# LOW COMPLEXITY TONALITY CONTROL IN THE *INTELLIGENT GAP FILLING* TOOL

*Konstantin Schmidt, Christian Neukam*

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

## ABSTRACT

When coding audio signals at low bitrates with a transform coder the most prominent artifacts are spectral holes resulting from spectral lines being quantized to zero. State of the art codecs circumvent this by Noise Filling [1] and Bandwidth Extension (BWE) [2]. Both methods have in common that they do not code parts of the waveform itself but code a coarse description of the signal. At decoder side a synthetic signal is generated and adjusted according to the coded parameters. The presented system called *Intelligent Gap Filling* (IGF) is a combination of both methods. Spectral holes are filled with random noise or with copied decoded signal components from lower frequency regions. In the latter case a control mechanism is required to adjust the tonality of the copied signal components to reach good audio quality. This paper describes the way of controlling the tonality of IGF. The presented approach is of low complexity and allows for selective application without producing additional algorithmic delay. IGF is part of the 3GPP standard Enhanced Voice Services (EVS) as well as MPEG-H standardized by Moving Picture Experts Group (MPEG).

***Index Terms—*** audio coding, bandwidth extension, tonality, whitening, parametric coding, EVS, MPEG-H

## 1. INTRODUCTION

Introduced by Makhoul in 1979 the extension of bandwidth by copying spectrum has a long tradition in audio coding [3]. It is applied in modern audio codecs like HE-AAC [4] (inside the SBR tool), [5] et al. A dedicated tool for controlling the tonality of the copied signal is essential for achieving good audio quality in these codecs. This paper focuses on the control of tonality in the IGF tool which operates in the Modified Discrete Cosine Transform (MDCT) domain as it is part of the transform domain codecs in EVS [6] and MPEG-H [7]. Nevertheless the presented approach could be used in any transform domain coding scheme. After a short overview of the IGF system in EVS the proposed tonality control is described and finally evaluated with a listening test.

## 2. OVERVIEW OF IGF IN EVS

To be able to code both speech and music with good quality the EVS codec is a switched codec that codes an audio frame with either a CELP-like time domain coder or an MDCT based transform domain coder called Transform Coded Excitation (TCX). Further details of the whole codec can be found in [6].
The IGF encoder is described in figure 1. First a power spectrum estimate $PS$ is calculated by adding the MDCT of the TCX codec with a Modified Discrete Sine Transform (MDST) of same size for each bin $i$:

$$PS_i = MDCT_i^2 + MDST_i^2 \tag{1}$$

Then the power spectrum estimate - above a certain start frequency - is grouped into parameter bands of about equal size on the Bark scale. For each parameter band the power spectrum energy is averaged over MDCT bins not coded by the core coder (further details are provided in [8]). These average energies are quantized in the logarithmic domain and coded with an arithmetic coder. Furthermore information about the temporal envelope, temporal tile shaping (TTS) (explained in [9]) and the tonality (as described below) are calculated and added to the bitstream.
At decoder side (see figure 2) the MDCT lines not coded by the core coder are filled - depending on the transmitted tonality information - with one of the following signals:

- copied MDCT lines from lower frequencies
- whitened MDCT lines (as explained in detail below)
- white random noise

This synthesized signal will be scaled to match the original signal energy and finally the temporal envelope will be adjusted (more in [8]).

## 3. CONTROLLING TONALITY OF THE SYNTHESIZED SIGNAL

As mentioned before the control of tonality is part of many state of the art BWE schemes. There are different approaches

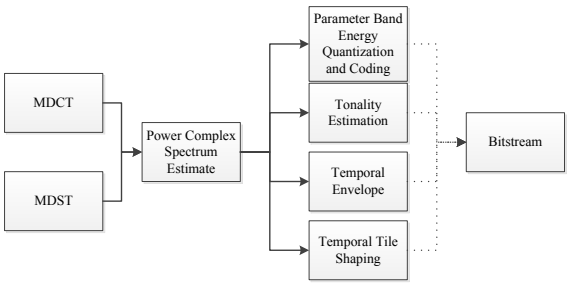


**Fig. 1**. IGF encoder overview

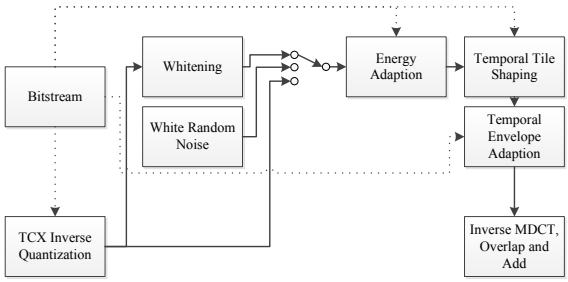


**Fig. 2**. IGF decoder overview

to synthesize a signal having less tonality. One is filtering the time line of filterbank sub-bands of adjacent frames with an Linear Prediction (LP) filter like in SBR [2]. This is feasible in the SBR tool since it operates on a filterbank with less frequency resolution (32 or 64 sub-bands) and higher time resolution. Since the presented IGF tool operates on the MDCT filterbank with up to 1200 sub-bands (although not all of them used for IGF) this approach would cause an intractable amount of computational complexity. Other approaches like the one presented in [10] are not able to remove any formant structure of the synthesized signal.

In order to control the tonality of the MDCT signal used to fill the gaps in the TCX coder, the following low complexity method is presented: Each MDCT bin is divided by the spectral envelope calculated by a moving average (MA) filter on the MDCT spectrum energy.

$$\hat{M}_i = \frac{M_i}{\sqrt{Env_i}} \tag{2}$$

where $M_i$ is the MDCT line $i$ and $Env_i$ is the MA filter calculated on the MDCT energy as

$$Env_i = \sum_{j=i-a}^{i+a} M_j^2 \tag{3}$$

and $2 \cdot a + 1$ is the length of the MA filter. The method in (2) however still needs a division and a square root per line, two very complex operations [1]. An unoptimized system might need more than 65 processor cycles for this whitening operation. To avoid this, both operations can be simplified in the log domain to:

$$\begin{aligned} \hat{M}_i &= \frac{M_i}{\sqrt{Env_i}} = M_i \cdot 2^{log_2(Env_i^{-\frac{1}{2}})} \\ &= M_i \cdot 2^{-0.5 \cdot log_2(Env_i)} \end{aligned} \tag{4}$$

This however still needs a logarithm and a power operation per line. By replacing the logarithm with a logarithm rounded to the next smaller integer the complexity is finally reduced to a minimum without harming the perceived audio quality. A logarithm rounded to the next smaller integer is on a fixed-point architecture the same as getting the highest active bit of a number and costs only one processor cycle. In addition the power operation reduces to a mere shift operation which also costs only one processor cycle.
The MA filter for bin $i$ can be implemented in a fixed-point architecture by initializing an accumulator with the sum of the first $2 \cdot a + 1$ lines:

$$acc = \sum_{j=i-a-1}^{i+a-1} M_j^2 \tag{5}$$

For line $i$ and each of the remaining lines these three steps have to be computed (shown in pseudo code):

$$\begin{aligned} acc &= acc + M_{i+a}^2; \\ acc &= acc - M_{i-a-1}^2; \\ Env_i &= acc; \end{aligned}$$

Neglecting a possible necessary scaling of the data the whole whitening operation can be carried out with only 8 operations per MDCT line:

- three operations for the MA filter
- one operation for getting the highest active bit (the rounded logarithm)
- one operation for negating the rounded logarithm
- two shift operations (one for the square root, one for inverting the logarithm)
- one multiplication.

This low complexity whitening operation will result in an MDCT signal that does not exhibit any formant structure. Furthermore, with an appropriate choice of the MA filter

[1] A division needs 32 processor cycles in BASOP for 32 bit variables. BASOP is part of the G.191 standard by the International Telecommunication Union (ITU) to count processor cycles

length, the perceived tonality will be lowered. This is important for signals with an unstable pitch - like speech signals. For these signals higher harmonics will spread over more bins than lower harmonics. And because in IGF MDCT bins are always copied from lower frequencies to higher frequencies this would introduce an unnatural tonality in the synthesized signal.

The optimal value of $a$ depends on the frequency resolution of the underlying transform. We found 3 to be a good value for $a$ having an MDCT with a frequency resolution of 25 Hz per bin. A smaller value would remove more tonal fine structure and make the signal more noisy. A larger value (10 or larger) would only remove formant structure and leave the tonal fine structure unchanged.

Figure 3 depicts the whitening operation. The uppermost figure shows the power spectrum of a very tonal signal. Below the envelope calculated by an MA filter is shown. The last figure shows the power spectrum divided by the envelope. The spectral peaks in the original spectrum are suppressed.

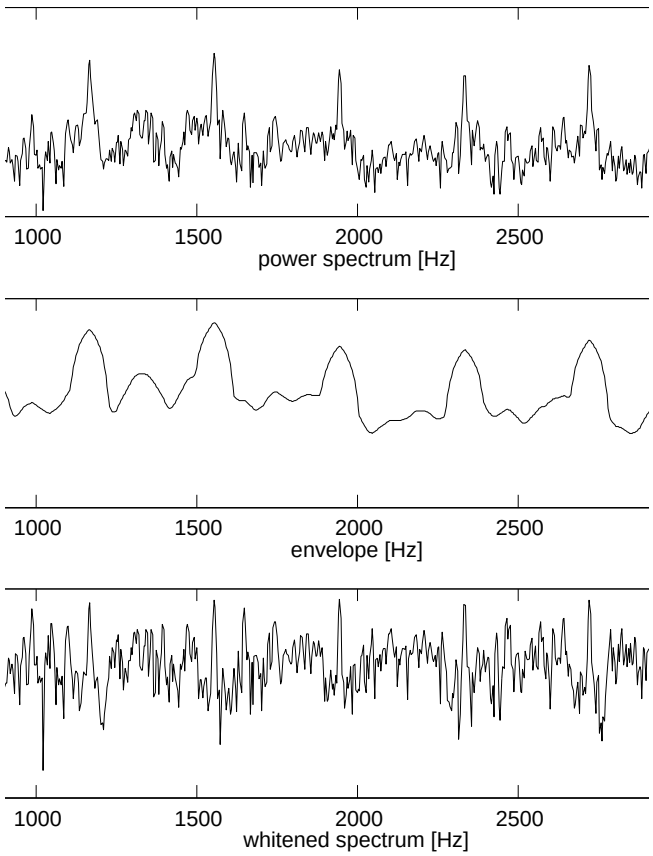


**Fig. 3**. Generation of the whitened spectrum (bottom) by dividing the power spectrum (top) by the envelope (middle)

Good audio quality requires a good decision about which of the three signals to use (*copied MDCT lines, whitened MDCT lines or white random noise* presented in figure 2). For this a feature $f$ is calculated and the selection is done by comparing the feature against certain thresholds. It is calculated on the power spectrum estimate of each frame at encoder side:

$$f_n = \frac{SFM_n}{CF_n} \tag{6}$$

where $SFM$ is the spectral flatness measure and $CF$ the crest factor of frame $n$:

$$SFM_n = \frac{2^{\sum_i^N log_2(PS(i))}}{\sum_i^N PS(i)} \tag{7}$$

and

$$CF_n = \frac{max(\sum_i^N log_2(PS(i))))}{\sqrt{\sum_i^N log_2(PS(i))^2}}, \tag{8}$$

where $max()$ picks the element $i$ with maximum value. Again the power spectrum estimate $PS$ is calculated by combining the MDCT with a MDST as in (1). For lowering the computational complexity of these features the logarithm is replaced by an integer-rounded logarithm as in (4).

A higher value of the feature means the signal is less tonal and a whitened MDCT spectrum is preferred as synthesized signal. In case the value is very high the signal is very noisy and it is beneficial to use pure white noise instead of copied MDCT signal. For certain signals, feature $f$ is seen to fluctuate a lot over consecutive frames, resulting again in an unstable synthetic signal due to an unstable decision. To circumvent this, the time line of the feature is filtered with a simple first order low-pass filter with the transfer function given by:

$$H(z) = \frac{0.5 + 0.5z^{-1}}{1 - 0.5z^{-1}} \tag{9}$$

In order to have a more selective control over the tonality, the parameter bands are combined to a coarser subdivision of the spectrum - called *tiles*. Depending on the codecs configuration the spectrum is divided in up to 4 tiles. In a typical configuration one tile is about 4 kHz wide. For each tile one distinct whitening level is added to the bitstream. The benefit of having more than one whitening level for the parametric representation of the signal will be shown in the next section. The thresholds for deciding which of the three signals to use in the decoder (*copied MDCT lines, whitened MDCT lines or white random noise*) are derived by creating a labeled data set containing a variety of signals ranging from very tonal to very noisy. The final thresholds are the ones minimizing the classification error.

## 4. EVALUATION

For evaluation of the proposed system several listening tests were conducted using the MUSHRA methodology [11]. Difference ratings with its means and confidence intervals were

calculated for every listener and every item. The statistical evaluation is based on Student's t-distribution. The audio signals used for the listening tests are described in table 1. All items are coded at 24 kbps.

| signal | description |
|---|---|
| a | female singing voice with music |
| b | radio jingle |
| c | speech with background music |
| d | classical music |
| e | harpsichord |
| f | rock music |

**Table 1**. Audio signals used for the listening test

First of all, the advantage of the selective tonality control is assessed by comparing the proposed system with three systems where tonality of the synthesized signal is not changed adaptively by any coded tonality information (see figure 4). It is either always random noise (STRONG), copied MDCT lines (OFF) or whitened copied MDCT lines (MID) for all frames. The results show first that over all files a system with adaptive tonality control is performing much better than a system without it. Looking at single results one sees that the presented system with tonality controlled by the presented feature is not always the best (b and f) but always very close to the best of the systems with fixed tonality. The biggest improvement of an adaptive tonality control is on classical music where the signal to be synthesized by IGF varies a lot in tonality.

To evaluate the influence of using the low complexity rounded logarithm a second listening test has been conducted where the proposed system is compared to a system which uses a floating point logarithm from the standard C library (see 5). Only one item was used for this test - the item where the presented whitening was used on most of the frames. One can see the audible influence of this optimization is quite small. In fact the optimized system is slightly better in quality but not significant.

## 5. FUTURE PROSPECTS

In principle it would be beneficial to derive the length of the MA filter from the pitch of the underlying signal so that the filter would cover a constant number of harmonics. This is because an MA filter covering many harmonics will not be able to change the tonality of the whitened signal. An MA filter covering only a single harmonic will make the signal very noisy. Although the pitch-lag is part of the TCX codec, there are often octave errors in the pitch estimation. Experiments using this pitch lag to derive the MA filter length have shown an unpleasant modulation in the whitened signal.

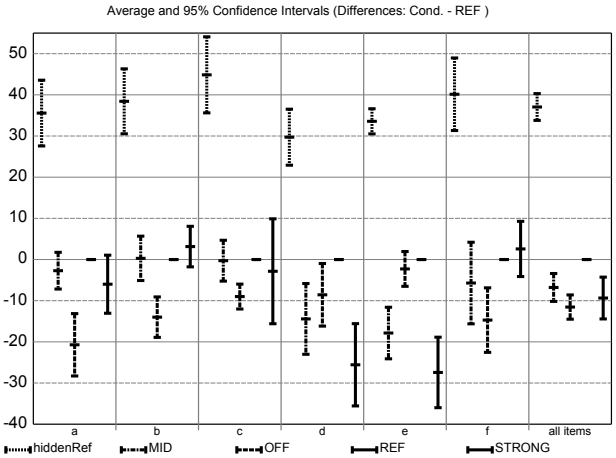


**Fig. 4**. Result of the listening test with 7 listeners. Box plots of differences to system with adaptive tonality control (REF). This system has a significantly better perceptual quality ($p < .05$) compared to the non-adaptive systems

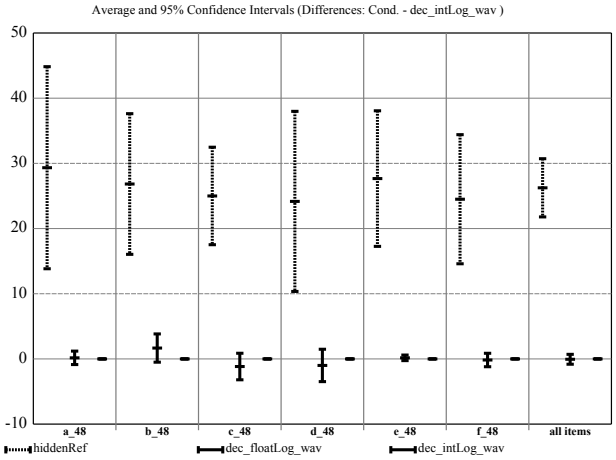


**Fig. 5**. Result of the listening test with 7 listeners. Box plots of differences to system with floating point logarithm instead of low complexity rounded logarithm don't show any significant difference

## 6. REFERENCES


[1] C.R. Helmrich, G. Marković, and B. Edler, "Improved low-delay MDCT-based coding of both stationary and transient audio signals," in *Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on*, May 2014, pp. 6954–6958.

[2] Martin Dietz, Lars Liljeryd, Kristofer Kjorling, and Oliver Kunz, "Spectral Band Replication, a Novel Approach in Audio Coding," in *Audio Engineering Society Convention 112*, Apr 2002.

[3] John Makhoul and Michael G. Berouti, "High-frequency regeneration in speech coding systems," in *IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP '79, Washington, D. C., USA, April 2-4, 1979*, 1979, pp. 428–431.

[4] Jürgen Herre and Martin Dietz, "MPEG-4 high-efficiency AAC coding [Standards in a Nutshell]," *Signal Processing Magazine, IEEE*, , no. Vol. 25, 2008, pp. 137–142, May 2008.

[5] M. Jelinek, T. Vaillancourt, and J. Gibbs, "G.718: A new embedded speech and audio coding standard with high resilience to error-prone transmission channels," *Communications Magazine, IEEE*, vol. 47, no. 10, pp. 117–123, October 2009.

[6] M. Dietz, M. Multrus, V. Eksler, V. Malenovsky, E. Norvell, H. Pobloth, Lei Miao, Zhe Wang, L. Laaksonen, A. Vasilache, Y. Kamamoto, K. Kikuiri, S. Ragot, J. Faure, H. Ehara, V. Rajendran, V. Atti, Hosang Sung, Eunmi Oh, Hao Yuan, and Changbao Zhu, "Overview of the EVS codec architecture," in *Acoustics, Speech and Signal Processing (ICASSP), 2015 IEEE International Conference on*, April 2015, pp. 5698–5702.

[7] Jürgen Herre, Johannes Hilpert, Achim Kuntz, and Jan Plogsties, "MPEG-H Audio - The New Standard for Universal Spatial/3D Audio Coding," *J. Audio Eng. Soc*, vol. 62, no. 12, pp. 821–830, 2015.

[8] C.R. Helmrich, A. Niedermeier, S. Disch, and F. Ghido, "Spectral envelope reconstruction via IGF for audio transform coding," in *Acoustics, Speech and Signal Processing (ICASSP), 2015 IEEE International Conference on*, April 2015, pp. 389–393.

[9] S. Disch, C. Neukam, and K. Schmidt, "Temporal Tile Shaping for spectral gap filling in audio transform coding in EVS," in *Acoustics, Speech and Signal Processing (ICASSP), 2015 IEEE International Conference on*, April 2015, pp. 5873–5877.

[10] K. Tsujino and K. Kikuiri, "Low-complexity Bandwidth Extension in MDCT domain for low-bitrate speech coding," in *Acoustics, Speech and Signal Processing, 2009. ICASSP 2009. IEEE International Conference on*, April 2009, pp. 4145–4148.

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