AUDIO COMPRESSION AT LOW BIT RATES USING A SIGNAL ADAPTIVE SWITCHED FILTERBANK

Deepen Sinha *James D. Johnston*

AT&T Bell Labs
Murray Hill, NJ 07974

ABSTRACT

A perceptual audio coder typically consists of a filterbank which breaks the signal into its frequency components. These components are then quantized using a perceptual masking model. Previous efforts have indicated that a high resolution filterbank, e.g., the modified discrete cosine transform (*MDCT*) with 1024 subbands, is able to minimize the bit rate requirements for most of the music samples. The high resolution *MDCT*, however, is not suitable for the encoding of non-stationary segments of music. A long/short resolution or "window" switching scheme has been employed to overcome this problem but it has certain inherent disadvantages which become prominent at low bit rates (< 64 *kbps* for stereo). We propose a novel switched filterbank scheme which switches between a *MDCT* and a *wavelet* filterbank based on signal characteristics. A tree structured *wavelet* filterbank with properly designed filters offers natural advantages for the representation of non-stationary segments such as attacks. Furthermore, it allows for the optimum exploitation of perceptual irrelevancies.

1. INTRODUCTION

Compression of wideband audio signals to very low bit rates is desirable for a number of applications, *e.g.*, transmission and storage of digital audio, multimedia applications, etc. The compressed bit rates of 64 *kbps* for stereo signals and 32 *kbps* for single channel (mono) audio (compression factors of about 22-25) are particularly attractive for some of these applications A number of *perceptual* audio coding algorithms have been proposed in recenty ears [1-5], which claim to provide transparent compression in the range of 128-256 kbps (i.e., compression factors in the range 5-12). The AT&T Perceptual Audio Coder (PAC) [1,2], a well known algorithm for high quality compression of one or more channels of audio, leads to nearly transparent coding at approximately 128 kbps for stereo signals. It also offers promise for higher compression ratios (> 20). PAC is a perceptually driven adaptive subband compression algorithm based on the Modified Discrete Cosine Transformation (*MDCT*) filterbank (also known as a modulated lapped transform or *MLT*). Bit allocation is determined by an elaborate perceptual masking model which hides quantization noise in the signal through the exploitation of auditory masking properties of the ear.

In this paper we present a novel coding scheme based on switched filterbanks that can be used in PAC (or similar subband coding scheme) to improve the quality of compressed audio at low bit rates. This scheme utilizes a signal adaptive switched filterbank for analysis and synthesis. Specifically, it switches between a high spectral resolution *MDCT* and a non-uniform (tree structured) wavelet filterbank (*WFB*) based on the time-varying characteristics of the signal. As explained below, the switched filterbank scheme enhances the quality of attacks (i.e., signal segments containing rapid changes in the signal energy level, e.g., castanets, triangles, drums, etc.). Distortion of attacks is a particularly noticeable artifact at the low bit rates. Therefore improved coding of attacks leads to significant improvement in the subjective quality of a number of audio signals.

The organization of this paper is as follows. In section 2, we present an overview of the proposed coding algorithm. In section 3, we discuss the justification for employing a switched filterbank structure. Following this, in section 4 we address the design issues involved in the development of the switched filterbank scheme. The quality enhancement has been verified with the subjective listening tests summarized in section 5.

2. BASIC ENCODER STRUCTURE

The overview of an encoder based on the switched filterbank idea is illustrated in Figure 1. Briefly, the encoder resembles a generic perceptual subband coding scheme. The signal is broken into its subband components using a filterbank. The subband components are then quantized adaptively using a perceptually generated quantizer step size (threshold). The perceptual model for this encoder is similar to PAC [1] with suitable enhancements for the wavelet filterbank as discussed below. In PAC, the threshold in each critical band is based on the energy level in that and nearby critical bands, a measure of *tonality* (i.e., tone or noise like nature of the signal), and a model for the spread of masking across critical bands. Quantized subband coefficients are further compressed using a noiseless coding scheme and the thresholds are adjusted with the help of a rate loop to satisfy bit rate constraints.

The unique component of the proposed coder is that a signal adaptive *switched* filterbank is employed. The analysis filterbank is "normally" a *MDCT* but in the event of non-stationarity it switches to a *WFB* structure. The switching decision is updated about every 25 *msec* based

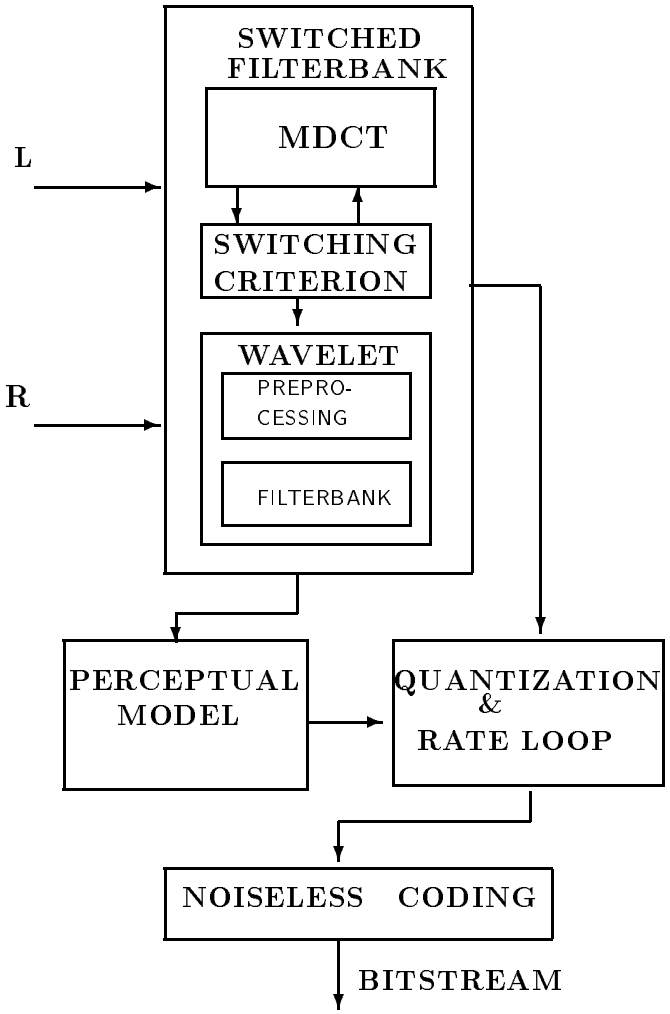

Figure 1. Block diagram of the Switched Filterbank Audio Encoder

on the characteristics of the signal. The exact form of the *WFB* and the switching details will be presented in Section 4. In the following section we discuss the rationale behind the switched filterbank algorithm and its advantages in coding.

3. ADVANTAGES OF THE SWITCHED *MDCT* /*WFB* ANALYSIS FILTERBANK

These may be summarized as follows.

- Both the MDCT and the wavelet filterbank are particularly suitable for encoding a different class of signals. A high resolution *MDCT* (e.g., with 1024 subbands or frequency lines in PAC) leads to a very compact representation for stationary signals (most of music - instrumental as well as vocal - falls into this category). However, signals that contain transients or sharp attacks (e.g., castanets, triangles, etc.) cannot be represented compactly in the *MDCT* filterbank. These signals require a higher time resolution at high frequencies both for compact representation and for optimal exploitation of perceptual irrelevancies. Wavelet filterbanks are quite attractive for the encoding of such signals [5]. Besides the fact that wavelet representation of such signals is more compact than the representation derived from a high resolution *MDCT* (as measured by the *coding gain*), wavelet filters have desirable temporal characteristics. In a wavelet filterbank the high frequency filters (with a suitable moment condition as discussed below) typically have a compact impulse response. This prevents excessive time spreading of quantization errors during synthesis (perceptible as the so-called "pre-echo" problem). To reduce pre-echo *MDCT* based encoders employ conservative masking thresholds whereby the quantizer step sizes in a segment containing attacks are constrained by the steps in the previous segment. This prevents full utilization of the masking potential of the signal and leads to an excessive bit demand during attacks. Wavelet representation of sharp attacks is therefore better suited for the exploitation of perceptual irrelevancies for such signals.

  The signal dependent switched (*MDCT* /*WFB*) filterbank scheme allows us to combine the advantage of these filterbanks for the respective class of signals.

- The switched filterbank algorithm offers important advantages over other previously reported schemes for handling sharp attacks [1,4,6]. In the first scheme used in PAC, a "window switching" algorithm is employed. Here the typical "long" *MDCT* is replaced by a "short" *MDCT* (i.e., a *MDCT* with $1/8^{th}$ frequency resolution) during periods of non-stationarity in the signal. The disadvantage of this approach (i.e., switching *MDCT* resolution) is that the resulting time resolution is uniformly higher for all frequencies. In other words one is forced to increase the time resolution at the low frequencies to increase it to the necessary extent at higher frequencies. An ideal filterbank for sharp attacks is a non-unform structure whose subband match the critical band division of frequency axis (i.e., the subbands are uniform on the bark scale). Moreover it is desirable that the high frequency filters in the bank be proportionately shorter. To achieve this, some coding schemes utilize a hybrid or cascade structure [4,6]. These consist of a (uniform or non-uniform) filterbank as the first stage. Each of the subbands may be further split using uniform filterbanks (to increase frequency resolution for stationary signals). The problem with this approach is that one is forced to use the hybrid structure for stationary signals as well. This has disadvantages in comparison with *MDCT* in terms of resulting frequency response of the filters as well as the implementation cost.

  The switched filterbank algorithm is a good compromise as the desired non-uniform filterbank for attacks is used only when necessary. The *MDCT* filter is used most of the time (i.e., for stationary segments) ensuring implementation and coding efficiencies for such signals.

- The switched filterbank structure adds relatively small additional computational burden to a *MDCT* based

coding scheme. Also fast implementation of the WFB is possible using well known techniques.

4. SWITCHED FILTERBANK DESIGN ISSUES

A number of issues specific to the switched filterbank coding algorithm are addressed below. These pertain to the design of the tree structured *wavelet* filterbank, design of transition filters, and adaptation of the psychoacoustic model to the *wavelet* filterbank

A. Design of an appropriate wavelet filterbank

As alluded to previously, the frequency split provided by the non-uniform filterbank should approximately match the critical band or bark scale. In addition the support of higher frequency filters should be proportionately smaller. We employ a tree structure to meet these two goals. A tree structure has the natural advantage that the effective support (in time) of the subband filters is progressively smaller with increasing center frequency. This is because the critical bands are wider at higher frequency so fewer cascading stages are required in the tree to achive the desired frequency resolution (cascading increases the effective support of the corresponding subband filter). Additionally, proper design of the prototype filters used in the tree decomposition ensures (see below) that the high frequency filters in particular are compactly localized in time.

The decomposition tree is based on sets of prototype filterbanks. These provide two or more bands of split and are chosen to provide enough flexibility to design a tree structure that approximates the critical band partition closely. The three filterbanks were designed by optimizing parametrized paraunitary filterbanks using standard optimization tools and an optimization criterion based on weighted stopband energy [7]. In this design, the *moment* condition plays an important role in achieving desirable temporal characteristics for the high frequency filters. An M band paraunitary filterbank with subband filters $\{H_i\}_{i=1}^{i=M}$ is said to satisfy a P^{th} order moment condition if $H_i(e^{j\omega})$ for $i = 2, 3, ..M$ has a P^{th} order zero at $\omega = 0$. [7]. For a given support for the filters, K, requiring $P > 1$ in the design yields filters for which the "effective" support decreases with increasing P. In other words most of the energy is concentrated in an interval $K' < K$ and K' is smaller for higher P (for a similar stopband error criterion). The improvement in the temporal response of the filters occurs at the cost of an increased transition band in the magnitude response. However, requiring at least a few vanishing moments yields filters with attractive characteristics.

The impulse response of a high frequency *wavelet* filter (in a 4 band split) is illustrated in Figure 2. For comparison the impulse response of a filter from a modulated filterbank with similar frequency characteristics is also shown. It is obvious that the *wavelet* filter offers superior localization in time.

B. Switching Mechanism and a Switching Criterion

The MDCT is a lapped orthogonal transform. Therefore, switching to a wavelet filterbank requires orthogonalization

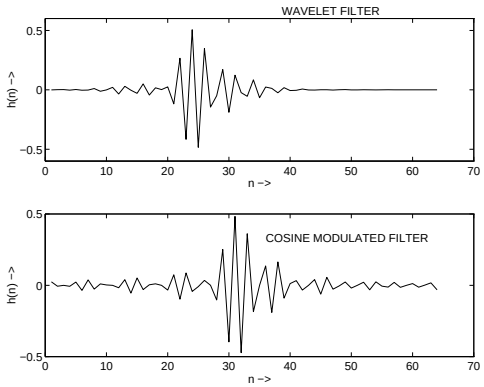

Figure 2. High Frequency wavelet and cosine-modulated Filters

in the overlap region. While it is straightforward to setup a general orthogonalization problem, the resulting transform matrix is inefficient computationally. The orthogonalization algorithm can be simplified by noting that a MDCT operation over a block of $2 * N$ samples is equivalent to a symmetry operation on the windowed data (i.e., outer $N/2$ samples from either end of the window are folded into the inner $N/2$ samples) followed by an N point orthogonal block transform Q over these N samples. Perfect reconstruction is ensured irrespective of the choice of a particular block orthogonal transform Q. Therefore, Q may be chosen to be a DCT for one block and a wavelet transform matrix for the subsequent or any other block. The problem with this approach is that the symmetry operation extends the wavelet filter (or its translates) in time and also introduces discontinuities in these filters. Thus it impairs the temporal as well as frequency characteristics of the wavelet filters. In the present encoder this impairment is mitigated by the following two steps: (i) start and stop windows are employed to switch between $MDCT$ and WFB (this is similar to the window switching scheme in PAC), (ii) the effective overlap between the transition and wavelet windows is reduced by the application of a new family of smooth windows [8]. The resulting switching sequence is illustrated in Figure 3.

The next design issue in the switched filterbank scheme is the design of a $N \times N$ orthogonal matrix Q^{WFB} based on the prototype filters and the chosen tree structure. To avoid circular convolutions we employ transition filters at the edge of the blocks. Given a subbband filter, c_k, of length K a total of $K_1 = (K/M) - 1$ transition filters are needed at the two ends of the block. The number at a particular end is determined by the rank of a $K \times (K_1 + 1)$ matrix formed by the translations of c_k. The transition filters are designed through optimization in a subspace constrained by the pre-determined rows of Q^{WFB}.

Finally, to make the switching mechanism work effectively it is important to use a reliable criterion for switching between the $MDCT$ and WFB filterbanks. The criterion should detect the attacks accurately yet not mis-identify any. An undetected attack, if encoded in the $MDCT$ filterbank will result in annoying distortions, especially at lower bit rates. On the other hand, coding a stationary signal with the wavelet filterbank results in significant wastage of

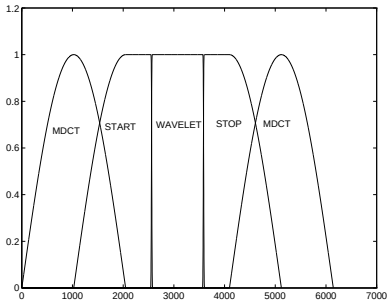

Figure 3. A Filterbank Switching Sequence

bits and processing power. We have experimented with a *perceptual entropy (PE)* [1] and an energy-based criterion. The *PE* based criterion is more reliable but computationally more costly. The decision to switch filterbank is made about once every 25 $msec$.

C. Computation of Perceptual Threshold for Wavelet Coefficients

The thresholds for the quantization of wavelet coefficients are based on an estimate of time-varying *spread* energy in each of the subbands and a tonality measure which is estimated as in PAC. The spread energy is computed by considering the spread of masking across frequency as well as time. In other words, an inter-frequency as well as a temporal spreading function is employed. The shape of these spreading functions may be derived from the cochlear filters [10]. The temporal spread of masking is frequency dependent and is roughly determined by the (inverse of) bandwidth of the cochlear filter at that frequency. This can last from 10s of $msec$ at low frequencies to less than a $msec$ at higher frequency. Post-masking seems to last longer at all frequencies. We use a fixed temporal spreading function for a range of frequencies (subbands). Naturally, the shape of spreading function becomes narrower with higher frequencies. The coefficients in a subband are grouped in a *coder* or *scalefactor* band and one threshold value per coderband is used in quantization. The coderband span ranges from 10 $msec$ in the lowest frequency subband to about 2.5 $msec$ in the highest frequency subband. The thresholds are quantized on a log scale and are further compressed using a Huffman code.

5. SUBJECTIVE QUALITY EVALUATION

The performance of the switched filterbank scheme on coder quality was assessed with listening tests involving 12 subjects (including 5 "experts"). For each encoded sample, the subject was presented with a stimulus of the format A-B-C-A, where A is the original music, one of B and C is a hidden reference and the other encoded music. The subject was asked to rate B and C on a 5.0 point Mean Opinion Score (MOS) scale (5 being perceptually indistinguishable and 1 being very annoying). They were further asked to award a 5.0 to the one they believe to be the hidden reference. At the bitrate of 64 $kbps$ stereo, the MOS scores for signals with attacks (like castanets, triangles, breaking glasses, etc.) were 0.4-0.6 higher when encoded with the $MDCT/WFB$ switched filterbank scheme. Here the reference encoder was the standard PAC algorithm (which employs the standard long/short window switching).

6. CONCLUSIONS

In this paper we presented a novel signal adaptive switched filterbank scheme for the compression of audio signals. By switching between a high resolution $MDCT$ filterbank and a higher time resolution wavelet filterbank, the technique achieves high coding gain as well as optimum exploitation of perceptual irrelevancies. The switched filterbank structure offers advantages over the conventional window switching and hybrid schemes in terms of filterbank characteristics and in some cases computational complexity. The improvement in coding quality was verified with the help of subjective listening tests.

REFERENCES

[1] J. D. Johnston, D. Sinha, S. Dorward, and S.R. Quackenbush. "The AT&T Perceptual Audio Coder (PAC)," Presented at the AES convention, New York, Oct. 1995.

[2] J.D. Johnston and A. J. Ferreira, "Sum-difference stereo transform coding," *In proceedings of IEEE, ICASSP,* , pp. II:569-572, April, 1992.

[3] R. N. J. Veldhuis. "Subband Coding of Digital Audio Signals Without Loss of Quality," *In proceedings of IEEE, ICASSP,* Glasgow, Scotland, pp. 2009-2012, May 1989.

[4] K. Brandenburg, G. Stoll et. al. "The ISO-MPEG-Audio Codec: A Generic-Standard for Coding of High Quality Digital Audio," *Journal of the Audio Engineering Society.* Vol. 42, No. 10, October 1994

[5] D. Sinha and A. H. Tewfik. "Low Bit Rate Transparent Audio Compression using Adapted Wavelets," *IEEE Transactions on Signal Processing,* vol. 41, no. 12, pp. 3463-3479, Dec. 1993.

[6] J. Princen and J.D. Johnston "Audio Coding with Signal Adaptive Filterbanks," *In proceedings of IEEE, ICASSP,* Detroit, 1995.

[7] P. P. Vaidyanathan. "Multirate Digital Filters, Filter Banks, Polyphase Networks, and Applications: A Tutorial," *Proceedings of the IEEE,* vol. 78, no. 1, pp. 56-92, January 1990

[8] D. Sinha "A New Family of Smooth Windows," In Preparation.

[9] J. D. Johnston. "A Method of Estimating the Perceptual Entropy of an Audio Signal," *In proceedings of IEEE, ICASSP,* 1988.

[10] J. B. Allen, editor, "The ASA edition of Speech Hearing in Communication." Acoustical Society of America, Woodbury, New York, 1995.

AUDIO COMPRESSION AT LOW BIT RATES USING A SIGNAL ADAPTIVE SWITCHED FILTERBANK

Deepen Sinha and James D. Johnston
AT&T Bell Labs
Murray Hill, NJ 07974

A perceptual audio coder typically consists of a filterbank which breaks the signal into its frequency components. These components are then quantized using a perceptual masking model. Previous efforts have indicated that a high resolution filterbank, e.g., the modified discrete cosine transform ($MDCT$) with 1024 subbands, is able to minimize the bit rate requirements for most of the music samples. The high resolution $MDCT$, however, is not suitable for the encoding of non-stationary segments of music. A long/short resolution or "window" switching scheme has been employed to overcome this problem but it has certain inherent disadvantages which become prominent at lower bit rates (< 64 *kbps* for stereo). We propose a novel switched filterbank scheme which switches between a $MDCT$ and a *wavelet* filterbank based on signal characteristics. A tree structured *wavelet* filterbank with properly designed filters offers natural advantages for the representation of non-stationary segments such as attacks. Furthermore, it allows for the optimum exploitation of perceptual irrelevancies.