TEMPORAL TILE SHAPING FOR SPECTRAL GAP FILLING IN AUDIO TRANSFORM CODING IN EVS

Sascha Disch, Christian Neukam, Konstantin Schmidt

Fraunhofer Institute for Integrated Circuits (IIS)
Am Wolfsmantel 33
91058 Erlangen, Germany
sascha.disch@iis.fraunhofer.de

ABSTRACT

At low bitrates, next generation audio coders apply waveform preserving transform coding only for the perceptually most relevant parts of the signal. The resulting spectral gaps are filled in the decoder through techniques like Intelligent Gap Filling (IGF). IGF is currently being standardized in MPEG-H 3D-Audio and also in 3GPP Enhanced Voice Service (EVS). In IGF processing, spectral tiles are copied from a spectral source location into a target location and subsequently adapted by parameter steered post-processing to best match relevant properties of the original signal. Important properties include the spectral and temporal envelope. Since IGF operates on Modified Discrete Cosine Transform (MDCT) spectra of rather long time blocks, temporal envelope shaping is not trivial. In this paper, Temporal Tile Shaping (TTS) is presented. TTS is based on linear prediction in the MDCT domain for shaping the temporal structure of the gap filling signal in the target tiles with subblock granularity. A listening test demonstrates the advantage of the proposed method.

Index Terms— Audio Coding, Noise Filling, Temporal Noise Shaping, Intelligent Gap Filling, Enhanced Voice Service

1. INTRODUCTION

Storage or transmission of audio signals is often subject to strict bitrate constraints. To accommodate this, next generation audio codecs like 3GPP Enhanced Voice Service (EVS) [1, 2, 3] comprise transform coding functionality by TCX, alongside with tightly integrated parametric tools such as noise filling or spectral Intelligent Gap Filling (IGF), as initially standardized within MPEG-H 3D-Audio [4].

These parametric techniques recreate the signal parts that have been zeroed by the encoder, either a-priori on purpose or by simply falling into the dead-zone of the quantizer, in order to meet the low bitrate constrains. While noise filling inserts random noise in the spectrum, IGF harvests spectral tiles from the remaining non zero signal spectral parts and applies parameter guided post-processing [5] to fill said gaps in the spectrum.

In EVS, noise filling and IGF are directly integrated into the Modified Discrete Cosine Transform (MDCT) domain of TCX. Since low bitrate TCX operates on a fixed time block duration of 20 ms, additional parametric tools like Temporal Noise Shaping (TNS) and its extension, Temporal Tile Shaping (TTS), are needed to provide a sufficient time resolution in order to avoid pre- and post-echo artifacts [6], especially for these parts of the spectrum that are reconstructed through IGF. This paper describes the functionality of TNS/TTS and supports the benefits of TTS for transients reproduction by listening test data.

2. RELEVANT PRIOR WORK

In transform coders [7], quantization noise is spread all over the temporal duration of a transform block, inducing audible pre- and post-echoes if their energy is above the temporal masking threshold. Many transform codecs utilize different block sizes and perform a signal adaptive block switching to confine quantization noise below the temporal masking threshold. Switching between different block sizes requires the insertion of transition blocks.

In EVS, for bitrates equal or greater than 48 kbps, the transform coder is indeed able to switch block size [8]. Nevertheless, at bitrates lower than 48 kbps, the core coder signal adaptively switches between Algebraic Code Excited Linear Prediction (ACELP) speech coding and fixed block size TCX transform coding. Since ACELP can not provide suitable transition blocks, an immediate switching between ACELP and a block-switched transform coder would not be possible without introducing additional look-ahead delay. Furthermore, short block length inevitably reduce the transform coding gain, which is especially unfavorable at low bitrates. Therefore, to satisfy the requirements of low delay and high perceptual quality for bi-directional communication applications, the EVS codec is bound to using a fixed 20 ms block size for low bitrates.

Other techniques rely on high time resolution filter banks e.g. a Quadrature Mirror Filterbank (QMF) having approximately 1.2 ms temporal resolution. Some examples are audio bandwidth extension [9, 10] that uses the high time resolution to define a signal adaptive sub-framing, and Guided Envelope Shaping (GES) [11] in MPEG Surround [12] that directly applies a temporal gain curve to the signal in the QMF domain. Again, such a domain is not available within the EVS codec.

Additionally, or alternatively, temporal envelope shaping techniques have been used to confine noise or other additive signal components below the temporal envelope of the original signal. Temporal Noise Shaping (TNS) [13, 14, 15, 16] is a standard technique and part of Advanced Audio Coding (AAC) [17, 10]. TNS can be considered as an extension of the basic scheme of a perceptual coder, inserting a forward prediction filter (FIR type) in frequency direction as an optional processing step between the analysis transform and quantization stage of the encoder. Exploiting the Fourier correspondence of spectral autocorrelation and squared Hilbert envelope, this filter is effective to temporally flatten the signal that is contained in a transform time block. In the decoder, an inverse filter operation (IIR type) is performed prior to the synthesis transform that is effective to temporally shape the signal alongside with its added noise.

Thereby the quantization noise gets masked by the transient.

In this paper, the extension of the idea of quantization noise shaping by TNS towards shaping of the parametrically generated signal components by IGF is proposed. Another novelty of the proposed technique is to efficiently integrate transform coding, spectral gap filling and temporal shaping of parametrically recreated gap filling signal portions within a single low temporal resolution domain such as MDCT.

3. TEMPORAL TILE SHAPING

3.1. Principle

To fulfill the audio bandwidth requirements of EVS at low bitrate settings, the transmitted waveform coded fullband TCX spectra in MDCT domain typically exhibit large spectral gaps containing zeroes. Through application of spectral gap filling by IGF [1], these gaps are subsequently filled with transposed and parameter controlled post-processed so-called spectral tiles. Figure 1 and Figure 3 show a simplified signal flow graph of the transform coding (TCX) path within the EVS codec.

3.2. Encoder

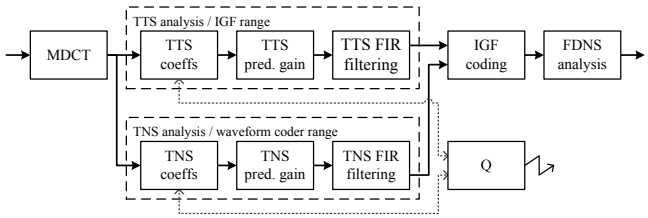

Fig. 1. Block diagram of the TCX encoder signal path within EVS. Only the signal path between MDCT and FDNS analysis is shown.

In the encoder depicted in Figure 1, after transformation of the audio signal into the MDCT domain, the TNS and TTS linear prediction (LP) coefficients are derived from the spectral autocorrelation of bandpass regions.

TNS and TTS filter coefficients can be identical if derived and applied on a shared bandpass region of the spectrum. In this case, the joint coefficients represent the temporal envelope of a rather wide highpass band, covering both TCX core and IGF reconstructed regions. Consequently, TTS usage is equivalent to extending the application of the TNS filter of the core coder into spectral regions predominantly reconstructed by IGF.

In other encoder configurations, TNS and TTS filters are defined to be completely separate filters, each of which reflecting the properties of its associated individual spectral bandpass region. In this case, the TTS filter coefficients solely represent the target temporal envelope of spectral regions predominantly covered by IGF. Table 1 lists the TNS/TTS filter configurations defined within the EVS codec.

To obtain the coefficients, the Toeplitz system of equations has to be solved [1]:

$$\sum_{j=1}^{N} a\left(j\right) r_{xx}\left(\left|i-j\right|\right) = -r_{xx}\left(i\right), \forall i \in [1, N] \tag{1}$$

where a corresponds to the LP coefficients to be calculated, r_{xx} is the normalized autocorrelation of one specific spectral bandpass region and N is the number of LP coefficients. For solving this set of equations the so-called Levinson-Durbin recursion is used [1]:

$$\begin{aligned}
&E(0) = r_{xx}(0)\\
&\texttt{for } i = 1 \ldots N\\
&\quad k_i = -\frac{r_{xx}(i) + \sum_{j=1}^{i-1} a_j^{(i-1)} r_{xx}\left(i-j\right)}{E\left(i-1\right)}\\
&\quad a_i^{(i)} = k_i\\
&\quad \texttt{for } j = 1 \ldots i-1\\
&\quad\quad a_j^{(i)} = a_j^{(i-1)} + k_i a_{i-j}^{(i-1)}\\
&\quad \texttt{end}\\
&\quad E(i) = \left(1 - k_i^2\right) E\left(i-1\right)\\
&\texttt{end}
\end{aligned}$$

where E corresponds to the residual error energies and k corresponds to the reflection coefficients. The final solution for the required LP coefficients is then given by the following equation [1]:

$$a_j = a_j^{(N)}, \forall j \in [1, N] \tag{2}$$

Next, the encoder calculates for each processing range the prediction gain g. This prediction gain is later used to determine if TNS/TTS processing should be applied in each particular spectral range. It is defined as follows [1]:

$$g = \frac{n}{E(N)} \tag{3}$$

with n being the number of subdivisions used for calculation the autocorrelation of each spectral region. If the linear prediction gain is above a given threshold, the filter coefficients that are output from the Levinson-Durbin recursion in the form of partial autocorrelation (PARCOR) coefficients k, are quantized into 4 bit values using an arcsine non-uniform quantizer shown in Figure 2. The quantized PARCOR coefficients are then provided for being multiplexed into the bit stream.

In the TNS/TTS filtering stages the quantized PARCOR coefficients are de-quantized and transformed to LP coefficients. These LP coefficients are used as finite impulse response (FIR) filter coefficients in the corresponding filtering stage. Although TNS can determine the filter direction [17], in the EVS context the TNS/TTS filters are always applied in upward direction.

bitrate [kbps]	bw	TCX20	TCX10	TCX5
24.4, 32	SWB	1 (600 Hz-16 kHz)	n/a	n/a
48, 96, 128	WB	1 (600 Hz-8 kHz)	2 (800 Hz-4.4 kHz, 4.4 kHz-8 kHz)	1 (600 Hz-8 kHz)
48, 96, 128	SWB	2 (600 Hz-4.5 kHz, 4.5 kHz-16 kHz)	2 (800 Hz-8.4 kHz, 8.4 kHz-16 kHz)	1 (800 Hz-16 kHz)
24.4, 32	FB	1 (600 Hz-20 kHz)	n/a	n/a
48, 96, 128	FB	2 (600 Hz-4.5 kHz, 4.5 kHz-20 kHz)	2 (800 Hz-10.4 kHz, 104 kHz-20 kHz)	1 (800 Hz-20 kHz)

Table 1. TNS/TTS number of filters and spectral range.

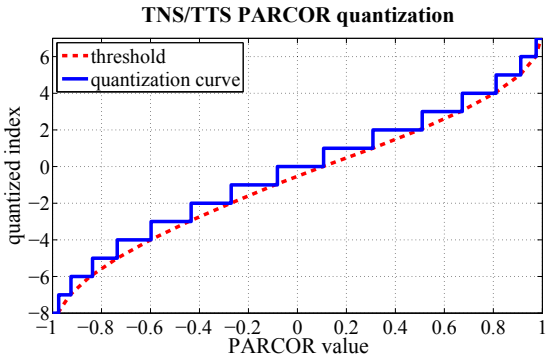

Fig. 2. TNS/TTS quantization curve. The PARCOR values are mapped to the quantization index (solid curve).

Subsequent to the TNS/TTS analysis, the main IGF processing takes place. In order to fit the given bit rate requirements, IGF modifies the spectral content by inserting gaps of zeroes into the MDCT spectrum. In addition, a set of gap filling parameters is calculated and prepared to be multiplexed in the final bit stream as well.

The last step in the shown processing sequence in Figure 1 is the spectral flattening through Frequency Domain Noise Shaping (FDNS) which is used for quantization noise shaping in the frequency domain [8]. Afterwards, the TCX coder encodes the waveform content up to Nyquist frequency, i.e. the waveform coder range plus the remaining spectral content in the IGF range.

3.3. Decoder

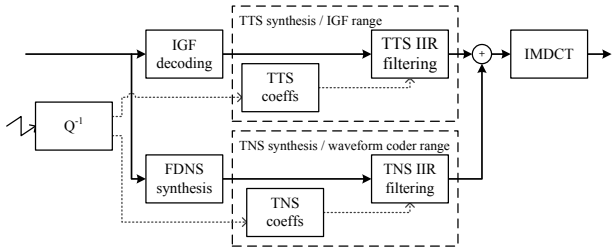

Fig. 3. Block diagram of the TCX decoder signal path within EVS. Only the signal path between IGF decoding / FDNS synthesis and IMDCT is shown.

In the decoder shown in Figure 3, after fullband TCX waveform decoding, IGF restores the spectral gaps introduced in the IGF encoder. IGF uses pre-defined frequency tiles which are either transposed or inserted into the spectral gaps. The spectral gap filling and its post-processing are steered by the transmitted gap filling parameters. In a parallel signal path the inverse FDNS restores the spectral shape of the noise filled waveform coded signal and shapes thereby the introduced TCX quantization noise [8]. On each signal path TNS and TTS synthesis filter are applied to their appropriate spectral regions.

First, the transmitted TTS/TNS PARCOR coefficients are de-quantized and converted into LP coefficients in the same way as done in the analysis stages at the encoder. Later, these de-quantized LP coefficients are used as infinite impulse response (IIR) filter coefficients in the filtering stage to finally shape either the quantization noise in the waveform coder range or the gap filled signal in the IGF range. After combining the gap filled signal with the waveform coded signal, the inverse MDCT (IMDCT) provides the output audio signal for further processing.

It has to be noted that on decoder side the IGF processing takes place in the spectral flattened domain before the FDNS synthesis calculation of TCX. Moreover, noise filling and additional spectral flattening of the frequency content in the IGF range are also part of the IGF post-processing. Due to these enhanced noise filling processing steps, transient signal portions inserted by spectral gap filling exhibit considerable temporal dispersion with respect to the fullband TCX waveform coded signal portions, leading to pre- and post-echo artifacts. To remedy the dispersion, IGF is applied in the temporally flattened TNS residual signal domain and the IGF generated signal portions are subjected to TTS synthesis filtering in order to shape their temporal envelope to match the envelope of the waveform coded content. The synthesis filter used for TTS in a certain spectral region is identical to the quantized TNS synthesis filter calculated by the encoder for that particular region.

3.4. Evidence for TTS benefits

Figure 4 displays the spectrogram of an original castanet recording signal for reference. Figure 5 shows a typical pre-echo effect in front of a transient onset from a castanet hit due to IGF generated signal portions as described in subsection 3.3. Finally, Figure 6 demonstrates the reduction of the pre-echo effect through the application of TTS filtering on the IGF generated signal portions.

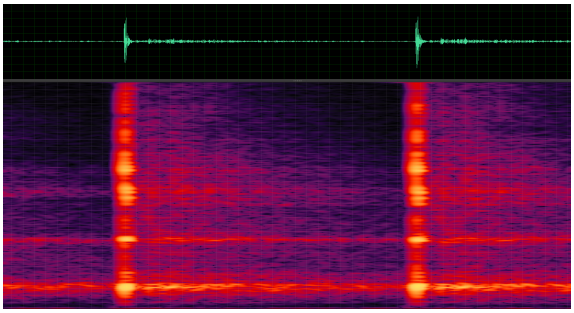

Fig. 4. Original castanet signal. Top: time domain, bottom: DFT spectrogram; x-axis: time, y-axis: amplitude/frequency respectively.

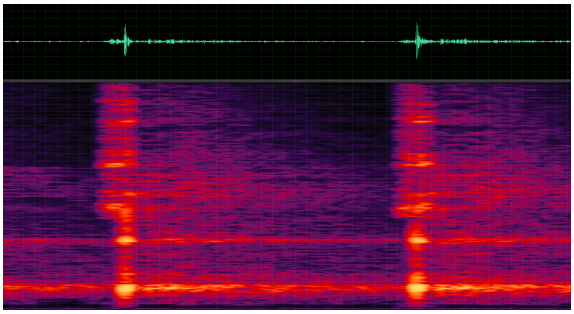

Fig. 5. Coded castanet signal without using TTS. TNS is active in the low frequency range. Top: time domain, bottom: DFT spectrogram; x-axis: time, y-axis: amplitude/frequency respectively.

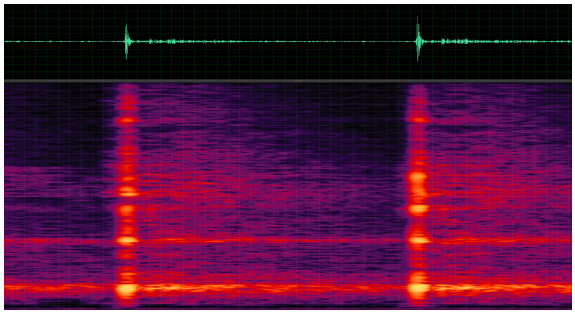

Fig. 6. Coded castanet signal using TTS. TNS is also active in the low frequency range. Top: time domain, bottom: DFT spectrogram; x-axis: time, y-axis: amplitude/frequency respectively.

When stringing together consecutive spectral IGF tiles, spectral correlation at the tile borders will be corrupted and the temporal envelope of the audio signal will be impaired by dispersion. Hence, another benefit of performing the IGF tile filling in the TNS residual domain is that, after application of the TNS/TTS common shaping filter, tile borders are again seamlessly correlated, resulting in a more faithful temporal reproduction of the signal.

4. LISTENING TEST AND RESULTS

4.1. System under test

For quantifying the perceptual quality contribution of TTS, a listening test was set up such that the TTS performance can be evaluated within the TCX core coder signal path as standardized for EVS.

Since the genuine EVS coder might switch in dependency of the input signal between the ACELP core, where TTS is not applicable, and the TCX core, where TTS is operative, TCX was forced for generating the TTS evaluation listening test stimuli.

This way, the listening test compared the EVS codec in its default setting, where TTS usage is controlled by the encoder, to a modified EVS codec, where TTS has been deactivated throughout.

The bitrate was chosen to be 24.4 kbit/s for coding of Super Wide Band (SWB) signals. The sample rate of all items is 32 kHz.

4.2. Listening test

11 expert listeners participated in the listening test, which was conducted using the MUSHRA methodology [18]. The test consists of 10 items, 7 music items, 1 speech item, 1 voiced speech item and 1 mixed item (speech + music), listed in Table 2. The test items were presented together with the original signal and one low-pass filtered anchor with a cut-off frequency of 3.5 kHz. For reproducing the sound Stax headphones and amplifier were used.

Difference ratings with its means and confidence intervals were calculated for every listener and every item pairwise between the two conditions "TTS on" and "TTS off". Figure 7 presents the results. The statistical evaluation is based on a non-parametric Wilcoxon signed-rank test since a standard normal distribution of the ratings cannot be assumed.

The results show that the perceptual quality obtained by using TTS is significantly better ($p < .05$) than without using TTS for 8 out of 10 items. For the other 2 items there is no degradation. In terms of mean values, albeit not statistical significant, TTS rather tends to improve the quality also on these signals. Outstanding improvements of up to 20 MUSHRA points are observed for items containing strong transient pulses, like e.g. castanets (si02) and low pitch electronic vocoder music (fatboy). In the overall rating, TTS significantly improves the perceptual quality of the IGF equipped EVS system over a modified system without TTS.

item	description
a2s3	speech over music
a6s3	pop music
AOI	percussion
dora	speech
es01	voiced speech
fatboy	electronic music
Music	rock music
no1	pop music
si02	castanets
sm02	glockenspiel

Table 2. List of tested items

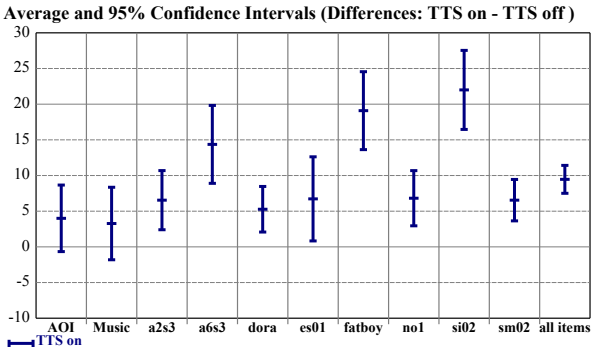

Fig. 7. Listening test results: Box plots of differences of individual ratings. TTS improves the perceptual quality significantly ($p < .05$) in 8 of 10 cases (significant positive values) and overall means.

5. CONCLUSION

In the EVS codec, Intelligent Gap Filling (IGF) is used to fill spectral gaps within the waveform coded signal of the transform coder (TCX) path. Due to a fixed time block length of 20 ms at low bitrates, special precautions have to be taken to prevent possible pre- and post-echoes in the IGF generated signal parts.

In this paper, a technique called Temporal Tile Shaping (TTS) is presented, which has been included into the EVS codec to remedy said potential echo artifacts. The TTS algorithm is demonstrated to successfully shape the temporal envelope of IGF time-frequency tiles with sub-block precision similar to the quantization noise shaping capability of TNS. The conducted MUSHRA listening test shows that TTS clearly improves the perceptual audio quality of IGF processed signals. While none of the test items is degraded, the vast majority (8 out of 10) of test items were graded significantly better in comparison to a system that does not use TTS.

In summary, TTS enables the efficient integration of perceptual transform coding using a fixed block length, spectral gap filling and an appropriate temporal shaping of the spectral gap filling generated signal portions in the MDCT domain.

6. REFERENCES

[1] 3GPP, *TS 26.445, EVS Codec Detailed Algorithmic Description; 3GPP Technical Specification (Release 12)*, 2014.

[2] Martin Dietz and al., "Overview of the EVS Codec Architecture," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Brisbane, Australia*, April 2015.

[3] Stefan Bruhn and al., "Standardization of the new 3GPP EVS Codec," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Brisbane, Australia*, April 2015.

[4] ISO/IEC, *MPEG N14747, Text of ISO/MPEG 23008-3/DIS 3D Audio*, JTC1/SC29/WG11, 2014.

[5] C. Helmrich, A. Niedermeier, S. Disch, and F. Ghido, "Spectral Envelope Reconstruction via IGF for Audio Transform Coding," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Brisbane, Australia*, April 2015.

[6] M. Erne, "Perceptual audio coders what to listen for," in *111th AES Convention*, 2001.

[7] Jürgen Herre and Sascha Disch, *Perceptual Audio Coding*, pp. 757–799, Academic press, Elsevier Ltd., August 2013.

[8] G. Fuchs, C. Helmrich, G. Marković, M. Neusinger, E. Ravelli, and Moriya T., "Low delay LPC and MDCT-based Audio Coding in EVS," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Brisbane, Australia*, April 2015.

[9] ISO/IEC, *MPEG 14496-3:2001/AMD1:2003, Bandwidth Extension*, JTC1/SC29/WG11, 2003.

[10] ISO/IEC, *MPEG 14496-3:2005 Information Technology, Coding of Audio-Visual Objects, Part 3: Audio, Third Edition*, JTC1/SC29/WG11, 2005.

[11] Johannes Hilpert and Sascha Disch, "The MPEG Surround Audio Coding Standard [Standards in a nutshell]," *Signal Processing Magazine, IEEE*, , no. No. 1, Vol. 26, 2009, pp. 148–152, January 2009.

[12] ISO/IEC, *MPEG 23003-1:2007, MPEG-D (MPEG Audio Technologies), Part 1: MPEG Surround*, JTC1/SC29/WG11, 2007.

[13] Jürgen Herre and James D. Johnston, "Enhancing the performance of perceptual audio coders by using temporal noise shaping (tns), preprint 4384," in *101st AES Convention*, 1996.

[14] Jürgen Herre and James D. Johnston, "Exploiting both time and frequency structure in a system that uses an analysis/synthesis filterbank with high frequency resolution,," in *103rd AES Convention*, 1997.

[15] J. Herre and James D. Johnston, "Continuously signal-adaptive filterbank for high-quality perceptual audio coding," in *IEEE ASSP Workshop on Applications of Signal Processing to Audio and Acoustics*, 1997.

[16] Jürgen Herre, "Temporal noise shaping, quantization and coding methods in perceptual audio coding: A tutorial introduction," in *17th AES International Conference on High Quality Audio Coding*, 1999.

[17] ISO/IEC, *MPEG International Standard ISO/IEC 13818-7, Generic Coding of Moving Pictures and Associated Audio: Advanced Audio Coding*, JTC1/SC29/WG11, 1997.

[18] ITU-R, *Recommendation ITU-R BS.1534 Method for the subjective assessment of intermediate quality level of audio systems*, Geneva, 2014.