A HARMONIC BANDWIDTH EXTENSION METHOD FOR AUDIO CODECS

Frederik Nagel

Fraunhofer Institute for Integrated Circuits (IIS)
Am Wolfsmantel 33
91058 Erlangen, Germany

Sascha Disch

Leibniz Universität Hannover
Schneiderberg 32
30167 Hannover, Germany

ABSTRACT

Today's efficient audio codecs for low bitrate application scenarios often rely on parametric coding of the upper frequency band portion of a signal while the lower frequency band portion of the same is conveyed by a waveform preserving coding method. At the decoder, the upper frequency signal is approximated from the lower frequency data using the upper frequency band parameters. However, commonly used methods of bandwidth extension almost inevitably suffer from a sensation of unpleasant roughness, which is especially present for tonal music items. In this paper we expose the origin of the roughness and propose a bandwidth extension method, which does not introduce roughness into the reconstructed audio signal. A listening test demonstrates the advantage of the proposed method compared to a standard bandwidth extension.

Index Terms— Audio coding, Audio systems, Vocoders

1. INTRODUCTION

Audio bandwidth extension (BWE) is a standard technique within modern audio codecs to efficiently code wide-band audio signals at low bitrates. In the past, codec bitrate constraints have been accounted for by simply lowpass filtering the audio. BWE instead relies on a parametric representation of the high-frequency band (HF) which is estimated from the low-frequency band (LF) signal. Even though high frequency content of the audio material is preserved in such audio codecs, this sometimes comes at the price of undesired auditory artifact perceptions, such as roughness, as will be described later.

Audio codecs utilizing BWE functionality, of which the most well known is HE-AAC [1], are applied in contemporary mobile multimedia players, mobile phones and digital radio services.

In this paper, we will briefly review existing methods of BWE that can be found in the literature. Then we will identify prominent sources of roughness artifacts induced by BWE. Subsequently we propose an algorithm, which circumvents these problems. To prove the effectiveness of our solution, we present listening results using an enhanced HE-AAC v2 codec [2] at bitrates of 12 and 16 kbit/s for monophonic audio in two versions - the fully HE-AAC V2 compliant 'Spectral Band Replication' (SBR) [3, 4] and a modified version implementing the proposed method. Finally, we conclude with the discussion.

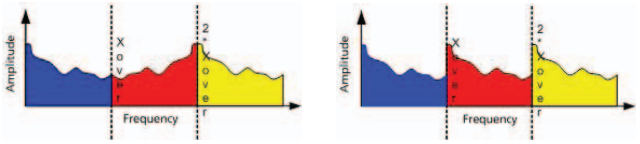
2. BACKGROUND

2.1. Bandwidth extension principle

Besides unguided bandwidth extension methods [5], commonly used BWE utilizes a parametric representation of the original HF content to synthesize an approximation of the HF from the LF data. One step of this processing is typically a frequency translation operation, which derives the HF 'raw' spectrum from the LF spectrum prior to parametric post-processing.

As such, the LF portion can for instance be simply copied within a filterbank representation to the HF location as done in SBR [3, 4]. Alternatively, higher frequencies can be generated by either non-linear processing [5] or upsampling [6]. Figure 1 shows the different so called patching methods for BWE; a comprehensive overview can be found in [7].

The spectral shape of the upper frequencies can be post-processed for instance with the help of scale factors, LPC or cepstral coefficients [3, 4, 7]. Subsequently, tonality is adapted and, if necessary, missing sinusoids are added.



(a) Upsampling (Mirroring) (b) Copying or modulation

Fig. 1. Amplitude spectra resulting from two different BWE patching methods.

2.2. Auditory roughness and timbre

Audio signals that are treated by BWE sometimes suffer from severe artifacts, such as auditory roughness and unpleasant

timbre.

Roughness is the perception of rapidly changing amplitude of a tone. A sound consisting of two sinusoids is perceived as rough, if the difference in frequencies between the two tones is between 30 and 600 Hz, i. e. if there is an amplitude modulation with a frequency between 15 and 300 Hz. The intensity of the perceived roughness also depends on the spectral position of the tones; it is maximum at 1 kHz [8]. Timbre describes the characteristic sound of an instrument or voice [9].

2.3. Bandwidth extension artifacts

BWE artifacts occur particularly at low bitrates when only a small LF bandwidth can be afforded. The primary source of roughness artifacts was found to be the patching operation (see subsection 2.1), which translates the LF to raw HF. A simple copy operation corresponds to a spectral shift and does not preserve the harmonic relations of tonal components of the signal.

In addition, in the boundary region between LF and HF undesired beating effects between LF and synthesized HF can occur if tonal peaks from each section are placed in spectral vicinity to each other due to the copying or mirroring operation. This leads to the perception of auditory roughness. Figure 3(b) visualizes this problem caused by traditional SBR processing.

Furthermore, since the width of critical bands increases with the frequency [8], sinusoidal peaks, originally located in different critical bands in the LF region, now occupy one critical band in the HF part and are thus resolved differently by human auditory perception: the sinusoids fuse into one tone exhibiting temporal amplitude modulation. This is also perceived as a rough and unpleasant auditory sensation.

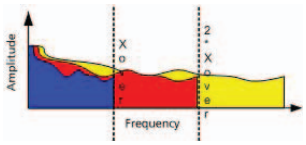
3. METHOD

3.1. Spectral stretching

Our new approach avoids the aforementioned problems resulting from mirroring or copying operations. By stretching the LF spectrum as seen in Figure 2, harmonic continuation in the HF is ensured intrinsically. Therefore we refer to this method as 'Harmonic Bandwidth Extension' (HBE).

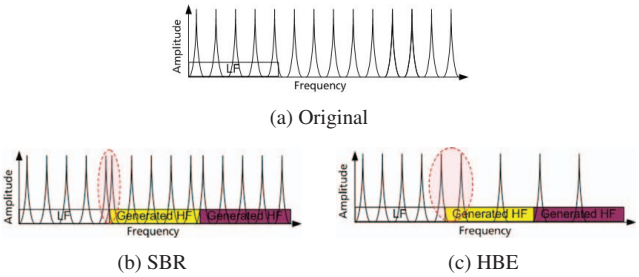
The spectral stretching is implemented using phase vocoders [10, 11]. At the phase vocoder, grains of length Γ are taken from the signal using an analysis hop size Λ and transformed to frequency domain. In a next step, all DFT phases are multiplied by the stretching factor Ψ. After IDFT synthesis, the grains are re-combined by overlap-add with a different synthesis hop size $\Psi\Lambda$. This effects in a time dilatation or compression of the original signal. A final decimation by factor Ψ results in a signal having a stretched spectrum at unchanged temporal duration.

We suggest using a block length of $\Gamma = 1024$ samples which is equivalent to 32 ms at 32 kHz sampling rate and a hop size of $\Lambda = 128$ samples (4 ms). The DFT analysis window is a flat top window with cosine roll-off of 170 samples. The synthesis window is of the same size and is calculated as the quotient of a flat top window with a cosine square roll-off of 256 samples and the analysis window. New investigations have shown possible advantages of different analysis windows having a better nearby sidelobe attenuation, such as the hamming or the bartlett window. The question of an optimal analysis window for the phase vocoder has thus to be addressed precisely in further investigations.



Stretching (Phase Vocoder)

Fig. 2. Amplitude spectra resulting from the phase vocoder method.



(a) Original

(b) SBR

(c) HBE

Fig. 3. Amplitude spectra of strictly harmonic original signal and bandwidth extended versions.

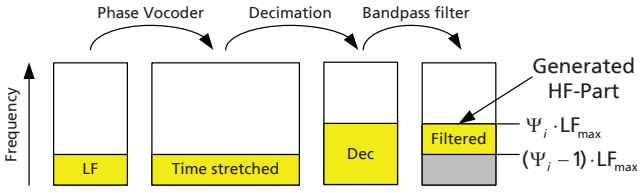


Fig. 4. Steps of harmonic bandwidth extension. HBE consists of the application of a phase vocoder plus decimation. Additionally a bandpass filter is applied in order to achieve only the desired part of the spectrum.

3.2. Harmonic bandwidth extension (HBE)

The HBE method employs multiple phase vocoders operating in parallel in order to obtain the final HF patch. Figure 4

illustrates the processing steps inside one of these phase vocoders. The LF signal with the maximum frequency LF_{max} is time stretched by different integer factors Ψ_i, downsampled by Ψ_i and subsequently bandpass filtered to the range $[(\Psi_i - 1)\mathrm{LF}_{max} : \Psi_i \mathrm{LF}_{max}]$. The highest stretching factor Ψ_{max} is determined by the desired frequency f_{max} to be synthesized. In a last step, the contributions from all vocoders are summed up to form the patch which is used to substitude the traditional patch of standard SBR. The entire processing scheme of HBE is displayed in Figure 5.

After the HF part of the signal has been generated, the spectrum is shaped for instance by scale factors, LPC, or cepstral coefficients. Additionally, the tonality is adapted according to original tonality and, if indicated, missing sinusoids are added like done in standard SBR. For our experiments, we used an enhanced SBR (eSBR), which utilized all tools from SBR, but replaced its patching algorithm by the phase vocoder technique.

The spectrum is hence spread with increasing frequency, i. e. at higher frequencies it is less dense compared to lower frequencies. Due to the integer stretching factors, the spectrum is always harmonic. In particular, no unwanted roughness sensation due to beating effects can emerge at the border between LF and HF and between different HF parts. This is depicted in Figure 3(c).

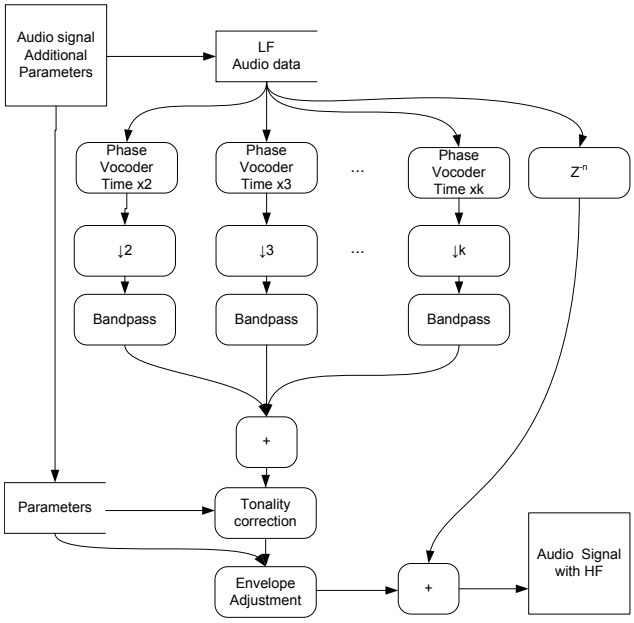


Fig. 5. Block diagram of processing data with HBE.

4. EXPERIMENT

The proposed method is compared with SBR, a well-known BWE method used in HE-AAC. We used an enhanced HE-AAC codec [2] and replaced the standard SBR patching procedure, which is basically a copy operation within a QMF representation, by an HBE time domain processing as described in subsection 3.2. The HE-AAC core coder had a bandwidth of LF_{max} = 4 kHz, which was extended by either SBR or HBE to 12 kHz for 12 kbit/s and 13.5 kHz for 16 kbit/s; input sample rate was 32 kHz.

18 listeners participated in the listening test, 17 male and one female with a mean age of 27 (SD=4). 10 listeners were expert listeners. Ratings were given within a paired comparison blinded test according to ITU-R BS.1284 on a 7-point comparison scale (-3 to 3). The test consisted of 7 items, 1 voiced item, 4 music items and 2 items with mixed content (speech + music). The items were presented coded at bitrates of 12 kbit/s and 16 kbit/s mono using SBR and HBE, respectively, alongside the original. All trials were replayed twice in randomized order, which was kept constant for all listeners. Thus each listener had to rate 28 trials in total. The items were replayed from a fanless computer equipped with a professional sound card. Stax headphones and amplifier were used.

5. RESULTS

The arithmetic means from the two ratings for SBR and HBE were calculated for each listener and both bitrates 12 and 16 kbit/s. Positive values indicated that HBE was rated higher than SBR.

The HBE items were rated highly significantly better ($p = .01$) than the SBR for five of seven items for both bitrates as can be seen in Figure 6(a). HBE is also evaluated significantly better overall as Figure 6(b) shows. Two items, however, were not determined as better: one singing solo voice (es01) and one pizzicato guitar item (Music_3). For 12 kbit/s, both items are barely significantly worse on a 99%-level of significance. No significant differences ($p = .05$) were found between expert and non-expert listeners and no dependency on age could be observed ($\rho = .06; p > .33$).

6. DISCUSSION AND CONCLUSION

The new bandwidth extension scheme HBE showed improved performance compared to SBR. 5 of 7 items gained from the new method and the overall mean was highly significantly positive, indicating the advantage of HBE. Two items, however, were rated worse or at least not better. For speech items and voice, the current version of HBE appears not to be the ideal BWE method. An additional experiment with many speech items showed that SBR is better suited for speech. This might be explained by the exact spectral periodicity of voiced speech signals due to the glottal pulse train type of excitation. By application of HBE, many of the resulting harmonics are removed which leads to a changed timbre, which listeners apparently do not like (Figure 3). So signal adaptive

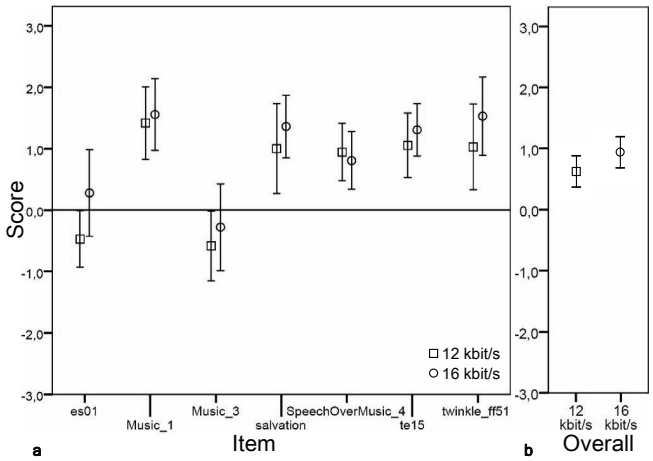


Fig. 6. Listening test results with 12 kbit/s and 16 kbit/s coded items. (a) HBE outperforms SBR significantly ($p = .01$) in 5 of 7 cases (significant positive values). (b) HBE outperforms SBR overall.

switching between SBR patching for voiced speech and HBE patching for music is a viable option [2].

The bad rating for the Music_3 item can be explained by a lack of exactness of the onsets of the guitar notes. In a straightforward implementation of a phase vocoder, transient events loose their original vertical coherence. This can result in pre- and post-echoes and 'phasiness' artifacts. Several techniques for transient handling in a phase vocoder exist [12, 13, 14], which could therefor be incorporated in this new method. Alternatively, HBE could solely be used for signal parts in which no transients occur.

Recently, a new type of vocoder has been presented, which could be used alternatively for the transposition stages in HBE [15]. It uses a signal decomposition into subband carriers and their modulation components and offers an intrinsic envelope preservation property. Using this type of vocoder inside HBE will be a potential topic for future research.

In general, even without transient handling, HBE showed a highly significantly improved performance compared to SBR and can thus improve new audio codecs, such as presented in [2].

7. REFERENCES

[1] ISO/IEC 14496-3:2005, *Information technology - Coding of audio-visual objects - Part 3: Audio*, ISO, Geneva, Switzerland.

[2] M. Neuendorf, P. Gournay, M. Multrus, J. Lecomte, B. Bessette, R. Geiger, S. Bayer, G. Fuchs, J. Hilpert, N. Rettelbach, R. Salami, G. Schuller, R. Lefevbre, and B. Grill, "Unified speech and audio coding scheme for high quality at low bitrates," *IEEE Int. Conf. on Acoustics, Speech and Signal Proc.*, 2009.

[3] Martin Dietz, Lars Liljeryd, Kristofer Kjörling, and Oliver Kunz, "Spectral band replication, a novel approach in audio coding," in *112th AES Convention*, Munich, 2002.

[4] Stefan Meltzer and Reinhold Böhm, "SBR enhanced audio codecs for digital broadcasting such as 'Digital Radio Modiale' (DRM)," in *112th AES Convention*, Munich, Germany, 2002.

[5] Erik Larsen, Ronald M. Aarts, and Michael Danessis, "Efficient high-frequency bandwidth extension of music and speech," in *112th AES Convention*, Munich, 2002.

[6] Bruno Bessette, Redwan Salami, Roch Lefebvre, Milan Jelinek, Jani Rotola-Pukkila, Janne Vainio, Hannu Mikkola, and Kari Järvinen, "The adaptive multirate wideband speech codec (amr-wb)," *IEEE Transactions on Speech and Audio Processing*, vol. 10, no. 8, pp. 620–636, 2002.

[7] Erik Larsen and Ronald M. Aarts, *Audio bandwidth extension : application of psychoacoustics, signal processing and loudspeaker design*, John Wiley and Sons Ltd, Hoboken, 2004.

[8] Eberhard Zwicker and Hugo Fastl, *Psychoacoustics: facts and models*, Springer series in information sciences. Springer, Berlin, 2. edition, 1999.

[9] William A. Sethares, *Tuning, timbre, spectrum, scale*, Springer, Berlin, 2. ed edition, 2005.

[10] J. L. Flanagan and R. M. Golden, "Phase vocoder," *Bell System Technical Journal*, pp. 1493–1509, 1966.

[11] Mark Dolson, "The phase vocoder: A tutorial," *Computer Music Journal*, vol. 10, no. 4, pp. 14–27, 1986.

[12] A. Röbel, "A new approach to transient processing in the phase vocoder," in *6th Conference on Digital Audio Effects (DAFx-03)*, London, 2003, pp. 344–349.

[13] J. Laroche and M. Dolson, "Improved phase vocoder time-scale modification of audio," *IEEE Transactions on Speech and Audio Processing*, vol. 7, no. 3, pp. 323–332, 1999, 1063-6676.

[14] Chris Duxbury, Mike Davies, and Mark B. Sandler, "Improved time-scaling of musical audio using phase locking at transients," in *112th AES Convention*, Munich, 2002, Audio Engineering Society.

[15] Sascha Disch and Bernd Edler, "An amplitude- and frequency modulation vocoder for audio signal processing," *Proc. DAFX-08 Conference on Digital Audio Effects*, 2008.