BEST WAVELET-PACKET BASES FOR AUDIO CODING USING PERCEPTUAL AND RATE-DISTORTION CRITERIA

Markus Erne, George Moschytz, Christof Faller

Swiss Federal Institute of Technology

Signal and Information Processing Laboratory
Sternwartstrasse 7
CH-8092 Zurich, Switzerland
erne@isi.ee.ethz.ch

ABSTRACT

This paper presents a new approach to the adaptation of a wavelet filterbank based on perceptual and rate-distortion criteria. The system makes use of a wavelet-packet transform where each subband can have an individual time-segmentation. Boundary effects can be avoided by using overlapping blocks of samples and therefore switching bases is possible at every tree-level without affecting other subbands. A modified psychoacoustic model using perceptual entropy can control the switching of the wavelet filterbank and the individual time-segmentation of every subband allows to take advantage of temporal masking. Additionally a rate-distortion measure can control the filterbank for lossless audio coding applications or in cases where large coding gains can be achieved without using perceptual criteria. The weight of the perceptual measure as well as the weight of the rate-distortion measure can be selected individually, enabling to trade lossless coding versus perceptual coding.

1. INTRODUCTION

In the last few years, many high quality audio compression algorithms have been developed. Some make use of uniform polyphase filterbanks and others are based on modified discrete cosine transforms [1], using window switching. Although window switching will help to minimize blocking artifacts such as pre-echoes, spectral distortion at the frame-boundaries cannot be avoided. Some algorithms use lapped orthogonal transforms [2], [3], [4] and many proposals for wavelet-based audio coding schemes [5], [6], [7] have been published recently. Uniform polyphase filterbanks can be implemented efficiently, they do not approximate the human auditory system well and they do not offer large coding gains in a rate-distortion metric. Transform coders use block-based processing and show spectral distortion at the block-boundaries as well as pre-echo phenomena. The variety of existing musical instruments such as castanets, harpsichord or pitch-pipe exhibiting various coding requirements due to their completely different temporal and spectral fine-structure, suggests to use a filterbank with variable time-frequency resolution. Wavelet-filterbanks are known for a flexible time-frequency tiling but most wavelet-based audio coding algorithms are focussed to mimic the response of the human auditory system. Although a frequency resolution of the filterbank equal to the human auditory system will allow to apply frequency-domain masking accurately because every critical band has a dedicated quantizer, such a system does not optimize the coding gain in a rate-distortion sense. Additionally this approach implies that the signal energy is spread over the full audio-bandwidth and therefore does not allow to allocate subband resources for better spectral or temporal resolution in case of momentarily bandlimited input signals. Best-basis search algorithms in a rate distortion sense for wavelet-packet transforms have been published for a fixed time-segmentation [8] as "single-tree" algorithm as well for variable time-segmentation over all subbands as "double-tree" algorithm [9]. We extended these techniques to a variable time-segmentation in every subband [10]. This framework allows to individually switch nodes of the wavelet-packet tree at completely different locations in time without affecting other nodes of the tree. The approach is well adapted to musical notation. In order to track each individual note, a flexible time-segmentation of every subband must be achieved and the position and the width of the subband in terms of pitch must be altered as well.

2. SIGNAL ADAPTIVE WAVELET-FILTERBANK

2.1. Boundary Conditions

Tree-structured wavelet-packets provide a set of orthonormal bases in $L^2(\mathbf{R}^N)$. A wavelet-decomposition can be written in matrix form, using infinite matrices:

$$\mathbf{y}_\infty = \mathbf{A}_\infty \mathbf{x}_\infty$$

With the infinite matrix $\mathbf{A}_\infty$:

$$\begin{pmatrix}
\ddots & & & & & & & & \\
& l[n] & l[n-1] & \dots & l[1] & & & & \\
& h[n] & h[n-1] & \dots & h[1] & & & & \\
& & & l[n] & l[n-1] & \dots & l[1] & & \\
& & & h[n] & h[n-1] & \dots & h[1] & & \\
& & & & & \ddots & \ddots & \ddots & \ddots \\
& & & & & & & l[n] & l[n-1] \dots l[1] \\
& & & & & & & h[n] & h[n-1] \dots h[1] \\
& & & & & & & & \ddots
\end{pmatrix}$$

For finite length signals, we selected the finite length sub-matrix $\mathbf{A}$ to the form:

$$\begin{pmatrix} l[n]\, l[n-1] \ \ldots \ \ l[1] & & & & \\ h[n]h[n-1] \ \ldots \ \ h[1] & & & & \\ & l[n]\, l[n-1] \ldots l[1] & & & \\ & h[n]h[n-1] \ldots h[1] & & & \\ & & \ddots \ \ddots \ \ddots \ \ddots & & \\ & & & l[n]\, l[n-1] \ldots l[1] & \\ & & & h[n]h[n-1] \ldots h[1] & \end{pmatrix}$$

For the decomposition of a signal-vector of length k, $\mathbf{A}$ has k columns and $k-(N-2)$ rows for a given filterlength N. The decomposed signal $\mathbf{y}$ therefore can be written in matrix form: $\mathbf{y} = \mathbf{A}\mathbf{x}$ For the reconstruction, the synthesis matrix $\mathbf{B}$ is chosen to be a submatrix of $\mathbf{A}$. $\mathbf{B}$ has $k-(N-2)$ columns and $k-2(N-2)$ rows.

$$\begin{pmatrix} l^*[n]\, l^*[n-1] \ \ldots \ \ l^*[1] & & & & \\ h^*[n]h^*[n-1] \ \ldots \ \ h^*[1] & & & & \\ & l^*[n]\, l^*[n-1] \ldots l^*[1] & & & \\ & h^*[n]h^*[n-1] \ldots h^*[1] & & & \\ & & \ddots \ \ddots \ \ddots \ \ddots & & \\ & & & l^*[n]\, l^*[n-1] \ldots l^*[1] & \\ & & & h^*[n]h^*[n-1] \ldots h^*[1] & \end{pmatrix}$$

The reconstructed output vector $\hat{\mathbf{x}}$ is of the form: $\hat{\mathbf{x}} = \mathbf{B}\mathbf{y}$

$$\hat{\mathbf{x}} = \begin{pmatrix} \hat{\mathbf{x}}[1] \\ \hat{\mathbf{x}}[2] \\ \hat{\mathbf{x}}[3] \\ \vdots \\ \hat{\mathbf{x}}[k-2(n-2)] \end{pmatrix} \text{ and } \mathbf{AB} = \begin{pmatrix} 0 \ldots 0 \ 1 & \ldots 0\, 0 \ldots 0 \\ \vdots \ \ \vdots & 1 \quad \vdots \ \vdots \\ & \ddots \\ 0 \ldots 0 \ldots & 1\, 0 \ldots 0 \end{pmatrix}$$

and it can be shown [10] that $k-2(N-2)$ samples out of the k samples of $\mathbf{x}$ can be reconstructed perfectly. Such a framework allows to process overlapping blocks of input samples in order to reconstruct the signal perfectly. In contrast to windowing-methods used in transform-coders or implementations using boundary-wavelets, no spectral distortion at the frame-boundaries can occur. Extending this method to the full wavelet-packet-tree of depth L, a common time-measure for switching the basis at all nodes of the wavelet-packet tree can be defined.

2.2. Switching Bases

By implementing a sophisticated memory management, a framework can be realized which allows to up- and down-switch the basis at every level of the tree. It is evident that for tree-levels near the root, the basis can be switched more frequently which matches a requested high temporal resolution in the upper frequency bands whereas at the lowest tree-level with narrow subbands, a lower temporal switching-resolution due to fewer available samples can be tolerated. All nodes of the wavelet-packet tree can be switched individually and the filterbank can fully adapt to the signal, depending on different criteria.

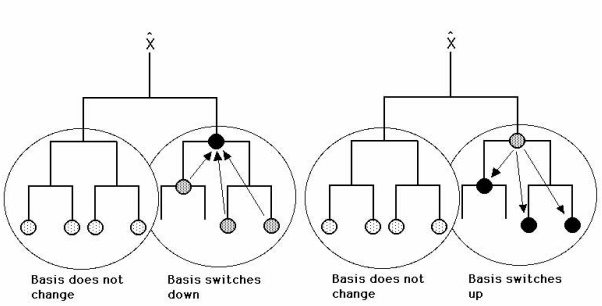

Figure 1: Up- and Down-Switching between different nodes of the wavelet-packet-tree

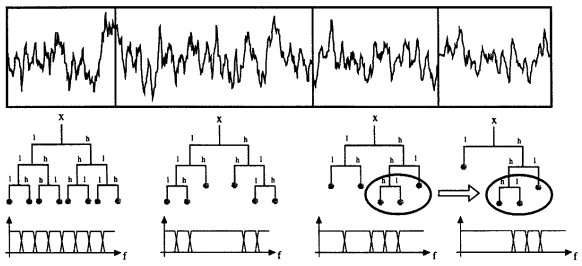

Figure 2: Individual switching of the filterbank, depending on the input signal

2.3. Choice of the wavelet

The length and the choice of the wavelet are not only important for the frequency selectivity and the time-resolution of the wavelet-filterbank but as we have shown, the length N of the filter will influence the switching of the filterbank directly. N should be chosen as small as possible in order to guarantee high time-resolution and a maximum of possible up- and down-switching positions per input-block. Additionally, the wavelet should have a maximum of vanishing moments. In contrast, tree-structured filterbanks tend to have a limited frequency separation of the individual subbands due to the iteration of the filter within the tree. Although wavelet-transforms can have perfect reconstruction, care has to be taken in the case of a perceptual audio coder. A wavelet-based perceptual audio coder will require some quantizing of the wavelet coefficients and therefore unmasked, aliased quantization noise may appear in sidelobes of the subband filters. Therefore sufficient stopband attenuation of the subband filters is required and longer FIR-filters are needed. A compromise between the requirement for high frequency separation between adjacent bands and high temporal resolution has to be found and it turned out that Daubechies wavelets of length $N = 20$ and Beylkin wavelets of similar lengths are valid candidates.

3. BEST BASIS SEARCH

Having developed a framework for the individual switching of each node of the wavelet-packet tree, a measure on how to find the best basis for each signal interval has to be evaluated.

3.1. Best basis search using a rate-distortion measure

Best basis search algorithms have been published [11] and some of them make use of a least mean square error or a one-sided entropy metric. The momentary entropy in subband j at level i of the wavelet-packet tree is:

$$entropy_{ij}[k] = \frac{1}{N}\sum_{n=1}^{N} -\log_2(p_{ij}[k][quantized_{ij}[k-n+1]])$$

The reason we have chosen a common time measure for the up- and down-switching of every node now becomes obvious. In order to compare the entropy in every subband, we need to scale the entropy according to the number of samples in each subband. The scaled entropy in each subband is computed using a sliding window and a forgetting-factor for past samples before becoming part of a cost-function for every subband. The overall costs are compared for the parent node and both children nodes and depending on the result, the basis is switched up or down accordingly. The

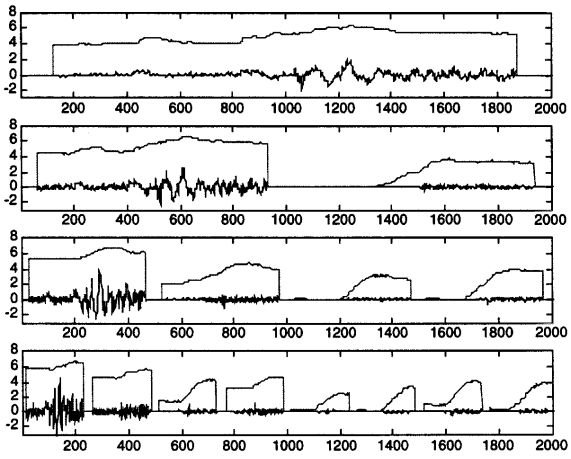

Figure 3: The scaled entropy in each subband is computed

same principle can be used if the scaled energy in every subband is used as a reference for switching the basis. Although the one-sided metrics such as entropy and energy do work well for fixed quantizers, they are not optimal in a rate-distortion sense. In [12], a method has been presented which jointly finds the optimal basis and the optimal quantization using the Langrangian cost function:

$$J(\lambda) = D + \lambda R$$

It can be shown that R-D optimality can be achieved when all leaves of the wavelet packet tree operate at a constant slope on their R-D curves.This approach will give best results in a rate-distortion sense, but it does not take any perceptual criteria into consideration.

3.2. Best basis search using a perceptual measure

For a perceptual measure, masking effects of the human auditory system become very important. In frequency domain masking, a strong noise or a strong tone masker will mask the noise or a tone of the maskee [13]. All signals which are below the masking

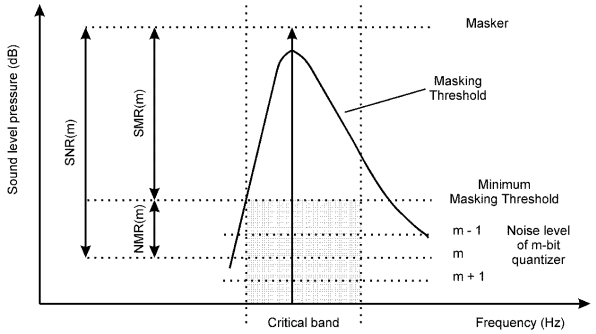

Figure 4: Frequency domain masking showing the masking threshold

threshold will not be perceived by the human auditory system and therefore quantization noise in every subband can be as high as the masking threshold permits. In a subband coding system, every subband has an individual quantizer. It may be an advantage to have a subband decomposition equal to the critical bands of the human auditory system in order to profit of in-band masking. But again a flexible frequency tiling will enable to take care of inter-band masking (e.g. masking across critical bands). Masking also occurs in the time domain. In the presence of abrupt signal transients, a listener will not perceive signals beneath the audibility threshold in the pre- and post-masking regions. Only a few available percep-

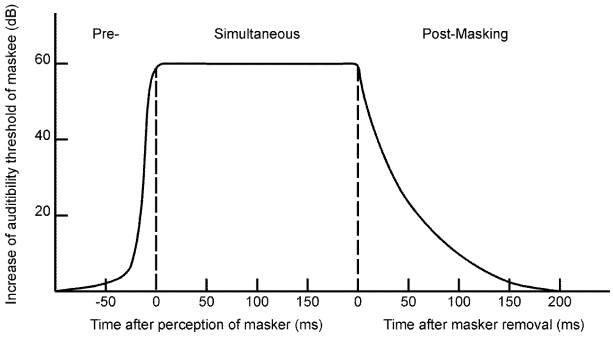

Figure 5: Temporal masking

tual coders take advantage of temporal masking. A flexible, signal adaptive filterbank allows to analyze the temporal structure of the signal in every individual subband and to adapt the filterbank by taking profit of all masking effects in the frequency domain and in the time domain. With the current filterbank framework, pre-echoes can be avoided due to the individual segmentation in the time-domain for every subband. Similarly to the rate-distortion measure, we can define a perceptual measure for searching the best basis in a perceptual sense. A useful metric for estimating the achievable perceptual coding gain is based on perceptual entropy [14]. Perceptual entropy is therefore an estimate to the lower bound of transparent coding although it does not take into account rate-distortion criteria and temporal masking effects.

4. WEIGHTED COST FUNCTION

As it has been pointed out in the introduction, audio signals can have completely different temporal and spectral structure. Combining the rate-distortion measure and the perceptual measure in a weighted cost-function enables to cover applications such as lossless audio coding for archiving and audio-on-demand applications on the Internet with the very same coding scheme. Depending on the weight of the individual measures, the filterbank will adapt either in a rate-distortion sense or alternatively in a perceptual sense. Care has to be taken because these measures are not additive in terms of overall costs. The rate-distortion measure will operate in every subband but for the perceptual measure, a more global analysis in terms of frequency domain masking and temporal masking is used. An additional input to the cost-function is based on the complexity of the algorithm. As pointed out in section 2.2, switching the basis will cause additional costs due to the redundant samples necessary for the reconstruction. If the complexity is to be kept as low as possible, switching the basis may be prohibited if the overall improvement in coding gain is rather small. Additionally, a "grid-function" for the switching can be set in order to avoid multiple up- and down-switching of the basis within a short segment of time.

5. RESULTS

The signal-adaptive wavelet-filterbank including the analysis based on a weighted cost-function has been implemented in MATLAB and C++. Several experiments and tests have been carried out in collaboration with the Swiss Broadcasting Company SRG. Although a simple uniform quantizer rather than the optimal quantizer resulting from the rate-distortion analysis have been used for these first tests, results are very promising. Artifacts such as pre-echoes completely disappeared when comparing with other coding schemes. The need for a signal-adaptive filterbank has been confirmed by a careful analysis of the switching activities of the filterbank. Further research activities will include the implementation of an adaptive quantizer and an entropy coding scheme.

6. CONCLUSIONS

A novel approach to a signal-adaptive filterbank for audio coding applications has been presented in this paper. In contrast to existing audio coding schemes, the algorithm allows individual time segmentation in every subband and every node of the wavelet-packet tree can be switched up- and down in order to increase the coding gain. A weighted cost function allows to optimize the filterbank based on a perceptual or a rate-distortion measure. This system can perform lossless compression, near-lossless compression or perceptual compression of audio signals, depending on the weights which have been selected for the cost function. The cost-function additionally takes other parameters such as computational complexity and overall coding delay into consideration.

7. REFERENCES

[1] Brandenburg K., Stoll G., "The ISO/MPEG-Audio Codec: A Standard for Coding of High Quality Digital Audio", *AES Convention Preprint 3336*, March 1992

[2] Princen J., Johnston J.D., "Audio Coding with signal adaptive filterbanks", *Proceedings of ICASSP 95*, May 1995, pp. 3071-3073.

[3] Shlien S., "The modulated lapped transform, its time varying forms and its applications to audio coding standards", *IEEE Trans. on Speech and Audio Processing*, Vol. 5, No. 4, July 1997, pp. 359-366

[4] Malvar H.S., "Signal Processing with Lapped Transforms", *Artech House*, Norwood, 1992.

[5] Sinha D., Tewfik A., "Low Bit Rate Transparent Audio Compression using Adapted Wavelets", *IEEE Trans. on ASSP*, Vol. 41, No.12, December 1993, pp. 3463-3479.

[6] Kudumakis P, Sandler M., "On the Performance of Wavelets for Low Bit Rate Coding of Audio Signals", *Proceedings of ICASSP 95*, May 1995, pp. 3087-3090.

[7] Srinivasan P., Jamieson L.H., "High-Quality Audio Compression Using an Adaptive Wavelet Packet Decomposition and Psychoacoustic Modeling", *IEEE Trans. on Signal Processing*, Vol. 46, No. 4, April 1998, pp. 1085-1093

[8] Wickerhauser M.V., "Adapted Wavelet Analysis from Theory to Software", *IEEE Press*, 1994

[9] Herley C., Kovacevic J,. Vetterli, M., "Tilings of the time-frequency plane: Construction of arbitrary orthogonal bases and fast tiling algorithms", *IEEE Trans. on Signal Process.*, Vol. 41, December 1993, pp. 3341-3359

[10] Faller C., Erne M., Moschytz G.S., "Wavelet Based Audio Compression", *Semesterarbeit an der ETH-ZŠrich*, February 1998

[11] Coifman R., Wickerhauser M.V. "Entropy-Based Algorithms for Best Basis Selection", *IEEE Trans. on Information Theory*, Vol. 38, No. 2, March 1992 pp. 713-718

[12] Ramchandran K., Vetterli M., "Best Wavelet Packet Bases in a Rate-Distortion Sense", *IEEE Trans. on Image Processing*, Vol. 2, No. 2, April 1993, pp. 160-175

[13] Zwicker E., Fastl H., "Psychoacoustics Facts and Models", *Springer Verlag*, 1990

[14] Johnston J., "Estimation of Perceptual Entropy Using Noise Masking Criteria", *Proceedings of ICASSP 88*, May 1995, pp. 2524-2527