CANDIDATE PROPOSAL FOR ITU-T SUPER-WIDEBAND SPEECH AND AUDIO CODING

Bernd Geiser, Hauke Krüger, Heinrich W. Löllmann, Peter Vary [1]
Deming Zhang, Hualin Wan, Hai Ting Li, Li Bin Zhang [2]

[1] Institute of Communication Systems
and Data Processing (ind)
RWTH Aachen University, Germany
geiser@ind.rwth-aachen.de

[2] Core Network Research Dept.
Huawei Technologies Co., Ltd.
Beijing, P.R. of China
zhangdeming@huawei.com

ABSTRACT

This paper describes the speech and audio codec that has been submitted to ITU-T by Huawei and ETRI as a candidate for the upcoming super-wideband and stereo extensions of Rec. G.729.1 and G.718. The core codec in the current implementation is G.729.1 and the encoded frequency range is increased from 7 kHz to 14 kHz. Therefore, the maximum bit rate is raised from 32 kbit/s to 64 kbit/s by adding five bitstream layers. A comprehensive overview of the codec is presented with a focus on the mono coding components. The results of the listening tests that have been conducted during the ITU-T qualification phase are summarized. The proposed codec passes all quality requirements for mono input signals.

***Index Terms*—** Speech and audio coding, ITU, standardization

1. INTRODUCTION

Standardized speech and audio codecs that offer a very high or nearly transparent quality while remaining compliant with tight conversational requirements (e.g. delay constraints in two-way communication) are only recently emerging. Typically, such codecs offer a high acoustic bandwidth (more than 10 kHz) together with stereophonic encoding. Potential applications are mainly found in high-quality conferencing scenarios and Voice-over-IP telephony but they also include other audio services such as streaming, e-learning or remote monitoring. Still, the deployment of new codecs in existing system environments and networks often turns out to be a time-consuming and costly procedure. A more timely and economic approach is to apply *embedded coding techniques* that are built on top of existing (and preferably widely deployed) codecs [1]. This concept facilitates a smooth migration towards the new service quality.

The demand for backwards-compatible *and* high-quality speech and audio codecs for conversational services has been identified and addressed within ITU-T SG16 by launching corresponding standardization processes. It was required to enhance both ITU-T G.729.1 [2, 3] and ITU-T G.718 [4, 5] with "super-wideband" (SWB, acoustic bandwidth of 0.05–14 kHz) and stereo capabilities. Thereby, backwards compatibility with the existing codecs should be maintained. Because both codecs have a sufficiently similar structure, it has later been decided to join the efforts and a combined work item was defined in Q23/SG16. Though, for the qualification phase, only G.729.1 had to be considered as the base layer codec while the adaptation to G.718 is scheduled for the next phase. In the present contribution, we describe the proposal that has been submitted to ITU-T by Huawei (China) and ETRI (South Korea) as a candidate for qualification.

2. CODEC OVERVIEW

In compliance with the applicable "Terms of Reference" [6], the submitted algorithm constitutes an embedded and fully backwards compatible extension of ITU-T Rec. G.729.1. The key features of the submitted codec proposal are:

- Input/output sampling rates of 8, 16 and 32 kHz.
- Extension of the encoded bandwidth from 0.05–7 kHz (Wideband, WB) to 0.05–14 kHz (Super-Wideband, SWB), starting at a bit rate of 36 kbit/s.
- Stereophonic coding for WB audio signals (starting at 40 kbit/s) and SWB audio signals (starting at 48 kbit/s).
- Concealment of frame erasures.

To realize these features, the bit rate of G.729.1 is increased to 64 kbit/s by adding five bitstream layers. Here, the G.729.1 "core layer" with a rate of 32 kbit/s is referred to as **L1**. The "enhancement layers" are **L2a**, **L2b** (4 kbit/s each) and **L3**, **L4**, **L5** (8 kbit/s each). Layer L2a contains one or two bits to specify the input sampling rate and the number of channels: 32 kHz mono (0), 16 kHz stereo (10) or 32 kHz stereo (11). In the following, we will focus on the 32 kHz mono mode of the codec.

For the mono case, the high level structure of the codec proposal is shown in Fig. 1. The input signal (32 kHz sampling rate) is split into two critically sampled (16 kHz) subband signals $s_{\mathrm{wb}}(n)$ and $s'_{\mathrm{swb}}(n)$ by means of an infinite impulse response quadrature mirror filter bank (IIR QMF). The signal $s_{\mathrm{wb}}(n)$ is fed into the G.729.1 core codec and encoded as described in [2] with the two exceptions that optionally an IIR QMF analysis filter bank can be used in G.729.1 for the 4 kHz split ("IIR QMF mode"), and that, for 32 kHz input, the 3 kHz lowpass pre-processing in the 4–8 kHz path is bypassed before entering the TDAC (Time Domain Aliasing Cancellation) transform codec. In the 8–16 kHz path, the signal $s'_{\mathrm{swb}}(n)$ is pre-processed by spectral mirroring, delay or phase compensation and 6 kHz lowpass filtering to obtain $s_{\mathrm{swb}}(n)$. This signal is then processed and encoded jointly with the 0–8 kHz transform coefficients from the TDAC module of the core codec.

On the *decoder* side, first, the G.729.1 layer L1 is decoded as described in [2] with the exception that additional transform coefficients can be passed to the TDAC decoder. For normal mode, the further processing in the G.729.1 core coder is identical to [2]. In "IIR QMF mode", the synthesis filter bank is replaced by an IIR QMF solution with integrated phase equalization (PE) [7]. The high frequency band is synthesized by the SWB decoder. The 8–16 kHz time domain signal $\tilde{s}_{\mathrm{swb}}(n)$ is either processed by a phase compensation (for "IIR QMF mode") or suitably delayed (for normal mode). Finally, the 0–8 kHz and 8–16 kHz bands are combined by IIR QMF synthesis with integrated phase equalization.

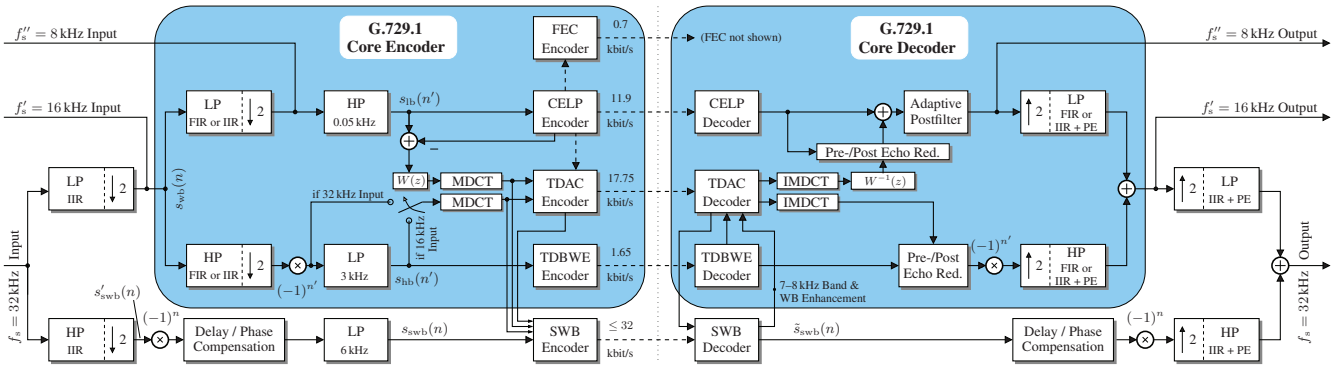

Fig. 1. Encoder and decoder overview (mono only)

3. SWB MONO ENCODER

This section describes the SWB mono encoder. The respective signal flow chart is shown in the left part of Fig. 2.

3.1. Adaptive Temporal Envelope (ATE) and FEC Information

In the encoder, first, the temporal structure of the signal $s_{\mathrm{swb}}(n)$ is analyzed and each frame is classified as either "transient" or "stationary". This decision is communicated to the decoder using one bit. For *stationary* frames, a logarithmic *gain* is determined and quantized with 5 bits. For *transient* frames, a *temporal envelope* consisting of 8 logarithmic subframe gains (2.5 ms subframes) is encoded by separate differential quantization for even and odd subframes. The quantized temporal information enables an efficient frame erasure concealment (FEC) in the decoder (see Sec. 4.3). Therefore, the information is either transmitted redundantly (for stationary frames) or distributed across neighboring frames (for transient frames), i.e., we have:

- A mode bit (transient/stationary) for the current frame and a (repeated) mode bit for the previous frame (FEC),
- 5 bits for the gain of the current frame if it is classified as stationary and 5 bits for the gain of the previous frame if this frame has been classified as stationary (FEC),
- 17 bits for the differentially coded subframe gains (odd subframes) if the current frame is classified as transient and 17 bits for the differentially coded subframe gains (even subframes) if the previous frame has been classified as transient.

The ATE and FEC bits are transmitted in L2a. The total number of bits is 12 for consecutive stationary frames, 24 for transient-stationary transitions and 36 for consecutive transient frames. From the quantized ATE, a "temporal gain function" (TGF) is constructed by overlap-add of scaled Hann windows. The input signal $s_{\mathrm{swb}}(n)$ is then normalized using the TGF, resulting in the signal $s^{\mathrm{T}}_{\mathrm{swb}}(n)$. The TGF exhibits a pronounced low-pass characteristic such that spectral leakage is largely avoided.

3.2. Spectral Envelope

The temporally normalized signal $s^{\mathrm{T}}_{\mathrm{swb}}(n)$ is transformed to the frequency domain using a modified discrete cosine transform (MDCT, 40 ms / 640 samples Kaiser-Bessel window, 20 ms / 320 samples frame shift). The 240 MDCT coefficients for the 8–14 kHz band are jointly processed with the 40 TDAC coefficients that describe the 7–8 kHz band. The encoder computes a *spectral envelope* in terms of 18 logarithmic subband gains ($1 \cdot 8 + 17 \cdot 16$ MDCT coeffs.). This envelope is quantized in a multi-stage approach: First, spherical vector quantization (VQ, cf. Sec. 3.4) is applied to a reduced 16-dim. vector (the last 4 subbands are merged into 2). The bit allocation is 4 bits for the vector mean and 34–37 bits for the VQ depending on the bit budget in L2a. If necessary, the quality of the spectral envelope is improved by scalar quantization of the VQ error and entropy coding of the quantizer indices (Huffman codes). Additionally, three control bits indicate which refinements are sent. This multi-stage quantization scheme efficiently captures outliers that stem from an insufficient VQ encoding. Entropy coding naturally entails a variable bit rate, but the encoder restricts these bits to L2a and L2b.

3.3. Parametric Fine Structure Encoding

Here, the harmonic structure of the 8–14 kHz MDCT signal is analyzed. Since the MDCT is a strongly phase-sensitive transform and thus inappropriate for concise signal analysis, we use the so-called "pseudo-spectrum" from [8] instead and apply an additional spectral whitening (tilt compensation). From this representation, first, a correlation based tonality value is computed and quantized with 3 bits. Secondly, a flag indicates whether the 8–14 kHz fine structure is similar to the WB fine structure. If this is the case, no additional information is transmitted. If dissimilarity is detected, the harmonic structure of the 8–14 kHz signal is explicitly transmitted. This explicit description consumes up to 14 bits (depending on the tonality) and consists of a fractional harmonic grid and an offset. The tonality value and the mode flag are always transmitted in L2a. The description of the harmonics is transmitted in L2a if enough bits are available. Otherwise, these bits are moved to L2b.

3.4. MDCT Coefficient Quantization

For logarithmic spherical vector quantization [9] of the MDCT subbands, a number of GLCVQs (Gosset Low Complexity VQs) of dimension 8 or 16 with bit rates up to 2 bits per sample has been designed. For this gain-shape concept, the gain component is already available through the spectral envelope from Sec. 3.2. For the shape component, an adaptive bit allocation procedure assigns a certain number of bits to each of the 18 MDCT domain subvectors based on the quantized spectral envelope. The selected quantizers are applied to the respective MDCT domain subvectors and the obtained GLCVQ indices fill the bitstream up to the 64 kbit/s limit (L3–L5).

The GLCVQ codebooks are composed of vectors that are located on spherical shells of the 8-dim. *Gosset Lattice* E_8. As a generalization, our approach can handle arbitrary dimensions by using the E_n lattice. The quantization algorithm is based on the representation of all codevectors as *permutation codes* in analogy to [10]. Here, a very low computational complexity and memory consumption has been achieved by grouping the *classleader vectors* from the original algorithm into a low number of *classleader root vectors* based on separate handling of signs and magnitudes for each vector coordinate.

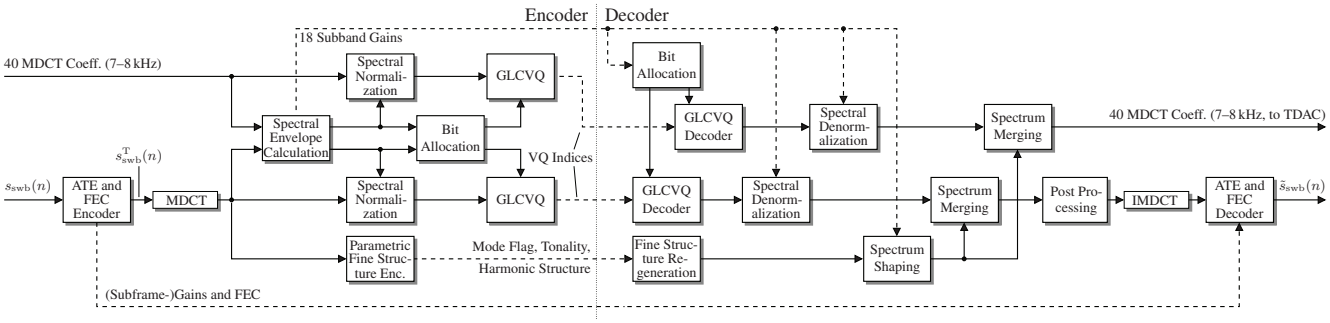

Fig. 2. SWB encoder and decoder. (ATE: Adaptive Temporal Envelope, GLCVQ: Gosset Low Complexity Vector Quantization)

In the quantization routine, instead of searching all classleader vectors, the optimal codevector is found by comparing the classleader root vectors with a sign-removed version of the input vector. The final lattice indexing is performed with the method from [4].

3.5. Wideband Enhancement Encoding (not shown in Fig. 2)

For certain signals, especially some tonal signals, the G.729.1 output (WB) still exhibits an insufficient quality. For example, the TDAC bit allocation algorithm may allocate *zero* bits to a significant number of spectral subbands. Therefore, our codec proposal implements a wideband enhancement (WBE) module which transmits additional (or previously insufficiently encoded) TDAC subbands using vector quantization. The respective bit allocation is based on a fixed bit budget and the order of perceptual importance from the TDAC module is taken into account. The WBE information is available in bitstream layer L2b and/or L3.

4. SWB MONO DECODER

The SWB mono decoder is illustrated in the right part of Fig. 2. Its operation, as described below, depends on the received bit rate.

4.1. Fine Structure Regeneration and Spectrum Shaping

This module is responsible for the regeneration of a spectral fine structure in the MDCT domain based on the parametric description from Sec. 3.3. If the transmitted mode flag indicates similarity between low and high band, the fine structure in the 7–14 kHz range is derived from the WB frequencies by spectral replication. Additionally, the tonality of the replicated signal is adjusted according to the received tonality value by "peak sharpening" or "noise mixing". If the mode flag indicates dissimilarity between low and high band, the fine structure in the 7–14 kHz range is generated as a mixture of pseudo random noise and synthetic harmonics which are placed according to the fractional harmonic grid and the offset parameters. The energy ratio of noise and harmonic components is controlled by the tonality value. Yet, concise signal synthesis in the MDCT domain is not straight-forward. Therefore, the individual harmonics are synthesized by imitating the MDCT domain behavior of time domain sinusoids [8].

The generated MDCT coefficients are spectrally shaped by multiplication with a gain correction factor which is the ratio of the desired subband gain (spectral envelope) and the measured gain of the generated fine structure. For tonal subbands, an interpolated gain factor is applied, instead. This is done to mitigate potential artifacts due to misplaced harmonics, in particular for high-pitched signals (e.g. violin). Note that even disharmonic signals (e.g. triangle) can be regenerated reasonably well since unwanted harmonics can be suppressed due to the relatively fine-grained spectral envelope.

4.2. Quality Refinement by GLCVQ and Post-Processing

If the complete spectral envelope is received, the same bit allocation as in the encoder can be computed. Then, for each received index of the GLCVQ, the parametric signal from Sec. 4.1 is replaced by the decoded MDCT subvector of dimension 8 or 16, respectively. This vector is denormalized by the spectral envelope.

As in [2], post-processing is applied in the MDCT domain to improve the subjective quality, whereby an adaptive subband decomposition has been used here. Then, the inverse MDCT is applied to the 8–16 kHz components (coefficients above 14 kHz are set to zero). The 7–8 kHz subband is passed to the TDAC decoder of G.729.1.

4.3. Adaptive Temporal Denormalization and FEC

Based on the 8–16 kHz time-domain signal, temporal denormalization by multiplication with the TGF from Sec. 3.1 is performed. This denormalization effectively suppresses pre-echo artifacts that are particularly strong in the scenario where no quantized MDCT coefficients are available (purely parametric signal, see Sec. 4.1). Nevertheless, due to the stationary/transient distinction, spectral details in stationary signal segments can be preserved.

If a *frame erasure* is signaled to the decoder, concealment is performed. For the 7–8 kHz band, the MDCT coefficients from the previous frame are copied and the FEC scheme of the G.729.1 TDAC is reused. In the 8–14 kHz band, for the *first lost frame*, the previous coefficients are repeated, but during temporal denormalization the transmitted FEC information is used to find the type of the missing frame (transient/stationary). For stationary frames, the FEC information contains the overall gain factor which is used for denormalization. In transient frames, the FEC information contains the differentially coded subframe gains of subframes with *odd* indices. Hence, an interpolated TGF with 5 ms resolution can be reconstructed and used for denormalization. In case of *bursty frame erasures*, i.e., if the current frame is lost and in addition the previous frame was lost, the currently missing frame is assumed to be stationary, and the (averaged) gain factor from the previous (reconstructed) frame is decreased by one quantization step before applying denormalization. For *recovery frames*, i.e., for the first frame after a frame erasure, in principle the complete information is available, i.e., the decoding can proceed as usual. However, there is one exception which occurs for transient recovery frames. The frame type (transient) is known, but the subframe gains of the odd subframes are unavailable since they should have been received in the previous (erased) frame. Again, the complete TGF cannot be reconstructed and only the subframe gains with *even* indices are used to form the TGF.

4.4. Wideband Enhancement Decoding (not shown in Fig. 2)

Here, following the WBE bit allocation according to Sec. 3.5, received VQ indices are decoded and the quantized MDCT subbands are inserted into the WB spectrum of the TDAC decoder.

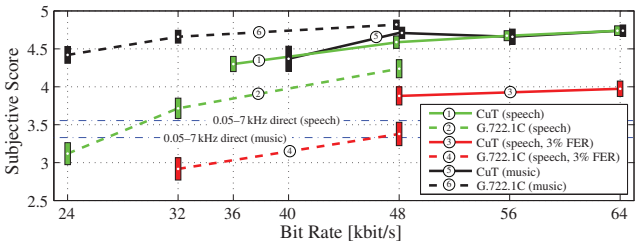

Fig. 3. Test results for Experiment 1a (mono clean speech, with and without frame erasures, conducted at Huawei listening lab) and Experiment 3a (mono music, conducted at Dynastat listening lab).

5. CHARACTERIZATION AND TEST RESULTS

This section characterizes the candidate codec in terms of computational complexity, algorithmic delay and subjective quality.

5.1. Computational Complexity and Algorithmic Delay

During the codec design—the implementation was done in floating point arithmetic—complexity considerations have been very important, e.g. critical subsampling, GLCVQ or parametric signal analysis. For mono input at 32 kHz sampling rate, this has led to a solution with a worst-case complexity of only 14.6 weighted million floating point operations per second (excluding the G.729.1 complexity)[1]. Owing to the IIR QMF filter bank with phase equalization [7], the algorithmic delay of G.729.1 (48.9375 ms) is only increased by 2.1875 ms which is considerably lower than for a competitive FIR QMF solution. If, in addition, the G.729.1 FIR QMF bank is replaced by the IIR solution, the total delay increase is merely 0.59 ms.

5.2. Subjective Listening Test

Within the ITU-T qualification phase, extensive subjective listening tests have been conducted in order to compare the proposed "codec under test" (CuT) with the existing ITU-T super-wideband codec G.722.1 Annex C [11, 12]. The applied test methodology was the so called "triple stimulus/hidden reference/double blind" test method, in short "Ref-A-B" [13]. In this method, the "reference" is the unprocessed signal and the samples A and B are, in random order, the test sample and again the (hidden) reference. "Ref-A-B" is designed as an "expert listener" method. Therefore, this test is able to assess the "listening expertise" of the subjects by evaluating their ability to identify the hidden reference. Hence, unreliable votes can be excluded from the evaluation. Finally, the votes of 24 subjects, ranging from 1.0 (very annoying impairment) to 5.0 (imperceptible impairment), have been accepted and used for each test condition. For the mono codec, the following conditions have been tested:

- Clean speech with and without frame erasures (Experiment 1a)
- Noisy speech (Experiment 2a)
- Music (Experiment 3a).

In these experiments, the proposed codec passed all requirements as defined in the "Terms of Reference" [6]. Concrete results for clean speech and music are presented in Fig. 3 where mean scores and the associated 95% confidence intervals are shown. For clean speech, the proposed coder is clearly better than G.722.1C at comparable bit rates. This also holds for frame erasure conditions. For music at 48 kbit/s, it is not worse than G.722.1C at the same bit rate. Yet, for higher bit rates, there is apparently a quality saturation and it can be reasoned that the bit budget set aside for wideband enhancement (Sec. 3.5) is still too limited. An integrated bit allocation both for 7–14 kHz (MDCT coefficients) and 0–7 kHz (G.729.1 TDAC error) based on a joint bit budget promises further improvement. Still, in contrast to G.722.1C, the proposed codec can offer a stable quality for varying sources, i.e., for both speech and music. In Exp. 2a, this finding could also be confirmed for *noisy speech* (mean subjective score of CuT@48 kbit/s with office noise: 4.76, music noise: 4.54).

[1] Taking into account an estimated "floating point to fixed point conversion penalty" (factor 1.2), the estimated *additional fixed point* complexity is 17.5 WMOPS. The fixed point complexity of the G.729.1 core codec is 36 WMOPS. WMOPS: Weighted million operations per second

6. CONCLUSIONS

This paper has introduced the Huawei/ETRI candidate for ITU-T super-wideband speech and audio coding. The presented results for mono input signals confirm that the proposed codec can offer high performance with low computational complexity. In particular the low bit rates of 36 and 40 kbit/s (L1+L2a/L2b) provide compact but comprehensive information to synthesize additional frequency content. The developed techniques also show potential for an application to other audio coding tasks.

However, the stereo test results for both ITU-T candidate codecs indicate that a high-quality and robust solution for 2-channel encoding under the given constraints [6] remains an item for further study.

7. REFERENCES

[1] B. Geiser, S. Ragot, and H. Taddei, "Embedded Speech Coding: From G.711 to G.729.1," in *Advances in Digital Speech Transmission*, R. Martin, U. Heute, and C. Antweiler, Eds. Chichester, UK: John Wiley and Sons, Ltd., Jan. 2008, ch. 8, pp. 201–247.

[2] ITU-T Rec. G.729.1, "G.729 based embedded variable bit-rate coder: An 8-32 kbit/s scalable wideband coder bitstream interoperable with G.729," 2006.

[3] S. Ragot *et al.*, "ITU-T G.729.1: An 8-32 kbit/s scalable coder interoperable with G.729 for wideband telephony and Voice over IP," in *Proc. of ICASSP*, Honolulu, Hawai'i, USA, Apr. 2007.

[4] ITU-T Rec. G.718, "Frame error robust narrowband and wideband embedded variable bit-rate coding of speech and audio from 8-32 kbit/s," 2008.

[5] T. Vaillancourt *et al.*, "ITU-T EV-VBR: A robust 8-32 kbit/s scalable coder for error prone telecommunications channels," in *Proc. of EUSIPCO*, Lausanne, Switzerland, Aug. 2008.

[6] "Terms of reference and time schedule for the joint qualification phase of SWB extension of G.EV-VBR and G.729.1," ITU-T SG16 temp. doc. Q23/16 TD358-WP3, Annex Q23B and Annex Q23C, Apr. 2008.

[7] H. W. Löllmann and P. Vary, "Design of IIR QMF Banks with Near-Perfect Reconstruction and Low Complexity," in *Proc. of ICASSP*, Las Vegas, NV, USA, Mar. 2008, pp. 3521–3524.

[8] L. Daudet and M. Sandler, "MDCT analysis of sinusoids: exact results and applications to coding artifacts reduction," *IEEE Transactions on Speech and Audio Processing*, vol. 12, no. 3, pp. 302–312, May 2004.

[9] H. Krüger, R. Schreiber, B. Geiser, and P. Vary, "On Logarithmic Spherical Vector Quantization," in *Proc. of International Symposium on Information Theory and its Applications (ISITA)*, Auckland, New Zealand, Dec. 2008.

[10] J. P. Adoul, C. Lamblin, and A. LeGuyader, "Baseband speech coding at 2400 bps using spherical vector quantization," in *Proc. of ICASSP*, San Diego, CA, USA, Mar. 1984.

[11] ITU-T Rec. G.722.1, "Low complexity coding at 24 and 32 kbit/s for hands-free operation in systems with low frame loss," 2005.

[12] M. Xie, D. Lindbergh, and P. Chu, "ITU-T G.722.1 Annex C: A new low-complexity 14 kHz audio coding standard," in *Proc. of ICASSP*, Toulouse, France, May 2006.

[13] ITU-R Rec. BS.1285, "Pre-selection methods for the subjective assessment of small impairments in audio systems," 1997.

Acknowledgment — The authors would like to thank Hervé Taddei from Huawei and the colleagues from ETRI for their contributions to this work.