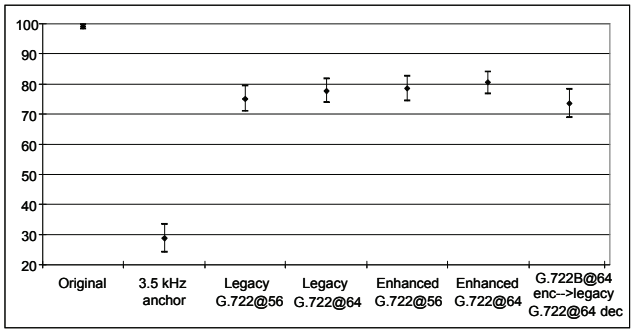


Fig. 4. Informal MUSHRA test results for WB conditions

together with the higher bit rate modes at 80 and 96 kbit/s in 3 experiments (clean speech, noisy reverberant speech, music and mixed content). Each experiment was run by two listening laboratories (A and B). The requirements for G.722B at 64, 80 and 96 kbit/s were to be not worse than G.722.1 C at 24, 32 and 48 kbit/s [9] respectively for all kinds of signals. The objectives for the G.722B at 64 and 80 kbit/s were to be not worse than G.722.1C at 32 and 48 kbit/s respectively, and for G.722B at 96 kbit/s to be better than G.722.1C at 48 kbit/s.

Selected results extracted from the G.722B Characterization test report [10] are summarized in Fig. 3a-c by means of the difference scores (comparing the mean difference of the coded outputs to the originals). G.722B at 64 kbit/s not only passes all the requirements and objectives for its own bitrate but also it could almost pass all the requirements set for the highest 96 kbit/s bitrate (except for music in laboratory A). For clean speech even the objectives for G.722B at 96 kbit/s could be passed by G.722B at 64 kbit/s (i.e. better than G.722.1C at 48 kbit/s).

4.2. G.722B backward compatibility in wideband

Informal listening test results are provided in Fig. 4 to illustrate the quality enhancement of G.722 core in G.722B and the impact of G.722 bitstream watermarking. The MUSHRA test methodology [11] was used with 9 items (3 clean speech, 6 music and mixed content) and 8 expert listeners. Legacy G.722 encoder and the enhanced G.722 encoder in G.722B were compared at 56 and 64 kbit/s. A last condition tested the scenario where the G.722B bitstream at 64 kbit/s is decoded by the legacy 64 kbit/s G.722 decoder. Note that all these conditions used the legacy G.722 decoder. It can be observed that thanks to noise shaping the enhanced G.722 encoder quality at 56 kbit/s is equivalent to legacy G.722 at 64 kbit/s. Furthermore when the "watermarked" SWB 64 kbit/s bitstream is accidentally decoded as a legacy G.722 bitstream the quality is equivalent to legacy G.722 at 56 kbit/s. Therefore, the SWB extension at 64 kbit/s can be decoded by the legacy G.722 decoder without annoying degradation.

4.3. Delay and complexity of G.722B at 64 kbit/s

The algorithmic delay is 12.3125 ms. The observed encoder and decoder complexity worst-cases, measured with basic operators specified in the ITU-T Software Tool Library STL2009 [12], are 7.926 and 10.613 WMOPS, respectively.

5. CONCLUSION

This paper described the 64 kbit/s mode of G.722B superwideband extension of G.722. This new ITU-T SWB standard provides state-of-the-art performance for SWB speech and music with very low delay and complexity while being backward compatible with G.722.

ACKNOWLEDGEMENTS

The authors would like to thank Simão Campos, Hervé Taddei, Yusuke Hiwasaki, Catherine Quinquis, Paolo Usai, Wu Wenhai, Xu Jianfeng, Lang Yue and Alain Le Guyader for their kind help during the ITU-T standardization process of the G.722 superwideband codec.

REFERENCES

[1] ETSI TS 102 527-1 (CAT-iq 1.0), New Generation DECT; Part 1: Wideband speech
[2] ETSI TS 102 527-3 (CAT-iq 2.0), New Generation DECT; Part 3: Extended wideband speech services
[3] ITU-T Rec. G.722, "7 kHz audio-coding within 64 kbit/s, " November 1988
[4] X. Maitre, "7 kHz audio coding within 64 kbit/s," IEEE Select. Areas. Com., vol. 6, no. 2, pp. 283-298, Feb. 1988.
[5] ITU-T Rec. G.722 Amendment 1 (ex G.722-SWB) (pre-published), 7 kHz audio-coding within 64 kbit/s: New Annex B with superwideband embedded extension, July. 2010.
[6] L. Miao et al., "G.711.1 Annex D and G.722 Annex B – New ITU-T superwideband codecs, " accepted to ICASSP 2011, Prague, Czech Republic
[7] B. Kovesi, S. Ragot, "A Low Complexity Packet Loss Concealment Algorithm for ITU-T G.722," ICASSP 2008, Las Vegas, Nevada, USA
[8] Rec. ITU-R BS.1116-1, "Methods for the subjective assessment of small impairments in audio systems including multichannel sound systems," 1997.
[9] C. Lamblin, C. Quinquis, P. Usai, "ITU-T G.722.1 annex C: the first ITU-T superwideband audio coder, " IEEE Communication Magazine, vol. 46, N°10, October 2008.
[10] ITU-T TD289 GEN/16, "LS on G.722-SWB and G.711.1-SWB extension and G.718 post-processing", source: ITU-T Q7/12 rapporteurs, Geneva, Switzerland, July 2010.
[11] Rec. ITU-R BS.1534, "Method for the subjective assessment of intermediate quality levels of coding systems," 2003.
[12] Rec. ITU-T G.191, "Software tools for speech and audio coding standardization," March 2010.
[13] H. Ding, "Wideband audio over narrowband low-resolution media," ICASSP 2004, vol. 1, pp 489-492, Montreal, Quebec, Canada.