Superwideband Extension of G.718 and G.729.1 Speech Codecs

Lasse Laaksonen\textsuperscript{1}, Mikko Tammi\textsuperscript{1}, Vladimir Malenovsky\textsuperscript{2}, Tommy Vaillancourt\textsuperscript{2}, Mi Suk Lee\textsuperscript{3}, Tomofumi Yamanashi\textsuperscript{4}, Masahiro Oshikiri\textsuperscript{4}, Claude Lamblin\textsuperscript{5}, Balazs Kovacs\textsuperscript{5}, Lei Miao\textsuperscript{6}, Deming Zhang\textsuperscript{6}, Jon Gibbs\textsuperscript{7}, Holly Francois\textsuperscript{8}

\textsuperscript{1}Nokia, Finland \hspace{1cm} \textsuperscript{2}VoiceAge, Canada \hspace{1cm} \textsuperscript{3}ETRI, Korea \hspace{1cm} \textsuperscript{4}Panasonic Corp., Japan \hspace{1cm} \textsuperscript{5}France Telecom, France \hspace{1cm} \textsuperscript{6}Huawei Technologies, China \hspace{1cm} \textsuperscript{7}Motorola, United Kingdom

Abstract

This communication presents the recently standardized superwideband (SWB) extensions of ITU-T G.718 and G.729.1. These extensions were standardized as G.718 annex B and G.729.1 annex E. The SWB functionality is implemented using embedded scalable layers on top of the wideband (WB) core codecs, and it extends the bit rate of the codecs to 48 and 64 kbit/s for the G.718 and G.729.1, respectively. The main technology is a two-mode SWB coding method of the high frequencies. In addition, the G.729.1 SWB extension enhances the lower frequency range. The codec performance is illustrated with some listening test results extracted from the ITU-T Characterization phase.

Index Terms: speech coding, audio coding, embedded coding, scalable coding, superwideband, bandwidth extension, ITU

1. Introduction

ITU-T G.718 \textsuperscript{[1]} and G.729.1 \textsuperscript{[2]} are two fairly recent speech and audio coding standards. They are both embedded scalable wideband (WB) codecs that employ transform coding based on the Modified Discrete Cosine Transform (MDCT) on top of a Code-Excited Linear Prediction (CELP) coding. The highest bitrate for these codecs is 32 kbit/s and lower operating points can be obtained by truncation of the bitstream. G.718 and G.729.1 have 5 and 12 bit rates, respectively.

An ITU-T activity was initially launched to develop a joint superwideband (SWB) and stereo extension for G.718 and G.729.1. This work later evolved into two successive phases of SWB mono and WB/SWB stereo development. The SWB phase has recently been concluded.

This paper presents the SWB mono extension for G.718 and G.729.1 as standardized by ITU-T. It is based on a joint effort from Ericsson, ETRI, France Telecom, Huawei Technologies, Motorola, Nokia, Panasonic, Texas Instruments, and VoiceAge.

The paper is organized as follows. In Section 2, we present a general description of the SWB extension. In Sections 3 and 4, the encoder and decoder functionalities are described. Selected test results are discussed in Section 5.

2. General description of the extension

The SWB [50 – 14000 Hz] extensions of G.718 and G.729.1 share most of their key technologies. The main difference between the two extensions is that the G.729.1 SWB includes WB improvements which are not included in G.718 extension. This difference is shown in Table 1, which presents the codec bit rates and main technologies in each embedded extension layer. The layers are denoted as 6mo through 10mo, where ‘mo’ refers to SWB mono operation.

The G.718 SWB extension comprises 3 layers extending the codec to total bit rates of 36, 40, and 48 kbit/s. An option to omit WB Layer 5 of G.718 lowers the operational bit rates to 28, 32 and 40 kbit/s. This option is also available for the AMR-WB interoperable core at 12.65 kbit/s of G.718 thus providing AMR-WB interoperable SWB speech coding starting from 28 kbit/s.

The G.729.1 SWB extension consists of 5 layers with total bit rates of 36, 40, 48, 56, and 64 kbit/s. SWB Layers 9mo and 10mo enhance the lower frequency (LF) content (below 7000 Hz). A further difference between the two extensions is introduced in Layer 7mo, which in the case of G.729.1 implements an adaptive bit allocation between SWB coding and WB improvements.

Table 1. Main technologies of G.718/G.729.1 SWB.

<table>
<thead>
<tr>
<th>Layer</th>
<th>Bit rate</th>
<th>G.718</th>
<th>G.729.1</th>
</tr>
</thead>
<tbody>
<tr>
<td>6mo</td>
<td>4 kbit/s</td>
<td>Generic / Sinusoidal</td>
<td>Generic / Sinusoidal</td>
</tr>
<tr>
<td>7mo</td>
<td>4 kbit/s</td>
<td>Sinusoids</td>
<td>Sinusoids / WB impr.</td>
</tr>
<tr>
<td>8mo</td>
<td>8 kbit/s</td>
<td>Sinusoids</td>
<td>Sinusoids</td>
</tr>
<tr>
<td>9mo</td>
<td>8 kbit/s</td>
<td>-</td>
<td>WB improvement</td>
</tr>
<tr>
<td>10mo</td>
<td>8 kbit/s</td>
<td>-</td>
<td>WB improvement</td>
</tr>
</tbody>
</table>

For both cores, the extension accepts SWB signals sampled at 32 kHz. Signals sampled at 8 or 16 kHz bypass the SWB extension and are processed using the core codec only. The input signal is processed in frames of 20 ms each. The additional algorithmic delay of the SWB extensions is 6.75 ms and it is introduced by the sampling rate conversions and post-processing.

3. Encoder overview

Figure 1 presents a high-level block diagram of the SWB encoder. The high frequency (HF) encoding is performed in the MDCT domain for the frequency range of 7000 – 14000 Hz. For both core codecs, the codec exploits a two-mode approach \textsuperscript{[3]}: based on the outcome of tonality estimation either Generic mode coding or Sinusoidal mode coding is used in the first SWB layer. Higher layers add sinusoidal components, which further improve the quality of the HF.
content. In addition, the G.729.1 SWB extension improves the LF content.

3.1. Tonality estimation

The coding mode in the first SWB layer is defined based on the tonality of the input signal. Tonal signals are very periodic in nature and a special coding mode has been designed for them. The coding mode is selected using tonality estimation, performed in the MDCT domain by means of peak-by-peak correlation analysis between the current and previous log-energy spectra. The algorithm consists of four steps.

First, a spectral floor is calculated and subtracted from the log-energy MDCT spectrum (of length 280) of the high-frequency part by applying a moving-average (MA) filter. The spectrum is then smoothed in amplitude and compressed by two in frequency in the leading and trailing parts – for the central part only amplitude smoothing is applied. This results in a compressed spectrum with a length of 147 bins.

In the second step, local minima of the compressed spectrum are searched. The third step then consists of calculating a correlation map and a long-term correlation map, which are again peak-wise operations. Thus, for every peak (defined as a local maximum) in the current compressed spectrum, normalized correlation is calculated w.r.t. the previous compressed spectrum. The correlation operation considers all bins of the peak between two consecutive minima. The correlation map of the current frame is used to update the long-term correlation map, set initially to zero.

As a final step, all bins of the long-term map are summed together. The tonality decision is done by comparing this correlation sum to an adaptive threshold. If the sum is higher than the threshold, the frame is considered tonal, otherwise it is non-tonal. The adaptive threshold is decreased when the decision is correct and increased when the decision is incorrect. It thus acts as a hangover.

3.2. Generic mode coding

The Generic mode coding is used for all those frames, which are considered non-tonal. In this mode, the HF band is divided into four non-overlapping subbands and for each of them the most similar match is searched from the coded and envelope normalized low-frequency content. The most similar match is then scaled with two scaling factors to obtain the synthesized HF content.

First, the locally synthesized low-frequency content is normalized in logarithmic domain. This step is needed because the MDCT domain envelope of the LF content often evolves rapidly, which makes it difficult to find good matches for the HF subbands. In addition, there may be large differences in dynamics of the signal, which make the quantization of the scaling factors more difficult.

The subband search is performed for four subbands of widths 40, 70, 70, and 100 bins in the MDCT domain. For even subbands the search area is fixed, while for odd subbands the area adaptively varies depending on the best match position in the preceding subband.

The best match is searched in each subband \( j = 0, \ldots, 3 \) by maximizing the similarity measure of the form

\[
S_j(k) = \frac{\text{corr}(k)}{\text{ene}(k)},
\]

where \( \text{corr}(k) \) denotes the correlation at index \( k \) and \( \text{ene}(k) \) is the corresponding energy. Two scalings are then performed to obtain as good perceptual match to the original subband as possible. The first scaling operates in the linear domain and is used to match the high amplitude peaks of the spectrum. The scaling factor is given by

\[
a_l(j) = \frac{\text{corr}(k'_j)}{\text{ene}(k'_j)},
\]

where \( k'_j \) is the index of the best match for subband \( j \). The signs and absolute values of the four scaling factors \( a_l(j) \) are quantized separately.

The second scaling operates in the logarithmic domain and is used to provide better match with the energy and the logarithmic domain spectral shape. Thus, the logarithmic domain representations are needed both for the original and the so-far synthesized HF content. The second scaling factor is defined as

\[
a_s(j) = \frac{\text{corr}_\text{LOG}}{\text{ene}_\text{LOG}},
\]

where \( \text{corr}_\text{LOG} \) is the correlation calculated for the even samples of the logarithmic domain subband from which the maximum value of the said subband has been subtracted, and \( \text{ene}_\text{LOG} \) denotes the corresponding energy. Only the even samples are considered to reduce complexity. If \( \text{ene}_\text{LOG} \) is zero, \( a_s(j) \) is set to 1, and additionally only non-negative values are allowed. Both the first and second scaling factors are quantized using vector quantization (VQ). The bit allocation of the Generic mode coding is presented in Table 2.

In addition, two sinusoidal components are added to the spectrum in this Generic mode layer.

The next subsection describes the methods used for the coding of the sinusoidal components.
3.3. Sinusoidal mode coding

The Sinusoidal mode coding is used in frames that are classified as tonal. In this mode, the HF content is obtained by transmitting a finite set of sinusoidal components.

In Layer 6mo, 10 sinusoids are searched and added to the high frequencies. The first eight sinusoids are searched using a track structure defining all the allowed positions for each of the sinusoids. This makes the search efficient and reduces the required bit rate. The remaining two sinusoids are selected more freely and they cover the highest parts of the HF region.

Each sinusoid is characterized by its position, sign, and amplitude. The position of each sinusoid is selected based on the absolute difference between the original and synthesized HF content. In the first SWB layer the synthesized content is initially all zero, and the sinusoids are thus placed at original signal maxima as found within their allowed search ranges.

The sinusoid signs are transmitted separately. The amplitudes are quantized in the logarithmic domain using VQ. Table 3 presents the bit allocation for the Sinusoidal mode coding.

### Table 3. Bit allocation of Layer 6mo, Sinusoidal.

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Bits</th>
</tr>
</thead>
<tbody>
<tr>
<td>Extension type</td>
<td>1</td>
</tr>
<tr>
<td>Coding mode</td>
<td>1</td>
</tr>
<tr>
<td>Sinusoid positions</td>
<td>51</td>
</tr>
<tr>
<td>Sinusoid signs</td>
<td>6</td>
</tr>
<tr>
<td>Sinusoid amplitudes</td>
<td>21</td>
</tr>
</tbody>
</table>

3.4. Additional sinusoids

In Layers 7mo and 8mo, additional sinusoidal components are selected and added to the HF spectrum. For G.718, Layer 7mo adds 10 sinusoids and Layer 8mo 20 sinusoids. For G.729.1, the bit allocation of Layer 7mo is adaptive (see section 3.5) and results in transmitting 0, 2, 4, 6, 8, or 10 sinusoids depending on the available bit budget.

In Layer 7mo of G.718, the 10 added sinusoids are selected in 5 track structures of 2 sinusoids each. For Generic mode frames, the exact positions of the tracks are determined based on the energy of the synthesized HF content. For this purpose, the HF part is divided into 8 subbands of equal width. The five highest energy subbands are selected as the track positions. For Sinusoidal mode frames, the track positions are pre-determined such that they complement the coding of Layer 6mo.

In Layer 8mo, the 20 sinusoids are selected in predetermined track structures for both coding modes. However, the track positions differ between the modes to take into account the coding of the previous layers.

In the additional sinusoid layers, majority of the bit rate is used to transmit the positions of the sinusoids. This is illustrated in Table 4, which presents the Layer 7mo bit allocation for G.718 (and G.729.1 in case of 10 sinusoids). One bit in Layer 7mo is reserved for the Layer 4 option of G.718. In G.729.1 SWB this bit is not used. Layer 8mo follows the same principle as Layer 7mo, using twice the bit rate. In G.729.1, Layer 7mo may use 0, 19, 30, 49, 60, or 79 bits for additional sinusoids depending on the adaptive bit allocation.

### Table 4. Bit allocation of Layer 7mo (G.718).

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Bits</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sinusoid positions</td>
<td>50</td>
</tr>
<tr>
<td>Sinusoid signs</td>
<td>5</td>
</tr>
<tr>
<td>Sinusoid amplitudes</td>
<td>24</td>
</tr>
<tr>
<td>G.718 Layer 4 option</td>
<td>1</td>
</tr>
</tbody>
</table>

3.5. WB improvements

In addition to coding the HF part of the signal, the G.729.1 SWB extension improves the perceptual quality of the LF signal in Layers 7mo, 9mo and 10mo, where the last two are not available for G.718.

As already mentioned, in G.729.1, Layer 7mo implements an adaptive bit allocation that first considers the bits allocated to each low-frequency subband between 4000 – 7000 Hz by G.729.1 WB coding. For each of the subbands, 9 bits from a maximum of 80 bits are allocated when the corresponding core codec bit allocation for that subband is zero. These bits are then used to quantize the wideband MDCT coefficients according to the same process as in Layers 9mo and 10mo. The bits remaining after the adaptive allocation for wideband improvements are used to encode sinusoidal components. Layers 9mo and 10mo enhance the wideband signal by coding the difference between the Time Domain Aliasing Cancellation (TDAC) input and the locally synthesized G.729.1 output. This wideband MDCT error signal is divided into 18 subbands as in G.729.1: all subbands have 16 coefficients except the last one which comprises 8 coefficients.

First, the envelope of the wideband MDCT error envelope is coded. The envelope is defined as the root mean square of the sinusoids in logarithmic domain. Based on the spectral envelope of the G.729.1 coding output and the number of bits allocated during that step, a prediction is computed. The predicted spectral envelope is then subtracted from the wideband error envelope. The resulting envelope is further rounded and coded using a 2-bit/subband linear quantizer that consumes a total of 36 bits.

The subbands of the wideband error signal are ordered by decreasing perceptual importance computed in function of the quantized envelope. This ordering is the basis for bit allocation and multiplexing of VQ indices. As the perceptual importance can be determined also in the decoder, no side information is needed.

The quantization method of each subband depends on its dimension (the number of coefficients). If the dimension is 8, spherical vector quantization (SVQ) is used. This process is similar as in the G.729.1 standard. If the dimension is 16, Gosset low complexity vector quantization (GLCVQ) is exploited. The total bit rate for the subband coefficients combining Layers 9mo and 10mo is 284 bits. In addition, 18, 27, 45, 54, or 72 bits can be allocated for the wideband enhancements in Layer 7mo.

4. Decoder overview

The SWB decoder performs decoding of the high-frequency content based on the core codec synthesis and the SWB bitstream to provide a 32-kHz output signal. Most operations at the decoder correspond to local decoding performed at the encoder. The specific parts of the decoder are frame erasure...
cases, switching, and post-processing such as music enhancement.

4.1. Frame erasure concealment and switching

The frame erasure concealment of the wideband signal is taken care of by the core codec. The HF signal error concealment mainly consists of reducing overall energy of the HF content in a controlled way. When a frame is lost, the HF content from the previous frame is copied and scaled down. The sinusoidal components are however preserved longer. Their positions are saved in each frame and the previous frame sinusoidal components are repeated during a lost frame. Their energy is scaled down slightly less than the overall HF energy.

When switching between SWB and WB outputs, the bandwidth fluctuation of the high frequencies may be disturbing. The decoder therefore utilizes a zero-bit bandwidth extension, when a drop in bit rate removes the actual SWB coding [4]. The bandwidth extension has two modes: active processing for predictive approximation of the HF content based on wideband signal analysis, and stand-by processing during which the energy predictions are updated.

4.2. Music enhancement

A music enhancer is used in the decoder to reduce inter-tone quantization noise thus increasing music performance. A signal classification based on the signal stability estimation [5] is designed so the speech quality will not be affected.

The music enhancement is carried out as frequency-domain post-processing based on perceptual masking. An attenuation factor is applied to each frequency component in the wideband signal. In this way frequencies with insufficient quality can be reduced to a level below the estimated masking threshold.

The attenuation factors are estimated based on three parameters: local masking magnitude (which can be seen as the local perceptual loudness, and is a weighted sum of the spectral components around each bin), local masked magnitude (which can be considered as the perceptual error floor, and is another weighted sum of the spectral components), and overall average magnitude. The estimated attenuation factors are further smoothed by combining them with past values and additional smoothing factors that consider, e.g., the voicing and spectral sharpness of the wideband signal. As the last step, the attenuation factors are applied to the low-frequency signal.

5. Performance evaluation

Quality evaluation of the G718 and G729.1 SWB extensions was performed during the ITU-T Characterization of the proposed codecs in fall 2009. Using the Ref-A-B test methodology [6] three different experiments were carried out: clean speech, noisy reverberant speech, and music. In total, 5 laboratories participated in testing and all experiments were tested in two different laboratories such that both speech experiments were tested with two different languages. All the clean speech requirements and objectives were passed in both laboratories, all the noisy speech requirements but one were passed in one laboratory with one close fail in another laboratory, and all the music requirements but one were passed in one laboratory with two fails in another laboratory. Figures 2 and 3 present selected clean and noisy speech results from the Characterization phase, respectively. ITU-T G722.1C was used as a reference codec in all experiments.

The worst-case complexity of the G718 SWB extension is 79 WMOPS (weighted millions of operations per second). Using the AMR-WB interoperable mode, it is reduced to 70 WMOPS. For G729.1 SWB, the worst-case complexity of the encoding and decoding is 61 WMOPS. All of the complexity figures include both the core coding and the SWB extension.

6. Conclusion

We have presented a new embedded scalable SWB extension. It has been used to provide superwideband extensions to G718 and G729.1, approved in March 2010 by ITU-T as G718 annex B [7] and G729.1 annex E [8]. The overall codec structure was described, and selected listening test results of the extension were presented. This new SWB extension provides state-of-the-art performance for SWB speech.

7. Acknowledgements

The authors would like to thank Jari Hagqvist, Anssi Rämö, Henri Toukoumaa, Adriana Vasilache, Milan Jelinek, Vaclav Eksler, Heesik Yang, Hiroyuki Ehara, Takuya Kawashima, David Virette, Stéphane Ragot, Pierre Berthet, Jianfeng Xu, Herve Taddei, Wenhai Wu, Yue Lang, Qi Zhang, Chang Liu, Zexin Liu, Haiting Li, Jim Ashley, Tenkasi Ramabadr, Mark Jasiuk and Udar Mittal.

8. References