AUDIO CODING WITH SIGNAL ADAPTIVE FILTERBANKS

J. Princen and J. D. Johnston

AT&T Bell Laboratories,
600 Mountain Ave.,
Murray Hill, NJ, 07974.

ABSTRACT

In this paper we present a high quality audio coding system based on a novel nonuniform modulated filterbank coupled with time-varying cosine modulated filterbanks in a cascade architecture. The system makes use of psychoacoustic thresholds in a natural way to adapt the resolution of the filterbank to achieve high coding gain on a wide range of signal types. Results show that the system provides excellent quality at 64 kb/s and good quality at 48 kb/s for monophonic coding.

1. Introduction

In the last few years the development of useful high quality audio compression schemes has enabled new consumer electronics devices [1, 2] and applications such as Digital Audio Broadcasting [3] and Digital Surround Sound for movies [4]. Though the performance of these schemes is impressive - eg for the MPEG Layer II algorithm [5] it is generally accepted that excellent quality can be achieved at around 256 kb/s for a stereo pair, while for more sophisticated algorithms such as MPEG Layer III and AT&T's PAC [6], excellent quality can be achieved at 128-160 kb/s for stereo - there is still a desire to drive these bitrate requirements down even further, with the eventual goal of achieving excellent quality on a wide range of material at << 128 kb/s for a stereo pair.

Most current high quality audio compression algorithms are based around the system paradigm shown in figure 1. Compression is achieved by the combination of a filterbank, PCM quantization and possibly lossless (eg Huffman) coding. The coders make use of sophisticated psychoacoustic models which determine the distribution of bits and hence noise in time and frequency, in an attempt to minimize perceptual distortion. The main aim of this paper is to discuss this audio coding system paradigm and in particular the filterbank element. We will argue that the resolution of this filterbank must vary with time if the system is to achieve optimum performance on a wide variety of signal types. A particular coding architecture which makes use of psychoacoustic thresholds in a natural way to drive a signal adaptive filterbank is presented.

2. Optimum Time-Frequency Decompositions

All currently published psychoacoustic models make use of the short-term signal energy distribution as a function of frequency. From this energy distribution and the tone-like or noise-like characteristics of the signal as a function of frequency a set of thresholds, representing the just noticeable noise levels are calculated [7, 5]. Assuming that the introduced quantization noise is below these levels, signal coding will be transparent. In practice the models are not perfect, nevertheless they offer a significant improvement in performance over subband coders which assign quantization noise (or bits) based on minimum sum squared error.

The time-frequency (TF) resolution of the psychoacoustic analysis should match the TF resolution of the auditory system, which is known to be largely determined by the behavior of the basilar membrane. These resolution characteristics are reflected in the critical band scale which indicates that the frequency resolution in the psychoacoustic model should vary from 100 Hz at low frequencies to around 4 Khz at high frequencies (ie a 40:1 change in resolution) [8]. This implicitly suggests that the temporal resolution should increase by a factor of about 40:1 from low to high frequencies. Most psychoacoustic models use a very low uniform temporal resolution (eg the models suggested in the MPEG standard [5]). A lack of temporal resolution at high frequencies has little effect on the thresholds calculated for stationary signals, however, the thresholds calculated for dynamic signals will be inaccurate, and this could lead to audible distortions. This behavior can be corrected by using a nonuniform filterbank to calculate energy distributions.

Figure 2 shows the thresholds calculated from two distinctly different signals: An attack from a castanet and the relatively stationary signal from a piccolo. To get a feel for what happens when different filterbank resolutions are used for coding we make the simplifying assumption that the filterbank TF resolution can be represented by a tiling of this plane by a set of rectangles. The size of a tile roughly represents the spread of the associated basis function in time and frequency. To ensure transparent coding the minimum threshold within any tile should be used. Hence, if the threshold changes dramatically over a tile we are paying a very high price for transparency. For example, if we were to use a filterbank with low temporal resolution on a signal like castanets a huge bitrate demand would be generated around the attack, because of the need to make sure the threshold requirements before and after the attack are met. In contrast filterbanks with low temporal resolution can be used to code piccolo without any penalty because the thresholds are relatively stationary.

These threshold maps may suggest that a critical band-like filterbank is optimum for coding, since such a filterbank will naturally meet the psychoacoustic threshold constraints without introducing high bitrate demand. This is not so.

J. Princen is now with Silicon Graphics Inc., Mountain View, CA.

0-7803-2431-5/95 $4.00

In fact it is clear from both the threshold plots that there are potentially better filterbanks. For piccolo, since thresholds are relatively constant, it is possible to use a filterbank with a very high frequency resolution (and hence low temporal resolution) without paying any penalty to meet the psychoacoustic thresholds. Such a filterbank will always be better than a critical band system since it will have much higher coding gain. Analogously for the castanets signal, the threshold is relatively constant across a band of upper frequencies and it is possible to use a higher temporal resolution filterbank (which has low frequency resolution) without paying any penalty, and this filterbank will always have higher coding gain than a critical band system for this signal. These arguments serve to illustrate that *there is no optimum filterbank resolution in this system.* For any particular signal the optimum decomposition depends primarily on the TF energy distribution and on the signal characteristics as determined by the psychoacoustic analysis.

3. A cascade approach

It is difficult to design an effective time varying filterbank which can have arbitrary resolution, so we must take the practical step of designing one with a finite set resolutions which cover the range of signal possibilities. A further practical requirement is that we should have some method of mapping the psychoacoustic thresholds onto quantization step-sizes for the coder filterbank. One particular approach which is very attractive in this respect is the structure shown in figure 3. The design consists of a nonuniform filterbank cascaded with time-varying cosine modulate uniform filterbanks [9].

3.1. Nonuniform filterbank

There are very few techniques available for the design of critically sampled nonuniform filterbanks. The most popular method is to make use of uniform designs in a cascade structure. Cascade structures in general produce designs which have a poor channel isolation for a given impulse response length. Direct form near perfect reconstruction modulated nonuniform designs based on the adjacent channel aliasing cancellation principles first suggested by [10] and [11] are possible [12]. The essential idea is to form a nonuniform filterbank using uniform sections "joined" by transition filters. An example of such a filterbank, which is used in our current coder implementation, is shown in figure 4. This filterbank matches the critical bands approximately at low frequency but has higher resolution at high frequencies. The filterbank is divided into 4 sections (8 bands in 0-750 Hz, 4 bands in 750-1500 Hz, 12 bands in 1500-6000 Hz, 24 bands in 6000-24000 Hz) with a total of 48 bands. All of the filters within a section are modulated from a single real lowpass prototype, and the transition filters are modulated from a complex lowpass prototype. If there are a small number of sections there is a significant computational advantage over an unstructured design. The structure also simplifies the design process [12].

3.2. Time-varying cosine modulated filterbank

For each uniform section a different resolution switchable uniform filterbank is used at the filter outputs. The end result is that in its highest frequency resolution mode the cascaded filterbank is approximately uniform with 512 bands, while in its highest temporal resolution mode it is close to a critical band system. Time-varying cosine modulated filterbanks were first used in [13] for the case where the basis function had a length limited to twice the number of channels (ie $2M$). They have subsequently been used in many audio coding systems [5, 6, 4]. The extension of this idea to longer basis functions has recently been proposed in [9]. The advantage of longer basis functions is that better stopband rejection can be obtained, yielding higher coding gain for stationary signals.

In all designs switching from one resolution to another requires a set of transition windows (windows are the lowpass prototypes used to generate the basis functions) which are specifically designed to maintain orthogonality and give perfect reconstruction. In the case of length $2M$ basis functions a single transition window is required, while for longer basis functions the number of transition windows is greater [9], ie the transition takes place over several time-steps. The design of transition windows is simplified by the lattice representation of cosine modulated filterbanks [14].

In the present system the time-varying filterbanks are switched between a no decomposition state and a decomposition state using the adaptation strategy outlined in the next section. A typical set of transition windows which achieve this switch for length equal to $4M$ with $M = 8$ are shown in figure 5. Note that the switch from decomposition to no decomposition can be achieved by time-reversing this sequence of windows.

4. Adaptation strategy

In the example of figure 3 and 4, four uniform sections make up the nonuniform filterbank and, although it would be possible to allow each of the cascaded filters to vary with time independently, we generally vary each of the filters within a section together. This cascade structure allows a very natural adaptation strategy. The nonuniform filterbank is used as a part of the coding system and also to determine the psychoacoustic thresholds (though in calculating the thresholds it is oversampled, while for coding only critical sampling is used). The psychoacoustic thresholds are used to drive the resolution selection process. If a particular section is in high resolution mode the cascaded filterbank basis functions cover a region of thresholds calculated using the nonuniform filterbank. The minimum threshold over the region is used to ensure transparent coding. Once the thresholds are known it is possible to calculate the Perceptual Entropy (PE) [7, 15], for each filterbank resolution and the resolution which minimizes PE is chosen. The PE is well correlated with bitrate and therefore this adaptation strategy minimizes the bitrate required for transparent coding.

In the current system the coding decisions are made on a frame basis. A frame is a sequence of filterbank outputs which is long enough to ensure that a transition from one state to another can be accomplished (1024 samples in the current implementation). Hence frames can be thought of as being either: low resolution, high resolution, low− >high transition, or high− >low transition. Since a decision on the type of the current frame determines the possibilities for future frames look-ahead should be used. In our current system we use one frame look-ahead. The sequence of frame types is also restricted so that it is not possible to have two transitions following each other directly. The frame types which are possible depend on the decomposition used in the previous frame. As an example, if the previous frame was low resolution possibilities for the next two frames are: low, low− >high; low− >high, high; low, low. In each case the perceptual entropy [15] is calculated for the two frames,

and the choice which minimizes PE is used to determine the decomposition of the current frame.

The foregoing discussion has implied that a single frame type decision is made, however, the decomposition can be chosen independently for each section.

5. Quantization and coding

For each frame we have a set of thresholds and a set of quantized samples which must be coded. Both these are represented on a time-varying, possibly nonuniform time-frequency sampling pattern.

5.1. Threshold coding

The number of thresholds depends on the filterbank state. In high resolution modes approximately one threshold per critical band per frame is used. For high temporal resolution modes a larger number of thresholds are required to represent dynamic temporal activity, particularly at high frequencies. Each threshold is differentially coded and compressed using lossless Huffman coding.

5.2. Sample coding

Filterbank samples are quantized using linear PCM. The 2D map of quantized samples is divided into regions, and these regions are ordered. The ordering is done independently for each filterbank section in a way that depends on the filterbank state. A small number of different Huffman codebooks have been designed which are distinguished by the maximum absolute value (MAV) of the samples which they can code. These codebooks are identical to those used on the PAC algorithm [6]. Each region is assigned a Huffman codebook according to MAV. Side information is required to signal the codebook for each region and the amount of side information can be traded against the efficiency of sample coding by adaptively merging regions.

6. Results

In our preliminary experiments with the above coding architecture we have used the 48 band nonuniform filterbank shown in figure 4 coupled with time-varying systems which lead to a highest resolution of 512 uniform bands over the complete frequency range. The system has been tested on the complete set of signals from the ISO test set. The codec is monophonic, so only the right channel from each signal was used in the tests. The quality at fixed bit rates of 48 and 64 kb/s was evaluated informally by headphone listening. Results show that the system provides excellent quality on most signals at 64 kb/s, though transparency is not achieved on all signals. At the lower rate of 48 kb/s audible distortions are apparent on almost all signals, though the quality remains very good. Typical distortions present are: granularity for noise likes signals (eg Cabasa and fricatives in speech); Some post echoes due to imperfect switching decisions and tonal distortion which is audible on one signal (Cello). An informal comparison with a ISO/IEC MPEG Layer III codec (which also provides excellent quality at 64 kb/s and good quality at 48 kb/s) has shown that the proposed coder provides better quality on almost all signals.

It is likely that with further work optimizing the choice of filterbanks resolutions, lossless coding and the coding of side information, that the results at 48 kb/s can be improved significantly. It should also be possible to simplify the nonuniform filterbank (possibly to 3 rather than 4 sections) without compromising quality.

7. References

[1] K. Tsutsui, H. Suziki, O. Shimoyoshi, M. Sonohara, K. Akagiri, and R. Heddle, "ATRAC: adaptive transform acoustic coding for minidisc," in *Proc. Audio Engineering Society Convention*, (San Francisco), 1992.

[2] G. C. Wirtz, "Digital compact cassette: Audio coding technique," in *Proc. Audio Engineering Society Convention*, (New York), 1991.

[3] N. S. Jayant, "The AT&T DAB system," in *Proc. 2nd Int. Symp. on Digital Audio Broadcasting*, (Toronto), March 1994.

[4] C. C. Todd, "AC-3: Flexible perceptual coding for audio transmission and storage", high quality digital audio," in *Proc. Audio Engineering Society Convention*, 1992.

[5] K. Brandenburg and G. Stoll, "The ISO/MPEG audio codec: A generic standard for coding of high quality digital audio," in *Proc. Audio Engineering Society Convention*, 1992.

[6] J. D. Johnston and A. J. Ferreira, "Sum-difference stereo transform coding," in *Proc. IEEE Int. Conf. ASSP*, pp. II–569–II–572, April 1992.

[7] J. D. Johnston, "Estimation of perceptual entropy using noise masking criteria," in *Proc. IEEE Int. Conf. ASSP*, pp. 2524–2527, April 1988.

[8] B. Scharf, "Critical bands," in *Foundations of Modern Auditory Theory* (J. V. Tobias. ed.), (New York), Academic, 1970.

[9] R. L. De Queiroz and K. R. Rao, "Time-varying lapped transforms and wavelet packets," *IEEE Trans. on Signal Proc.*, vol. 41, pp. 3293–3305, December 1993.

[10] H. J. Nussbaumer, "Pseudo QMF filter bank," *IBM Technical Disclosure Bulletin*, vol. 24, pp. 3081–3087, Nov. 1981.

[11] J. H. Rothweiler, "Polyphase quadrature filters - a new subband coding technique," in *Proc. IEEE Int. Conf. ASSP*, (Boston, MA), pp. 1280–1283, April 1983.

[12] J. Princen, "The design of non-uniform modulated filterbanks," in *IEEE Int. Symp. on Time-Frequency and Time-Scale Analysis*, (Philadelphia, PA), Oct. 1994.

[13] B. Edler, "Codierung von audiosignalen mit uberlappender transformation und adaptiven fensterfunktionen," *Frequenz*, September 1990.

[14] H. S. Malvar, *Signal Processing with Lapped Transforms*. ARTECH House, 1992.

[15] J. D. Johnston, "Transform coding of audio signals using perceptual noise criteria," *IEEE Journal on Selected Areas in Commun.*, vol. 6, pp. 314–323, February 1988.

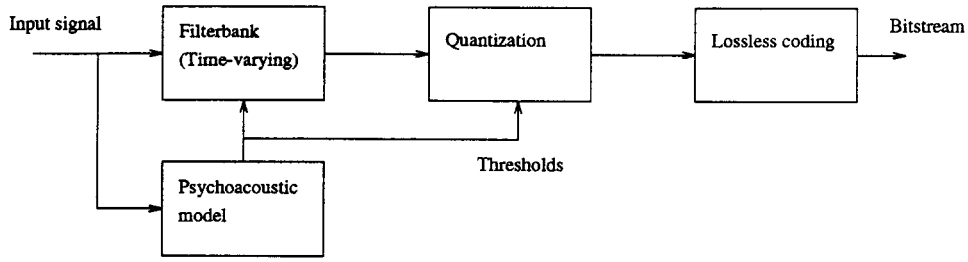

Figure 1: Audio coding system paradigm.

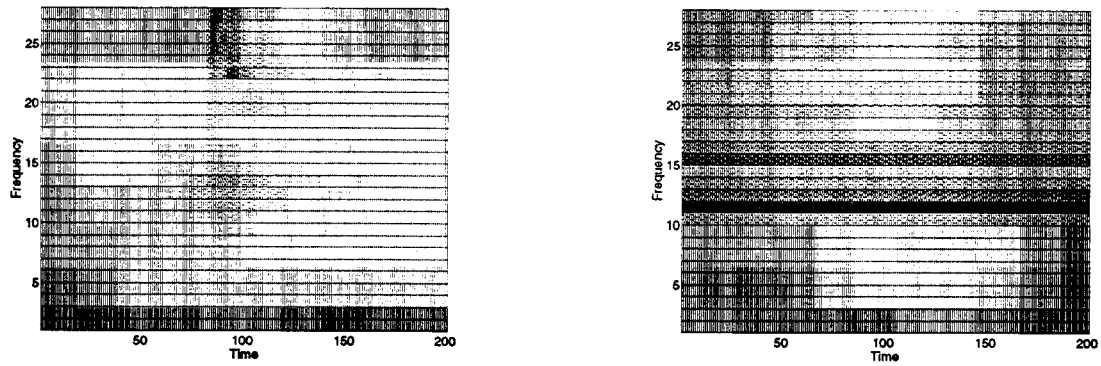

Figure 2: Thresholds for a) Castanets. b) Piccolo. Higher thresholds are darker. Temporal resolution is uniform with each sample representing 1/6 mSec. Frequency range is from 0 to 24Khz approximately on a Bark scale.

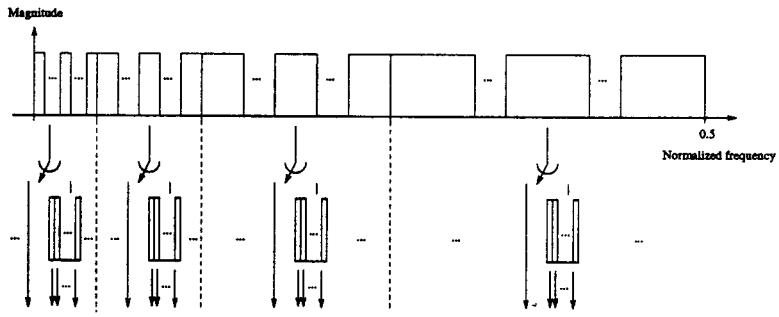

Figure 3: Switched filterbank.

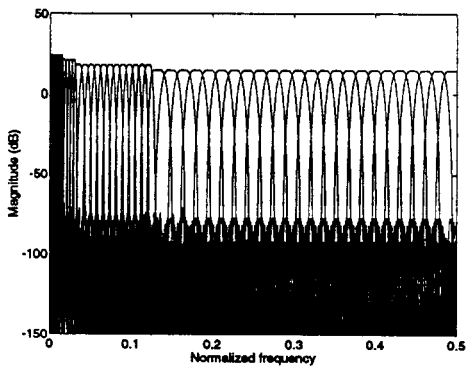

Figure 4. 48 band nonuniform filterbank

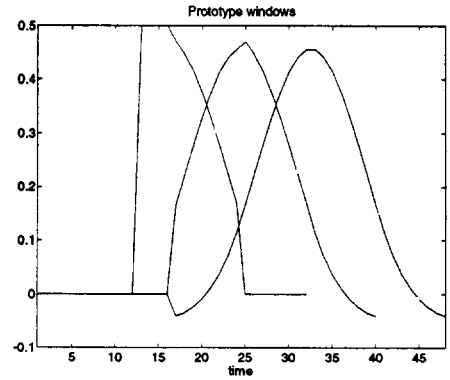

Figure 5. Time-varying windows