Subband instantaneous-frequency analysis to determine masking with high temporal resolution for use in audio codecs

Nils Koppaetzky[1], Stephan D. Ewert, Birger Kollmeier, Volker Hohmann[2]

[1] *Carl von Ossietzky Universität Oldenburg, Germany, Email: nils.koppaetzky@uni-oldenburg.de*

[2] *Carl von Ossietzky Universität Oldenburg, Germany, Email: volker.hohmann@uni-oldenburg.de*

Introduction

Algorithms for perceptual audio coding use psychoacoustical masking models to compress audio data with minimal impact on the perceived audio quality. Latest psychoacoustical research emphasizes the excellent spectro-temporal resolution of the auditory system. Most common audio codecs, however, apply classical spectral masking models which generally reach the required high spectral resolution at the expense of an insufficient temporal resolution. An application of a current psychoacoustical model which simulates the high spectro-temporal resolution of the auditory system, might offer an increase of the coding efficiency or of the perceptual coding quality, respectively. To analyze the general effect of an enhanced time resolution, a current nonlinear auditory filterbank model with instantaneaous-frequency conrol was tested in this contribution. The applicability of the model for the classification of the tonality of the signal was examined. The tonality model as well as the resulting perceptual coding quality which can be achieved by a combination of the model with a simple perceptual coder are presented.

Motivation

The main principle behind perceptual audio coding is reducing the bitrate by reducing the quantisation accuracy, so that the error induced by the increased quantisation noise (q-noise) is imperceptible. This can be achieved by hiding the q-noise below the threshold of hearing in quiet and the masked threshold: When presenting a narrowband quantisation noise within one auditory frequency band (critical band, CB) with a level at or below the masked threshold within that critical band, the noise will be imperceptible. We distinguish two major masking patterns.

- Tone to noise masking:
  A tone masks a noise within a CB about 24 dB signal-to-noise ratio (SNR)
- Noise to noise masking:
  A noise masks a noise within a CB even at 6 dB SNR

Exploiting the noise-to-noise masking effect allows to increase the signal-to-quantitsation-noise ratio (SQNR) compared to the SQNR required for tone-to-noise masking without degrading perceptual quality. As the SQNR is proportional to the quantisation width in bits (6 dB SQNR per bit) this leads to a reducing of the bitrate depending on the signal tonality. The assumption underlying the present work is that the tonality estimate needs to be as fast as the temporal resolution of the auditory system, which is, according to recent auditory models, faster than what is normally achieved in audio coding schemes. To demonstrate the applicability of a fast tonality estimator based on a recent auditory model, it is combined with a simple audio coding scheme. In addition, it is proposed to shape the spectro-temporal characteristics of the quantisation noise according to the spectro-temporal resolution of the auditory system, which also leads to a variation of current audio coders, which basically shape the quantisation noise according to the spectro-temporal characteristics of the signal.

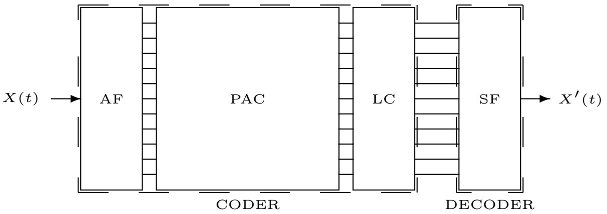

Figure 1: Block diagram of audio coder and decoder. CODER: Analysis Filterbank (AF) Perceptual audio coder (PAC) Lossless compression (LC). DECODER: Synthesis Filterbank (SF)

Audio Coder

The block diagram of audio coder and decoder used in this studs is shown in Fig 1. For the different processing blocks, the following algorithms were used:

- CODER
  - Modified discrete cosine transformation MDCT (AF)
  - Spectro-temporal block coding (blockcompanding) (PAC)
  - Huffman coding (LC)
- DECODER
  - Inverse-MDCT (SF)

Spectro-temporal block shaping

The MDCT is frequently used in audio coders and was chosen because it provides critical sampling and perfect reconstruction. The drawback is that it does not represent the spectro-temporal characteristics of the auditory system. The MDCT has a constant time and frequency resolution across the whole spectrum. The

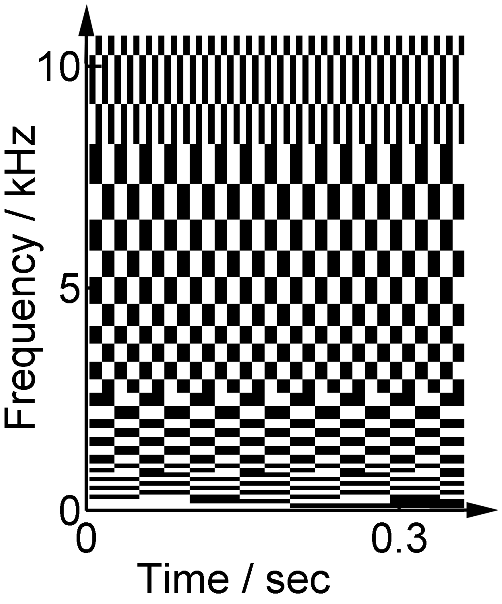

Figure 2: Logarithmic spectro-temporal-block-shaping: Processing blocks are alternately colored black and white to demonstrate the shape of the blocks.

auditory filter width is, like the filter width of an auditory filterbank like the Gammatone filterbank, increasing logarithmically with increasing center frequency. This leads to an increasing time resolution which is proportional to the filter width. To combine the constant (MDCT) and logarithmic (auditory filter) characteristics, it is proposed to shape the spectro-temporal blocks in which the MDCT spectrogram is quantized logarithmically. A possible shaping is shown in Fig. 2. At the low frequencies, blocks are narrow in frequency and long in the temporal dimension, i.e., cover only one or a few MDCT frequency bins, but several MDCT time frames. At high frequencies, blocks are wide in frequency, i.e., cover several MDCT frequency bins, but cover only one or a few MDCT time frames.

Instantaneous frequency based tonality estimation

Motivated by the auditory filterbank by Hohmann and Kollmeier[1], we propose to use an instantaneous frequency analysis at the outputs of Gammatone filters to determine the tonality of each subband signal. A gammatone filtered Signal $S'(t)$ can be assumed to be represented by an amplitude multiplied with a phase factor $S'(t) = A(t) \cdot e^{(i \cdot \phi(t))}$. The instantaneous frequency $f_{inst}(t) = \frac{d}{dt}(\phi(t)/2\pi)$ is the derivative of the phase with respect to time. As the instantaneous frequency is a time based measure it reaches a high temporal resolution. Important characteristics are:

- The variance of f_{inst} across time is high for noise signals and low for tonal siganls
- The mean of f_{inst} shows the principal frequency in each frequency subband.

Thus, it is proposed to use the variance of f_{inst} to estimate the signal tonality within an auditory filter. For this, the instantaneous frequency variance at the output of the respective gammatone filter is analysed. This allows to directly relate the bitrate (or SQNR) and the variance of f_{inst}. The SQNR is assumed to be proportional to the signal tonality as measured by $\left(\frac{var_\tau(f_{inst}(t))}{filterwidth}\right)^{-1}$ where the variance is taken over a windowlength of τ (it is reasonable to choose the time resolution of the MDCT for τ) and normalized with the filter width. This leads to the tonality-based bit assignment function which is shown in Fig. 3

Fig. 3 shows the relation between the variance of f_{inst} and the required SQNR which is proportonal to the quantisation width in bits used for the processing blocks. A variance of 0 represents the tone to noise masking paradigm and the variance at the root point of the function (black circle) represents the noise to noise masking (1 bit quantisation width). A linear function is chosen in between. The task that remains is to adjust the root point for correct tonality estimation.

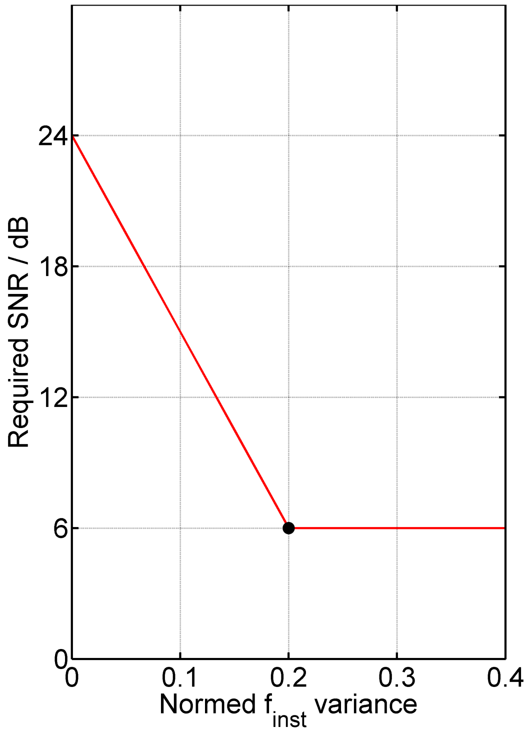

Figure 3: Tonality-based bit assignment fuction

Results

Fig. 4 shows the resulting coding scheme. The signal path contains MDCT, block companding using log-spaced specto-temporal tiles (STBC), Huffman coding and inverse MDCT. The tonality estimator contains auditory Gammatone filters (GFB) centered at the respective center frequencies of the tiles used for block-processing in the signal path. In each subband, the tonality is estimated (TE) from the subband instantaneous frequencies (IFA), and the number of bits used for the quantization of each block is determined according to the function in Fig. 3 . For the evaluation, a sampling frequency of 48kHz and 240 MDCT bands werde used. The time constants of the tonality estimator was set to 5 ms. As

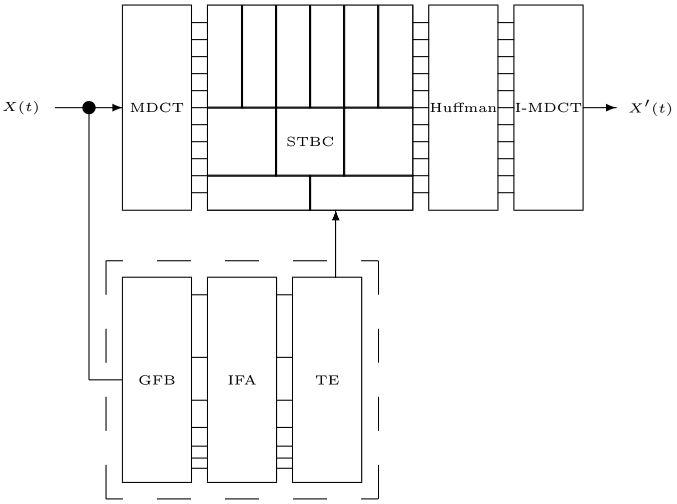

Figure 4: Block diagram of audio coder and decoder extended with tonality based bit assignment

a reference system, all blocks were coded with 5 bit, corresponding to an SQNR of about 30 dB. Blocks with levels below the threshold in quiet werde discarded (i.e., 0 bit were assigned to these blocks). Decreasing the root point for the tonality estimator starting at high values increases the influence of tonality estimation steadily. A 5s-sample of popular music (Way Back Into Love [Demo Version] performed by Drew Barrymore & Hugh Grant composed by Adam Schlesinger) was processed for several root points of the tonality function (Fig. 3). The perceptual quality was estimated for all processed samples by calculating the PEAQ measure[2], which assigns a quality index from 0 (imperceptible noise) to -4 (very annoying noise) by comparing the original and the coded signal. The result of this measurement is shown in Fig. 5.

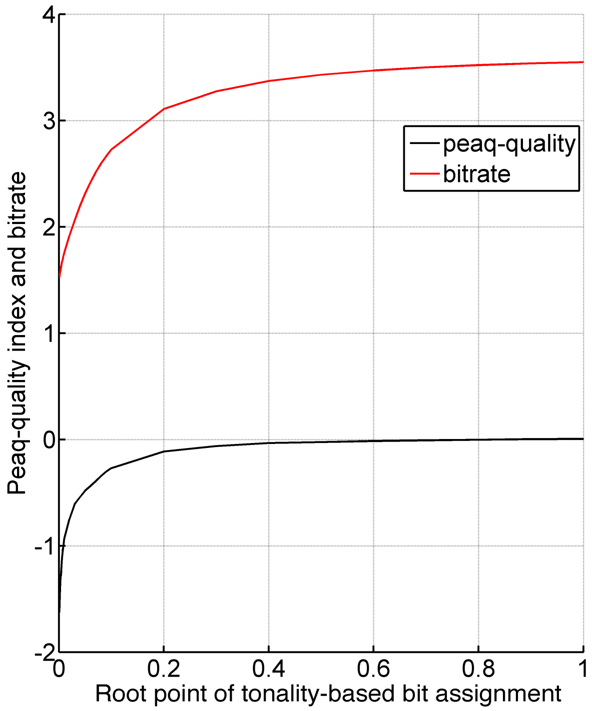

Figure 5: Bitrate and PEAQ quality index for various root points of the bit assignment function (shown in Fig. 3).

As we can observe, the bitrate (red curve) starts to decrease while the quality estimation (black curve) of PEAQ stays constant over a wide range. The point at which the quality curve starts to drop strongly is the root point to which we can reduce the bitrate by using tonality estimation without perceptual impact. If we allow a slight but not annoying perceptual impact (PEAQ index = -1) we can reduce the bitrate drastically against the bitrate without tonality-estimation-based bit assignment (reference coder, root point equals 1).
The histogram of the number of bits assigned to all blocks of the signal is shown in Fig. 6 for three different root points. The slope root points, bitrates (average over the whole signal) and quality indices are shown in Table 1 . We can observe that for decreasing slope point the maximum of the distribution shifts to the side with lower quantisation, as expected. This causes an decreasing average bitrate.

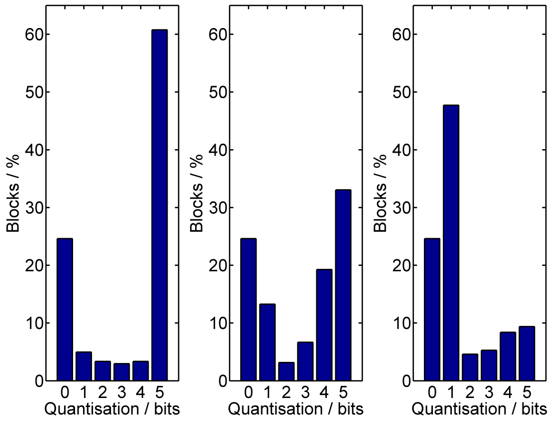

Figure 6: Histogram of bits assigned to all blocks of the 5s test signal for different root points (1,0.2,0.02).

	Root-Point	Bitrate (bit per sample)	PEAQ-Quality
Set1	1	3.2	−0.01
Set2	0.2	2.8	−0.22
Set3	0.02	1.8	−1.05

Table 1: Results

Summary

We have shown that mono audio coding with auditory system like spectro-temporal block shaping achieves an avarage bitrate of 3.3 bit per sample (compression rate of ≈ 1/5) The Instantaneous-frequenc-based tonality estimation employs a high temporal resolution and is capable to improve this result to an average bitrate of 3 bit per sample with no perceptual impact. If a minor perceptual impact (PEAQ quality index ≥ -1) is allowed, decreasing the avarge bitrate to 1.8 bps is possible which is nearly a compression rate of ≈ 1/10. This is most remarkable, because spectral masking across frequency bands was not included in the masking model. Extending the model towards instantaneous frequency based spectral masking estimation will be part of subsequent studies and should decrease of the bitrate further. Further studies with a larger database of audio samples is required to confirm the presented results.

References

[1] Hohmann, V., Kollmeier, B. (2006). A nonlinear auditory filterbank controlled by sub-band instantaneous frequency estimates. International Symposium on Hearing - ISH 2006, Cloppenburg, Germany, Springer.

[2] T. Thiede, W. C. Treurniet, R. Bito , T. Sporer, K. Brandenburg, C. Schmidmer, M. Keyhl, J. G. Beerends, C. Colomes, G. Stoll, B. Feiten PEAQ - der künftige ITU-Standard zur objektiven Messung der wahrgenommenen Audioqualität 20. Tonmeistertagung, Karlsruhe 1998. PEAQ