Non-uniform subband analysis-synthesis banks for the compression of audio signals

F.Argenti, V.Cappellini, E.Del Re, A.Fiorilli
Dipartimento di Ingegneria Elettronica, University of Florence
Via di Santa Marta, 3 - 50139 Florence
Italy

Abstract - Subband coding is widely used in the field of signal compression. A 32 uniform subbands bank is employed in the MPEG audio standard. In this work a mapping of the audio spectrum onto non-uniform subbands is analyzed, with the aim of increasing the perceptual quality of the reconstructed signal.

1 Introduction

Subband decompositions permit to represent a discrete signal by means of a set of subsequences, each of them related to a given interval of the input spectrum. The sensitivity of the human auditory system can be taken into account and the subbands that result to be less important from a subjective quality point of view can be more coarsely quantized (see [1][2] and the bibliography therein).

Uniform banks have been employed in the MPEG audio standard [3]: the input signal is divided into 32 subbands having the same width in the frequency domain. An optimal solution to the analysis/synthesis banks design problem should consider a model of the auditory system. In the models usually considered, the audio spectrum is divided into *critical bands*, being the human sensitivity approximately constant within a critical band. Critical bands have non-uniform width and, therefore, non-uniform banks are expected to exploit the psychoacoustical model better than uniform banks.

The problem of non-uniform banks has been recently considered in the literature [4]-[7]. The conditions that allow aliasing cancellation and perfect reconstruction are more complex when compared to the uniform case.

This work will analyze a suitable splitting of the input spectrum based on the psychoacoustic model used in MPEG Layer II; possible filter bank implementations of the identified splitting will also be discussed.

The MPEG audio standard is now briefly reviewed.

2 The MPEG audio standard

A block diagram of the MPEG audio standard is shown in Fig. 1.

The coder works on signals sampled at 32, 44.1 or 48 kHz, 16 bit/sample, and provides a compressed output at a bit rate varying from 192 to 32 kbit/s. Three layers of coding, having different efficiency and complexity are defined: in the following we will refer to the Layer II.

The filter bank used in MPEG provides a splitting of the audio spectrum into 32 uniform subbands: at a sampling frequency of 48 kHz, each subband refers to a 750 Hz wide frequency interval. Sets of 12 subband coefficients are generated from 384 input samples: for each set, a *scalefactor* is chosen from a table of predefined values.

Once the bit rate has been fixed, the available bits are distributed among the subbands. The bit allocation is based on a psychoacoustic model and is adapted to the input signal: in fact, it is computed every 1152 input samples (a *frame* of input data), i.e. it is maintained constant every 36 subband samples. Since the psychoacoustic model is fundamental in the derivation of the non-uniform

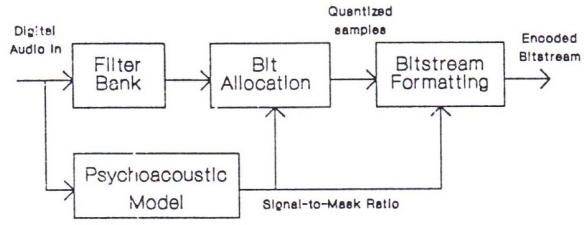

Figure 1: Block diagram of the MPEG coder

subband splitting, it is briefly described in the following

2.1 MPEG psychoacoustic model

The aim of the psychoacoustic model is the computing, for each frequency of the audio spectrum, of the *auditory threshold*: only spectral components with an energy higher than the threshold value are actually perceived by the human auditory system.

The psychoacoustic model is defined in the frequency domain. Experimental studies have shown that two effects must be taken into account:

- an *absolute threshold of hearing* defines the minimum energy that makes a sinusoidal component perceivable in the absence of other sounds;
- a high energy spectral component can mask adjacent components, i.e. it raises the auditory threshold in its proximity.

The steps that lead to the computation of the auditory threshold are now outlined in order to introduce the work that has been undertaken in this paper. The procedure is run every 576 input samples.

A.1 The FFT on 1024 Hanning windowed points, containing the 576 input samples in the central section, is computed.

A.2 For each component, the *energy* and an *unpredictability measure* is computed. Based on the latter, the spectral components can be classified, at the two extremes, as *tone-like* and *noise-like* components.

A.3 Energy and unpredictability are transposed in the domain of the *calculation partitions*. The calculation partitions are a refined version of the *critical bands*, that have been introduced in the study of the human auditory system as spectral intervals where masking effects can be considered as constant. The main incentive to study non-uniform banks relies on the non uniform width of the calculation partitions.

A.4 The effect of *masking* of high energy partitions is considered. Energy and unpredictability measure of each calculation partition are updated by applying a *spreading function* to the values of the adjacent ones.

A.5 The unpredictability measure of each calculation partition is converted into a *tonality index*. From the tonality index a Signal-to-Noise Ratio (SNR) is derived. Since the energy of each calculation partition is known, the *masking noise*, that is the noise tolerated without being perceived, is computed from the SNR.

A.6 The masking noise is converted again into the domain of the FFT components. These values are compared, for each frequency, with the absolute threshold of hearing and the maximum yields the masking threshold.

The masking thresholds are then translated into the domain of the subband intervals, (the *coder partitions*). Energy and masking thresholds are computed for each subband: from these two variables a *Signal-to-Mask Ratio* SMR_n, referred to the n-th subband, is obtained.

In the following a different subband splitting will be analyzed: the energy and masking thresholds will be mapped onto a new subband configuration and different values of the SMR_n will be obtained.

In the bit allocation procedure the number of bits available to code a frame of subband samples is first computed. For example, at the sampling frequency of 48 kHz, 1152 input samples are coded every 24 ms: if a bit rate of 192 kbit/s is aimed at, then the amount of available bits is 4608. A part of this quantity is assigned to code side information (the header, the bit allocation actually used for that frame, etc.); the remainder is used to code the scalefactors and the subband samples.

Let b_n be the number of bits assigned to the n-th subband: a Signal-to-Noise Ratio (SNR_n), due to the quantization at b_n bits, is tabulated. The quantity

$$MNR_n = SNR_n - SMR_n \qquad (1)$$

defines the Mask-to-Noise Ratio, that is a measure of how much the masking noise is superior to the quantization noise. The higher this value, the better the performance of the coder.

Eq. (1) is used to distribute the bit rate among the subbands. First, an allocation of zero bit per subband is supposed; from eq. (1) the band having minimum MNR is selected, its bit rate is posed equal to one and the number of bits necessary to code scalefactors and samples are subtracted from the amount of available bits. Then, the procedure is repeated: at each iteration, the minumum MNR subband is searched for, its bit rate incremented and the available bits reduced. The procedure stops when the available bits do not allow any further increment of the bit allocated to any subband.

A possible non-uniform subband splitting is now investigated.

3 Non-uniform subband splitting of the audio spectrum

The observation that the critical bands and the calculation partitions do not have a constant width raises the question if a better coder performance can be obtained by allowing a non-uniform splitting of the audio spectrum. For example, the masking threshold assumes at the low frequencies a variable behaviour: this information is lost when the masking threshold values are mappend onto an unique value and assigned to a subband.

Contrasting requirements need to be satisfied: on one hand, a higher number of subbands permits a full exploitation of the variability of the masking threshold; on the other hand, the complexity of the system increases as well as the load due to the scalefactors.

In this work we have considered subband structures similar to those resulting from a *wavelet packets* analysis. In [8] a tree structure based on iterative two-channel splitting is adapted to an input signal. The initial configuration consists of the input signal and two channel splitting is applied. A *cost function* (a measure of the *entropy* of the subbands) is defined: splitting is considered advantageous if the cost function value associated to the new profile is decreased. Then the procedure is iterated and stops when no further subband splitting decreases the cost function. So, a binary tree defines the minimum entropy wavelet packets representation of the signal.

In our case, the choice of a cost function must involve the psychoacoustic model and, more in particular, the masking thresholds computed for each FFT spectral component. The cost function that is to be defined must ascertain if band splitting yields an increment of the performance of the coder in terms of perceptual quality.

A value of MNR can be assigned to each subband. After splitting, two new MNRs can be associated to the split subbands. Since it is not clear how parent and split subbands MNRs can be compared between each other, another approach has been considered.

The bit allocation algorithm suggests that a nearly constant MNR is aimed at. Suppose a contant MNR is given: eq. (1) can be inverted as follows:

$$SNR_n = MNR + SMR_n \qquad (2)$$

Suppose a wavelet packets representation has been chosen. From the masking threshold values related to the FFT component, the SMR_n of each subband can be computed, while eq. (2) yields the values of the SNR_n due to the quantization: from

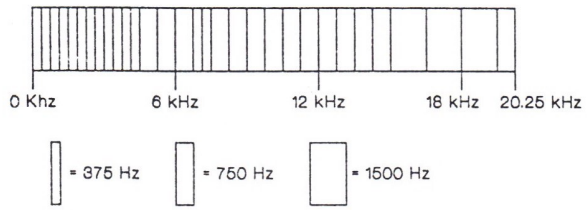

Figure 2: Optimal subband splitting obtained with MNR=22.5

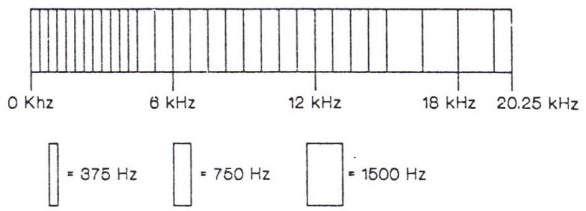

Figure 3: Subband splitting

SNR_n the number of bits that must be allocated to the n-th subband are derived. So, the new cost function used to evaluate a subband representation is the total number of bits (including scalefactors and the table of the bit allocation) used to code the input signal with a target MNR.

In [9] an experimental study using 20 audio signals originated by different sources and sampled at 48 kHz was conducted. Optimum splitting, i.e. yielding the minimum amount of code bits, was computed for each frame contained in the input data (each file contained about one hundred of frames). The subband profiles with higher occurrences were selected. No great differences were found in the tests obtained for different values of MNR, ranging from 20 to 26 dB.

In Fig. (2) the optimal profile obtained with MNR=22.5 is shown.

For a comparison, consider that the MPEG configuration consists of 32 uniform subbands having, at the sampling frequency of 48 kHz, a width of 750 Hz. The five subbands at higher frequency are not coded: so, only the frequency up to 20250 Hz are considered. The same limit has been imposed to the non-uniform representations.

For simplicity's sake, a slightly different structure, shown in Fig. 3, has been used in the following: the performance in terms of amount of code bits are very close to the optimum configuration.

The new subband splitting of the audio spectrum must be now evaluated in terms of increment of the MNR at a fixed bit rate. Therefore, a weighted mean of the MNR_n, taking into account the different subband widths, is introduced:

$$WMNR = \frac{1}{20250} \sum_n size_n MNR_n \qquad (3)$$

where $size_n$ is the width (in Hz) of the n-th subband and 20250 is the total width of the audio spectrum actually considered in coding.

The WMNR has been computed for each frame of the audio signals used in the experimental tests and an average value ($\overline{WMNR}$) has been obtained. By using the non-uniform splitting shown in Fig. (3) an increment of the $\overline{WMNR}$ has been encountered with respect to the MPEG configuration, even if of no great entity: in fact, $\Delta\overline{WMNR} = 0.324dB$.

4 Non-uniform banks

Different techniques [4]-[7] can be used to design non-uniform filter banks.

In MPEG the analysis and synthesis filter banks are obtained through cosine modulation of a 512-tap linear phase prototype. In the layer III these samples are furtherly processed with a Modified Discrete Cosine Transform (MDCT) to obtain further frequency resolution and to avoid pre-echo effects.

The subband decomposition considered in this work can be implemented by a tree structure. However, a bank composed by 512-tap filters would maintain the same characteristics of delay

introduced by the Layer II of MPEG. A possible solution would be to design three different prototypes and to obtain the analysis/synthesis bank through cosine modulation. A method to design cosine-modulated uniform FIR filter banks satisfying perfect reconstruction is given in [10]. In our case, particular care must be given to the fact that the three prototypes can not be designed independently of each other: first results of this study have given the conditions the prototypes must satisfy to cancel the most significant aliasing components; the design of actual filters is left to further study.

5 Conclusions

In this work audio signal compression by non-uniform subband decomposition has been analyzed. The masking threshold computed by the MPEG Layer II psychoacoustic model has been used to associate a cost function to a non-uniform splitting of the audio spectrum. The representations provided by the wavelet packets analysis have been taken into account. Experimental results have shown an increment of the Masking-to-Noise Ratio when the non-uniform configuration is used: the actual improvement of the quality of the reconstructed signal should be validated by subjective tests. The identified non-uniform bank can be implemented by a tree structure; the problem of the design of a filter bank having the same delay characteristics as in MPEG layer II is left to further study.

References

[1] N. Jayant, J. Johnston and R.Safranek, "Signal Compression Based on Models of Human Perception", *Proceedings of the IEEE*, Vol. 81, no. 10, pp. 1385-1422, Oct. 1993.

[2] P. Noll, "Wideband Speech and Audio Coding", *IEEE Commun. Mag.*, Nov. 1993.

[3] ISO/IEC JTC1/SC29, "Information technology - Coding of moving pictures and associated audio for digital storage media up to about 1.5 Mbit/s (Part 3, Audio)", DIS 11172, 1992.

[4] R.V. Cox, "The Design of Uniformly and Nonuniformly Spaced Pseudoquadrature Mirror Filters", *Trans. Acoust., Speech, Signal Processing*, Vol. ASSP-34, no. 5, pp. 1090-1096, Oct. 1986.

[5] P.Q. Hoang and P.P. Vaidyanathan, "Non-uniform Multirate Filter Banks: Theory and Design", in *Proc. Int. Symp. Circuits Syst.*, pp. 371-374, May 1986.

[6] K. Nayebi, T.P. Barnwell III and M.J.T. Smith, "Nonuniform Filter Banks: A Reconstruction and Design Theory", *Trans. Signal Processing*, Vol. 41, no. 3, pp. 1114-1127, Mar. 1993.

[7] J. Kovacevic and M. Vetterli, "Perfect Reconstruction Filter Banks with Rational Sampling Factors", *Trans. Signal Processing*, Vol. 41, no. 6, pp. 2047-2066, Jun. 1993.

[8] R.R. Coifman and M.V. Wickerhauser, "Entropy-based Algorithms for Best Basis Selection", *Trans. Inform. Theory*, Vol. 38, no. 2, part II (*special issue on wavelet transform and multiresolution signal analysis*, pp. 713-718, Mar. 1992.

[9] P. Fiorini, "Applicazione della teoria delle wavelets all'analisi e alla codifica del segnale audio", *Thesis for the degree in Electronics Engineering*, University of Florence, 1993.

[10] R.D Koilpillai and P.P. Vaidyanathan, "Cosine-Modulated FIR Filter Banks Satisfying Perfect Reconstruction", *Trans. Signal Processing*, Vol. 40, no. 4, pp. 770-783, Apr. 1992.