Audio Engineering Society

Convention Paper 10249

Presented at the AES 159th Convention
2025 October 23–25, Long Beach, CA, USA

This paper was peer-reviewed as a complete manuscript for presentation at this convention. This paper is available in the AES E-Library (http://www.aes.org/e-lib), all rights reserved. Reproduction of this paper, or any portion thereof, is not permitted without direct permission from the Journal of the Audio Engineering Society.

Perceptually Controlled Selection of Alternatives for High-Frequency Content in Intelligent Gap Filling

Steven van de Par[2,4], Sascha Disch[1,3], Andreas Niedermeier[1], and Bernd Edler[1,3]

[1]*Fraunhofer IIS, Erlangen, Germany*
[2]*Acoustics Group, Cluster of Excellence "Hearing4All", University of Oldenburg, Oldenburg, Germany*
[3]*Friedrich Alexander University (FAU), International Audio Laboratories Erlangen, Erlangen, Germany*
[4]*Fraunhofer HSA, Oldenburg, Germany*

Correspondence should be addressed to Steven van de Par (steven.van.de.par@uni-oldenburg.de)

ABSTRACT

When encoding audio at very low bitrates using a transform coding technique like AAC, the encoder may not be able to encode the full audio bandwidth with good quality. For this reason, bandwidth extension techniques have been proposed, like Intelligent Gap Filling, that copy low-frequency spectral bands to high frequencies and adjust their properties by a set of parameters. Although such methods provide a significant quality improvement, the selected bands and their interaction with neighboring bands may cause artifacts mainly due to misaligned tonal components. In this contribution, a new method is proposed to generate high-frequency content for Intelligent Gap Filling. It is based on creating non-linear distortions in the time-domain decoded core signal. For tonal signals, this will create a harmonic spectrum with potentially better tonal alignment. As a non-linearity, sign- and RMS-preserving squaring of the input signal between pairs of consecutive zero crossings is used. A perceptually controlled dynamic selection mechanism is investigated to chose between the alternatives of no extension, copy-up and non-linear distortion. Listening tests show that for some tonal items the dynamic selection obtains an improvement in audio quality.

1 Background

Perceptual audio encoding methods such as MPEG 1/2 Layer 3 (MP3) and MPEG AAC [1] provide excellent methods to reduce the bit-rate for representing audio signals. Where PCM audio is typically represented with 16 bits per sample or more, AAC provides good audio quality already at 2 bits per sample or less.

The main approach in such perceptual audio encoding methods is to transform the PCM input signal into a time-frequency representation e.g. using Modified Discrete Cosine Transforms (MDCT) followed by a reduced bit-depth representation of the resulting time-frequency samples. The bit-depth is regulated in such a way that the error signal that is produced is inaudible in the presence of the encoded signal. Here the inaudibility is determined by a psycho-acoustic masking model [2].

Such methods provide near to transparent quality at rather low bit rates of 48 kbit/sec which is equivalent to 1 bit per sample at a sampling rate of 48 kHz for a mono audio signal. Further reduction in bit-rate would prohibit the encoding of all frequency components, and as a consequence, certain frequency regions will

dynamically need to be set to zero to allow a lower than 48 kbit/sec bit rate.

It was found that such dynamic on- and off-switching of frequency bands can create very noticeable musical noise artifacts. This is especially thought to happen when the on- and off-switching pattern is unrelated to the inherent modulations in the encoded audio signal. To avoid such musical noise artifacts, it often is preferred to permanently limit the encoding bandwidth even considering this limitation itself can be audible.

As a means to overcome the reduced bandwidth at very low encoding bit rates, several parametric tools have been proposed such as Spectral Bandwidth Replication (SBR) [3] and Intelligent Gap Filling (IGF) [4] where frequency content from a low-frequency range is copied up in the range where no waveform encoded signal is present. In both methods, frequency 'source' lines from the low frequencies are copied up to fill the missing frequency content, and a spectral shaping is applied to ensure a similar spectral envelope as in the original audio signal. Various parametric tools are available to further improve quality, including temporal noise shaping and insertion of incidental tonal components that cannot be copied up from the source bands. Such methods allow to encode the full audio bandwidth at very low bit rates with good audio quality using a signal-preserving encoding at low frequencies, and a parametric representation at high frequencies.

2 Alternatives for High-Frequency Content Generation

Despite the good improvement in audio quality that is obtained with e.g. the IGF method, in some cases the source band does not provide the spectral structure that is required at high frequencies to perceptually match the original audio signal. In some cases the inserted high-frequency components sound too grainy or rough, signaling that a too high degree of temporal modulation exists in the high frequencies. In some other cases rich structures of overtones can not be properly generated from the source band.

For this reason, in this contribution, an alternative method is investigated for generating the missing high-frequency spectral lines. It consists of a non-linear transformation applied on the time-domain representation of the decoded MDCT coefficients. This non-linear transformation creates distortion products at high frequencies that can be spectrally shaped to match the original audio signal as well as possible. Besides this non-linear transformation, a small degree of reverberation prior to spectral shaping is applied which reduces the temporal modulation.

Initial experiments showed that the alternative method for generating the high-frequency substrate does provide an improved audio quality in many cases. For some signals, however, it creates a substrate that sounds too noisy. Thus, it was considered that it would be desirable to dynamically choose between the native IGF method created by copying up frequency content, and the alternative method based on non-linear distortion.

We used a perceptual model [5] that is able to rate the perceptual quality resulting from the two alternative substrates on a coarse spectro-temporal scale, to allow to dynamically switch between both substrates. In addition, the perceptual model also assessed whether simply not inserting any components in the missing bands would lead to better audio quality.

In the following, the newly proposed perceptually steered IGF method will be described in detail (sections 3 and 4) together with a short outline of the psychoacoustic model (section 5) that was used. Then a listening test will be presented that evaluated the quality of the proposed method in comparison to the existing IGF method (section 6).

3 Extensions to Intelligent Gap Filling

Waveform encoders, at very low bitrates, do not allow to encode the full frequency range available in the original audio material. In Intelligent Gap Filling [4], the missing frequency content, in the form of spectral lines within the MDCT domain, is obtained from source tiles of MDCT coefficients that are encoded by the encoder and copied to the spectral location of target tiles. A flexible tiling is possible with the IGF method. The MDCT coefficients that are copied from the core bands to the upper bands, are temporally shaped with prediction filters and adapted in level per scale-factor band to match the original signal.

However, the source band may not always provide a spectral fine structure that is suitable to parametrically represent the missing frequency bands. The source

band may provide a structure that is too strongly modulated in the temporal domain, creating a too strong perceived roughness. In addition, at the transition between different tiles, modulation artifacts may be created.

To overcome some of these problems, an alternative method based on creating a non-linearly transformed version of the waveform encoded signal may be used applied to the time domain signal. It would have the advantage to not have tonal components in neighboring copied bands that could create modulation artifacts. The non-linearity will also tend to create an increasingly richer pattern of overtones towards high frequencies, thereby potentially remediating the roughness artifacts that can occur due to a spectrum that is composed of a too strongly structured set of partials.

The method presented here creates missing frequency content using a number of steps. First all zero crossings of the decoded (band-limited) time-domain waveform are detected. In between each pair of consecutive zero-crossing, the samples are squared with preservation of their sign. Then, the signal obtained this way is equalized in root-mean-square value to match the original time signal that was present between the two consecutive time samples. Before subjecting the signal to this non-linear transformation, a high-pass filter at 2000 Hz is used. This emphasizes the mid- and high-frequency content of the audio signal. Subsequently, the distorted signal is added to the input signal before filtering. Finally, the resulting signal is convolved with an exponentially decaying white noise tail to generate artificial reverberation. The T60 time is set to 10 ms. By having this artificial reverb, strong modulations in the non-linear signal are reduced to some degree. The proposed alternatives may provide an advantage over native IGF, but may not do so in all cases. Therefore, based on informal listening, it was deemed advantageous to allow for dynamic switching on a per-frame basis between the proposed new method, the native IGF method, and the entire omission of bandwidth extension. To allow this dynamic switching, a steering mechanism is used which is based on a psycho-acoustic model that has already been used successfully for making encoding decisions in different contexts in the past [6, 7].

4 Alternative High-Frequency Content Selection in Encoder

To steer the dynamic selection between the alternatives of high-frequency content generation, the encoder has to be adapted. In Fig. 1 an outline is given of the encoding framework proposed. The top half represents the encoder side of the proposed method. First the audio signal (left side) is AAC encoded resulting in a set of MDCT coefficients. These are presented to the three different blocks, rightward of the AAC encoder. In the upper block, native IGF (IGF copy up) is used to generate the missing frequency regions. In the middle block, the new method (IGF non-linear) described in section 3 is applied. Because the newly proposed method requires the presence of a time-domain signal from the waveform encoder, an AAC decoder is used first. The lower block (no IGF), represents the option that no signals are inserted in the empty frequency regions created by the AAC encoder.

Each of the three options, IGF copy up, IGF non-linear, and no IGF, are decoded to the time domain and fed to the psycho-acoustic model (section 5). Within this psycho-acoustic model, a comparison is made with the original signal for each of the options to get a prediction of the strength of perceivable differences in the form of a perceptual distortion measure. This perceptual distortion measure is defined per short time interval and per critical band. From this, as described in detail in the next section, a decision can be derived on a frame-by-frame basis which encoding option leads to the lowest predicted perceptual distortion.

In the decoder, lower part of Fig. 1, the three methods used in the encoder are applied to the waveform decoded signal. Based on the decisions made by the psycho-acoustic model in the encoder, dynamically the best fitting high-frequency content generation method is selected on a frame-by-frame basis, and per frequency region to compose a decoded signal. The signal selection is done within the MDCT domain.

5 Psycho-acoustic Model

The psycho-acoustic model is based on an earlier model of Dau et al. [8] that was extended and adapted for use in combination with audio encoding methods [6, 7]. The model is implemented in the time domain and first has a gamma-tone filter bank representing the cochlear filtering process. These filters are implemented to create complex-valued output, from which the Hilbert envelope is created to be used further as filter output. Each of these filter outputs is processed by a sequence of adaptation loops, that provide an adaptive gain control representing adaptation processes in the auditory

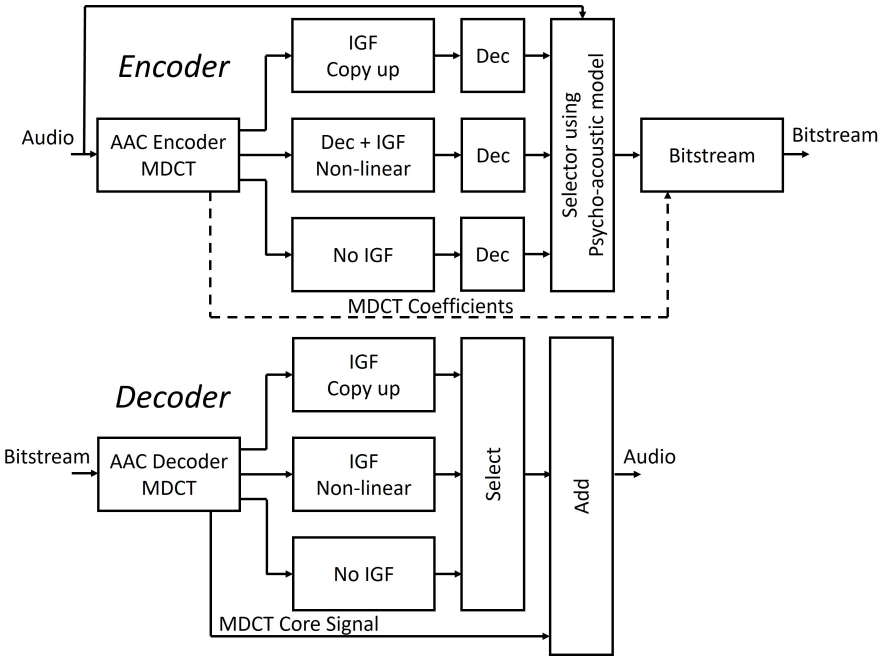


Fig. 1: A schematic overview of the processing steps in the encoder, upper part, and the decoder. In the upper part three options regarding parametric modeling the high frequency content are made available to a psycho-acoustic model to chose the best option.

system. These adaptation loops create a gain reduction for steady-state signals, and the resulting reduced sensitivity to input signals remains for some time after a steady-state signal has been switched off. This allows to account for the psycho-acoustic phenomenon of forward masking [9]. The adaptation loops proposed in Dau et al. [10] have been modified to limit the large overshoot response at the onset of steady-state signals.

Subsequent to the adaptation loops, a modulation filter bank is used with a Q-factor of 4 ranging up to center frequencies of 243 Hz. This modulation filter bank represents the auditory processing of modulations and allows to detect modulation artifacts in tonal audio components with better sensitivity than in noisy audio components.

The resulting output is a set of time domain signals for each modulation filter, and auditory filter. Together, these signals are termed the internal representation. In the original Dau et al. [8] model, this internal representation was considered to be subjected to additive noise, which allowed to incorporate the basic limitations of auditory processing in the model. By all these processing steps, the internal representation should provide a perceptually relevant transformation of the input signal, that contained only the perceivable information of the input signal.

To apply this model in the context of low-bitrate audio coding, internal representations of original and encoded signals are compared using a simple subtraction operation leading to a so-called internal difference representation (IRd).

In subsequent studies [6, 5, 7], some further modifications were made to the internal representation that are also used in the current study. First of all, a stage was incorporated to model the reduction in auditory masking when different frequency regions across the basilar membrane are co-modulated. This is known as Comodulation Masking Release (CMR) [11], and is modeled here based on a magnification of differences in the IRd when comodulation was measured using a correlation measure.

A further extension is very specific for audio coding. It emphasizes errors that are due to the addition of modu-

lations. The reasoning here is that added modulations, for example due to spectral holes, are often not related to the original signal, and may therefore be likely to stand out as separate auditory objects, unrelated to the original signal, and for that reason being perceptually more salient than other types of signal modifications.

Finally, the perceptible distortion is determined via a summation of the squared values of the IRd. To get more reliable estimates of perceptible distortions averaging was done across time and frequency. For this reason, the full bandwidth of the audio signal was partitioned in four regions, each covering an equal range of auditory filters. Note that in the encoder settings that were used in the listening tests (See Section 6), only the highest region is relevant because it covered the high-frequency range where parametric representations could be used. In order to get stable decisions about the preferred encoding method, also temporal averaging was performed incorporating 67 ms frames before and after the current frame of interest. Thus based on these averaging operations across time and frequency, a frame-by-frame decision was made for the four bandwidth ranges, as to which encoding option led to the least predicted perceptible distortion. This option was then chosen to serve as the signal used in the decoding process.

6 Listening Test

A listening test was conducted that evaluated the perceptual quality of the newly proposed method against a standard conform implementation of IGF. For this purpose, a MUSHRA test [12] was conducted, using the original signal as reference and hidden reference, and further included a standard conform implementation of IGF method, the newly proposed method, and an Anchor, created by low-pass filtering the original signal with a 3.5 kHz cut-off.

The wave-form preserving AAC audio encoder was set at a very high bit-rate and its encoding bandwidth was limited to 25% (6 kHz) of full bandwidth at a sampling rate of 48 kHz. This high bit-rate was used to ensure that the waveform encoder essentially provided transparent quality such that the listening-test results only reflected the influence of the parametric encoding stages.

Excerpts were selected out of a larger set of commonly used excerpts for evaluation of low bitrate audio coding methods with emphasis on excerpts containing tonal content. All items were encoded and an analysis was made regarding the degree to which the psycho-acoustic model decided to use the newly proposed non-linear distortion method to generate the decoded signal. This analysis showed that for some signal types, such as speech, the newly proposed method was essentially not selected. For the listening test, specifically excerpts were selected where a good proportion of the time, the newly proposed method was used (> 25% of the time). Excerpts with a lower, or even near to zero percentage, would not differ substantially from the original IGF method and would not be interesting to include in the listening test.

Table 1: Audio excerpts with percentage of times the newly proposed method for generating high frequency content was selected by the psycho-acoustic model, and percentage of time the high frequency band was left empty. The upper part of the table shows excerpts selected for the listening test, the lower part shows a few representative excerpts often used in low-bitrate audio coding

Excerpt	Percentage New Method	Percentage Empty
Violin	28.4	0
Glockenspiel	29.1	31.4
Harpsichord	26.6	0
Organ	31.9	0.4
Pitchpipe	68.0	0
Trumpet	35.9	0
Brahms	2.6	0
Castanets	2.3	10.3
German male	0.3	0
Susan Vega	0.8	0

Table 1 shows the list of selected excerpts in the top part including the percentages of time when the new method for generating high-frequency content was selected by the psycho-acoustic model. Percentages exceed 25% and go up to even 68%. In the lower half of the table, a few excerpts are shown that are often used in listening tests for low-bit-rate audio coding. For these items, rarely the new method is selected. It appears that - according the the psycho-acoustic model - the newly proposed method is most advantageous

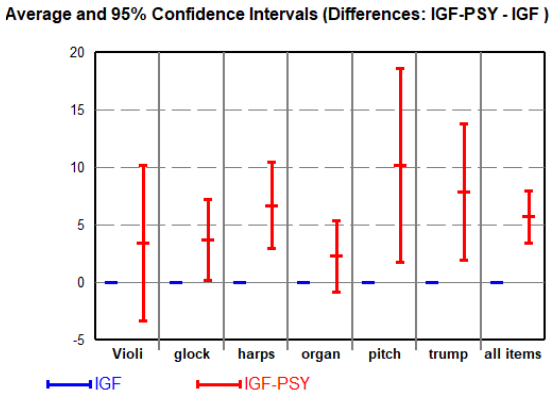


Fig. 2: Results of the MUSHRA listening test in terms of differences to the reference method termed IGF method. The IGF-PSY data show improvements relative to the reference IGF. The horizontal axis labels represent different audio excerpts, the right most data show mean results. Error bars represent 95% confidence intervals.

for tonal excerpts. In the rightmost column, the percentages of times are listed, where the psycho-acoustic model decided that neither IGF, nor the newly proposed method was deemed beneficial and that instead no substitution of high-frequency content was supposed to be better. It seems that especially for excerpts with clear transients, this occurred often. Possibly, the psycho-acoustic model observed that this way pre-echoes could be avoided, an artifact to which this model is known to be sensitive.

In Fig. 2, results are shown for the 6 excerpts used in this listening test based on averaged scores of 9 expert listeners. The results are shown in terms of the improvement obtained relative to the reference method (IGF). Error-bars represent 95% confidence intervals. As can be seen, an average improvement of 6 points on the MUSHRA scale was found. Specifically the excerpts for which a larger percentage of times the newly proposed method was used for generating high-frequency content, show best improvements (Pitch pipe, and Trumpet).

7 Discussion

A new method for generating high-frequency content, to be used in combination with a low-bit rate waveform codec was presented. The method can be used to complement the Intelligent Gap Filling method [4] which copies-up frequency lines from the decoded low-frequency source-band to the high-frequencies. In the newly proposed method, high-frequency content is generated by applying a non-linear distortion to the decoded low-frequency signal. Essentially, the non-linear distortion is applied to segments of the time-domain signal in-between two subsequent zero-crossings by a squaring operation while preserving the sign and rms-value. For both methods, the spectral envelope of the original audio signal is reinstated on a frame-by-frame basis for each scale-factor-band.

Since this newly proposed method for generating high-frequency content does not always provide an improvement in audio quality, a psycho-acoustic model [6, 7] is used to decide whether it is beneficial to use the IGF method, the newly proposed method, or to not create any high-frequency content. The model is based on the modulation filter-bank model of Dau et al. [8] and allows to detect artifacts specifically noticeable in tonal components, such as added modulations, or added noisiness. An analysis showed that the psycho-acoustic model specifically favored the new method to be used for audio excerpts with dominant tonal components. Based on the listening test results, it seems the model is successfull in only selecting the newly proposed method when it does lead to an improvement in audio quality.

Listening test results showed that an improvement of 6 MUSHRA points on average could be obtained by allowing the audio encoder to use the newly proposed method. It is important to note that this new method is not suitable as a full replacement of IGF, but that, using the psycho-acoustical model, the encoder could be steered at each time frame to use the best method for creating the high-frequency content. Due to using the psycho-acoustically controlled encoding, an overall improvement could be obtained for highly-tonal excerpts. In other types of excerpts, the psycho-acoustic model tends to not use the newly proposed method and quality should therefore be at the level of the existing IGF method; therefore such items were not included in the listening test.

8 Conclusion

The newly proposed method for selectively generating high-frequency audio content based on non-linear

distortion products generated from low-frequency audio components was shown to provide improved audio quality compared to using an implementation of the established Intelligent Gap Filling (IGF) method [4]. To achieve this improvement, a psycho-acoustic model is used that decides when it is advantageous to use either the non-linear distortion method, the existing IGF method based on copying up signal components, or to not use any additional high-frequency content. Since it became apparent from informal listening that using non-linear distortion exclusively for generating high-frequency content provided no general quality improvement over IGF, it was successfully demonstrated with a listening test that the psycho-acoustic model was able to make appropriate decisions about when to use the non-linear distortion method, to resort to legacy IGF, or to refrain from adding any high frequency content. For selection of the non-linear method, the psycho-acoustic model seemed to favor highly tonal audio excerpts over more dynamic audio signals such as speech.

References

[1] Herre, J. and Dietz, M., "MPEG-4 high-efficiency AAC coding [Standards in a Nutshell]," *Signal Processing Magazine, IEEE*, (Vol. 25, 2008), pp. 137–142, 2008.

[2] Painter, T. and Spanias, A., "Perceptual coding of digital audio," *Proc. IEEE*, 88(4), pp. 451–513, 2000.

[3] Zhong, H., Villemoes, L., Ekstrand, P., Disch, S., Nagel, F., Wilde, S., Chong, K.-S., and Norimatsu, T., "QMF Based Harmonic Spectral Band Replication," in *Proceedings, 131st AES Convention*, p. Conv.Paper8517, 2011.

[4] Disch, S., Niedermeier, A., Helmrich, C. R., Neukam, C., Schmidt, K., Geiger, R., Lecomte, J., Ghido, F., Nagel, F., and Edler, B., "Intelligent Gap Filling in Perceptual Transform Coding of Audio," *Journal of the Audio Engineering Society*, (9661), 2016.

[5] van de Par, S., Disch, S., Niedermeier, A., Burdiel Pérez, E., and Edler, B., "Temporal Envelope-Based Psychoacoustic Modelling for Evaluating Non-Waveform Preserving Audio Codecs," in *Audio Engineering Society Convention 147*, 2019.

[6] Disch, S., van de Par, S., Niedermeier, A., Burdiel, P. E., Berasategui, C. A., and Bernd, E., "Improved Psychoacoustic Model for Efficient Perceptual Audio Codecs," *Journal of the Audio Engineering Society*, (10029), 2018.

[7] van de Par, S., Disch, S., Niedermeier, A., and Bernd, E., "Informed Postprocessing for Auditory Roughness Removal for Low-Bitrate Audio Coders," *Journal of the Audio Engineering Society*, (10534), 2021.

[8] Dau, T., Kollmeier, B., and Kohlrausch, A., "Modelling auditory processing of amplitude modulation. I. Detection and masking with narrow-band carriers," *J. Acoust. Soc. Am.*, 102, pp. 2892–2905, 1997.

[9] Fastl, H. and Zwicker, E., *Psychoacoustics: Facts and Models*, Springer, 2007.

[10] Dau, T., Püschel, D., and Kohlrausch, A., "A quantization model of the 'effective' signal processing in the auditory system. I. Model structure," *J. Acoust. Soc. Am.*, 99, pp. 3615–3622, 1996.

[11] Buus, S., "Release from masking caused by envelope fluctuations," *J. Acoust. Soc. Am.*, (78), pp. 1958–1965, 1985.

[12] ITU-R, *Recommendation BS.1534-1 Method for subjective assessment of intermediate sound quality (MUSHRA)*, Geneva, 2003.