# PERCEPTUAL AUDIO CODING WITH ADAPTIVE NON-UNIFORM TIME/FREQUENCY TILINGS USING SUBBAND MERGING AND TIME DOMAIN ALIASING REDUCTION

*Nils Werner and Bernd Edler*

International Audio Laboratories Erlangen

## ABSTRACT

In this paper, we investigate the coding efficiency of perceptual coding using an adaptive non-uniform orthogonal filterbank based on MDCT analysis/synthesis and time domain aliasing reduction. We compare its performance to a system using a traditional adaptive uniform MDCT filterbank with window switching. The comparison is performed using a listening test at two different quantization settings. The statistical evaluation shows that the percetpual quality of the non-uniform filterbank significantly out-performs that of the uniform filterbank by 5 to 10 MUSHRA points.

***Index Terms—*** TDAC, MDCT, Perceptual Coding, Time-Frequency Transform

## 1. INTRODUCTION

In perceptual coding, entropy and thus bitrate is commonly reduced by discarding redundant and perceptually irrelevant information. This is achieved using a filterbank and quantization. This filterbank, a quantizer and a psychoacoustic model are used together to shape the quantization noise so it is as close to the masking threshold as possible, as to maximize the coding efficiency and perceptual quality of the overall system [2].

During synthesis, quantization noise will be shaped in time and frequency by the spectral and temporal shape of the filterbank's impulse and frequency response. It follows that, to allow fine-grained control of the quantization noise-shape, it is desirable to use a filterbank with an impulse response compact in both time and frequency.

The most commonly used critically sampled filterbank with these properties is the modified discrete cosine transform (MDCT), a filterbank which has a uniform time/frequency resolution in all bands.

However, the human auditory system exhibits a non-uniform time/frequency resolution [3], resulting in different masking threshold shapes for different frequencies. It is thus expected that a non-uniform filterbank with compact impulse responses will be able to follow the masking threshold more closely in both high and low frequencies. This allows, without introducting audible artifacts, the introduction of more quantization noise, thereby allowing for a lower bitrate than a uniform filterbank.

In our previous work, we were able to show that a non-uniform orthogonal filterbank based on cascading two MDCTs and time domain aliasing reduction (TDAR) was able to achieve impulse responses that were compact in both time and frequency [1]. A similar approach was published in [4], albeit without overlap in frequency, and limited to a sine-window in time.

In this work, we will evaluate the perceptual quality of such a non-uniform filterbank in an audio coder, and compare it to the performance of a uniform filterbank with window switching as used in current coders, such as Advanced Audio Coding (AAC) [5].

## 2. CODING SYSTEM

The evaluation system models a simple perceptual coder, with an analysis filterbank, a psychoacoustic model [6, 7], quantizer, perceptual entropy estimation [8], and a synthesis filterbank. In the two competing systems, the filterbank was either a uniform MDCT with window-switching [9] (WS), or a non-uniform MDCT with subband-merging and TDAR [1] (denoted by SM).

The relevant filterbank-parameters — window-switching boundaries for the uniform MDCT, or mergefactors and TDAR boundaries for the non-uniform MDCT — were adaptively and optimally chosen to minimize the overall remaining entropy. No additional post-processing steps or coding-tools were used.

### 2.1. Filterbank Parameters

The window switching filterbank uses an MDCT with the usual AAC frame lengths: long frames of 1024 samples or 8 short frames of 128 samples and appropriate transition windows between them. The cosine window was used.

The subband merging filterbank uses an initial MDCT with frame length 1024, and then divides the spectrum into 8 *mergefactor-bands* of 128 coefficients each. Each

N. Werner and B. Edler are with the International Audio Laboratories Erlangen, a joint research institute between the Friedrich-Alexander University Erlangen-Nürnberg (FAU) and Fraunhofer Institute of Integrated Circuits, IIS, Erlangen, Germany (e-mail: nils.werner@audiolabs-erlangen.de)

mergefactor-band may then be merged with a series of MDCTs with identical frame lengths $N \in \{1, 2, 4, 8, 16, 32\}$, called a *mergefactor*. As per design of the system, during analysis the optimal choice in mergefactors was not known yet, and each mergefactor-band does not know the mergefactor of any of its neighbors. Thus, the windows at the mergefactor-band edges were chosen to always be asymmetrical, and steep enough to accomodate the steepest possible neighbor mergefactor, see Figure 1.

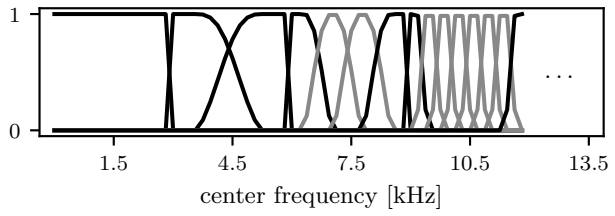


**Fig. 1**. Example window choices in four mergefactor-bands. Steep transition windows at the mergefactor-band edges are highlighted in black.

This design choice limits the overall flexibility of the filterbank and introduces less-than-ideal temporal ripples for these asymmetric windows [1], but offers a way to efficiently and independently optimize the mergefactor for each mergefactor-band.

The cosine window was used as the transform window, and a Kaiser-Bessel-derived window with $\beta = 5.1$ was used as the merge window.

Finally, quantization step sizes are controlled using a real valued distortion parameter $q$, which multiplicatively lowers or raises the estimated masking threshold from the perceptual model by the constant factor $q$. After quantization, the perceptual entropy estimator calculates a theoretical bitrate $r$, which naturally is dependent on $q$. For $q = 1.0$, the psychoacoustic model predicts transparent coding without any audible artifacts, for larger values $q > 1.0$, quantization stepsize increases, the bitrate $r$ drops, and the perceived quality of the resulting decoded signal is expected to deteriorate.

### 2.2. Parameter Optimization

To perform optimal parameter tuning, each signal was transformed and quantized using all possible parameter combinations, and the perceptual entropy of each frame for each parameter was estimated. Among all of the output coefficients, an optimal combination of parameters that minimizes the overal perceptual entropy was computed, and the output signal was then synthesized using these parameters.

To find optimal filterbank parameters, each mergefactor-band in each frame (a *merge-tile* of 128 coefficients) was quantized and its entropy was calculated. The graph of all parameters of all merge-tiles in one mergefactor-band then forms a trellis, where each transition weight equals the entropy of the following merge-tile.

As previously noted, not all parameter combinations and transitions will allow perfect reconstruction during synthesis, e.g. when switching from long to short frames, an asymmetric start window must be used inbetween. Similar rules apply for the use of TDAR in the non-uniform filterbank [1]. To prevent these illegal parameter transitions, the transition weights in the trellis were multiplied with a mask that encodes all legal and illegal transitions, i.e. 1 for legal and $\infty$ for illegal transitions. For window switching this method is described in great detail in [10].

Afterwards, a minimum-weight path through the trellis was computed using dynamic programming, resulting in an overal optimal parameter path in each individual mergefactor band that also guarantees perfect reconstruction.

This approach requires multiple encoding passes, a very large lookahead, and is thus not suitable for an actual on-line coder. However it guarantees that both methods performed at their highest possible efficiency at all times. For online encoding, methods for decoding such trellis diagrams under latency constraints exist [11].

Both systems assumed simple and uncompressed transmission of necessary side information: for window switching, 1 bit was used for each frame to signal long- and short-frames ($\lceil \log_2(2) \rceil = 1$). For subband merging, 29 bits were used per frame to signal mergefactor and TDAR flag (8 mergefactor-bands with 6 mergefactors and 2 TDAR values each, $\lceil \log_2((6 \times 2)^8) \rceil = 29$). Scale factors or masking thresholds were known at the decoder side.

## 3. GENERAL OBSERVATIONS

After running the encoding/decoding process, one can observe the following properties:

In the highest two to three mergefactor-bands, ranging from 15 kHz–24 kHz, the coder almost always chose a mergefactor of 1, disabling merging. In the midsection, mergefactor-bands 2–5 or frequency range between 3 kHz–15 kHz, the coder mainly chose either mergefactor 1 or 32. In the lower mergefactor-band, ranging from 0 kHz–3 kHz, the coder mostly chose mergefactors 1 and 2. Mergefactors 4, 8, and 16 were rarely chosen, see Figure 2.

This observation agrees with basic assumptions about the auditory system: the high frequencies exhibit a very high threshold in quiet, so effectively almost everything is quantized to zero, making the choice in mergefactor irrelevant. In the mid-range frequencies the auditory system has a high temporal resolution, while in the lower frequencies, the human ear has a higher frequency resolution.

Secondly, one notices that for any chosen distortion parameter $q$, the corresponding bitrate of the subband merging filterbank is below that of the window switching filterbank.

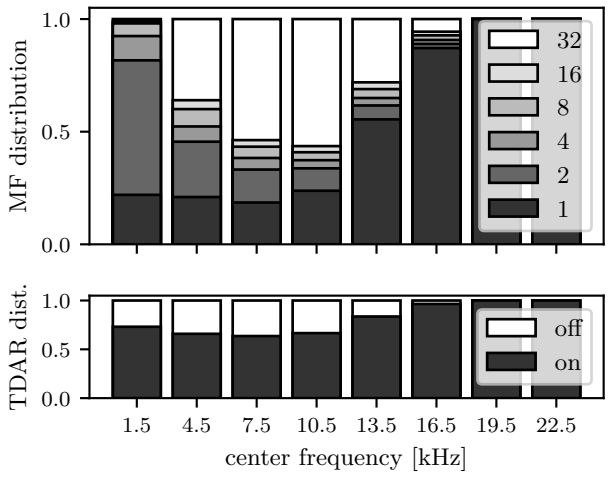


**Fig. 2**. Distributions of Mergefactor (MF) and Time Domain Aliasing Reduction (TDAR) choices made by the coder.

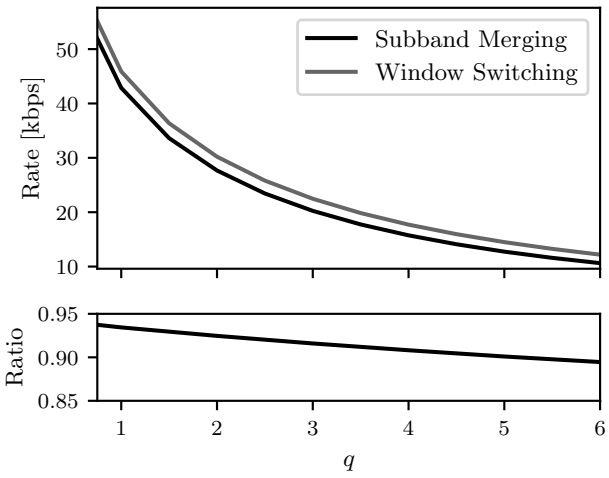


**Fig. 3**. Average bitrates of the two systems for different distortion parameters $q$ over 39 test items. Second plot shows the ratio between the two average bitrates.

On average, the non-uniform system required 5–13% fewer bits per sample to code the signals, see Figure 3.

## 4. LISTENING TEST SETUP

While the lower average bitrate achieved in Section 3 in itself is a nice result, it doesn't say anything about the perceptual quality of the system. What remains to be shown is that the two systems performed at a comparable perceptual quality level, or that we can use the bitrate difference to improve the perceptual quality of the more efficient system.

To test this, three different quality settings at different quantizer stepsize coefficients and thus average bitrates were considered: high quality (HQ), medium quality (MQ) and low quality (LQ), see Table 1.

| | $q$ | avg. Rate |
|---|---|---|
| High Quality (HQ) | 1.0 | ~ 46 kbps |
| Medium Quality (MQ) | 2.75 | ~ 26 kbps |
| Low Quality (LQ) | 4.0 | ~ 18 kbps |

**Table 1**. Quality settings and their distortion parameter $q$ and resulting average bitrate.

As per design of the perceptual model, for HQ no audible artifacts were expected [7]. And indeed, during small-scale ABC/HR (ITU-R BS.1116–3) [12] listening tests, expert listeners could not discern significant differences between either method and the reference signal. As conducting such a listening test on a larger scale is very exhausting but unlikely to reveal any meaningful results, it was skipped in favor of the two remaining quality settings MQ and LQ.

For MQ and LQ, the distortion parameter $q$ of the window switching filterbank system was chosen such that its output bitrate matched that of the subband merging filterbank. This means that the distortion parameter $q$ for the subband merging filterbank was lower than for the window switching filterbank. By that, we show that we can achieve a higher perceived quality, while allowing the same bitrate as for the window switching filterbank. To test this, a listening test using the Multi-Stimulus Test with Hidden Reference and Anchor method (MUSHRA, ITU-R BS.1534–3) [13] was conducted.

## 5. TEST SIGNAL CORPUS

The test signals for this evaluation were taken from a set of signals used for audio coder development and tuning. It comprised male and female speech, and several music recordings containing both harmonic and percussive sounds. All conditions were loudness normalized to -30 dB LUFS using ITU-R BS.1770–4 [14]. See Table 2.

| ID | Name |
|---|---|
| s01 | Castanets |
| s02 | Suzanne Vega — Tom's Diner |
| s03 | German Male Speaker |
| s04 | English Female Speaker |
| s05 | *Unknown* — A Foggy Day |
| s06 | Tracy Chapman — Mountain O' Things |
| s07 | Ornette Coleman — In All Languages |
| s08 | Fools Garden — Lemon Tree |

**Table 2**. Test items

| | W | p |
|---|---|---|
| MQ | 0.90 | $< .001$ |
| LQ | 0.95 | $< .001$ |

**Table 3**. Results of Shapiro-Wilk test for normality for the pairwise MUSHRA score differences between the window switching filterbank (WS) and subband merging filterbank (SM) at medium quality (MQ) and low quality (LQ) settings. The parameter W denotes W-statistic and p denotes p-value.

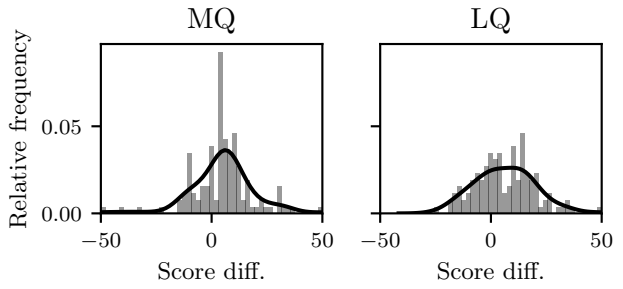


**Fig. 4**. Distributions and kernel density estimates of MUSHRA score differences between the window switching filterbank and subband merging filterbank at medium quality (MQ) and low quality (LQ) quality settings.

## 6. LISTENING TEST RESULTS

A total of N=16 expert listeners took part in the test.

First, a Shapiro-Wilk test was used to test the pairwise differences in MUSHRA scores between the two methods for normality. For LQ and MQ, the differences were significantly non-normal, see Table 3 and Figure 4. We therefore used a non-parametric Wilcoxon signed-rank test instead of the parametric paired t-test on all our conditions. A summary of all tests can be seen in Table 4.

A Wilcoxon signed-rank test was conducted to compare the perceptual quality of the two systems at MQ. There was a significant difference in the MUSHRA scores for the window switching filterbank and the subband merging filterbank, $p < .001$.

Secondly, a Wilcoxon signed-rank test was conducted to compare the perceptual quality of the two systems at quality setting LQ. There was a significant difference in the MUSHRA scores for the window switching filterbank and the subband merging filterbank, $p < .001$.

When comparing means and 95% confidence intervals of MUSHRA score differences for individual items, which are also commonly reported [13], the non-uniform filterbank consistently outperformed the window switching filterbank for all test items except one, see Figure 5.

| | | Median (IQR) | W | p |
|---|---|---|---|---|
| MQ | WS | 70.00 (27.50) | 2070.50 | $< .001$ |
| | SM | 80.00 (23.00) | | |
| LQ | WS | 46.50 (33.00) | 2181.00 | $< .001$ |
| | SM | 51.50 (25.25) | | |

**Table 4**. Median, Inter Quantile Range (IQR), and Wilcoxon signed-rank test results for the MUSHRA scores comparing the window switching filterbank (WS) and subband merging filterbank (SM) at medium quality (MQ) and low quality (LQ) quality settings. The parameter W denotes W-statistic and p denotes p-value.

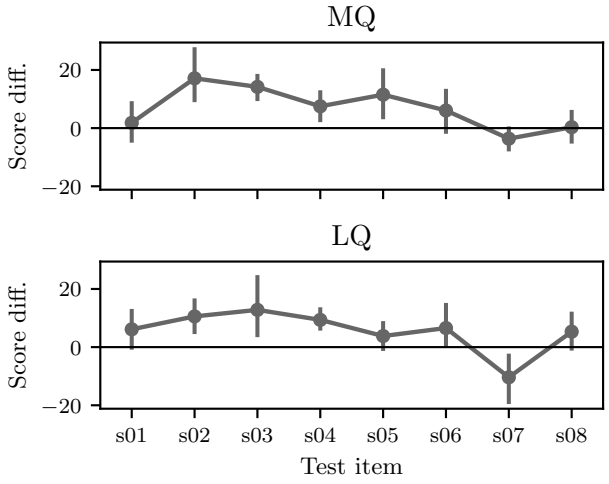


**Fig. 5**. Mean and 95% confidence intervals of MUSHRA score differences for individual items, window switching filterbank and subband merging filterbank at medium quality (MQ) and low quality (LQ) quality settings. Positive values favor subband merging over window switching.

## 7. CONCLUSION

In this paper, we presented a method of using a non-uniform orthogonal filterbank based on MDCT analysis/synthesis and TDAR in a simple audio coder. Its coding efficiency was then compared to a uniform window switching MDCT filterbank. On average, the non-uniform filterbank required 5–13% fewer bits per sample to code the test signals. This additional coding efficiency was subsequently used to improve the perceived quality of the coder at the same output bitrate. The improved perceived quality of 5 to 10 MUSHRA points was ascertained using a MUSHRA listening test and a subsequent statistical analysis. The difference in perceived quality was found to be statistically significant.

## 8. REFERENCES


[1] N. Werner and B. Edler, "Nonuniform Orthogonal Filterbanks Based on MDCT Analysis/Synthesis and Time-Domain Aliasing Reduction," *IEEE Signal Processing Letters*, vol. 24, no. 5, pp. 589–593, May 2017.

[2] Fernando C. Pereira and Touradj Ebrahimi, *The MPEG-4 Book*, Prentice Hall PTR, Upper Saddle River, NJ, USA, 2002.

[3] Brian C. J. Moore and Brian R. Glasberg, "Suggested formulae for calculating auditory-filter bandwidths and excitation patterns," *The Journal of the Acoustical Society of America*, vol. 74, no. 3, pp. 750–753, Sept. 1983.

[4] M. Purat and P. Noll, "A new orthonormal wavelet packet decomposition for audio coding using frequency-varying modulated lapped transforms," in *Proceedings of 1995 Workshop on Applications of Signal Processing to Audio and Accoustics*, Oct. 1995, pp. 183–186.

[5] Marina Bosi, Karlheinz Brandenburg, Schuyler Quackenbush, Louis Fielder, Kenzo Akagiri, Hendrik Fuchs, and Martin Dietz, "ISO/IEC MPEG-2 Advanced Audio Coding," *Journal of the Audio Engineering Society*, vol. 45, no. 10, pp. 789–814, Oct. 1997.

[6] Bernd Edler, Nicole Knölke, Jörn Ostermann, and Armin Taghipour, "Combination of Different Perceptual Models with Different Audio Transform Coding Schemes:Implementation and Evaluation," Nov. 2010, Audio Engineering Society.

[7] A. Taghipour, M. C. Jaikumar, and B. Edler, "A psychoacoustic model with Partial Spectral Flatness Measure for tonality estimation," in *2014 22nd European Signal Processing Conference (EUSIPCO)*, Sept. 2014, pp. 646–650.

[8] J. D. Johnston, "Estimation of perceptual entropy using noise masking criteria," in *ICASSP-88., International Conference on Acoustics, Speech, and Signal Processing*, Apr. 1988, pp. 2524–2527 vol.5.

[9] B. Edler, "Codierung von Audiosignalen mit überlappender Transformation und adaptiven Fensterfunktionen," *Frequenz*, vol. 43, pp. 252–256, Sept. 1989.

[10] V. Melkote and K. Rose, "Trellis-Based Approaches to Rate-Distortion Optimized Audio Encoding," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, no. 2, pp. 330–341, Feb. 2010.

[11] Mukund Narasimhan, Paul Viola, and Michael Shilman, "Online Decoding of Markov Models Under Latency Constraints," in *Proceedings of the 23rd International Conference on Machine Learning*, New York, NY, USA, 2006, ICML '06, pp. 657–664, ACM.

[12] ITU Radiocommunication Bureau, "BS.1116-3: Methods for the subjective assessment of small impairments in audio systems," *Recommendation ITU-R BS. 1116*, 2015.

[13] ITU Radiocommunication Bureau, "BS.1534-3: Method for the subjective assessment of intermediate quality level of coding systems," *Recommendation ITU-R BS. 1534*, 2015.

[14] ITU Radiocommunication Bureau, "BS.1770-3: Algorithms to measure audio programme loudness and true-peak audio level," *Recommendation ITU-R BS. 1770*, 2015.