A Qualified ITU-T G.729EV Codec Candidate for Hierarchical Speech and Audio Coding

Bernd Geiser, Peter Jax*, Peter Vary, Hervé Taddei†, Martin Gartner†, and Stefan Schandl‡

Institute of Communication Systems and Data Processing (ind)
RWTH Aachen University, Germany
{geiser|jax|vary}@ind.rwth-aachen.de

†Siemens AG, Munich, Germany
{herve.taddei|martin.gartner}@siemens.com

‡Siemens AG, Vienna, Austria
stefan.schandl@siemens.com

***Abstract*— We present an *embedded* and *hierarchical* 8–32 kbit/s speech and audio coding algorithm that has been successfully submitted to the ITU-T as a candidate [1] for ITU-T Rec. G.729.1 [2] (ex G.729EV). The submitting consortium consisted of *Siemens AG, Matsushita Electric Industrial Co., Ltd.,* and *Mindspeed Technologies, Inc.* This contribution gives a comprehensive overview of the proposed codec, describes the implemented algorithms, and states a detailed characterization as well as results of the official G.729EV qualification tests.**

I. Introduction

With the steadily increasing number of *Voice-over-IP* (VoIP) customers, a trend of the telecommunication *network infrastructure* towards packet switched techniques can be observed. The available bit-rates allow the transmission of toll to high quality *wideband* (50–7000 Hz) speech and audio signals.

However, the crucial factor for the large-scale deployment of wideband coding techniques in packet switched networks will be the interoperability with existing standards as well as with already installed infrastructure. This interoperability can be guaranteed by implementing a high quality speech and audio coding algorithm "on top" of widely deployed "legacy" *narrowband* (300–3400 Hz) standards. This approach is commonly known as *embedded* coding. A gateway to the "legacy" part of the network can simply discard the additional "high quality" bits and retain the bits related to the "legacy" codec. Thus, the bitstream interoperability is ensured.

Furthermore, apart from *bandwidth scalability*, *bit-rate scalability* is a desirable feature for future VoIP infrastructure, especially in highly heterogeneous networks. This bit-rate scalability can be achieved — in contrast to existing multi-mode coding standards like [3] (AMR-Wideband) — by a *hierarchical* bitstream concept, i.e., a "*layered*" bitstream format. Each additional layer successively improves the audio fidelity at the receiving terminal. The bit-rate scalability is obtained by appropriately *truncating the bitstream*, an operation that only introduces negligible complexity.

The hierarchical coding approach opens up a wide range of new applications. A few examples are given here: With a hierarchical codec, the network operator may, if desired, reduce the network load at every single node by reducing the transmitted bit-rate. Furthermore, serving users with different connections and/or terminal equipments within a telephone or a video conference becomes possible without major transcoding overhead. Some users will only understand the core bitstream of the hierarchical codec but others can decode a signal of higher quality. If the codec is designed such that the computational complexity decreases with a decrease of the bit-rate, hierarchical coding could become interesting for saving battery life in mobile devices. Finally, for storage applications, users could, for example, listen to their voice mail box from different kinds of terminals and still receive messages with the best quality possible.

An *embedded* and *hierarchical* codec as described above is in the focus of the ITU-T Q10/16 standardization body under the acronym *G.729EV* (EV stands for "embedded variable bit-rate"). The embedded concept is realized based on a G.729 CS-ACELP [4] compatible codec acting as the *core layer* in a hierarchical framework. Several additional bitstream layers gracefully increase the speech or audio quality.

The remainder of this paper introduces the codec candidate of *Siemens AG, Matsushita Electric Industrial Co., Ltd.,* and *Mindspeed Technologies, Inc.* which has been submitted to the ITU-T for qualification (cf. [1]). Therefore, Sec. II summarizes the "Terms of Reference," i.e., the official *requirements and objectives* for the G.729EV competition. Then, Sec. III gives a comprehensive overview of the submitted codec, whereas Sec. IV goes into algorithmic details. A characterization including the quality evaluation in terms of official ITU-T qualification test results conclude the paper.

II. Terms of Reference

This section briefly lists the ITU-T requirements which applied for the development of G.729EV proposals.

The digitally sampled wideband input signal must be segmented into frames of 20 ms length. These frames are then fed into the encoder which should produce a bitstream with a bit-rate of 32 kbit/s. This bitstream must be hierarchically organized with a 8 kbit/s core layer that is completely interoperable with the ITU-T G.729 recommendation [4]. Wideband capability is targeted for bit-rates of at least 14 kbit/s. Over the whole bit-rate range of 8–32 kbit/s, the quality should increase

*Peter Jax is now with Thomson Corporate Research, Hannover, Germany.

TABLE I

G.729EV REQUIREMENTS AND OBJECTIVES FOR CLEAN SPEECH AND ERROR-FREE CONDITIONS ($\not<$: NOT WORSE THAN, $>$: BETTER THAN).

bit-rate [kbit/s]	Requirement	Objective
8	$\not<$ G.729A	$>$ G.729A
12	$\not<$ G.729E	$>$ G.729E
14	$\not<$ G.722.2@8.85 kbit/s	$\not<$ G722.2@12.65 kbit/s
24	$\not<$ G.722@48 kbit/s	$>$ G.722@48 kbit/s
32	$\not<$ G.722@56 kbit/s	$>$ G.722@56 kbit/s

gracefully. An excerpt from the important quality requirements is shown in Tab. I. The algorithmic delay of the codec is required to stay below 60 ms (the objective is 45 ms). The computational complexity should be low to moderate, i.e., a maximum of 40 WMOPS[1] is acceptable.

III. OVERVIEW

Here the *encoder* and *decoder structure* of our codec candidate is introduced. Algorithmic details are given in Sec. IV.

A. Encoder

Our candidate encoder is shown in Fig. 1. The input signal $s(k)$, sampled at $f_s = 16$ kHz, is fed in parallel into three signal flow branches.

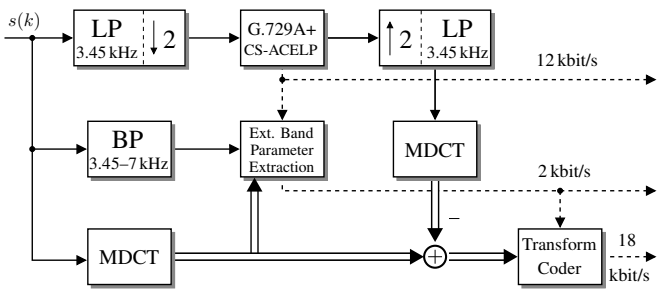

Fig. 1. Encoder Overview — Single solid lines: Time domain signal flow, double solid lines: MDCT domain signal flow, dashed lines: Parameters

The first branch encodes the narrowband part of the signal. This is done by a slightly modified G.729A [6] encoder and by an enhancement stage (here labeled "G.729A+"). Refer to Sections IV-A and IV-B for details. The output is a 8 kbit/s bitstream that is interoperable with the G.729 recommendation plus additional bits which allow for a certain quality enhancement as will be shown in the test results. In total, the sent bit-rate of this codec branch is $(8 + 4)$ kbit/s $= 12$ kbit/s.

The second branch takes care of encoding the *extension band* (EB) frequencies between 3.45 kHz and 7 kHz by means of a rather coarse parametric description. Sec. IV-C explains which parameters are extracted. The bit-rate for the parametric description is 2 kbit/s.

In the third and final encoder branch, a *transform coder* is implemented and the processing is done in the MDCT domain [7], see Sec. IV-D for details. To reach the required 32 kbit/s, up to 18 kbit/s are produced.

[1] WMOPS: Weighted Mega Operations per Second, a measure according to the ITU-T fixed point arithmetics library from recommendation G.191 [5]

B. Decoder

Since the encoder produces a hierarchical bitstream, the decoder operation, as illustrated in Fig. 2, depends on the received bit-rate.

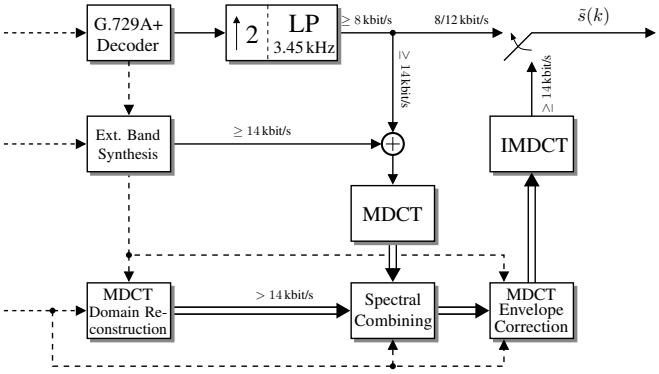

Fig. 2. Decoder Overview — The signal flow branches are labeled with the bit-rates at which they are active. — Single solid lines: Time domain signal flow, double solid lines: MDCT domain signal flow, dashed lines: Parameters

At rates of 8 and 12 kbit/s, only the top decoder branch is active. It includes the G.729A+ decoder.

At a received rate of 14 kbit/s, additionally the "Extension Band Synthesis" block is active which regenerates the respective frequency components. The output of this module is added to the G.729A+ output in order to obtain a wideband signal.

For bit-rates higher than 14 kbit/s, the intermediate wideband output signal is refined in the MDCT domain using the received MDCT envelope and coefficients.

MDCT domain post-processing ("envelope correction") is done for all bit-rates between 14 and 32 kbit/s.

IV. ALGORITHMIC DESCRIPTION

This section describes the algorithms which are used in the most important functional units of the proposed codec.

Sec. IV-A starts with the description of the codec's core component: The embedded G.729A compatible codec. Sections IV-B, IV-C, and IV-D are dedicated to the *narrowband enhancement layer*, the *wideband extension layer*, and the *refinement layers* respectively. The algorithmic description concludes with a short discussion of the implemented *pre-echo reduction scheme* (Sec. IV-E), the *post-processing* algorithms (Sec. IV-F), and the *treatment of bandwidth switchings* (Sec. IV-G). The latter is essential because of the inherent *bandwidth scalability* of the whole codec.

A. Core Layer: Enhanced G.729A CS-ACELP

The core layer of our hierarchical codec produces a narrowband signal $s_{\mathrm{nb}}(k')$ with a cut-off frequency of $f_c = 3.45$ kHz. The respective processing can thus be carried out with a reduced sampling frequency of $f'_s = 8$ kHz. Low-pass filtering and decimation is applied before entering this layer.

The core codec is fully interoperable with the CS-ACELP decoder from the ITU-T G.729A recommendation [6]. However, some modifications have been issued to increase the quality. In

particular, the fixed codebook (FCB) search is orthogonalized to the adaptive codebook (ACB) vector, i.e., the FCB search is performed such that the filtered FCB vectors are orthogonal to the filtered ACB vectors (cf. [8]). Thus, the ACB and FCB vectors are jointly optimized.

B. Narrowband Enhancement Layer: Cascade CELP

The 12 kbit/s "narrowband enhancement layer" is also based on ACELP techniques. It implements an additional FCB search procedure based on the 17 bit FCB of the G.729A codec.

Therefore, the encoding error of the 8 kbit/s layer in the residual domain is chosen as the new target signal. Within the additional FCB search procedure, the algebraic pulses are convoluted with a *dispersion pattern* that has been obtained through offline training [9].

The gain factor for this additional codebook is quantized with 3 bits using a predictive scalar quantizer. Together with the 17 bits from the additional FCB search, this yields the desired bit-rate for this layer: $20\,\text{bit}/5\,\text{ms} = 4\,\text{kbit/s}$.

After signal synthesis in the (local) decoder, upsampling to 16 kHz is performed in order to combine $s_{\text{nb}}(k')$ with the extension band components $s_{\text{eb}}(k)$.

C. Wideband Layer: Time Domain Bandwidth Extension

The "wideband layer" is available for bit-rates of 14 kbit/s and above. It is responsible for the encoding of the "extension band (EB) frequencies" such that, together with the 12 kbit/s narrowband part, a wideband speech signal of good quality can be produced at the decoder. The respective technique has been labeled "Time Domain Bandwidth Extension" (TD-BWE).

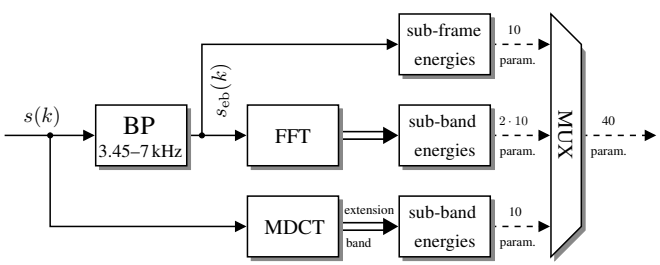

Fig. 3. Parameter Extraction for the Extension Band.

The TD-BWE encoder isolates the EB signal components of $s(k)$ by band-pass filtering and performs a parameter extraction. The parameter set comprises a "time envelope" and *two* "frequency envelopes" of the EB signal components as shown in Fig. 3. The sub-frame energy is computed every 2 ms, the 10 FFT domain sub-band energies every 10 ms, and the 10 MDCT domain sub-band energies for the EB frequencies every 20 ms. In total this yields 40 parameters per frame. The determined time and frequency envelopes are jointly quantized and encoded. The parameter quantization with a rate of 2 kbit/s is done using split vector quantization in a transformed domain, cf. [10].

The TD-BWE decoder synthesizes the EB frequencies. Therefore, first, an artificial EB "excitation signal" $s_{\text{exc}}(k)$ with a consistent pitch structure is produced based on the parameters of the 8–12 kbit/s layers. Then, the time and frequency envelopes of $s_{\text{exc}}(k)$ are consecutively shaped by gain manipulations and filtering operations to match the transmitted parametric description.

An in-depth description of the implemented algorithms has been presented in [10] and [11]. In the former, the application to the GSM Enhanced Fullrate Codec [12] has been investigated, whereas the latter discusses the application to the G.729A+ codec from Sec. IV-A and IV-B in more detail.

D. Enhancement Layers: Transform Coding

With the output of the "wideband layer" described in the preceding section, a wideband signal of good quality is already available at the decoder. Now the task of the "enhancement layers" of the G.729EV candidate is a gradual quality improvement when the received bit-rate increases. For the highest bit-rate (32 kbit/s) general audio capability is targeted. In order to meet the requirements w.r.t. the granularity of the hierarchical bitstream, a transform coding technique proved to be suitable and thus a "classical" MDCT domain encoding (e.g., [13]) with a window length of 40 ms has been chosen.

Here the MDCT domain difference signal $S(\mu) - S_{\text{nb}}(\mu)$, where $S_{\text{nb}}(\mu)$ is the transformed G.729A+ output, is subdivided into 20 sub-bands of equal bandwidth (cf. [14]). The *MDCT envelope* (sub-band energies) is now encoded using *entropy coding*. Thereby, the vector quantized MDCT sub-band energies from the TD-BWE module (Sec. IV-C) are used as a prediction for the extension band MDCT envelope. Thus, by means of the TD-BWE vector quantization, a good rate-distortion trade-off is achieved for "typical" EB envelopes, whereas remaining outliers are "caught" by the subsequent entropy coding of the residual quantization error. This procedure yields a "refined" MDCT envelope for the extension band at bit-rates higher than 14 kbit/s.

The quantized MDCT envelope is used to normalize the spectral coefficients. The normalized spectral coefficients are quantized with a *spherical vector quantizer* whose *bit-allocation* is controlled by an energy based criterion. Since the sub-bands have equal bandwidth, i.e., the perceptually relevant *bark scale* is not considered, a weighting of the sub-band energies according to the number of bark bands per MDCT sub-band (bark/sub-band) is applied first. In total, a maximum bit-rate of 18 kbit/s is produced by the transform coder.

The decoder side reconstructs the MDCT coefficients in the sub-bands which could be decoded at the received bit-rate. For MDCT sub-bands in the narrowband frequency range, the decoded and denormalized coefficients are added to the MDCT transform of the G.729A+ output. In the extension band, the TD-BWE output is replaced band-wise.

Finally, the decoder implements an *envelope correction* for sub-bands which are produced by the TD-BWE decoder but were not allocated any bits by the transform coder. Here a smoothed version of the refined extension band envelope is applied before the inverse MDCT is performed.

E. Pre-Echo Reduction

Pre-echoes may be produced by the enhancement layers (Sec. IV-D) due to quantization effects and the large block-size (40 ms). If applicable, the implemented pre-echo reduction takes advantage of the high time resolution of the energy envelopes in the CELP and TD-BWE layers. Thus, pre-echo artifacts which stem from the large block size are reduced noticeably. Moreover, no additional information needs to be transmitted for this scheme.

F. Post-Processing

The narrowband layers at 8 kbit/s and 12 kbit/s use time-domain short-term and long-term post-filters, similar to traditional CELP-type codecs, for perceptual quality improvements of the decoded signals.

For the wideband layers from 14 kbit/s up to 32 kbit/s, the MDCT domain spectral coefficients are post-processed to achieve a similar effect.

G. Treatment of Bandwidth Switchings

Hierarchical coding can nicely benefit from "Unequal Error Protection " (UEP). This means that the more important layers (e.g., the core layer) could be better protected and thus less subjected to packet losses than other layers.

Although a stable communication link can be guaranteed with such an UEP scheme, packet losses may result in very fast switching of the decoded signal's bandwidth between wideband and narrowband. This leads to a major loss of subjective speech quality.

The G.729EV candidate implements an algorithm to tackle such bandwidth switchings. Its goal is to produce a signal with constant bandwidth despite any short-term bit-rate variation of the decoded bitstream (e.g., between 12 kbit/s and 14 kbit/s). After a drop of the decoded bandwidth from wideband to narrowband has been detected, the bandwidth rendering stays as large as 7 kHz for a time of $T_{\text{hold}} = 100$ ms. This value is chosen to be larger than the expected length of packet loss bursts. If no consecutive wideband frames can be decoded within T_{hold}, the output is switched (or gracefully faded) to narrowband. In analogy, it is beneficial to let the desired bandwidth rise with a certain gradient in case of an increase of the decoded bandwidth.

In order to produce the output signal with the desired bandwidth of 7 kHz, *bandwidth extension techniques* (e.g., [15]) are used. That is, the missing frequency range is filled by synthetically generated signal components that are estimated by exploiting information from the decoded narrowband signal and possibly from adjacent (previously decoded) wideband signal frames. Therefore, a simple (low-complexity) estimator for the time and frequency envelope parameters of the TD-BWE codec layer (Sec. IV-C) has been designed. The estimation especially benefits from previously decoded (and thus perfectly known) wideband frames. The described scheme yields a significantly better performance than, e.g., simple envelope repetition.

V. Characterization

The proposed speech and audio coding algorithm has been characterized and tested by ITU-T SG-16 within the G.729EV qualification test phase. [1] and [16]–[21] document the characterization and the respective test results, which shall be summarized here for our codec candidate: "CuT D."

A. Bit Allocation

The bit allocation per 20 ms speech frame of our codec is presented in Table II.

TABLE II
Bit Allocation per 20 ms Frame for the G.729EV Candidate

Parameter	1st Sub-Frame (10 ms)	2nd Sub-Frame (10 ms)	Total (20 ms)
G.729 Core Layer			
Line Spectrum Pairs	18	18	36
ACB Delay	$8+5$	$8+5$	26
Pitch-Delay Parity	1	1	2
FCB Index	$13+13$	$13+13$	52
FCB Sign	$4+4$	$4+4$	16
CB Gains (Stage 1)	$3+3$	$3+3$	12
CB Gains (Stage 2)	$4+4$	$4+4$	16
Σ	$40+40$	$40+40$	**160**
NB Enh. Layer			
FCB Index	$13+13$	$13+13$	52
FCB Sign	$4+4$	$4+4$	16
FCB Gains	$3+3$	$3+3$	12
Σ	$20+20$	$20+20$	**80**
WB Layer			
TD-BWE VQ	—		**40**
Enh. Layers			
Envelope Quant.	Entropy coding, varying rate		R
MDCT Coeff. Quant.	—		$360-R$
Σ	—		**360**
Σ			**640**

B. Algorithmic Delay

The total algorithmic codec delay of our G.729EV candidate is 48.75 ms. This splits into the following contributions: Framing (20 ms), filter delays ($3 \cdot 1.25$ ms $= 3.75$ ms), G.729A look-ahead (5 ms), and MDCT windowing with look-ahead (20 ms).

C. Algorithmic Complexity Estimation

The algorithmic complexity (according to [5]) for the main modules of our codec candidate is estimated as follows:

- 17.8 WMOPS for the ACELP core layer (Sec. IV-A) and narrowband enhancement layer (Sec. IV-B),
- 4.5 WMOPS for the wideband layer (Sec. IV-C), and
- 12 WMOPS for the enhancement layers (Sec. IV-D).

Allowing for a good complexity margin for the remaining modules and the glue code, the total complexity is estimated well below 40 WMOPS.

D. ITU-T Qualification Test Results

Fig. 4 (partially) visualizes the subjective listening test results of the ITU-T G.729EV qualification test for our codec candidate. The test results are obtained for clean speech and error-free transmission conditions at input levels of -16, -26 and -36 dBov ("dBov" is defined in [5]). The tested bit-rates are 8, 12, 14, 24, and 32 kbit/s. The results are given as subjective

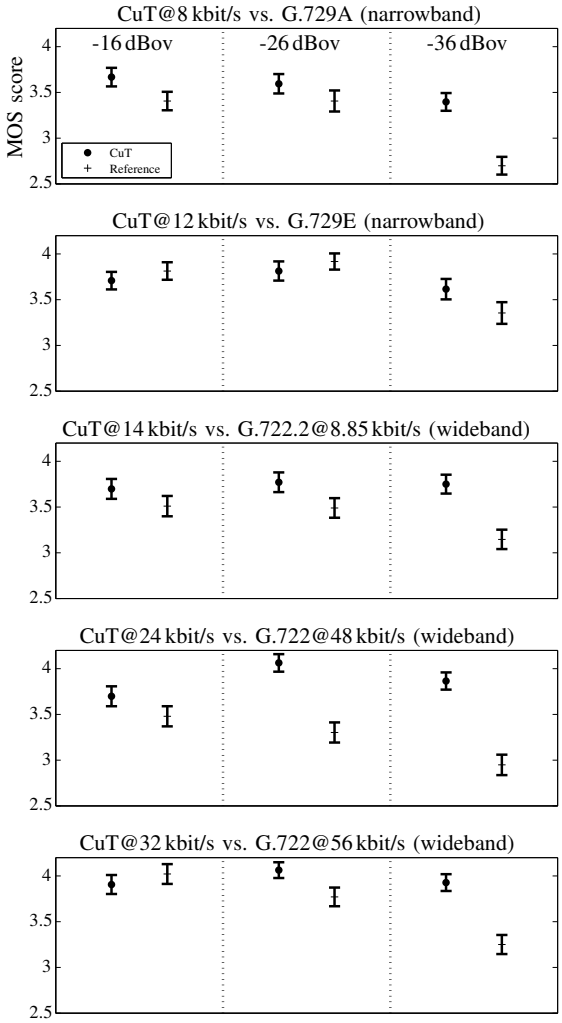

Fig. 4. Excerpt from the ITU-T qualification test results of the G.729EV proposal (CuT D) vs. the respective requirements (cf. Tab. I) for clean speech at various input speech levels and various bit-rates.

MOS scores ($\mathrm{MOS}_{\mathrm{CuT}}$ and $\mathrm{MOS}_{\mathrm{ref}}$) associated with their respective 95% confidence intervals $2\Delta_{\mathrm{CuT}}$ and $2\Delta_{\mathrm{ref}}$. The "not worse than"-*requirements* according to Tab. I are fulfilled if the following relation holds:

$$\mathrm{MOS}_{\mathrm{CuT}} + \Delta_{\mathrm{CuT}} \geq \mathrm{MOS}_{\mathrm{ref}} - \Delta_{\mathrm{ref}}$$

For the cases shown in Fig. 4, the requirements are fulfilled. This is also true for the G.729EV requirements not listed here. It can be observed that the speech quality increases gracefully with an increasing bit-rate and that the obtained quality is relatively independent from the input speech level.

VI. Conclusion

A G.729EV codec candidate that has been jointly developed by Siemens AG, Matsushita Electric Industrial Co., Ltd., and Mindspeed Technologies, Inc. has been presented in this paper. The respective characterization and the ITU-T listening tests certify full compliance of our algorithm with the "Terms of Reference" (Sec. II). All of the G.729EV *requirements* are met. Additionally, 17 out of 25 tested *objectives* are accomplished.

Acknowledgment

The authors would like to thank *Hiroyuki Ehara* from Matsushita Electric Industrial Co., Ltd. as well as *Eyal Shlomot* and *Yang Gao* from Mindspeed Technologies, Inc. for their invaluable contributions to this work.

References

[1] Siemens AG, "High level description of the scalable 8-32 kbit/s algorithm submitted to the qualification test by Matsushita, Mindspeed and Siemens," ITU-T SG16 delayed contribution T05-D214, July 2005.

[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] 3GPP TS 26.190, "AMR wideband speech codec; transcoding functions," Dec. 2001.

[4] ITU-T Rec. G.729, "Coding of speech at 8 kbit/s using conjugate-structure algebraic-code-excited linear-prediction (CS-ACELP)," 1996.

[5] ITU-T Rec. G.191, "Software tools for speech and audio coding standardization," 2005.

[6] ITU-T Rec. G.729 Annex A, "Coding of speech at 8 kbit/s using conjugate-structure algebraic-code-excited linear-prediction (CS-ACELP). Annex A: Reduced complexity 8 kbit/s CS-ACELP speech codec," 1996.

[7] H. S. Malvar, *Signal Processing with Lapped Transforms*. Norwood: Artech House, 1992.

[8] M. Johnson and T. Taniguchi, "Pitch-orthogonal code-excited LPC," in *Proc. of GLOBECOM*, vol. 1, Dec. 1990, pp. 542–546.

[9] K. Yasunaga, H. Ehara, K. Yoshida, and T. Morii, "Dispersed-pulse codebook and its application to a 4 kb/s speech coder," in *Proc. of ICASSP*, vol. 3, Istanbul, Turkey, June 2000, pp. 1503–1506.

[10] P. Jax, B. Geiser, S. Schandl, H. Taddei, and P. Vary, "An embedded scalable wideband codec based on the GSM EFR codec," in *Proc. of ICASSP*, Toulouse, France, May 2006.

[11] ——, "A scalable wideband "add-on" for the G.729 speech codec," in *ITG-Fachtagung "Sprachkommunikation"*, Kiel, Germany, Apr. 2006.

[12] ETSI Rec. GSM 06.60, "Enhanced full rate (EFR) speech transcoding," version 8.0.1, release 1999, Nov. 2000.

[13] T. Painter and A. Spanias, "Perceptual coding of digital audio," *Proceedings of the IEEE*, vol. 88, no. 4, pp. 451–513, Apr. 2000.

[14] H. Taddei, D. Massaloux, and A. Le Guyader, "A scalable three bitrate (8, 14.2, and 24 kbit/s) audio coder," in *107th Convention of the Audio Engineering Society (AES)*, New York, NY, USA, Sept. 1999.

[15] P. Jax, "Bandwidth extension for speech," in *Audio Bandwidth Extension*, E. Larsen and R. M. Aarts, Eds. Wiley and Sons, Nov. 2004, ch. 6, pp. 171–236.

[16] "G.729EV qualification phase test results: subjective and objective (WB-PESQ) scores for experiment 1b wideband conditions," ITU-T SG16 temporary document Q10/16 TD68-WP3, July 2005.

[17] "G.729EV qualification phase test results: Experiment 5 (clean speech; wide band case, bit rate granularity)," ITU-T SG16 temporary document Q10/16 TD69-WP3, July 2005.

[18] "Qualification phase of G729EV: test results (exp 1-4)," ITU-T SG16 temporary document Q10/16 TD71-WP3, July 2005.

[19] "G.729EV qualification phase test results: Experiment 6 (clean speech; bit rate switching case)," ITU-T SG16 temporary document Q10/16 TD73-WP3, July 2005.

[20] "G.729EV qualification phase test results: Complexity evaluation," ITU-T SG16 temporary document Q10/16 TD76-WP3, July 2005.

[21] "G.729EV qualification phase test results: Frequency responses of G.729EV candidates," ITU-T SG16 temporary document Q10/16 TD81-WP3, July 2005.