SPECTRAL BAND REPLICATION TOOL FOR VERY LOW DELAY AUDIO CODING APPLICATIONS

Tobias Friedrich, Gerald Schuller

Fraunhofer Institute for Digital Media Technology
Ehrenbergstr. 29, 98693 Ilmenau, Germany
{friets,shl}@idmt.fraunhofer.de

ABSTRACT

In this paper a Spectral Band Replication (SBR) tool for low delay audio applications is presented. One goal of this enhancement tool is to reduce the needed bit rate for the representation of audio data using an arbitrary audio codec. Another goal is to keep the algorithmic delay as low as possible. A low coding delay is essential for instance for real time applications like distributed music production under live conditions or telephone conferencing. The low delay SBR approach proposed in this paper uses techniques developed for speech coding purposes and is associated with artificial bandwidth extension methods, particularly spectral folding. Further, the tool exclusively operates in the time domain using prediction methods and adaptive filters in order to avoid additional delay which can be caused by using a filter bank.

1. INTRODUCTION

Traditional audio coders using a psycho-acoustic model and subband coding come to a limit when reducing the bit rate. A further reduction of the bit rate is possible with using more redundancy reduction with parametric descriptions of the signal. A successful example is the Spectral Bandwidth Replication (SBR), which originates in speech coding and is used for instance in the MPEG HE-AAC Coder [1, 2]. It parametrically describes the high frequency portions of a signal, and can be used with different "core" coders.

The fundamental idea of SBR is to exploit large dependencies between the lower and upper spectral parts of an audio signal assuming that in most cases a spectral correlation exists [3]. The high frequency portions can be efficiently reconstructed by using the lower ones. Thus, transmission of the high frequency part is not necessary and the core coder can operate on the residual bandlimited signal. The underlying audio coder can be run with a comparatively high SNR, as it is only responsible for the lower frequencies. SBR recreates the high frequencies using only a small amount of transmitted side information. A significantly enhanced coding gain is the main reason for the use of SBR, since the high frequencies, which normally consume a significant amount of bits, do not need to be waveform coded anymore.

This paper is organized as follows. Section 3 describes the structure and functionality of our new low delay SBR approach. Section 4 provides a delay analysis of our new approach, and in Section 5 and 6 the resulting bit rate and subjective quality is compared with the MPEG SBR.

2. PREVIOUS APPROACHES

MPEG SBR has proven to be an attractive enhancement to audio coders for low bit rate coding. However, the delay introduced by the MPEG SBR amounts to 961 samples (30 ms at 32 kHz sampling frequency) [4], which mainly results from its 64-channel QMF used for spectral estimation and computation of the SBR data. Hence, MPEG SBR is not suitable to time critical applications like live productions using multiple wireless microphones and simultaneous in-ear monitoring, since they require a lower delay. Artificial bandwidth extension methods developed for narrowband speech in telephony ($300-3400$ Hz) can be divided into two classes. The first class uses a rectified upsampled narrowband as substitution for the highband, for instance RELP [5, 6], which does not have the requirement for a high quality reconstruction. In contrast, our goal is to obtain a high quality audio reconstruction. The second class performs a so-called spectral folding [7], where in the decoder, the narrowband speech signal is upsampled by a factor of two. Due to upsampling, a mirror image appears in the upper spectral part which is then used as the highband signal. This highband signal is shaped by a shaping filter with fixed coefficients and finally level adjusted by means of gain parameters [8]. The performance of this method is strictly dependent on the characteristics of the shaping filter. Since different signals have different spectral characteristics, a shaping filter with fixed coefficients cannot provide optimal results.

3. NEW APPROACH

Our low delay SBR tool is based on the artificial bandwidth extension technique using spectral folding as mentioned in Section 2. But unlike the latter technique, our low delay SBR tool is designed for wideband signals and can handle speech as well as music signals and uses a signal-adaptive shaping filter.

The block diagram of our low delay SBR tool with an arbitrary perceptual audio core codec, which is denoted simply as core codec in the following, is given in Figure 1.

3.1. Encoder

An analysis filter bank splits the input PCM audio signal $x(n)$ into two critically sampled subband signals having equal bandwidth, $y_{LB}(n)$ and $y_{HB}(n)$. The lowband $y_{LB}(n)$ is conventionally coded by the core codec, whereas the highband is passed through the envelope estimator which approximates the spectral envelope of the downsampled (and as a result mirrored) highband signal using linear predictive coding (LPC). The basic structure of the envelope

978-1-4244-1619-6/07/$25.00 ©2007 IEEE

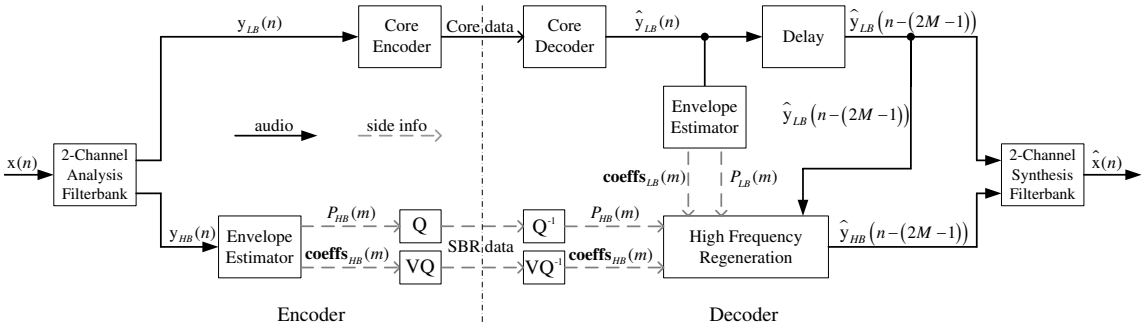

Figure 1: *Block diagram of our SBR tool with a core codec.*

estimator is shown in Figure 2. Since the envelope estimator works in the encoder as well as in the decoder, operating on the highband and the lowband respectively, for generality no indices for the high- and the lowband are given in Figure 2. The envelope estimator works as follows: First the input signal y(n) is blocked into blocks containing $M = 64$ samples and windowed with 50% overlap. For the windowing a sine window with a length of $2M = 128$ samples is used. Thus, the vector $\mathbf{y}(m)$ contains 128 time samples, where m denotes the block number. $\mathbf{y}(m)$ is used to calculate the autocorrelation function $\mathbf{ACF}(m)$. Then the spectral envelope is approximated via the Levinson-Durbin algorithm on the basis of $\mathbf{ACF}(m)$, generating the prediction error power $P(m)$ and a set of LPC coefficients $\mathbf{coeffs}(m)$.

After obtaining the spectral information for the highband, $P_{HB}(m)$ and $\mathbf{coeffs}_{HB}(m)$ are quantized using logarithmic quantization and a vector quantizer respectively. This information is parameterized as scaling factors (prediction error powers) and indices of vector quantizer (VQ) codebook entries and transmitted as side information to the decoder. The VQ codebook entries contain the reflection coefficients describing a spectral envelope and the scaling factors are needed for the energy adjustment of the regenerated highband. For each block m, 14 bits are allocated to the side information for our SBR tool, which provides a fixed bit rate. 8 bits represent the VQ index and 6 bits represent the quantized scaling factors.

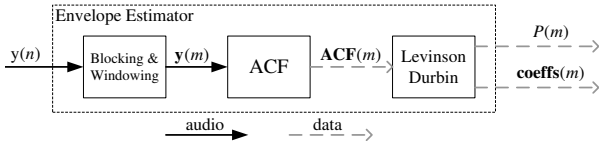

Figure 2: *Building blocks of the envelope estimator.*

3.2. Decoder

On the decoding side, the core decoded lowband $\hat{y}_{LB}(n)$ is used for the regeneration of the highband $\hat{y}_{HB}(n)$. The lowband is analyzed by the envelope estimator. The obtained set of LPC coefficients $\mathbf{coeffs}_{LB}(m)$, the prediction error power $P_{LB}(m)$, and the low pass signal, delayed by $2M - 1$ samples are passed to the high frequency regeneration unit, which also uses the parametric envelope description for the highband provided by the dequantized SBR side information.

The high frequency regeneration unit is shown in Figure 3. Notice that m denotes the block index and n the sample index. The subscript indices *LB* and *HB* stand for lowband and highband respectively. First the core decoded lowband is passed through an FIR prediction filter. The FIR filter (see Figure 3) "whitens" the signal by filtering $\hat{y}_{LB}(n)$ with its inverse spectral envelope represented by the LPC filter coefficients. In this context "white" means that the prediction residual $e_{LB}(n)$ has a more flat spectrum. The white prediction residual of the lowband is then fed into the IIR shaping filter to regenerate the highband. For this shaping, the transmitted LPC coefficients $\mathbf{coeffs}_{HB}(m)$ describing the spectral envelope of the original highband are used. The quotient between the transmitted prediction error power of the highband $P_{HB}(m)$ and the prediction error power of the lowband $P_{LB}(m)$ is used for a power adjustment of $e_{LB}(n)$. The denominator $P_{LB}(m)$ scales the prediction residual of the lowband to variance 1 so that $e_{LB}(n)$ can be assumed as an almost white random signal with variance 1. The nominator $P_{HB}(m)$ scales the excitation signal to the variance of the original highband. The output of the shaping filter provides an artificial signal having similar properties as the original highband concerning spectral shape and energy.

Finally, the two-channel synthesis filter bank performs upsampling and filtering resulting in a SBR enhanced wideband signal $\hat{x}(n)$. Note that the upsampling operation leads to a mirrored copy in the upper spectrum. However, this mirrored spectral part is a sufficient approximation of the original highband, since the envelope estimation in the SBR encoder is performed on the mirrored highband. The filter algorithm of the FIR and IIR filters uses a lattice structure. The essential benefit of a lattice filter compared to a Direct Form II filter is the possibility of interpolating the filter coefficients without the filter becoming unstable. Every block (i.e. 64 samples) a set of LPC coefficients and one scaling factor are updated. Between the mentioned updates a samplewise interpolation is used. Thus, the lattice filter is supplied with a set of reflection coefficients and a scaling factor for each sample.

4. DELAY ANALYSIS

This section focuses on the "algorithmic delay" which is the delay caused by the algorithm alone. Delays caused by the limited speed of a hardware implementation or a transmission channel are excluded. Delay sources of our low delay SBR tool are:

- 2-Channel Filter bank delay $= 10$ samples
- Encoder analysis over one frame $= 2 \cdot 127$ samples

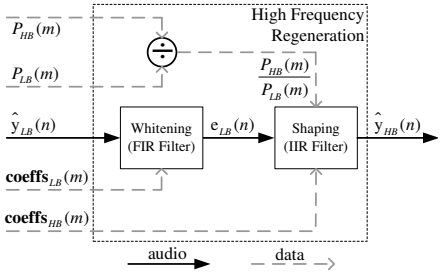

Figure 3: *Basic structure of the high frequency regeneration process in the SBR decoder.*

- Decoder analysis over one frame $= 2 \cdot 127$ samples

The two-channel symmetric filter bank uses elliptic filters of order 16 having a stopband attenuation of 100 decibels with a transition bandwidth of $1/160$ of the sampling frequency. The delay for the analysis and synthesis filter bank cascade amounts to 10 samples (0.2 ms for 48 kHz sampling frequency).

For block based processing, a certain amount of time has to pass to collect the samples belonging to one block [9]. Since the SBR sine window has a length of 128 samples, the resulting delay is 127 samples. Due to downsampling, the delay doubles to $2 \cdot 127$ samples (5.3 ms for 48 kHz sampling frequency). The blocking operation is also performed on the core coded lowband in the decoder, again producing $2 \cdot 127$ samples delay. Since the core encoder usually already has a higher delay than needed for the SBR tool, the SBR tool only introduces an additional delay in the decoder. Table 1 lists the "additional delay" in the decoder resulting from different SBR versions. A more detailed description of how the delay of the regular MPEG SBR and of a new modified SBR version for AAC-ELD is calculated can be found in [4].

SBR	samples	time [ms]		
		32 kHz	44.1 kHz	48 kHz
MPEG SBR	961	30.0	21.8	20.0
AAC-ELD SBR	577	18.0	13.1	12.0
low delay SBR	264	8.3	5.9	5.5

Table 1: *Delay of the different SBR tools in samples and ms.*

5. BIT RATE COMPARISON

The minimum and maximum side information bit rates of the regular SBR version and the fixed bit rate of our low delay SBR version are given in Table 2. Our low delay SBR produces a higher bit rate than MPEG SBR (factor 2.36 for 32 kHz and 2.49 for 44.1 kHz [10]). This is the result of the very small block size used for our SBR to obtain a low delay.

6. LISTENING TEST RESULTS

A MUSHRA [11] listening test was performed to asses the performance of the low delay SBR version using the test items listed in Table 3. The MPEG SBR tool was set to use the same crossover frequency as our low delay SBR tool. This is necessary to provide comparable results between MPEG SBR and our low delay SBR. Usually the MPEG SBR crossover frequency is set by the core coder. In our listening tests we used the low passed original signal instead of a core coder, because the goal was to only evaluate the SBR tool. Furthermore, the MPEG SBR was operated with maximum reconstruction quality, i.e. with the maximum bit rate. The test conditions for the subjective listening tests are listed in Table 4.

	32 kHz kbps		44.1 kHz kbps	
	min	max	min	max
MPEG SBR	1.04	1.89	1.39	2.50
low delay SBR	3.42		4.71	

Table 2: *Minimum and maximum bit rates for our low delay SBR and the regular MPEG SBR.*

Test File	Description
es01	Suzanne Vega
es02	Male German Speech
es03	Female English Speech
sc01	Haydn Trumpet Concert
sc02	Classical Orchestral Music
sc03	Contemporary Pop Music
si01	Harpsichord
si02	Castanets
si03	Pitchpipe
sm01	Bagpipe
sm02	Glockenspiel
sm03	Plucked Strings

Table 3: MPEG verification test files of 1997

	System under Test	Condition
1	Original	no Processing
2	Anchor 3.5 kHz	Lowpass Filtering at 3.5 kHz
3	Anchor 7 kHz	Lowpass Filtering at 7 kHz
4	Anchor 10 kHz	Lowpass Filtering at 10 kHz
5	low delay SBR	Time Domain SBR
6	regular SBR	Frequency Domain SBR

Table 4: Systems under test for each test signal.

Figure 4 shows the results of the MUSHRA test and reveals that in general our low delay SBR performs similar in comparison to the MPEG SBR. The MPEG SBR has problems with si02 and si03. si02 is a sequence of transients (Castanets), and si03 a tonal signal of strong stationarity (Pitchpipe). The bad performance of MPEG SBR for the transient signal can be traced back to the fact of using longer frames. Better results are achieved by our low delay SBR version due to its better time resolution. It is not surprising that low delay SBR is only rated as "good" for strong tonal signals like si03, whereas the rating in the same range for the regular SBR is surprising. The low delay SBR introduces some kind of modulation artifacts in the high frequency parts. These modulation artifacts are assumed to be originating from a harmonic component exactly located at the SBR crossover frequency. After adding the low- and the highband, the harmonics at the spectral borders can interfere with each other. Furthermore, the test results reveal some problems of low delay SBR dealing with voiced speech (es01).

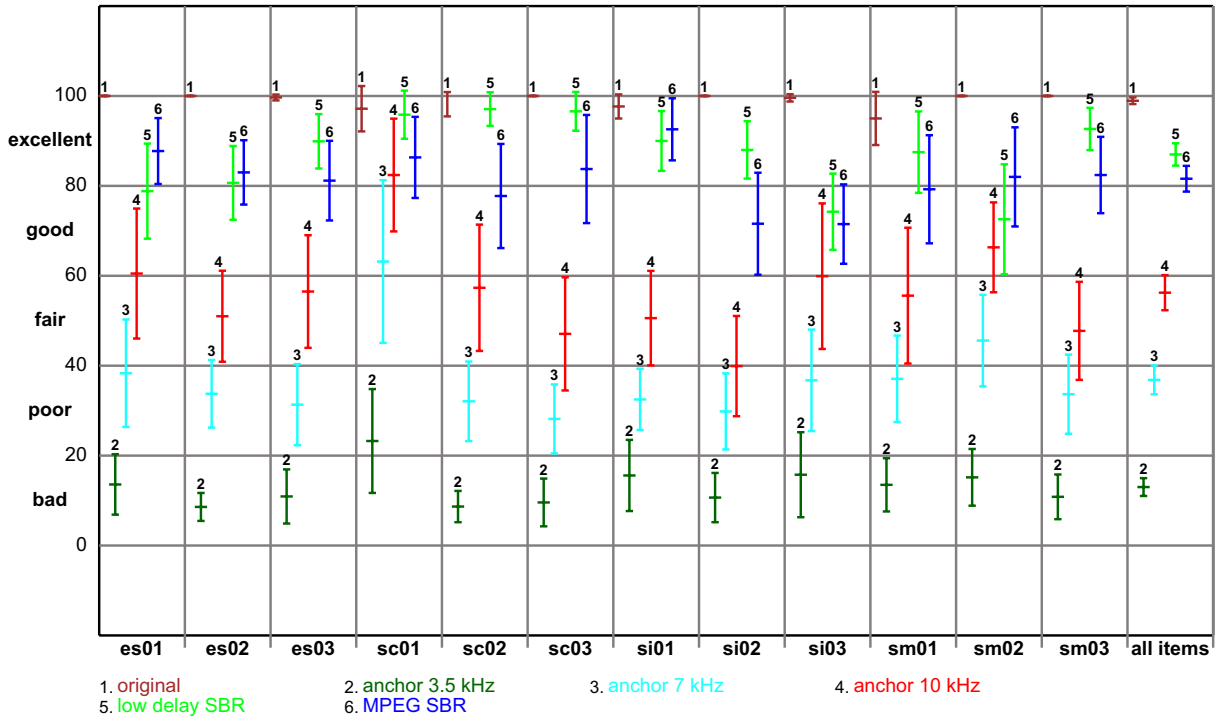

Figure 4: *Mean Subjective Scores of 12 listeners with 95 % confidence intervals for the* 32 *kHz test session.*

7. CONCLUSIONS

In this paper an SBR tool is presented producing a very low algorithmic delay while maintaining a comparable audio performance to a regular SBR as used in HE-AAC. Although the resulting bit rate of the low delay SBR tool is about twice as high as those produced by the regular SBR tool, it is not prohibitively high for practical applications, and produces the same or even better reconstructed audio quality, as shown in a subjective listening test. Further, the new SBR approach has an inherent low computational complexity as well as providing a constant bit rate, which is a benefit for many real-time or streaming applications. For more details see [10].

8. REFERENCES

[1] ISO/IEC, "Coding of audio-visual objects - part 3: Audio (mpeg-4 audio, 3rd edition)," *ISO/IEC Int. Std. 14496-3:2005*, 2005.

[2] 3GPP, "Spectral band replication (sbr) part (3gpp ts 26.404 version 6.0.0 release 6)," *ETSI TS 126 404 V6.0.0 Technical Specification*, September 2004.

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

[4] M. Schnell, R. Geiger, M. Schmidt, M. Jander, M. Multrus, G. Schuller, and J. Herre, "Enhanced mpeg-4 low delay aac - low bitrate high quality communication," *AES 122th Convention, Vienna, Austria*, May 2007.

[5] C. K. Un and D. T. Magill, "The residual-excited linear prediction vocoder with transmission rate below 9.6 kbit/s," *IEEE Transactions on Communications*, vol. 23, no. 12, pp. 1466–1474, December 1975.

[6] J. R. Deller, J. G. Proakis, and J. H. L. Hansen, *Discrete-Time Processing of Speech Signals*, J. Griffin, Ed. Macmillan Publishing Company, 1993.

[7] L. Kallio, "Artificial bandwidth expansion of narrowband speech in mobile communication systems," Master's thesis, Helsinki University of Technology, Department of Electrical and Communications Engineering, Laboratory of Acoustics and Audio Signal Processing, December 2002.

[8] H. Yasukawa, "Quality enhancement of band limited speech by filtering and multirate techniques." *IEEE Proc. of International Conference on Spoken Language Processing*, pp. 1607–1610, 1994.

[9] E. Allamanche, R. Geiger, J. Herre, and T. Sporer, "Mpeg-4 low delay audio coding based on the aac codec," *106th AES Convention, Munich, Germany*, May 1999.

[10] T. Friedrich, "Spectral band replication tool for very low delay audio coding applications," Master's thesis, Technische Universität Ilmenau, February 2007.

[11] ITU-R, "Recommendation bs. 1543-1: Method for the subjective assessment of intermediate sound quality (mushra)," *International Telecommunication Union, Geneva, Switzerland*, 2001.