SCALABLE SUPERWIDEBAND EXTENSION FOR WIDEBAND CODING

Mikko Tammi, Lasse Laaksonen, Anssi Rämö, and Henri Toukomaa

Nokia Research Center, Tampere, Finland

ABSTRACT

Recent trends in speech and audio codec standardization include scalability and extending the signal bandwidth beyond wideband (WB) to superwideband (SWB). In this paper we introduce a SWB extension for the ITU-T G.718 WB codec. In the SWB extension the high frequency content is generated utilizing the quantized MDCT domain coefficients of the WB core, which enables low additional delay. The proposed implementation is scalable with 4 kbps layers. In the first layer two different coding modes are used depending on the input signal type. The proposed SWB extension is evaluated with listening tests and complexity analysis.

Index Terms— Audio coding, superwideband extension, scalability.

1. INTRODUCTION

Wideband speech applications are slowly yet inevitably gaining ground over narrowband speech. This recent trend in the area of speech coding suggests increasing the quality and bandwidth of the coded signal instead of concentrating on improving the absolute compression efficiency. A more natural communications experience and a wide gamut of targeted applications from high-quality voice calls to multimedia streaming drive the change and extend speech coders to perform well with both speech and audio signals. While wideband (WB), typically defined as a 50 Hz – 7 kHz bandwidth, is generally considered sufficient for most voice services, other signals, such as music, demand even higher quality. Superwideband (SWB), with a bandwidth of 50 Hz – 14 kHz, and fullband (FB) audio (20 Hz – 20 kHz) answer these needs.

Scalability, often combined with backwards interoperability, is another key design aspect in state-of-the-art speech coding. ITU-T has recently standardized two scalable codecs, G.729.1 [1] and G.718 [2], with high-quality WB capability at 32 kbps and below. A joint SWB and stereo extension for these codecs is on the ITU-T roadmap for standardization in 2009.

Many aspects have to be considered when extending WB codec to SWB codec, especially in telecommunications applications in which both the delay and complexity should be maintained at a reasonable level, and at the same time the codec should preferably be scalable. Different bandwidth extension methods [3–5], are often employed especially in audio coding applications for coding the high frequencies. In these methods the coded low frequency content is utilized for coding the high frequencies. However, these methods may not meet the delay requirements, and in addition the scalability is not fully operational as the coded high frequency band typically cannot be improved with additional layers. This is because these methods are not waveform coders, i.e., they do not tend to maintain the original waveform shape of the high frequency content.

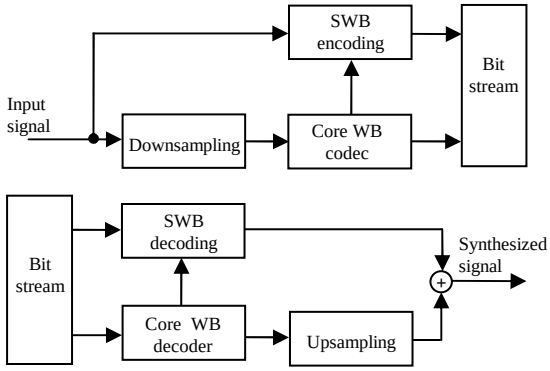

Figure 1. High level block diagram of the proposed SWB extension.

In this paper we introduce a SWB extension which like bandwidth extension methods utilizes the coded WB content, but at the same time maintains characteristics of a waveform coder. Thus the high frequency content can be easily updated with additional layers. This paper is organized as follows. In Section 2 we introduce the overall codec framework. Section 3 introduces the SWB coding method. The scalability and bit allocations of the proposed method are discussed in Section 4. Performance evaluation is provided in Section 5 and finally conclusions are drawn in section 6.

2. CODEC OVERVIEW

A high level block diagram of the proposed system is presented in Figure 1. The SWB extension utilizes the coded low frequency spectrum. Resampling is used for transforms between WB and SWB signals. In this work, the G.718 (EV-VBR) [2] codec, which represents state-of-the-art in WB speech and audio coding, is used as the core codec.

G.718 is an embedded scalable speech and audio codec designed for error prone communications channels and a wide array of applications including packetized voice, high quality audio/video conferencing, 3rd generation and future wireless systems, and multimedia streaming. It comprises five layers referred to as L1 (or core layer) through L5. Layers L1 and L2 are based on the ACELP technology, while layers L3 through L5 utilize transform coding (MDCT) to encode the error from L2.

The frame size of G.718 is 20 ms, and the algorithmic delay for wideband signals is 42.875 ms. However, the original EV-VBR baseline codec [6], which is closer to the development configuration used in this work, has a total algorithmic delay of 54.75 ms. This difference is mainly due to the window length used in the overlap-add operation in the transform coder.

The SWB extension we propose in this work was adopted as a starting point for the development that led to the joint candidate by

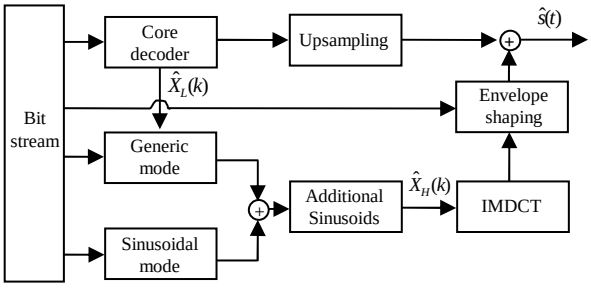

Figure 2. SWB decoder. Depending on frame type either generic of sinusoidal mode is used for synthesis

Ericsson, France Telecom, Matsushita/Panasonic, Motorola, Nokia, Texas Instruments, and VoiceAge for the ITU-T Study Group 16 Question 23 SWB/stereo extension for G.729.1 and G.718. For that development, the proposed SWB extension was implemented on top of the G.729.1 core and further enhanced in terms of quality improvements, as well as delay and complexity reductions. The work and results presented in this paper are, however, achieved in the context of the EV-VBR framework prior to finalization of the G.718 standard.

3. SWB EXTENSION

The proposed SWB extension method operates on the MDCT domain. This was selected because it is used also in the core codec and makes it possible to easily utilize the already quantized low frequency content in the frequency domain with only minor additional delay. The overview of the decoder of the proposed system is presented in Figure 2.

3.1 Generic mode

In many bandwidth extension methods such as in [3–5], it is commonly utilized that the low frequency spectrum can be directly transposed to the high frequencies and coarsely parameterized to follow the original high frequency envelope. This enables coding of the high frequency content with very low bit rate, and this kind of approach would be attractive also for SWB extension. However, we found that in the MDCT domain such approach did not lead to high synthesized signal quality, and we believe this is mainly due to the alias in the MDCT spectrum. On the other hand, we noticed that also in the MDCT domain it is rather easy to find similar shapes in the low frequency content as what there are in the high frequencies, which encourages on using the quantized low frequency content for coding the high frequencies, but in a manner which maintains the waveform shape. This kind of approach is introduced for example in [7], in which MDCT domain pitch filtering is used. However, the performance of that method is not quite at the aimed level.

Considering these findings, we ended up in a method in which the high frequency MDCT spectrum $X_H(k)$ is divided into m_b non-overlapping subbands $X_H^j(k)$, $j = 1, \ldots, m_b$, and for each subband the best match is searched from the already quantized and locally synthesized low frequency content $\hat{X}_L(k)$. The most similar match $\hat{X}_L^j(k)$ is obtained as $\hat{X}_L^j(k) = \hat{X}_L(k+d)$, where d is the instant which maximizes

$$S(d) = \left| \frac{\sum_{k=0}^{n_j-1} (X_H^j(k)\hat{X}_L(d+k))}{\sqrt{\hat{X}_L(d+k)^2}} \right| . \quad (1)$$

In (1) n_j is the length of the j[th] subband. This most similar match is scaled using two scaling factors $\alpha_1(j)$ and $\alpha_2(j)$. The first scaling factor operates on the linear domain to match the high amplitude peaks in the spectrum:

$$\alpha_1(j) = \frac{\sum_{k=0}^{n_j-1} (X_H^j(k)\hat{X}_L^j(k))}{\sqrt{\hat{X}_L(d+k)^2}} . \quad (2)$$

Notice that $\alpha_1(j)$ can get both positive and negative values. The second scaling factor $\alpha_2(j)$ operates on the logarithmic domain and is used to provide better match with the energy and the logarithmic domain shape. It is obtained as

$$\alpha_2(j) = \frac{\sum_{k=0}^{n_j-1} (\log_{10}(\alpha_1(j)\hat{X}_L^j(k)) - M_j)(X_H^j(k) - M_j))}{\sqrt{(\log_{10}(\alpha_1(j)\hat{X}_L^j(k)) - M_j)^2}} \quad (3)$$

where $M_j = \max_k(\log_{10}(|\alpha_1(j)\hat{X}_L^j(k)|))$. The overall synthesized subband $\hat{X}_H^j(k)$ is then obtained as

$$\hat{X}_H^j(k) = \zeta(k) 10^{\alpha_{2(j)}(\log_{10}(|\alpha_1(j)\hat{X}_L^j(k)|) - M_j) + M_j} , \quad (4)$$

where $\zeta(k)$ is –1 if $\alpha_1(j)\hat{X}_L^j(k)$ is negative and otherwise 1. The performance of these two scaling factors together is better than with only single factor. Removing the logarithmic domain scaling increases the roughness of the synthesized signal.

The above presented SWB coding method provides good quality for most signal types. However, it was noticed that for very periodic tonal signals, the proposed method may not be able to maintain the highly periodic signal shape with desired accuracy. Pitch pipe and bagpipes are classical examples of instruments producing such sounds. For this kind of tonal signals, a special coding method which we denote as sinusoidal mode was developed.

3.2 Sinusoidal mode

Strongly periodic signals can be detected by comparing the magnitude spectrums of two successive frames. As the magnitude information provided by MDCT is corrupted by the alias, some other transform should be used for reliable results. From the complexity point of view a good choice is Shifted Discrete Fourier Transform (SDFT) [8], as the MDCT coefficients can be obtained directly as the real part of the SDFT transform coefficients.

Let $Y_b(k)$ and $Y_{b-1}(k)$ be the complex SDFT coefficients of the current and previous frame, respectively. The similarity of the frames is measured using

$$S_b = \frac{\sum_{k=0}^{n_j-1} (|Y_b(k)| - |Y_{b-1}(k)|)^2}{\sum_{k=0}^{n_j-1} (|Y_b(k)|)^2} . \quad (5)$$

If S_b is smaller than a pre-defined limit δ, the sinusoidal coding mode is used instead of the generic mode.

It was found that for periodic signals the spectrum can be represented efficiently with a limited number of sinusoidal components. Only the spectrum components with highest amplitudes are modeled. $X_H(k)$ is divided into bands and for every band there is a predefined number of sinusoids available. The sinusoids are positioned into locations with highest absolute amplitude values. The sinusoids are characterized by their location (index), sign, and amplitude. In order to reduce the required bit rate, the bands were further divided into non-overlapping tracks such that a predetermined number of sinusoids could be selected per each track. This way, a very similar coverage of the total band is possible with a more limited number of possible locations for each sinusoid limiting thus also the number of bits required to transmit the exact locations.

3.3 Additional sinusoids

Both the generic and sinusoidal modes can be further enhanced with additional layers. This is an advantage in a scalable codec such as EV-VBR. In the proposed codec the quality improvement is achieved with additional sinusoids. The commonly used MDCT domain vector quantization was also considered, but in informal listening tests it was found that additional sinusoids give better results at low additional bit rates in our codec framework.

The additional sinusoidal layer operates similarly as the sinusoidal mode discussed above. The so-far synthesized spectrum $\hat{X}_H(k)$ is compared against the original spectrum $X_H(k)$ and utilizing tracks the sinusoids are positioned into frequencies where the absolute difference is largest.

3.4 Envelope shaping

Pre- and post-echoes are a typical artefact in low bit rate transform coding, and those were found also from the proposed SWB coding method. These can be efficiently detecting by performing local synthesis for $\hat{X}_H(k)$ in the encoder. The synthesized time domain signal $\hat{s}_H(t)$ is compared against the corresponding original signal $s_H(t)$, which is obtained by filtering the original input signal. We used a linear bandpass filter with passband from 7 kHz to 14 kHz.

Both signals, $\hat{s}_H(t)$ and $s_H(t)$, are divided into 2.5 ms subframes and their energies are compared. If the energy of the subframe is clearly higher in $\hat{s}_H(t)$, it is concluded that either a pre- or post-echo is present in the current subframe, and this information is transmitted to the decoder. In the decoder, the indicated subframes are attenuated in maximum 6 dB. The attenuation factor is interpolated between subframes to avoid sudden artificial energy changes.

4. SCALABILITY AND BIT ALLOCATION

Embedded scalability is one key design criterion in many recent speech codecs, as shown by the standardization of both G.729.1 and G.718. Scalability allows a smooth transition in perceived quality, and even signal bandwidth and the codec's multichannel capability, with varying bit rates.

The proposed SWB extension enables scalability. The core SWB layer provides a high-quality SWB extension and the SWB enhancement layers on top of the core extension layer further improve the synthesis quality and spectral match. We propose that further enhancement layers are selected with the WB core coding performance in mind to either enhance the lower frequency performance or to further improve the high-frequency match by exploiting the additional sinusoidal approach.

Table 1. Bit allocations.

	First layer		**Second layer**
	Generic mode	Sinusoidal mode	Additional sinusoids
SWB mode	1	1	-
Generic mode lag indices	32	-	-
Generic mode gains	28	-	-
Sinusoid positions	5	51	50
Sinusoid signs	1	6	5
Sinusoid amplitudes	4	21	24
Envelope shaping	8	-	-
Not used / Reserved	1	1	1

The bit allocations of the proposed SWB extension are presented in Table 1. The first SWB layer utilizes either the generic or sinusoidal mode and is coded with 4 kbps. Additional sinusoids can be flexibly added with almost any bit rate, but in this work we use 4 kbps layers. In the generic mode the high frequency spectrum is divided into four non-overlapping bands. The widths of the bands are 1, 1.75, 1.75, and 2.5 kHz covering the frequency band from 7 kHz to 14 kHz. For every band, the index from where the band is copied and two scaling factors are quantized. Vector quantization is used for joint coding of the scaling factors of adjacent bands. The quantization of the first scaling factor $\alpha_1(j)$ can be made more efficient by normalizing the spectral envelope of $\hat{X}_L(k)$ before initiating the coding, which decreases the dynamics of the scaling factor. In addition, one sinusoid is positioned in the lower part of the high frequency region. Its position, sign, and amplitude are sent. Scalar quantization is used for the amplitude value.

In the sinusoidal mode the high frequency region is divided into four bands. The first two of these bands cover the frequencies between 7 kHz and 10.2 kHz. They both have a total of 64 sinusoid positions divided into two tracks of 32 positions each such that one of the tracks covers the even positions while the other one covers the odd positions. Eight sinusoids in total are allocated to this region. Two more sinusoids are allocated to the higher frequencies: one in the 3rd band covering range 10.2–11.8 kHz and the other one in the 4th band between 11.8 and 12.6 kHz. The position of each sinusoid is transmitted along with the sign of the sinusoid. The amplitudes are jointly quantized using vector quantization.

The second layer further exploits the approach of the sinusoidal mode. A total of 10 sinusoids are placed in two pre-determined bands with four sinusoids in the 1st band covering frequencies 7–10.2 kHz and six sinusoids in the 2nd band between 10.2 and 12.6 kHz. The latter has three non-overlapping tracks instead of two. In this work, the same inner structure is used in the second SWB layer for both the generic and sinusoidal mode.

While switching between the SWB layers introduces no artifacts, switching from SWB to WB and back may annoy the listener as the bandwidth of the signal changes abruptly. It may therefore be beneficial to introduce artificial WB bandwidth extension for frames without real SWB content, and perform slow

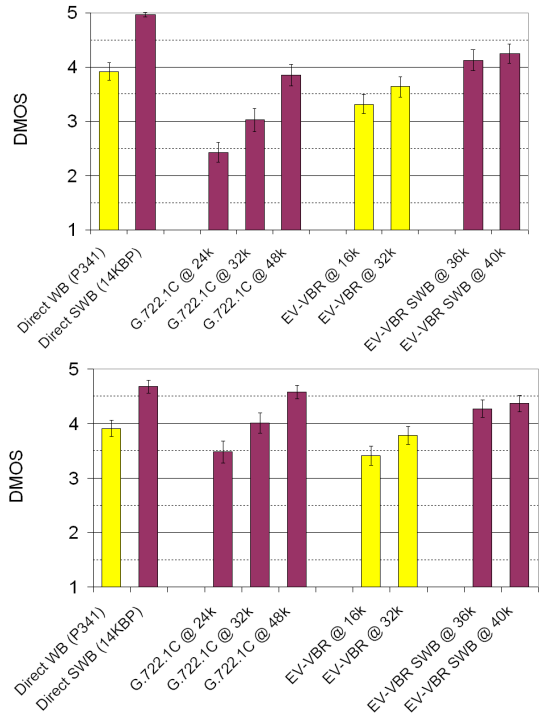

Figure 3. Listening test results for clean (upper picture) and noisy speech (lower picture).

ramping up or down of high frequency content when changing between WB and SWB modes. However, such functionality was not implemented in this work.

5. RESULTS

Two degradation category rating (DCR) listening tests were arranged to measure the performance of the proposed method. Figure 3 illustrate the results for both WB and SWB filtered experiments, tested with clean and noisy speech, respectively. EV-VBR (WB mode) and the G.722.1C SWB codec [9] were used as references. There were 10 and 12 listeners in clean and noisy experiments respectively. The results indicate good performance for both clean and noisy speech. In clean speech the performance of the proposed codec exceeds that of G.722.1C at 48 kbps already at the lowest SWB bit rate of 36 kbps. SWB clean speech was also preferred over WB speech by a wide margin. EV-VBR shows nice quality scalability with bit rate and band width. In noisy speech experiment there were both street and office noise at -20dB. The performance of the proposed method at 40kbps in noise is very near that of the original signal and statistically equivalent to G.722.1C at 48 kbps. Also with noisy speech experiment the SWB conditions were evaluated to be significantly better than WB conditions.

The additional complexity of the implementation was estimated to be about 35 WMOPS. The main sources of complexity are the resampling operations, SDFT and MDCT transforms, and the correlation based search used for matching bands from the low frequency content to the higher frequencies. The complexity of the proposed method has been reduced during the design of the joint 7-company ITU-T 16/Q.23 SWB/stereo extension candidate.

The additional delay caused by the SWB extension is below 2 ms. The extra delay originates from the two resampling filters (Figure 1), one in the encoder and the other in the decoder.

The proposed method is not limited for SWB coding only, and it can be easily utilized for example to extending WB signal to fullband. In this case, the number of extension bands is simply increased as a function of the coded bandwidth.

6. CONCLUSIONS

We have presented a SWB extension for WB coding, which can also be utilized for low bit rate fullband extensions. The method operates on the MDCT domain and utilizes the coded low frequency content by copying the most similar bands to high frequency content and scaling them with two scaling factors. Alternatively, very periodic signals are coded with a set of sinusoids. Assuming that the core codec operates on the MDCT domain as well, the proposed method can be implemented with reasonable additional complexity, small additional delay and very high flexibility. Scalability is enabled with a modular structure. Listening tests indicate good performance for both clean and noisy speech.

7. REFERENCES

[1] 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. ICASSP*, Honolulu, HI, USA, Vol IV, pp. 529-532, April 2007.

[2] T. Vaillancourt et al., “ITU-T EV-VBR: A Robust 8-32 kbit/s Scalable Coder for Error Prone Telecommunications Channels,” in *Proc.Eusipco*, Lausanne, Switzerland, August 2008.

[3] J. Mäkinen et al., “AMR-WB+: A New Audio Coding Standard for 3rd Generation Mobile Audio Services,” in *Proc. ICASSP*, Philadelphia, PA, USA, pp. 1109-1112, March 2005.

[4] T. Friedrich, and G. Schuller, “Spectral Band Replication Tool for Very Low Delay Audio Coding Applications,” in *Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, New Paltz, NY, USA, pp. 199-202, October 2007.

[5] M. Dietz, L. Liljeryd, K. Kjörling, and O. Kunz, ”Spectral band replication, a novel approach in audio coding,” in Proc. *AES 112th Convention*, Munich, Germany, Paper 5553, May 2002.

[6] M. Jelinek et al., “ITU-T G.EV-VBR Baseline Codec,” in *Proc. ICASSP*, Las Vegas, NV, USA, pp. 4749-4752, April 2008.

[7] M. Oshikiri, H. Ehara, and K. Yoshida, “Efficient Spectrum Coding for Super-Wideband Speech and Its Application to 7/10/15 kHz Bandwidth Scalable Coders,” in *Proc. IEEE ICASSP*, Montreal, Canada, Vol. 1, pp. 481-484, May 2004.

[8] Y. Wang, L. Yaroslavsky, M. Vilermo, and M. Väänänen, “Some Peculiar Properties of the MDCT,” in *Proc. IEEE ICSP2000*, Vol. 1, pp. 61-64, August 2000.

[9] M. Xie, D. Lindbergh, and P. Chu, “ITU-T G.722.1 Annec C: A New Low-Complexity 14 kHz Audio Coding Standard,” in *Proc. ICASSP*, Philadelphia, PA, USA, Vol. V, pp. 173-176, March 2005.