IMPROVED CODING OF TONAL COMPONENTS IN MPEG-4 AAC WITH SBR

Tomasz Żernicki, Marek Domański,

Poznań University of Technology, Chair of Multimedia Telecommunications and Microelectronics,
Polanka 3, 60-965, Poznań, Poland
phone: + (4861) 6653850, fax: + (4861) 6653899, email: tzernicki@multimedia.edu.pl, web: www.multimedia.edu.pl

ABSTRACT

Proposed is a new technique of effective regeneration of high-frequency tonal components in an augmented MPEG-4 AAC HE decoder. The basic idea is to synthesize the tonal components using the technique called synthetic sinusoidal coding that is already adopted in the MPEG-4 codecs in another context. Here, the idea is to mix this technique with standard Spectral Band Replication (SBR), i.e. to add some control information to a standard MPEG-4 AAC HE bitstream that is used to synthesize the high-frequency tonal components in a decoder. In that way, provided is proper synthesis of rapidly changing sinusoids as well as proper harmonic structure in the high-frequency band. The experiments show that the tool improves significantly the compression performance when added to an MPEG-4 AAC HE codec. This improvement has been confirmed by listening tests.

1. INTRODUCTION

Recently, audio compression came to a turning point when there appeared a family of advanced techniques that improved compression efficiency. The new audio coding systems related to the MPEG-4 AAC (Advanced Audio Coding) standard [1] continue the development of psychoacoustic codecs augmented by coding tools designed especially to deal with individual aspects of audio compression. In these new-generation codecs, significantly increased compression performance is obtained as a cumulative effect of application of many coding tools. One of them exploits the technique of spectrum replication from the family of audio bandwidth extension methods [2].

The High Efficiency Profile of the MPEG-4 AAC standard (MPEG-4 AAC HE) [1] includes a technique called Spectral Band Replication (SBR) [2-4]. The basic idea of SBR is based on the observation that the signal spectrum in the high-frequency bands is highly correlated with the signal spectrum in the low-frequency band. Therefore, it is possible to replace the high-frequency signal components with a modified version of the low-frequency band, avoiding the need to transmit the high-frequency signals at all. Additionally, the encoder transmits a very small amount of control information which the decoder uses to shape the spectrum in the high-frequency band. Moreover, the components representing sinusoids and noise that cannot be obtained by copying may be synthesized to limited extent at an SBR decoder. The bitstream of control parameters may be transmitted at very low bitrates of about 1-3 kbps. In a decoder, quite high subjective quality of the reconstructed audio signals may obtained because the human auditory system is relatively insensitive to signal distortions in high-frequency band (>6 kHz). 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 that the frequencies of the higher harmonics are shifted due to the process of band copying (see Fig.1). For some signals, such distortions may appear quite annoying.

In this paper, we propose a coding tool that reduces the above described distortions but also those produced in the case of quickly changing frequencies of harmonics that cannot be well tracked by the SBR decoder. The basic idea is to synthesize the tonal components using the technique called sinusoidal coding [5-9] that is already adopted in the MPEG-4 codecs in another context [1].

Here, the idea is to add some control information to a standard MPEG-4 AAC HE bitstream. That additional information will not interfere with the standard syntax and semantics of the MPEG-4 audio bitstream but it will be used in order to synthesize the high-frequency tonal components in a decoder. In that way higher quality of the reconstructed audio signal may be obtained thus shifting the rate-distortion curve of the encoder towards higher compression performance.

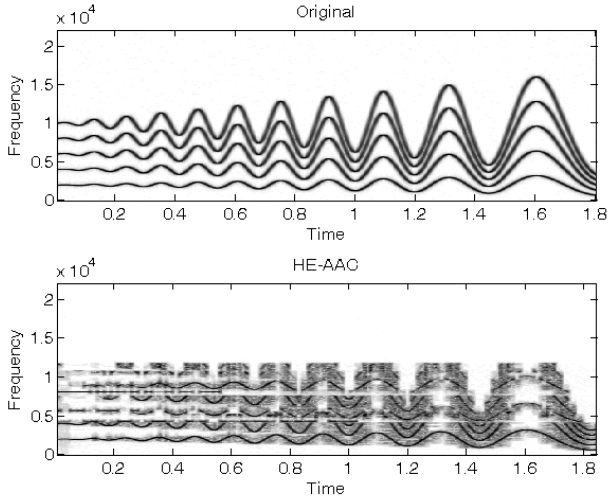

Figure 1. The spectrograms of an original tonal signal and its approximate reconstruction generated by the SBR decoder.

2. PROPOSED TECHNIQUE

The technique consists in augmenting the SBR tool by another tool based on sinusoidal modeling that is used to synthesize high-frequency harmonic components in a decoder. Parametric coding based on sinusoidal modeling has been already widely used in speech signals coding [5] but also in wideband audio coding according to MPEG-4 SSC (SinuSoidal Coding) [1]. The additional tool is used to process signal components above the f_{SBR} – the SBR cut-off frequency that typically is set with respect to the target bitrate. The signal components below f_{SBR} are encoded using classic perceptual coding (as described in MPEG-4 AAC) while the technique of SBR augmented by parametric coding of sinusoids is used for the frequencies above f_{SBR}.

The main task of the proposed tool is to eliminate the coding distortions caused by the SBR tool when fast frequency changes occur at high frequencies. Additionally, the proposed technique allows to keep the harmonic structure of tonal components, i.e. to preserve that frequencies of harmonics are integer multiples of the fundamental frequency from the low-frequency band.

The additional tool of sinusoidal modeling is implemented by two additional blocks, one in an encoder and the other in a decoder. These new blocks are linked by an additional bitstream of parameters of sinusoids (Fig. 2). Application of this tool does not affect the standard MPEG-4 AAC HE bitstream that is augmented only by the above mentioned stream of parameters.

In the encoder (Fig. 3), the above mentioned additional block identifies harmonic components in the spectrum above f_{SBR} and tracks them. Therefore, the parameters of sinusoids (A – amplitude, ω – radial frequency, φ - phase) are estimated in short windows in order to track their fast changes. The practical implementation employs the tracking technique [9] that does not presuppose any knowledge about the input signal, e.g. the structure of harmonics. The tracking technique should also deal properly with crossings of sinusoidal trajectories. It is important if there are many audio sources in the processed signal.

The parameters of all detected sinusoids (for the k-th sinusoid: A_k, ω_k, φ_k) are compressed for consecutive sample blocks. The compressed parameters are transmitted to the decoder where they are used to synthesize the sinusoids (Fig. 2).

The sinusoidal synthesis is also performed in the encoder – the synthesized signal s has to be subtracted from the input to the AAC encoder with SBR. In fact, the synthesized sinusoids s are used as a masker that masks the tonal components present in the input signal x (Fig. 2).

3. SINUSOIDAL MODELING

One of the main features of the proposed technique is application sinusoidal model [5,6] that represents a deterministic part of the signal as a sum of K quasi-sinusoidal components with time-variant parameters.

$$s_{\det}(t) = \sum\nolimits_{k=1}^{K} A_k(t) \cos\left(\varphi_k + 2\pi \int_0^t f_k(\tau)\, d\tau \right), \quad (1)$$

where $A_k(t)$, $\omega_k(t)$, $\varphi_k(t)$, k = 1, … ,K denote time-variant amplitude, radial frequency and phase, respectively.

The parameters $A_k(t)$, $\omega_k(t)$, $\varphi_k(t)$ are estimated in the process of sinusoidal analysis of the input signal x. Such an analysis has been already described in many papers [5-7,9]. The major steps of the analysis are shown in Fig. 3. The analysis starts with Short-Time Fourier Transformation (STFT). Then the tonal components must be identified (spectral peak detection and classification) and measured in order to estimate the parameters. In fact, the time-variant parameters are estimated in consecutive overlapping windows that contain N samples. In our implementation, there is $N = 1024$, and the windows overlap by $N/2$ samples.

The proposed technique exploits an extended version of the sinusoidal model from Eq. (1). Application of a first-order AM-FM model in the spectral analysis allows for estimation of the amplitude and frequency slope (modulation rate) according to the linear chirp model [8]

$$x(t) \cong \sum\nolimits_{k=1}^{K} A_k(t) \exp\left(\alpha_k t\right) \cos\left[\varphi_k + 2\pi \left(f_k t + \beta_k t^2\right)\right], \quad (2)$$

where α_k and β_k are the slope factors.

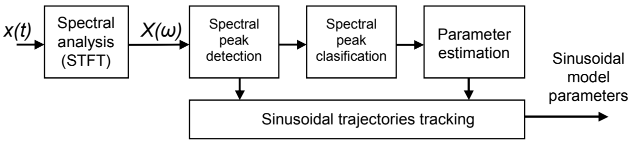
x(t)
Spectral analysis (STFT)
X(ω)
Spectral peak detection
Spectral peak clasification
Parameter estimation
Sinusoidal trajectories tracking
Sinusoidal model parameters

Figure 3. Sinusoidal analysis scheme.

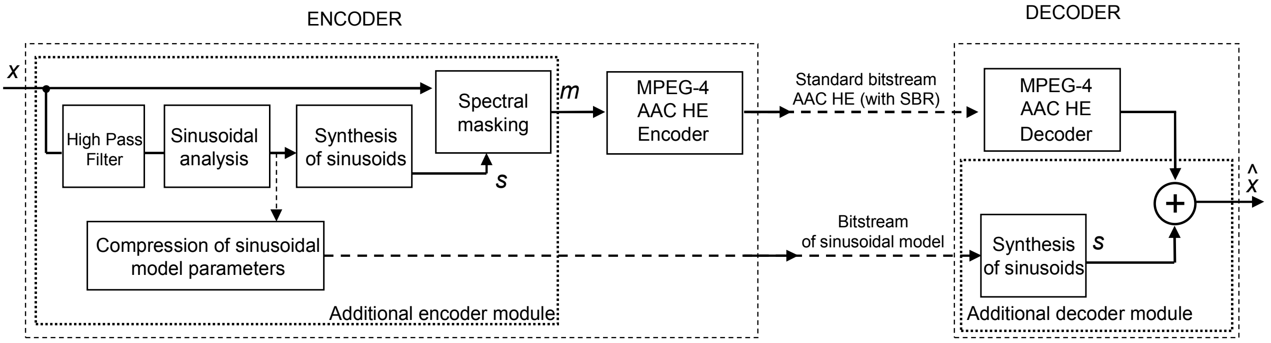

Figure 2. The modified MPEG-4 AAC HE codec with additional blocks related to sinusoidal modeling.

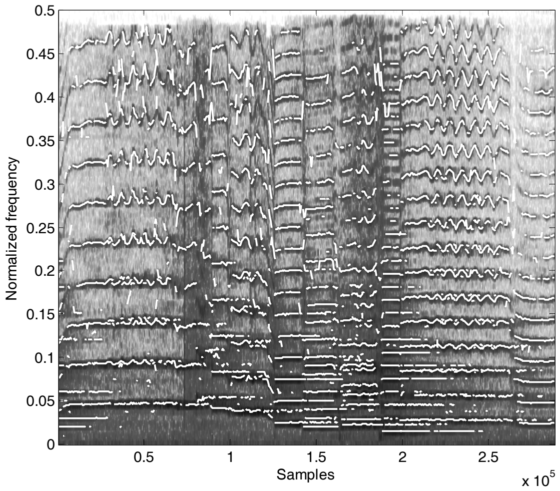

Figure 4. Typical sinusoidal trajectories shown on the background of an audio spectrogram.

In the next step, the sequences of parameters A_k, ω_k, φ_k are linked into sinusoidal trajectories that describe evolution of individual tones over time (Fig. 4). Tracking of sinusoidal trajectories has been already described in many papers, e.g. [5-7,9]. Here, in the implementation, a technique based on maximum likelihood is used.

Both in the encoder and in the decoder, the quasisinusoidal signals are synthesized from individual sinusoidal trajectories. For a given trajectory, values of frequency and amplitude are interpolated over time. In our implementation, the amplitude is linearly interpolated in the log (dB) domain, while the frequency is interpolated using cubic splines (Hermite function). Both in the encoder and in the decoder, the same signal *s* is synthesized that represents high-frequency harmonic components of the input audio signal *x*.

The number *K* (usually ≥ 10) of sinusoids tracked is adapted to the power of tonal components above f_{SBR}.

Next, the high-frequency tone components should be removed from the input to a classic MPEG-4 AAC HE encoder. In our codec, such removal is modeled by attenuation of the high-frequency tonal components in the input signal *x*. This attenuation is implemented as spectral masking by synthetic signal *s*.

$$M(k) = \frac{X(k)}{smooth\big(thr(|S(k)| \cdot \varepsilon)\big)}, \quad thr(x) = \begin{cases} x, & x > 1 \\ 1, & x \leq 1 \end{cases} \tag{3}$$

where:

M(k) – spectrum of the signal *m* with masked components,

S(k) – spectrum of masking (synthetic) signal *x*,

ε – masking threshold,

smooth(x) – convolution with a smoothing kernel.

The obtained signal *m* is encoded using the standard MPEG-4 AAC HE encoder, i.e. an encoder with SBR tool. In high frequencies, the signal m contains mainly noise components, which are less important, from the perceptual point of view, than tonal components. The noise-like components are very effectively encoded using the spectral band replication (SBR) technique. It needs to be outlined that SBR encoder does not transmit any information related to tonal components, e.g. scaling factors and parameters used for describing synthetic sinusoids. In this way the bitrate of SBR data stream is reduced.

The parameters of the sinusoidal model have to be transmitted to the decoder, therefore they also need to be compressed. In our implementation, uniform quantization, linear predictive coding (LPC) and Huffman coding has been used for that purpose. For the parameters of the sinusoidal model, Burg method of linear prediction is used [11]. The order of the predictor is 6 and there are 20 past samples used to estimate the predictor coefficients.

The above described technique improves reconstruction of the high-frequency tonal components as compared to the standard MPEG-4 AAC HE coding with the SBR tool (see Figs. 5, 6 and 7). For example, a trumpet produces a lot of high tones that are lost at an output of a standard MPEG-4 AAC HE decoder while being well reconstructed by the decoder proposed in this paper.

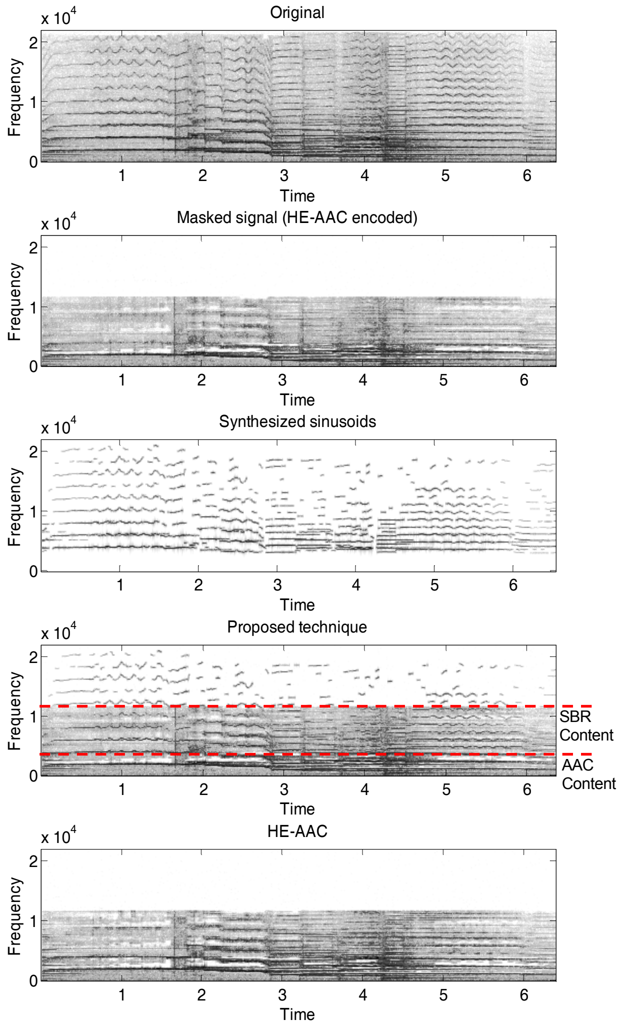

Figure 5. Exemplary spectrograms, from the top: 1) input signal, 2) masked signal *m* obtained from the spectral masking block, 3) synthesized sinusoids – signal *s*, 4) full reconstructed signal in the decoder – proposed technique, 5) full reconstructed signal in the decoder – MPEG-4 AAC HE.

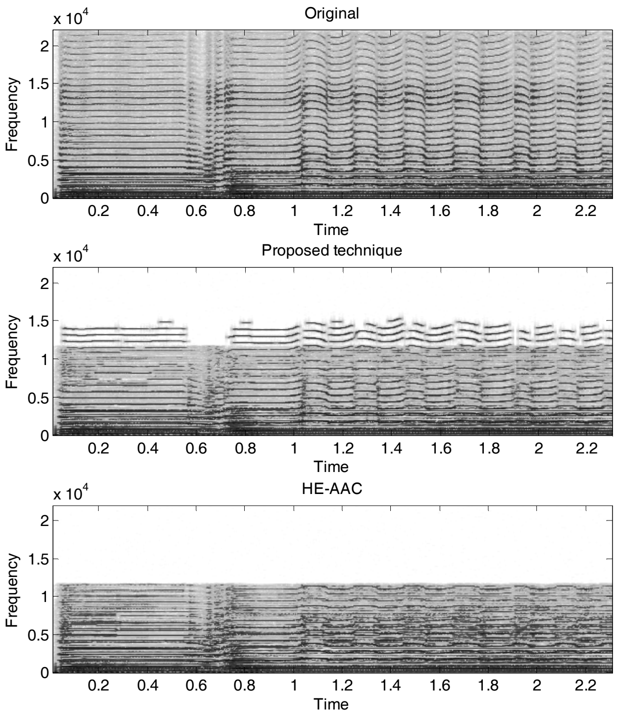

Figure 6. Spectrograms of a trumpet: 1) original, 2) decoded from the codec proposed, 3) decoded from standard MPEG-4 AAC HE codec with SBR.

In particular, Fig. 7. shows the reconstruction of the tonal components with quickly varying frequencies from Fig.1. The proposed codec outperforms clearly the standard MPEG-4 AAC HE (with SBR) – compare to Fig. 1, lower spectrogram). Please note that the SBR codec operates up to only 12 kHz at this low bitrate.

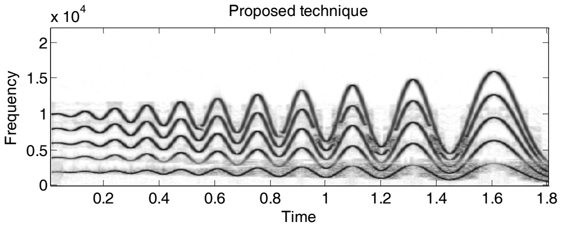

Figure 7. Spectrogram of reconstructed audio for the original from Fig. 1 (upper spectrogram).

4. EXPERIMENTS

In order to assess the efficiency of the technique, full encoder and decoder have been implemented as software that processes audio excerpts off-line. The experimental software was built on top of the 3GPP codec that is compliant with the MPEG-4 Audio standard, i.e. AAC High Efficiency Profile (with SBR). In the experimental software, the proposed additional blocks have been implemented by routines written in Matlab.

The experiments used wide range of audio excerpts, in particular from music records.

As described before, the new technique switches on when the audio analysis procedures identify strong tones above the SBR cutoff frequency f_{SBR}. In the absence of such tonal components in audio signal, the proposed technique has no impact on the efficiency of an MPEG-4 AAC HE audio codec. Therefore, the detailed listening tests have been performed for records of instruments with strong tonal elements in higher frequencies, e.g. the excerpts of violin, accordion, and trumpet from the standard EBU test material [12].

Due to results of these experiments it was established that only about 10 sinusoidal trajectories need to be coded in order to get higher audio quality than from MPEG-4 AAC HE (with SBR) at the same total bitrate. In the experimental implementation, encoded parameters of high-frequency sinusoidal trajectories needed about 3 – 6 kbps. These bitrates probably can be reduced by improving the compression technique used for these parameters.

The main objective of the listening tests is to compare the compression efficiency of the two audio codecs: the original MPEG-4 AAC HE (with SBR) and the same codec augmented with the tools proposed in this paper. The testing procedure was compliant with the ITU-R Recommendation BS.1534 [10] (MUSHRA – „Multi Stimulus test with Hidden Reference and Anchors"). This subjective quality methodology was chosen because it was developed for assessment of medium-quality audio sequences with significant distortions, e.g. resulting from compression. All subjective tests have been made for a group of 15 listeners who assessed 5 audio excerpts with sounds of musical instruments. Each excerpt was presented in 4 versions, i.e. decoded by the standard decoder and the proposed one, by 16 and 20 kbps in both cases. The original excerpts as well as their lowpass-filtered versions have been also heard by the listeners. Two types of lowpass-filtered excerpts have been used: with bandwidth limited to 3.5 kHz and 7 kHz, respectively. The results (Table 1) prove clearly that the proposed codec outperforms significantly the standard MEG-4 AAC HE codec for such musical sequences. Significant improvement was observed for all 5 excerpts and both bitrates.

Table 1. Results of subjective listening tests.

Signal	**MUSHRA score [%]**			
	16 kbps		**20 kbps**	
	HE-AAC	**Proposed technique**	**HE-AAC**	**Proposed technique**
accordion	53	79	60	80
brass	50	79	42	80
pipes	46	84	68	79
trumpet	81	84	76	75
violin	40	87	51	92
AVERAGE	54	83	60	81

The MUSHRA score is the statistical measure of subjective quality. 100% means imperceptible difference with the original signal.

The experiments also used hidden reference and the band-limited versions of the uncompressed excerpts (see Figs. 8-10): anchors 3.5 kHz and 7 kHz. The scores for the hidden reference are slightly below 100% but the level 100%

remains within the 95% confidence intervals. For the proposed technique, the average results are clearly better than those obtained from standard MPEG-4 AAC HE codec for all 5 excerpts and both bitrates (Fig. 8).

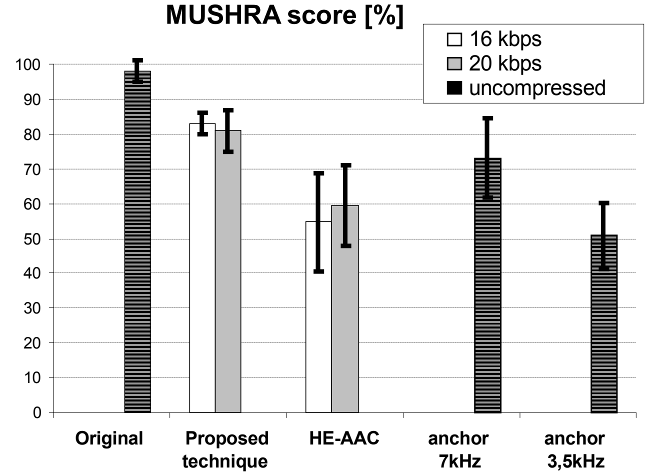

Figure 8. Average results of MUSHRA subjective listening tests for two bitrates of 16 and 20 kbps. The scores for the hidden reference (Original) and two anchors are given for comparison. The 95 % confidence intervals (15 listeners) are plotted for all cases.

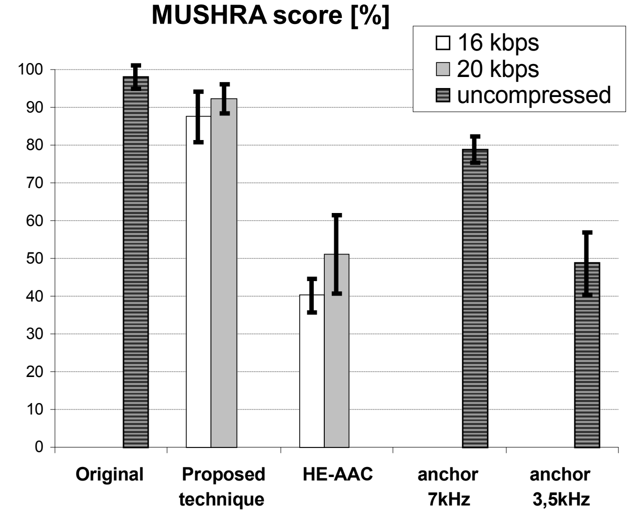

Figure. 9. Results of MUSHRA subjective listening tests for "violin" excerpt. The 95 % confidence intervals (15 listeners) are plotted.

Figure 9 proves that the proposed technique is well suitable for signals with strong, stable and quickly varying sinusoidal trajectories where pure MPEG-4 AAC HE is unable to reconstruct properly those signal components.

5. CONCLUSIONS

In this paper, a new compression technique has been proposed for applications in low-bitrate coding of wideband audio. The technique is proposed as an additional tool for MPEG-4 AAC and similar codecs with the SBR in order to improve their performance in higher frequencies. The improvement of the "rate-distortion" performance has been well proved using the standardized subjective listening tests.

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] D. Homm, T. Ziegler, R. Weidner, R. Bohm, „Bandwidth extension of audio signals by spectral band replication", *Proc. 1st IEEE Benelux Workshop on MPCA*, Louvain 2002.

[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. Lagrange, S. Marchand, J. B. Rault, "Enhancing the Tracking of Partials for the Sinusoidal Modeling of Polyphonic Sounds", *IEEE Trans. Audio, Speech and Language Proc.*, Vol. 15, July, 2007.

[8] M. Abe and J. O. Smith, "AM/FM rate estimation for time-varying sinusoidal modeling", in *Proc. Int. Conf. on Acoustics, Speech and Signal Processing*, 2005, pp. 201 - 204 (Vol. III).

[9] M. Bartkowiak, T. Żernicki, "Improved partial tracking technique for sinusoidal modeling of speech and audio", *Poznan University of Technology Academic Journals: Electrical Engineering*, No. 54/2007,
web: http://www.multimedia.edu.pl/publications/

[10] ITU-R, BS.1534, "Method for the subjective assessment of intermediate quality levels of coding systems", 2003.

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

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