Proc. IX European Signal Processing Conference (EUSIPCO-98), Sept. 1998, Rhodes, Greece, pp. 1245-1248

AUDIO SUBBAND CODING WITH IMPROVED REPRESENTATION OF TRANSIENT SIGNAL SEGMENTS

Jörg Kliewer
University of Kiel
Institute for Network and System Theory
Kaiserstr. 2, D-24143 Kiel, Germany
Email: jkl@techfak.uni-kiel.de

Alfred Mertins
University of Western Australia
Dept. of Electrical & Electronic Eng.
Nedlands WA 6907, Australia
Email: mertins@ee.uwa.edu.au

ABSTRACT

In this paper, we present a subband audio coding scheme with an attack-sensitive framing of the input audio signal, where the frame boundaries closely match both ends of the transient. Since each transient frame is processed as a symmetrically extended finite-length signal with a support-preservative MDFT filter bank, pre-echos can be almost completely avoided. Furthermore, the bit-allocation, which is calculated on a frame-by-frame basis, can be determined more accurately, because the calculation of the masking thresholds is carried out on signal segments with almost "stationary" energy distribution.

1 INTRODUCTION

In audio subband coding, the frame boundaries for partitioning the input signal are usually chosen independently of the content of the audio signal. For attack-like transients (e.g. castanets, triangles etc.) this results in audible pre-echos, since the duration of premasking in the human hearing system can be almost neglected [1].

In the following, a method for improving the quality of attacks in the reconstructed signal based on adaptive framing will be presented. By choosing the frame boundaries appropriately the attacks always appear at the beginning of a segment while the segments end in front of the next attack or in a stationary region. Since each transient frame is processed as an individual finite-length signal, pre-echos can be almost completely avoided. In contrast, for other methods based on adaptively changing the resolution of the time-frequency plane via switched filter banks [2, 3] or post-filtering of pre-echo corrupted frames [4], where the placement of the frame borders is only allowed on a fixed time grid, the transients may appear at any position within a frame.

2 TRANSIENT EXTRACTION AND ADAPTIVE FRAMING

Attacks can be regarded as signal segments, whose energy rapidly changes from a low to a high level. Thus, we choose an energy-based approach for transient extraction which uses the signal energy within two sliding rectangular windows. Let $x(n)$ denote the input audio signal, and let

$$E_L(n) = \frac{1}{L} \sum_{k=n-L}^{n-1} x^2(k), \quad E_R(n) = \frac{1}{L} \sum_{k=n+1}^{n+L} x^2(k),$$

denote the energies within length-L windows on the left- and the right-hand side of a center point n, respectively. We can now define a simple criterion $C(n)$ for measuring the transient character of $x(n)$ according to

$$C(n) = c \cdot \log\left(\frac{E_R(n)}{E_L(n)}\right) \cdot E_R(n), \quad \text{with } c \in \mathbb{R}.$$

The application to an attack segment of a castanet signal is shown in Fig. 1. The largest value of $C(n)$ is obtained when

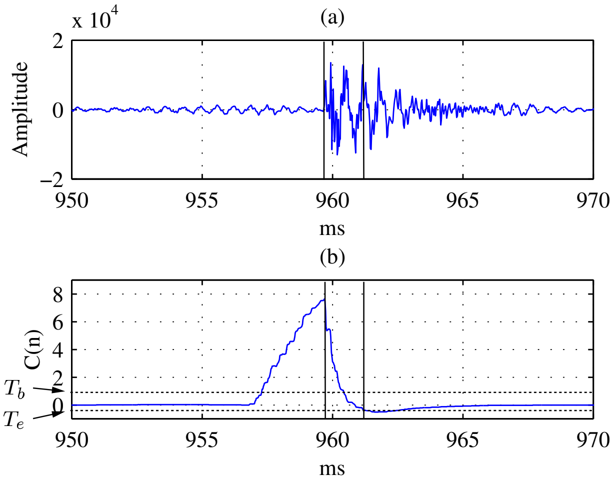

Figure 1: (a) Attack signal segment, (b) corresponding values of $C(n)$.

the center n is close to the "beginning" of an attack, where the ratio $E_R(n)/E_L(n)$ reaches its highest value. The search for the maximum starts when $C(n)$ is larger than a given threshold value T_b, which is marked with a dashed line in Fig. 1(b). The end of a transient is detected by searching for the largest value of $C(n)$ being smaller than some threshold T_e directly after the attack (see Fig. 1(b)).

Fig. 2 shows a simplified flowchart of the segmentation algorithm. In order to reduce the computational complexity, we use a two-step maximum search, where we first calculate a "subsampled" version of $C(n)$ with $n = K\,n'$, $n' \in \mathbb{N}$, on a coarser grid. As long as the increment K is smaller or equal to the width of the window L, we are still able to find the approximate position of a transient signal segment. If $C(n) > T_b$, a fine search between two window positions on the coarser grid is performed, where the new frame boundary

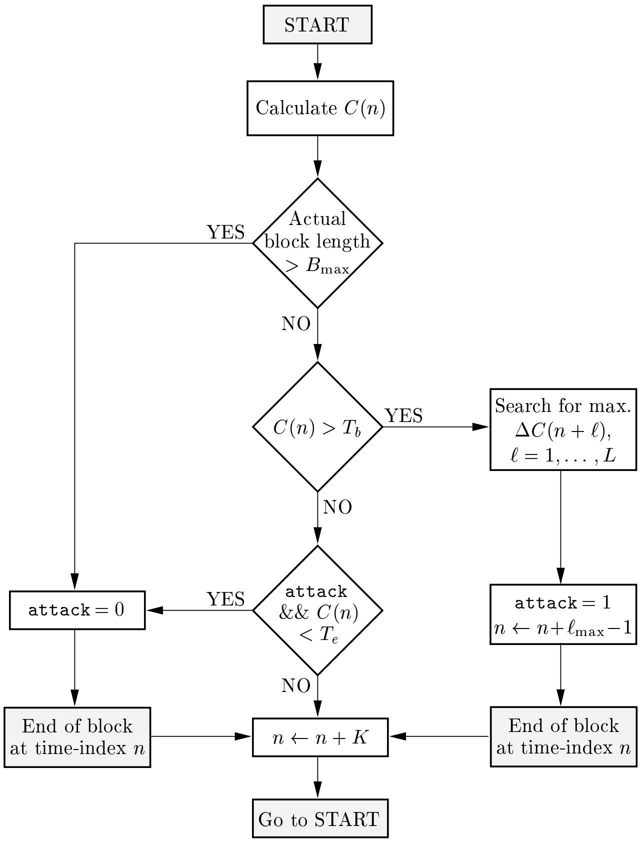

Figure 2: Simplified flowchart of the segmentation algorithm

is set at the time-index $n + \ell_{\max} - 1$. The index $\ell_{\max}$ denotes that value of ℓ, which leads to the maximal difference $\Delta C(n + \ell) = C(n + \ell - 1) - C(n + \ell)$. If $C(n)$ never exceeds the threshold T_b during the actual frame, the next segment starts after a maximal frame length of $B_{\max}$ samples.

Additionally, it is guaranteed that the segments are not shorter than a given minimal frame length $B_{\min}$. Thus for relatively stationary signals mainly long frames of length $B_{\max}$ are used (if T_b is chosen accordingly), whereas instationary parts necessarily lead to small frames of length $B_{\min}$.

Fig. 3 depicts the segmentation result for an excerpt of the castanet signal for $L = 128$ and $K = 64$. The beginnings of the transients exactly match the detected frame boundaries. Since the energy of the attack rapidly decreases, the lengths of these frames correspond to the minimal frame length $B_{\min}$ (here chosen as $B_{\min} = 384$). The next frame boundary is set after the maximal frame length of $B_{\max}$ samples (here $B_{\max} = 1536$), because this part can be regarded as almost stationary.

In order to avoid pre-echos, each transient frame must be processed as a finite-length signal, which requires the analysis and synthesis filter banks to be support preservative. As we can see in Fig. 3, we may have completely different signal behaviors at the left and the right boundary. Since in such cases circular convolution leads to noticeable boundary distortions after reconstruction when subband quantization is present [5], symmetric extension remains as the only applicable signal extension method when we want to process each frame without border distortions. This requires all analysis (and synthesis) filters to be linear-phase, which is satisfied

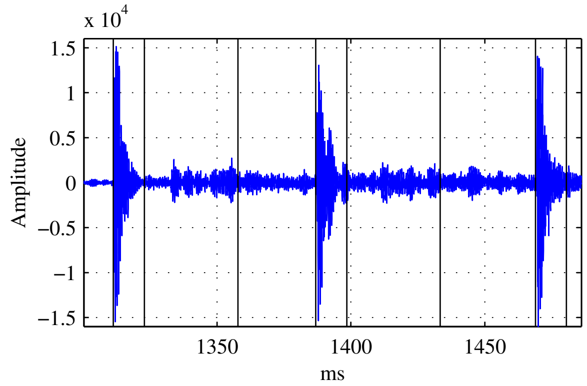

Figure 3: Castanet signal: Adaptive framing

by the Modified-DFT (MDFT) filter bank utilized in this paper.

3 MDFT FILTER BANKS

The MDFT filter bank is an M-channel complex modulated filter bank depicted in Fig. 4, where all linear-phase analysis and synthesis filters $H_k(z) = \mathcal{Z}\{h_k(n)\}$ and $F_k(z) = \mathcal{Z}\{f_k(n)\}$, resp., are obtained by complex modulation of a linear-phase prototype $p(n)$ with length L_p according to

$$h_k(n) = f_k(n) = p(n) \cdot e^{j2\pi/M \cdot k(n-(L_p-1)/2)},$$

with $k = 0, \ldots, M-1$, and $n = 0, \ldots, M-1$. Note that

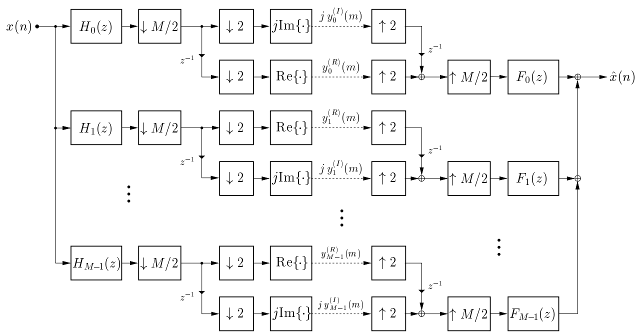

Figure 4: MDFT analysis and synthesis filter bank without subband modification

the linear-phase property does not hold for DCT-IV-type modulated filter banks, which are widely used in subband audio coders (i.e. in [6]).

It can be shown that due to the structural modifications in Fig. 4 (e.g. phase-shifts and alternated subsampling of real- and imaginary subband signals) the main aliasing components are canceled [7]. Thus, depending on the choice of the prototype, the MDFT filter bank can be designed for both almost perfect reconstruction and perfect reconstruction [8, 7]. For real-valued input signals $x(n)$, we have $y_k(m) = y^*_{M-k}(m)$ with $y_k(m) = y_k^{(R)}(m) + j\, y_k^{(I)}(m)$, $k = 1, \ldots, M-1$, where $y_k(m)$ denotes the k-th subband signal, such that the upper $M/2 - 1$ complex subbands are redundant and do not have to be transmitted to the synthesis side. In addition, the lowpass subband samples $y_0(m)$ are purely real and the highpass subband $y_{M/2}(m)$ is either real valued for L_p odd or purely imaginary for L_p even. Overall,

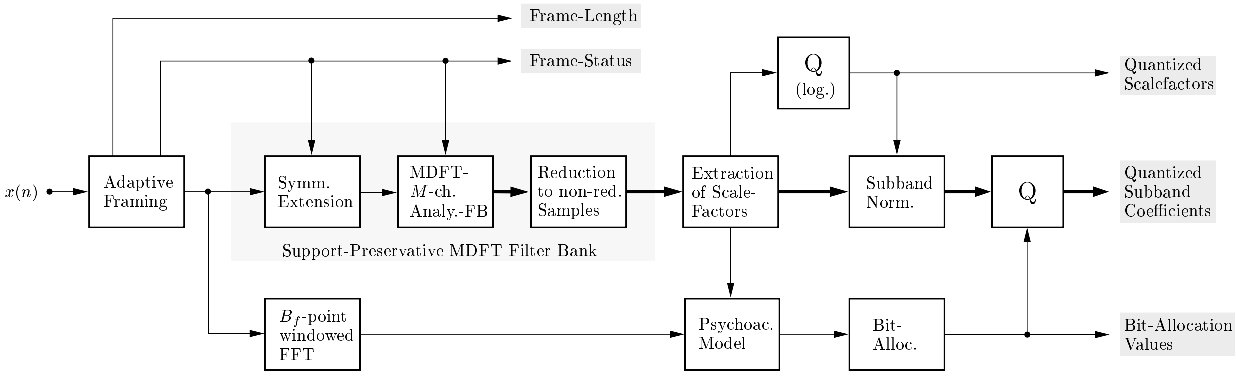

Figure 5: Encoder block structure

this leads to a critically subsampled filter bank for infinite-length signals.

The required support preservation for finite-length input signals can be achieved by applying the symmetric extension method in [9]. Note that due to the subsampling with and without a phase shift the linear-phase subband filters in Fig. 4 have different centers of symmetry, which excludes the utilization of standard symmetric extension techniques as in [5].

4 ENCODER STRUCTURE

The block diagram of the encoder, which is a modification of the MPEG Layer 1 encoder [6], is depicted in Fig. 5. Herein, all grey shaded parameters are sent to the decoder. After an adaptive partitioning of the input audio signal $x(n)$ (sampled at 44.1 kHz), the resulting (finite-length) signals are processed with the MDFT analysis filter bank. Since each frame-length has to be transmitted to the receiver, we restrict the frame-lengths to be multiples of $M/2$, which additionally leads to a convenient formulation of the symmetric extension method in [9]. The scale-factors (extracted from the M subband vectors) are logarithmically quantized, used for normalizing the subband signals, and then sent to the decoder. The bit-allocation values, which are used to linearly quantize all subband samples in the current block, are calculated via a psychoacoustic model on a frame-by-frame basis.

4.1 Processing of Finite-Length Signals

If all signal segments between two frames are processed individually as finite-length signals, blocking artifacts due to abrupt changes of the subband bit-allocation are likely to occur. One solution is a slight overlapping of two frames, which on the other hand leads to an increased bit-rate demand and thus to a loss of quality (especially for stationary signals, when the overall bit-rate is kept). This disadvantage can be avoided when the input signal is divided into regions according to Fig. 6. When an attack is detected, the following frame is processed as an individual signal via symmetric extension at the beginning and end of the frame. For almost stationary regions, we group several blocks to obtain a longer section, where the symmetric extension is only carried out at the beginning of the first frame and at the end of the last frame of this section. This results in an overlapped processing of the individual stationary frames except on the section

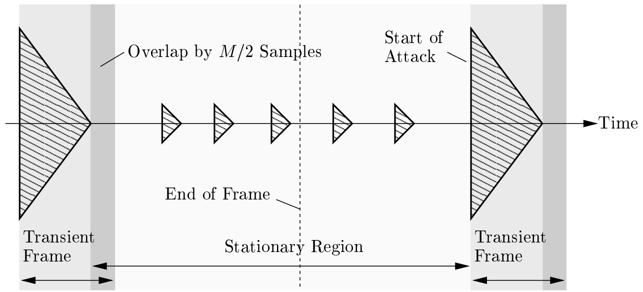

Figure 6: Partitioning of the input signal

boundaries. In order to avoid the blocking distortions, we still use an $M/2$-sample overlapping of two finite-length signal sections when the second section has no transient character (see Fig. 6). For attack-like frames, these distortions are assumed to be (pre-)masked by the attack. Note that due to the finite-length character of transient and stationary signal regions it is also possible to choose the filter bank parameters (i.e. prototype length, number of subbands) independently for each region.

4.2 Signal Adaptive Calculation of the Masking Threshold

In addition to the subband decomposition, an adaptive B_f-point windowed FFT is calculated from the partitioned input data, where $B_f = 2^L$ is the smallest power of two being larger than the actual input frame length $B_{\min} \leq B \leq B_{\max}$. The masking model utilized for calculation of the masking threshold is based on the MPEG psychoacoustic model 1 [6], but has been adapted to handle different FFT-window sizes. Note that since the number of FFT-points is chosen in accordance to the frame length, the masking thresholds are obtained with the appropriate time-frequency resolution. This corresponds to the behavior of the individual frames (stationary or transient). For example, for a transient frame of the castanet signal in Fig. 3, a lack of temporal resolution could lead to inaccurate masking thresholds and thus to noticeable distortions in the reconstructed signal.

Since the real and imaginary part of the subband samples $y_k(m)$, $k = 1, \dots M/2-1$ cover the same spectral range, both are quantized with the same bit-allocation, which is derived from the masking threshold.

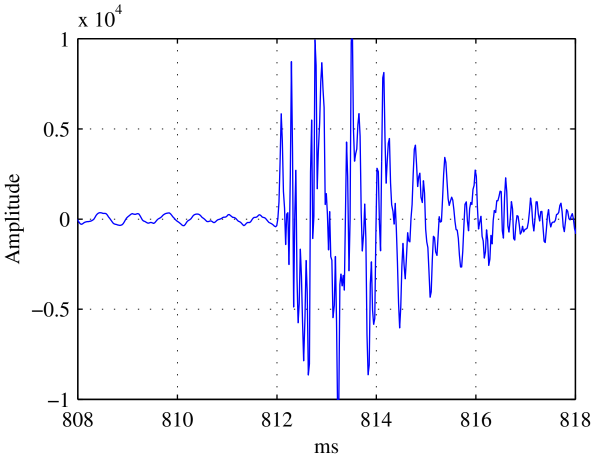

Figure 7: Original signal segment

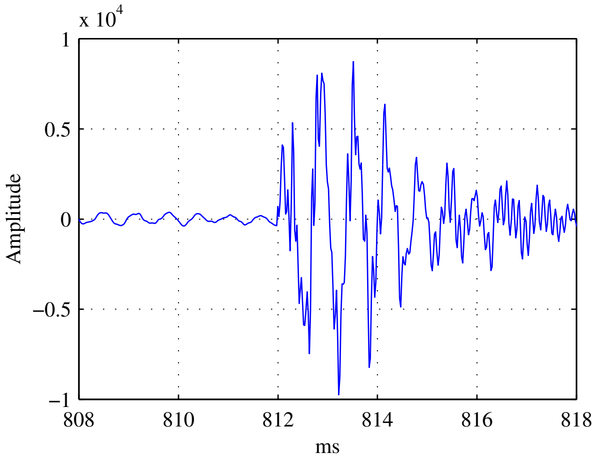

Figure 8: Reconstructed signal segment, coded at 80 kBit/s.

5 RESULTS

For the parameters $M = 64$, $B_{\min} = 384$, $B_{\max} = 1536$ and a prototype of length $L_p = 512$, the coder provides nearly transparent quality for most signals on the EBU SQAM-CD [10] with bit-rates in the range of 75-85 kBit/s for a single channel, where the lower value can be achieved for signals containing many "instationary" passages, as castanets and triangles. Especially for those signals we found a noticeable improvement on the quality of the reconstructed signal, when increasing the time-resolution of the analysis filter bank for non-stationary frames by switching to a shorter prototype with length $L_p = 256$. Further reduction of the bit-rate leads to noticeable lack of high frequency components for transient-like signals and to aliasing and noise-like distortions for stationary signals. When increasing the number of subbands, we also need to increase the minimum block length $B_{\min}$ in order to provide a sensible ratio of side- and subband coefficient information. However, this leads to a distorted representation of very sharp attacks as in the castanet signal.

The pre-echo suppression capabilities of the adaptive segmentation are visualized in Figs. 7 and 8. Fig. 7 shows the transient part of the castanet signal, whereas the reconstructed result is displayed in Fig. 8, after being coded at 80 kBit/s. Nearly no pre-echo artifacts are visible in the plot of the reconstructed signal.

6 CONCLUSION

In this paper, a method of avoiding the pre-echo problem for transient-like signals has been presented. In the proposed coder, the input signal frames obtained by an attack-sensitive separation of the audio signal are divided into stationary and transient regions, and the resulting signals are symmetrically extended and processed as finite-length signals with a support-preservative filter bank. Due to the adaptive framing of the input signal, the calculation of the psychoacoustic model can be carried out on signal segments with almost "stationary" energy distribution. Furthermore, we are able to use very long prototypes providing high stop-band attenuation and thus high coding gain, especially for almost stationary parts of the signal, without increasing the sensitivity to pre-echos.

7 ACKNOWLEDGEMENT

The authors would like to thank Mr. Christian Steinert for his invaluable help in programming and testing the simulation software.

8 REFERENCES

[1] E. Zwicker and H. Fastl. *Psychoacoustics*. Springer–Verlag, Berlin, 1990.

[2] J. Princen and J. D. Johnston. Audio coding with signal adaptive filterbanks. In *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, pages 3071–3074, Detroit, USA, 1995.

[3] D. Sinha and J. D. Johnston. Audio compression at low bit rates using a signal adaptive switched filterbank. In *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, volume 2, pages 1053–1056, Atlanta, USA, 1996.

[4] Y. Mahieux and J. P. Petit. High-quality audio transform coding at 64 kbps. *IEEE Trans. on Communications*, 42(11):3010–3019, November 1994.

[5] M. J. T. Smith and S. L. Eddins. Analysis/synthesis techniques for subband image coding. *IEEE Trans. on Acoust., Speech, Signal Processing*, ASSP-38:1446–1456, August 1990.

[6] International Organization for Standardization. *Coding of Moving Pictures and Associated Audio for Digital Storage Media at up to about 1.5 MBit/s, Audio Part (11172-3)*, November 1992.

[7] T. Karp and N. J. Fliege. MDFT filter banks with perfect reconstruction. In *Proc. IEEE Int. Sympos. Circuits and Systems*, Seattle, USA, May 1995.

[8] J. Kliewer. Simplified design of linear-phase prototype filters for modulated filter banks. In *Proc. European Signal Processing Conf.*, Signal Processing VIII: Theory and Applications, pages 1191–1194, Trieste, Italy, 1996.

[9] T. Karp, J. Kliewer, A. Mertins, and N. J. Fliege. Processing arbitrary-length signals with MDFT filter banks. In *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, pages 1479–1482, Atlanta, USA, May 1996.

[10] European Broadcasting Union, Geneva, Switzerland. *Sound Quality Assessment Material: Recordings for Subjective Tests*, 1988.