SINUSOIDAL SUBSTITUTION - AN INTEGRATED PARAMETRIC TOOL FOR ENHANCEMENT OF TRANSFORM-BASED PERCEPTUAL AUDIO CODERS

Sascha Disch[1], Benjamin Schubert[1]

[1]Fraunhofer Institute for Integrated Circuits IIS, Erlangen, Germany

ABSTRACT

Transform-based audio coders are the preferred technique for music data compression. However, at low bitrates, traditional coders based on Modified Discrete Cosine Transform are prone to strong warbling and roughness artifacts originating from sparsely coded tonal components. Parametric coders, in turn, suffer from an unpleasantly artificial sound and do not scale well up to perceptual transparency. Hybrid transform-based and parametric coding could potentially overcome the limits of the individual approaches. Yet, existing hybrid coders are hampered by the lack of integrative interplay between both techniques. We outline our ideas how to tightly integrate transform-based coding and parametric coding to obtain an enhanced perceptual quality and scalability. Also, we provide listening test results which demonstrate the benefits of our hybrid coder design.

Index Terms— Codecs, Parametric Audio Coding, Signal Synthesis

1. INTRODUCTION

Modern perceptual audio coders are required to deliver satisfactory audio quality at increasingly low bitrates. Most prominent examples are MPEG-2/4 Advanced Audio Coding (AAC) [1] and Unified Speech and Audio Coding (USAC) [2]. USAC comprises a switched core consistent of an Algebraic Code Excited Linear Prediction (ACELP) module primarily intended for low bitrate speech coding plus a Transform Coded Excitation (TCX) module [3] and, alternatively, an enhanced AAC module mainly intended for coding of music. Like AAC, also TCX is a transform-based coding method. While USAC supports efficient parametric coding of speech, the coding of music is still essentially left to the transform-based modules.

Most notably at low data rates, given the limited bit budget usually only few frequency lines of the transform spectra are coded to be non-zero. As a consequence, temporal modulation artifacts and so-called warbling artifacts are inevitably introduced into the coded signal, especially if the underlying transform is a Modified Discrete Cosine Transform (MDCT) [4]. Additionally, often the permissible latency is also very low, e.g. for bi-directional communication applications or distributed gaming, etc. A transform window shape that obeys strict delay constraints further emphasizes the warbling problem by inducing significant crosstalk between spectral lines due to an increased leakage effect. Most prominently, these types of artifacts are perceived in quasi-stationary tonal components.

Figure 1 shows spectrograms of 3 sinusoids with different frequencies, original (top panel) and with processing by MDCT, quantization with all spectral lines except the one with maximum absolute amplitude quantized to zero, and IMDCT (bottom panel). The warbling is clearly visible as slow temporal beating-like modulation.

For very low bitrates, fully parametric audio coders have been standardized, the most prominent of which are MPEG-4 Part 3, Subpart 7 Harmonic and Individual Lines plus Noise (HILN) [5] and MPEG-4 Part 3, Subpart 8 SinuSoidal Coding (SSC) [1][6].

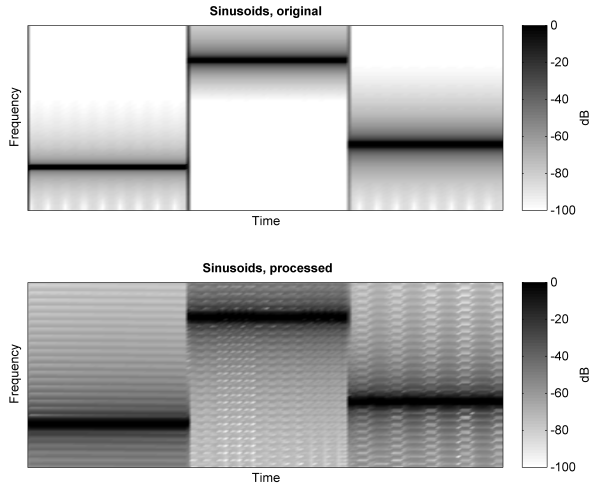

Fig. 1. Spectrogram of 3 sinusoids: original (top panel) and after MDCT, quantization and IMDCT processing with all spectral lines but the maximum line zeroed (bottom panel).

In [7], a hybrid of transform-based coding and parametric MPEG-4 SSC (sinusoidal part only) coding is proposed. In an iterative process, sinusoids are extracted and subtracted from the signal to form a residual signal to be coded by transform-based coding techniques. The extracted sinusoids are coded in

978-1-4799-2893-4/14/$31.00 ©2014 IEEE

a set of parameters and transmitted alongside with the residual. Also, the authors of [8] propose a hybrid coding approach that codes sinusoids and residual separately.

2. TRANSFORM-BASED CODER VERSUS PARAMETRIC CODER

At high or medium bitrates, transform-based coders are well-suited for coding of music due to their natural sound. Here the transparency requirements of the underlying psychoacoustic model are fully or almost fully met. However, at low bitrates, coders have to seriously violate the requirements of the psychoacoustic transparency model and in such cases transform-based coders are prone to warbling, roughness and musical noise artifacts.

Fully parametric audio coders are most suited for lower bitrates, but are known to sound unpleasantly artificial. Moreover, these coders do not seamlessly scale up to perceptual transparency at higher bitrates, since a gradual refinement of the rather coarse parametric model is not feasible.

Hybrid transform-based and parametric coding for music could potentially overcome the limits of the individual approaches, but is hampered in contemporary coder designs by a lack of interplay between the transform-based coding part and the parametric part of the hybrid coder. For example, problems relate to the signal division between parametric and transform-based coder part, bit budget stirring between transform-based and parametric part, efficient parameter signaling techniques, and seamless merging of parametric and transform-based coder output.

We present a way around the above mentioned limitations by application of sinusoidal substitution within a transform-based coder. In the proposed hybrid encoder, beginning from a lower cut-off frequency, local tonal regions that consist of groups of spectral lines, reaching between neighboring local minima and each encompassing a local maximum, are substituted by single so-called pseudo-lines, each having a similar energy as said regions to be substituted. This makes it possible that in the encoder the pseudo-lines can be handled by the subsequent regular psychoacoustic calculations, the quantizer and the noiseless coding unit just like any regular true spectral line. At the same time, this yields more sparsely populated spectra by confining the total energy of tonal components into a single stable spectral line, thereby avoiding the unstable spectral spreading of tonal energy in the MDCT that causes the above mentioned warbling artifacts. For synthesis, the pseudo-lines are erased from the spectrum and synthesized through insertion of t/f-adapted spectral tone patterns into the MDCT spectra prior to IMDCT calculation, as explained in detail in [9].

The proposed sinusoidal substitution can be integrated into existing transform-based coding schemes like AAC, TCX, etc. Stirring of the parameter quantization precision is implicitly performed by the existing rate control of the perceptual coder. This is very different from legacy hybrid coders containing sinusoidal modeling. In these coders, sinusoidal parameters are estimated and synthesized sinusoids are iteratively subtracted from the signal to obtain the residual which is transform coded. Here, parametric data and waveform data have their own bit budgets and the psychoacoustic model of the transform-based coder does not interact with the parametric coder (e.g. to determine number and quantization precision of the sinusoidal parameters).

3. PROPOSED HYBRID CODER

3.1. Sinusoidal Substitution - Encoder

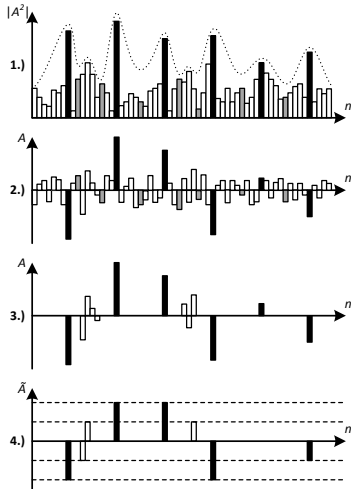

Fig. 2. Encoding process employing sinusoidal substitution.
Panel 1: calculation of a smoothed envelope (dashed line) based on the power spectra and determination of local centers of gravity (black bars) plus surrounding minima (gray bars).
Panel 2: insertion of scaled pseudo-lines (black bars) into the transform amplitude spectra containing true spectral lines (white bars).
Panel 3: deletion of areas bordering pseudo-lines from minima to minima.
Panel 4: quantization of hybrid spectra.

In sinusoidal substitution, local peaky areas of transform spectra are fully substituted by single sinusoids. Ideally, the replacement sinusoid has a frequency of the spectral local center of gravity (COG) and an energy of the original peaky spectral area. Suitable peaky areas, ranging between two local minima and encompassing a local maximum, are detected on a smoothed and slightly flattened spectral representation and are selected for substitution according to criteria such as peak height and peak shape. Such an analysis algorithm is described e.g. in [10]. The spectral resolution of such a peak detection can be higher than that of the transform-based coder if required to not exceed the Just Noticeable Difference

of Frequency Discrimination (JNDF) [11][12]. The replacement sinusoids are then represented by energy-scaled spectral pseudo-lines that are directly inserted into the spectra before quantization. The frequency of the replacement sinusoid is represented by the spectral position of the pseudo-line and can be further refined by fine-grain offset side information.

Figure 2 visualizes the encoding process of sinusoidal substitution. First, a smoothed envelope (dashed line) is calculated on the power spectra $\left|A^2\right|$ (panel 1). Next, in the smoothed envelope local centers of gravity and surrounding minima are determined leading to the black and gray bars visualized in Figure 2, respectively. In the transform spectra with amplitudes A (panel 2) that consist of true spectral lines (white bars), additionally scaled pseudo-lines (black bars) are inserted at spectral locations of gravity centers. Next, the surrounding areas of pseudo-lines are deleted from adjacent minima to minima (panel 3). Finally, the hybrid spectra are quantized by the perceptual coder into spectra with amplitudes $\widetilde{A}$ (panel 4). Please note that pseudo-lines are handled just like any other spectral line, and therefore may be also quantized to zero if deemed necessary by the perceptual model.

3.2. Sinusoidal Substitution - Decoder

Due to the removal of the entire spectral region governed by a tonal peak, the inserted pseudo-lines are always surrounded by zero-valued lines. Hence, in the decoder, pseudo-lines can be detected in a quantized spectrum by searching for isolated spectral lines. Nevertheless, especially for low bit rates, it might happen due to coarse quantization that also isolated lines that do not represent pseudo-lines occur. Thus additional information is transmitted to disambiguate lines and pseudo-lines through an array containing a binary value for each isolated line in the quantized spectrum. The flag array is generated on the actual quantized spectrum applying an iterative procedure, where the quantized spectrum, the resulting size of the flag array and, finally, the necessary side information are obtained. Additionally, fine-grain spectral offset information for each pseudo-line is needed in the decoder to generate the replacement sinusoid. This information is efficiently and conveniently conveyed through the polarity of the pseudo-lines, whereas a positive sign represents a half-bin offset (sinusoid is between two bins), and a negative sign indicates a zero offset (sinusoid is on-bin). Besides its simplicity, this approach does not cause additional bit consumption since quantizer and noiseless coding stage of transform-based coders are adapted to handle signed values. For synthesis, the pseudo-lines in the hybrid spectra are replaced by spectral tone patterns before calculating the inverse transform (IMDCT). Figure 3 visualizes the decoding process of sinusoidal substitution. First, in the transmitted hybrid spectra $\widetilde{A}$ (panel 1), true lines (white bars) and pseudo-lines (black bars) are identified. Subsequently (panel 2), all pseudo-lines are substituted

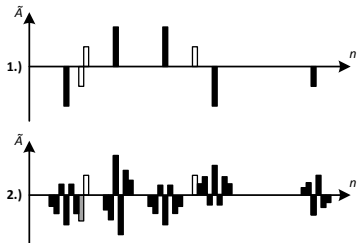

Fig. 3. Decoding process employing sinusoidal substitution.
Panel 1: identification of true lines (white bars) and pseudo-lines (black bars) in the transmitted hybrid spectra.
Panel 2: substitution of all pseudo-lines by spectral tone patterns (black bars), whereas true lines and tone patterns may overlap additively (gray bar).

by fine-grain offset adjusted, t/f-adapted and scaled sinusoidal spectral tone patterns (black bars). True spectral lines are either unaffected by sinusoidal substitution (white bars) or additively mixed with overlapping parts of tone patterns (gray bar). The substituted spectra are subsequently passed to the synthesis stage via IMDCT transform.

4. RESULTS

A MUltiple Stimuli with Hidden Reference and Anchor (MUSHRA) test [13][14] was conducted to evaluate the perceptual quality of the proposed hybrid coder using sinusoidal substitution. The test set consisted of 10 music items as listed in Table 1, spanning from classical orchestra to pop music. Each of these test items predominantly contains tonal instruments, e.g. bowed strings, solo brass or plucked guitar. A typical low bitrate transform-based coder scenario was chosen using a 256 band MDCT, a sampling frequency of 12.8 kHz, a bit rate of 13.2 kbps and a half-bin resolution (12.5 Hz) for the sinusoidal substitution. Sinusoidal substitution started above 800 Hz, and the maximum number of substituted sinusoids per time frame was 20. Tested conditions are listed in Table 2. «Plain core» denotes the transform-based coder outlined above, «core + LTP» is a similar core coder that additionally uses Long Term Prediction (LTP) which exploits inter-frame correlation via prediction of a common fundamental frequency of the audio content and is well-known especially in speech coding to improve perceptual quality of voiced (tonal) signals [15], and finally, «hybrid core» denotes the transform-based coder outlined above enhanced by the proposed sinusoidal substitution.

The perceptual quality of the items is rated on a scale ranging from excellent (100) to good, fair, poor and down to bad (0). 12 expert listeners participated in the test. The tests were performed in a dedicated listening test environment using high-quality electrostatic STAX headphones. Figure 4 and Figure 5 show the mean MUSHRA scores along with their 95% confidence intervals, assuming a Student's t-

	Item	Content
a	Björk (The Anchor Song)	Polyphonic brass
b	Brahms	Classic orchestra
c	Toni Braxton (Fairy Tale)	Classic guitar duo
d	Miles Davis	Trumpet jazz
e	Fiedel	Solo violin
f	Herbie Hancock	Trumpet jazz
g	Pitchpipe	Solo pitchpipe
h	Dire Straits (Your Latest Trick)	Saxophone pop
i	Susan Rosenberg	Solo harp
j	Steely Dan (Home at Last)	Funk pop

Table 1. Listening test items.

	Condition
1	hidden reference (original)
2	3.5kHz anchor
3	plain core
4	core + LTP
5	proposed hybrid core

Table 2. Listening test conditions.

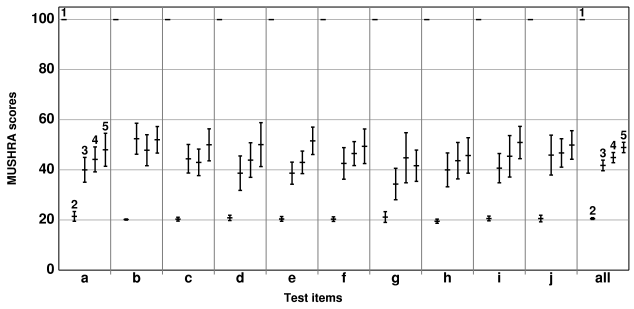

Fig. 4. Listening test results for 13.2 kbps mono coding; items of Table 1 in numbered conditions of Table 2 on x-axis; average MUSHRA scores and 95% confidence intervals on y-axis.

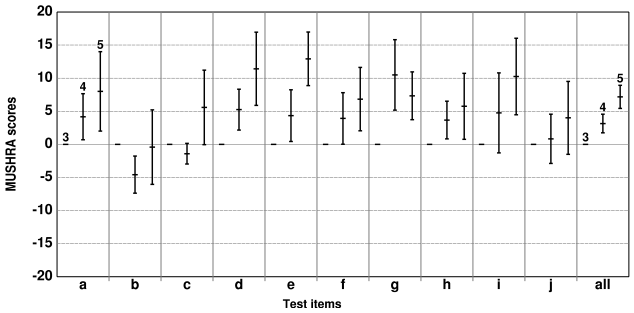

Fig. 5. Listening test results for 13.2 kbps mono coding; items of Table 1 in numbered conditions of Table 2 on x-axis; difference MUSHRA scores relative to condition «3» (plain core) and 95% confidence intervals on y-axis.

distribution. In the MUSHRA absolute scores, Figure 4, it can be seen in the first place that the low bitrate operation point of the coders results in an overall quality range situated between «fair» and «poor». However, up to more than 10 MUSHRA points can be gained by the proposed hybrid coder through sinusoidal substitution (items *d, e, i*). Averaged (*all* items), we also see a clear gain of approx. 7 MUSHRA points. In the MUSHRA difference scores wrt. «plain core» scores, Figure 5, there are many significant improvements visible (items *a, d, e, f, g, h*), no significant degradation, and averaged (*all* items) the significant improvement amounts to approx. 7 MUSHRA points. As main reason for the higher scores, listeners reported the successful reduction of modulation and warbling artifacts. The condition «core + LTP» also shows significant improvements, but considerably less compared to sinusoidal substitution. In mean, the significant improvement amounts to approx. 3 MUSHRA points. For one item (item *b*), the LTP leads to a significant degradation. This might be explained by the polyphonic nature of the item which does not match well with the LTP idea of predicting a common fundamental pitch.

5. CONCLUSIONS

We presented a novel hybrid audio coding scheme for low bitrate music coding that tightly integrates parametric techniques into transform-based coders. The proposed parametric extension consists of the replacement of natural tones, which would be otherwise badly coded and impaired by warbling artifacts, by sine tones that are perceptual similar, yet very efficient to code through parameters. The respective sinusoidal parameters are represented by spectral pseudo-lines that are directly inserted into the audio spectra at the transform-based encoder to substitute the original local signal content. The frequency of the sinusoid is represented by the spectral position of the pseudo-line and is refined by fine-grain spectral offset side information. Thereby, pseudo-lines can be further processed within the perceptual audio coder just like any other true spectral line. The common processing in the subsequent psychoacoustic calculations and in the quantizer establishes a tight interplay between both techniques. This hybrid coder design is opposed to traditional sinusoidal coders, which iteratively subtract synthesized sinusoids from a residual signal.

In a listening test, the perceptual quality of the hybrid coder was compared to that of a plain transform-based coder and a similar coder additionally using long term prediction. The low bitrate operation point of the tested coders results in an overall quality range situated between «fair» and «poor». All the more, the significant gain of approx. 7 MUSHRA points by the proposed hybrid coder utilizing sinusoidal substitution is extremely valuable, since it can make the difference between user «acceptance» or «annoyance» in terms of perceptual quality of a low bitrate application.

6. REFERENCES

[1] ISO/IEC 14496-3:2009, "Coding of Audio-Visual Objects, Part 3: Audio," Aug. 2009.

[2] Max Neuendorf, Markus Multrus, Nikolaus Rettelbach, Guillaume Fuchs, Julien Robilliard, Jeremie Lecomte, Stephan Wilde, Stefan Bayer, Sascha Disch, Christian Helmrich, Roch Lefebvre, Philippe Gournay, Bruno Bessette, Jimmy Lapierre, Kristofer Kjörling, Heiko Purnhagen, Lars Villemoes, Werner Oomen, Erik Schuijers, Kei Kikuiri, Toru Chinen, Takeshi Norimatsu, Chong Kok Seng, Eunmi Oh, Miyoung Kim, Schuyler Quackenbush, and Bernhard Grill, "MPEG Unified Speech and Audio Coding - The ISO/MPEG Standard for High-Efficiency Audio Coding of all Content Types," in *Audio Engineering Society Convention 132*, 4 2012.

[3] B. Bessette, R. Lefebvre, and R. Salami, "Universal Speech/Audio Coding using Hybrid ACELP/TCX Techniques," in *Acoustics, Speech, and Signal Processing, 2005. Proceedings. (ICASSP '05). IEEE International Conference on*, 2005.

[4] L. Daudet and M. Sandler, "MDCT Analysis of Sinusoids: Exact Results and Applications to Coding Artifacts Reduction," *Speech and Audio Processing, IEEE Transactions on*, vol. 12, pp. 302–312, 2004.

[5] H. Purnhagen and N. Meine, "HILN-the MPEG-4 parametric audio coding tools," in *Circuits and Systems, 2000. Proceedings. ISCAS 2000 Geneva. The 2000 IEEE International Symposium on*, 2000.

[6] Werner Oomen, Erik Schuijers, Bert Brinker, and Jeroen: Breebaart, "Advances in Parametric Coding for High-Quality Audio," in *Audio Engineering Society Convention 114*, 2003.

[7] N.H. van Schijndel and S. van de Par, "Rate-Distortion Optimized Hybrid Sound Coding," in *Applications of Signal Processing to Audio and Acoustics, 2005. IEEE Workshop on*, oct. 2005, pp. 235 – 238.

[8] Anibal.J.S. Ferreira, "Combined Spectral Envelope Normalization and Subtraction of Sinusoidal Components in the ODFT and MDCT Frequency Domains," *Applications of Signal Processing to Audio and Acoustics, 2001 IEEE Workshop on*, pp. 51–54, 2001.

[9] Sascha Disch, Benjamin Schubert, and Bernd Edler, "Cheap Beeps - Efficient Synthesis of Sinusoids and Sweeps in the MDCT Domain," in *Proceedings - ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing*, May 2013.

[10] Sascha Disch and Bernd Edler, "An Iterative Segmentation Algorithm for Audio Signal Spectra Depending on Estimated Local Centers of Gravity," in *Proceedings of the 12th International Conference on Digital Audio Effects*, September 2009.

[11] Eberhard Zwicker and Hugo Fastl, *Psychoacoustics - Facts and Models*, Springer, Berlin - Heidelberg, 2. edition, 1999.

[12] Craig C. Wier, Walt Jesteadt, and David M. Green, "Frequency Discrimination as a Function of Frequency and Sensation Level," *The Journal of the Acoustical Society of America*, vol. 61, no. 1, pp. 178–184, 1977.

[13] ITU-R BS.1534-1, *Method for the Subjective Assessment of Intermediate Quality Levels of Coding Systems*, ITU, Geneva, Switzerland, 2003.

[14] ITU-R BS.1116-1, *Methods for the Subjective Assessment of Small Impairments in Audio Systems including Multichannel Sound Systems*, ITU, Geneva, Switzerland, 1997.

[15] R.P. Ramachandran and P. Kabal, "Pitch Prediction Filters in Speech Coding," *Acoustics, Speech and Signal Processing, IEEE Transactions on*, vol. 37, no. 4, pp. 467–478, 1989.