Audio bandwidth extension by frequency scaling of sinusoidal partials

Tomasz Żernicki, Maciej Bartkowiak

Poznań University of Technology, Chair of Multimedia Telecommunications and Microelectronics,
Polanka 3, 60-965, Poland
tzernicki@multimedia.edu.pl, mbartkow@multimedia.edu.pl

ABSTRACT

This paper describes a new technique of efficient coding of high-frequency signal components as an alternative to Spectral Band Replication. The main idea is to reconstruct the high frequency harmonic structure trajectories by using fundamental frequencies obtained at the encoder side. Audio signal is decomposed into narrow subbands by demodulation based on the local instantaneous frequency of individual partials. High frequency components are reconstructed by modulation of the baseband signals with appropriately scaled instantaneous frequencies. Such approach offers correct synthesis of rapidly changing sinusoids as well as proper reconstruction of harmonic structure in the high-frequency band. This technique performs also a correct energy adjustment of sinusoidal partials. High compression efficiency has been achieved and confirmed by listening tests.

1. INTRODUCTION

Recently, audio compression has reached a turning point. A new class of advanced techniques has been developed that combine the concepts of waveform coding with parametric coding. One of them employs the technique of Spectral Band Replication (SBR) [1], from the family of audio bandwidth extension methods [2], [3]. The idea of SBR is based on the observation that the signal spectrum in the high-frequency range strongly depends on the signal spectrum in the low-frequency range. Such dependence allows for reconstructing of the high-frequency components of the signal from the low-frequency part (fig. 1). Therefore, it is possible to increase the efficiency of a traditional audio codec and avoid the need to transmit the high-frequency data, since the corresponding signal part may be re-synthesized from a modified version of the decoded low-frequency band (fig. 2). Nevertheless, some signals contain strong tonal components with many quite strong harmonics. Such harmonics may not be generated properly by the SBR tool. It happens quite often that the frequencies of the higher harmonics are shifted due to the process of band copying (fig. 3). For some signals such distortions may appear quite annoying. Perhaps the best examples of such situations are harmonically rich sounds with pitch changing due to glissandi or vibrato, such as violin, trumpet or opera voice. Higher overtones of these sounds cannot be properly reconstructed in the SBR decoder, because their spectra cannot be obtained by copying tonal parts whose frequency trajectories are not parallel in the TF plane. Therefore, offering a high quality programme at bit rates below 24kb/s is a real challenge, since it requires the bandwidth of the signal part handled by the core codec to be lowered to 5 kHz or less.

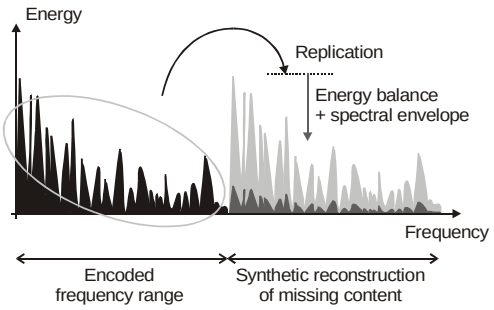

Figure 1. Reconstruction of high frequency content by spectral band replication.

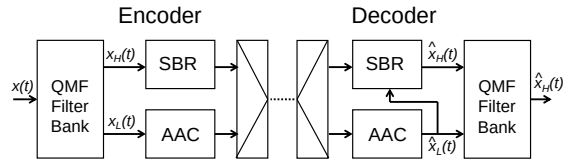

Figure 2. Low bit rate audio coding employing the SBR technique.

Other methods which deal with this problem use sinusoidal modeling of tonal components in the high frequency band [4]. These techniques give better frequency harmonic representation, but still there is a problem with proper energy concentration around particular sinusoidal partials. This effect can be observed as a "synthetic" sound.

We propose a coding tool that reduces the artifacts mentioned above as well as those produced by quickly changing frequencies of harmonics that cannot be properly reproduced by the SBR decoder. It is achieved by appropriate frequency scaling instead of frequency shifting.

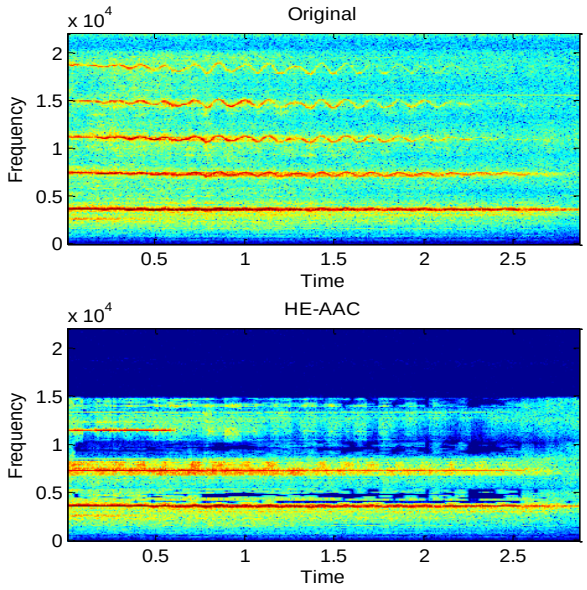

Figure 3. Spectrograms of vibratto signal encoded with HE-AAC with SBR.

2. PROPOSED TECHNIQUE

Our idea of improving the quality of reconstructed high frequency part of the spectrum comes from the observation that the most important auditory information is located in spectral peaks. Therefore, spectral peaks should be re-synthesized more truthfully. Each physical partial contains both tonal components and narrowband noise which vary with the local instantaneous frequency. Exact reconstruction of such variations is the key feature of the proposed technique.

Multiple fundamental frequencies of individual harmonic series are estimated and tracked in the proposed encoder (fig. 4). The audio signal is divided into several narrow subbands. Each subband covers an individual sinusoidal partial which varies with the fundamental frequency. In the decoder (fig. 5), high frequency components are reconstructed through frequency scaling of low frequency components. This scaling is achieved by modulation with integer multiples of individual instantaneous frequencies. For each subband, a correct energy is maintained w.r.t. the spectral energy envelope of the original signal. In the encoder (fig. 4), the regeneration of the harmonic structure is performed as well, and the reconstructed signal is subtracted from the original through spectral masking process. Obtained residual signal is treated as noise, modeled by linear prediction. A perceptually equivalent noise is re-synthesized in the decoder.

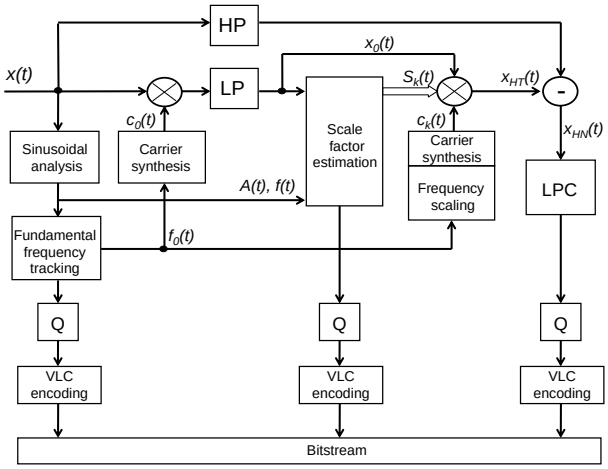

Figure 4. Encoder side of proposed technique.

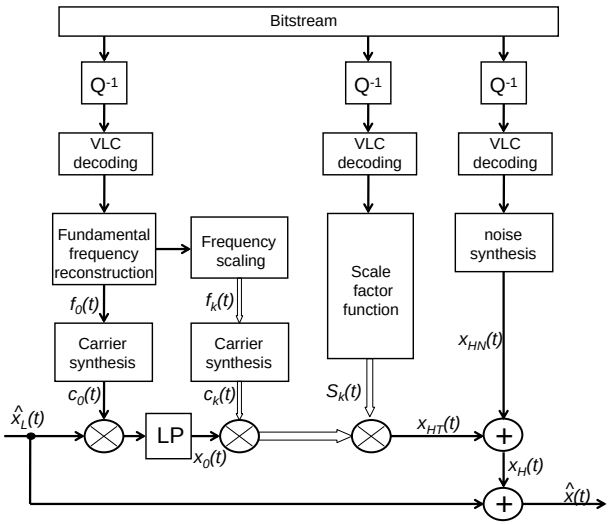

Figure 5. Decoder side of proposed technique.

The bitstream generated by the proposed coder comprises fundamental frequency trajectories, scaling factors of high frequency partials and LP coefficients describing noise components in the high frequency part of the spectrum.

3. REPRESENTATION OF HIGH FREQUENCY ENERGY

3.1. Signal analysis overview

In the proposed approach a fixed split frequency, f_{HFR} is determined, depending on the target bit rate. The low frequency part of the signal spectrum below f_{HFR} is encoded by a core codec which is out of the scope of this paper. The high frequency part is reconstructed by frequency scaling of tonal components.

For this purpose, the input signal is analyzed by a harmonic sinusoidal model (fig. 6) [5],[6] that provides the data on the harmonic series which are necessary in order to properly reconstruct the missing upper part of the spectrum. The input signal is analyzed on a frame basis using FFT in sufficiently long overlapping windows in order to maintain the required time – frequency resolution. Correct amplitude and frequency values are estimated using the QIFFT algorithm [7] in each frame and subsequently linked into frequency trajectories.

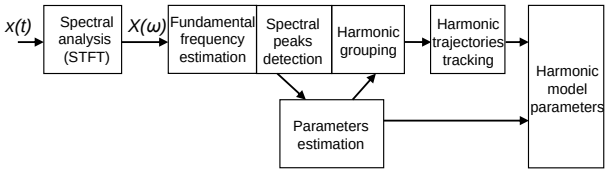

Figure 6. Sinusoidal analysis scheme.

We use this information to approximate the spectral envelope of each harmonic series and then calculate the energy scale factors.

3.2. Sinusoidal analysis

In the proposed approach we use sinusoidal analysis to achieve correct frequency and amplitude values of harmonic partials. Data obtained from this analysis is used for:

- Approximation of spectrum envelope, independently for every harmonic series encountered;
- Estimation of fundamental frequency trajectories of these harmonic series.

The process of frequency scaling is done by demodulation of the signal using its fundamental frequency followed by lowpass filtering. The baseband signal is subsequently modulated by several integer multiples of the fundamental frequency. This procedure is repeated for each fundamental frequency trajectory. The final signal is a sum of several narrow band signals.

Estimation of fundamental frequencies is needed for the demodulation procedure. We use harmonic modeling to estimate harmonic series in the signal, based on fundamental frequencies. The main idea of this technique is to achieve correct value of fundamental frequency and then calculate values of frequency and amplitude for every partial. The simplest way of achieving this process is integer multiplication of fundamental frequency.

In the next step, we group spectral peaks based on the information of harmonic signal structure. Then we track fundamental frequencies and harmonic peaks separately, forming sinusoidal trajectories.

To reduce the complexity of the algorithm and the amount of data that has to be transmitted, the maximum frequency value in each of fundamental frequency trajectories is determined first. If some trajectory includes frequency values above f_{HFR}, then this trajectory is discarded.

3.3. Frequency scaling

The modulation performing the above mentioned frequency scaling results in frequency shift by the instantaneous frequency of the modulating function. This frequency is calculated as an integer multiple of the fundamental frequency, so that an artificial overtone is generated.

This procedure is as follows:

- In order to achieve frequency shift that is continuously following the fundamental frequency of each of the harmonic series, the fundamental frequency estimated from the harmonic model is interpolated on a sample basis
- The demodulating function $c_0(t)$ is obtained as

$$c_0(t) = \exp\left(- j2\pi \int_{-\infty}^{t} f_0(t)dt \right) \quad (1)$$

- The demodulation product, $x_o(t)$ is subjected to low pass filtering. The baseband signal obtained in this way has slowly varying instantaneous frequency that is close to 0. This signal represents the fundamental tonal component together with some natural surrounding noise.

$$x_0(t) = \left(c_0(t) \cdot x(t)\right) * h_{LP}(t) \quad (2)$$

- The high frequency partials are reconstructed by modulation of the baseband signal $x_o(t)$ by a set of carrier functions, whose frequencies are integer multiples of the original $f_0(t)$.

$$c_k(t) = \exp\left(j2\pi \int_{-\infty}^{t} f_k(t)dt \right) \quad (3)$$

The obtained signal represents only partials from one harmonic series. Frequency scaling must be repeated for every trajectory of fundamental frequency.

$$x_{HT}(t) = \sum_{k=k\_start}^{k\_end} S_k \cdot c_k(t) \cdot x_0(t) \quad (4)$$

$$k\_start = \left\lceil \frac{f_{HFR}}{\max\left(f_0(t)\right)} \right\rceil \quad (5)$$

$$k\_end = \left\lfloor \frac{0.5 \cdot f_S}{\max\left(f_0(t)\right)} \right\rfloor \quad (6)$$

3.4. Energy scaling of replicated spectra

Apart from the proper character of the spectrum, the second key problem is to maintain a similar energy of the frequency scaled components as compared to the components in the original signal. In our implementation, additional scale factors are responsible for this energy preservation. They are calculated in the encoder, sent to the decoder and applied to the reconstructed signal.

For every harmonic series we calculate its spectral energy envelope. This envelope is approximated by a parametric curve which is used to determine the energy scale factors for individual subbands reconstructed by frequency scaling. As spectral envelope of an individual harmonic series often corresponds to the frequency response of instrument body, it is reasonable to employ a log-log approximation known as Bode diagrams. We use logarithmic frequency and amplitude scale, and approximate the resulting function with a low order (e.g. quadratic) polynomial.

The estimation of the spectral envelope is done for each harmonic series independently. Since spectra of individual sounds are not available even at the encoder side, we use amplitude values of harmonic partials delivered by the harmonic sinusoidal model. In case of many real music recordings, harmonic data from the sinusoidal model are incomplete, however. Therefore we first construct a vector which contains harmonic amplitudes A_{env} obtained from the sinusoidal analysis A_{sin}. This vector is supplemented by the data from the original spectrum A_{orig} at the positions of missing partials,

$$A_{env} = A_{\sin}(m,k) + A_{orig}(n,k) \quad (7)$$

where:

k – index of analysis frame; m – indices of frequency peaks from sinusoidal analysis n – indices of missing frequency peaks being multiplication of fundamental frequency,

$$\boldsymbol{n} = \boldsymbol{k} \setminus \boldsymbol{s} \quad (8)$$

where,

$\boldsymbol{n}$ - the set of indices of missing partials,

$\boldsymbol{k} = \{1,2, ...k\_end_x\}$,

$\boldsymbol{s} = \{s: s=[F_{sin}/f_0]\}$

F_{sin} denotes the set of frequencies of harmonic partials from the sinusoidal model.

We remove peaks whose frequencies are below the split frequency f_{HFR} from the obtained vector of spectral peaks. This vector represents spectral envelope of a single harmonic series.

A low-order polynomial is fitted to the envelope in log-log domain (fig. 7). The coefficients of the polynomial are quantized non-uniformly, using a quantizer operating in a nonlinear scale, and transmitted to the decoder.

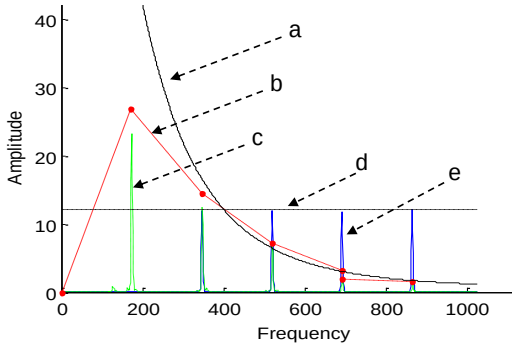

Figure 7. An example of energy scaling for single frame, a – polynomial approximation of original spectrum, b – spectral envelope of original signal, c –original signal spectrum, d – maximum amplitude of demodulated baseband signal, e – spectrum of frequency scaled baseband signal, before energy scaling.

3.5. Signal reconstruction

On the decoder side, the high frequency part of the signal is reconstructed from the bandlimited signal obtained from the core decoder, and using the parameters from the bitstream (fig. 2).

The fundamental frequencies of each of the encoded harmonic series are decoded from the bitstream. Their trajectories are used for demodulation of sinusoidal partials from the low frequency part of the spectrum in the same way as it is done in the encoder. It should be noted that the signal from core encoder and frequency trajectories are quantized and therefore laden with quantization error. This error can be minimized by proper set of the cutoff frequency of the low pass filter used for isolating the demodulated signal.

4. CODING DETAILS

4.1. Coding of fundamental frequency trajectories

Each trajectory of fundamental frequency has to be transmitted to the decoder, therefore it also needs to be efficiently encoded. In our implementation, linear predictive coding (LPC), uniform quantization and Huffman coding is used for this purpose. For encoding of the parameters of the trajectories, Burg variant of the linear prediction is used [8]. The order of the predictor as well as the number of past samples used to estimate the predictor coefficients depend on the time resolution of the trajectories. Good results are obtained for time resolution of 11.6 ms, where a predictor of 6th order operating on 20 past samples is adequate. The bit rate necessary to transmit the fundamental frequencies is about 0.5 to 1 kbps.

4.2. Coding of scaling factors

The polynomial coefficients which best fit the scaling curve are transmitted to the decoder for every frame. Coefficient values are converted to logarithmic domain and then uniformly quantized with quantization step equivalent to 0.5 dB. The bit rate required to transmit the Huffman encoded coefficients is about 2-4 kbps.

4.3. Representation of the residual

The residual signal represents the remaining noisy part of the upper part of the spectrum. It is obtained in the process of spectral masking, after reconstruction of tonal energy.

$$M(k) = \frac{X(k)}{smooth\left(thr\left(|S(k)| \cdot \varepsilon\right)\right)}, \qquad (9)$$

where $$thr(x) = \begin{cases} x, & x > 1 \\ 1, & x \le 1 \end{cases} \qquad (10)$$

and:

- $X(k)$ – FFT of the zero-padded original signal, $x(t)$
- $M(k)$ – FFT of the zero-padded residual signal $m(t)$
- $S(k)$ – FFT of the zero-padded reconstructed high-frequency partials,
- ε – masking threshold,
- $smooth(x)$ – a convolution with a smoothing kernel.

The LPC technique is used for noise modeling of the residual signal. The noise power spectrum envelope is represented by 16 PARCOR coefficients that are quantized in log scale with the resolution of 8 bits. The bit rate required to transmit the Huffman encoded LPC coefficients is about 1-2 kbps.

5. EXPERIMENTAL RESULTS

In order to verify the proper operation of the presented technique as well as to assess the compression efficiency, full encoder and decoder have been implemented as software that processes audio excerpts off-line. The experimental software has been implemented by routines written in Matlab. The experiments used wide range of audio material, in particular from classical and jazz music recordings, especially such containing parts played by instruments with strong tonal elements in high part of spectrum (e.g. violin, accordion, and trumpet) [9].

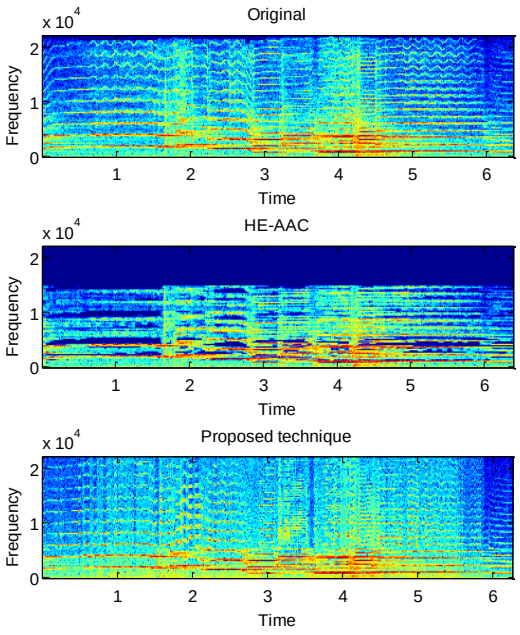

Figure 8. Example spectrograms of original signal (top), encoded with HE-AAC (middle) and proposed technique (bottom) at bit rate of 24 kbps. Please note that the operating frequency limit of the HE-AAC codec is fixed by the control mechanism and results from the low bit rate. Such limit is not present at the output of our codec.

In the experimental implementation, the total bit rate for the encoded parameters of high-frequency components is about 3 – 6 kbps which is comparable to the parametric data bitstream produced by the standard MPEG-4 HE-AAC encoder, employing the SBR technique. These figures may be probably reduced by improving the compression technique used for these parameters. Another way of bitstream reduction is to decrease the number of fundamental frequencies, leaving only those with most important energy from perceptual point of view. The most interesting conclusion is that the proposed technique is capable of offering a much improved the quality of the reconstructed programme w.r.t. the SBR technique operating at a similar bit rate.

A series of informal listening experiments was conducted with the participants being audio experts. In most of the experiments, the quality of the encoded music excerpts significantly surpassed that delivered by the SBR codec. This can also be clearly evident from the observation of spectrograms (fig. 8).

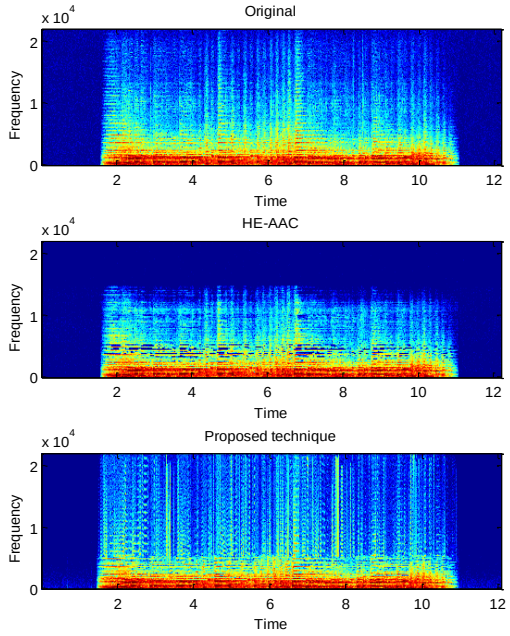

Figure 9. Example spectrograms of classical music (top), encoded with HE-AAC (middle) and proposed technique (bottom) at bit rate of 24 kbps.

6. CONCLUSIONS

A new approach to bandwidth extension for high quality low bit rate audio coding has been proposed in this paper. The main advantage of this technique is proper reconstruction of the energy of high frequency partials which varies in accordance with the sinusoidal trajectories. This approach results in a better quality of the reconstructed signal in comparison to existing methods such as SBR. On the other hand, the proposed

technique requires rather small amount of data to transmit, because part of the required data, such as amplitudes of partials, is estimated on the decoder side based on the low band signal derived from core encoder (e.g. AAC).

7. ACKNOWLEDGEMENTS

This work was supported by the research grant 3 T11D 017 30 of the Polish Ministry of Science and Higher Education.

8. REFERENCES

[1] ISO/IEC International Standard 14496-3: "Coding of Audio-Visual Objects – Part 3: Audio", 3rd Edition, 2005.

[2] E. Larsen, R. M. Aarts, "Audio Bandwidth Extension", J. Wiley & Sons, Chichester 2004.

[3] M. Dietz, L. Liljeryd, K. Kjörling, O. Kunz, "Spectral Band Replication, a novel approach in audio coding", *112th AES Convention*, Munich, May 2002.

[4] A. J. S. Ferreira, D. Sinha, "Accurate Spectral Replacement", *118th AES Convention*, Barcelona, Spain, May 2005.

[5] R.J. McAulay, T.F. Quatieri, "Speech analysis/synthesis based on sinusoidal representation", *IEEE Trans. on ASSP*., vol 34, no. 4, 1986

[6] X. Serra, "Musical sound modeling with sinusoids plus noise", in C. Roads et al (eds) *Musical Signal Processing*, Sweets & Zeitlinger, 1997, pp. 91-122.

[7] M. Abe and JO Smith III: "Design Criteria for the Quadratically Interpolated FFT Method (I): Bias due to Interpolation", *Technical Report STAN-M-114, CCRMA*, Stanford University, 2004.

[8] M. Lagrange, S. Marchand, M. Raspaud, J. B, Rault, "Enhanced partial tracking using linear prediction", *Digital Audio Effects (DAFX-03) Conference*, London, UK, 2003.

[9] European Broadcasting Union, "Sound Quality Assessment Material" compact disk, a part of an EBU publication Tech 3253.