Proceedings of the 5th WSEAS Int. Conf. on MULTIMEDIA, INTERNET AND VIDEO TECHNOLOGIES, Corfu, Greece, August 17-19, 2005 (pp142-148)

Efficient Signal Adaptive Perceptual Audio Coding

MUHAMMAD TAYYAB ALI, MUHAMMAD SALEEM MIAN
Department of Electrical Engineering,
University of Engineering and Technology,
G.T. Road Lahore,
PAKISTAN.
]

Abstract: -- In this paper we present an efficient high quality perceptual audio coding scheme with a novel dynamically switched filter bank design and an associated bit allocation strategy. An additional processing block is added prior to the decomposition subband filter bank, where a decision for appropriate time/frequency resolution selection is taken based on the nature of the audio signal. A block-by-block analysis is then carried out to identify the prominent spectral peaks (PSPs), with a view to efficiently utilize their masking potential in bit rate reduction. A particular consideration is given to critical band boundaries while formulating uniform subbands, to prevent the injected coding noise becoming perceptible. A compact representation is achieved by selection of either a critical, uniform or non-uniform non-critical subband configuration depending on minimization of Perceptual Entropy (PE) estimates. An optimum bit allocation algorithm with a rate-distortion loop allots lesser bits to subbands with PSPs as compared to the actual bit demand. Test results based on ITU-R recommendations show that our coding scheme performs considerably better as compared to MPEG-1 Layer-3 (the de-facto standard for audio interchange on the Internet) for a majority of signal types. The paper concludes with a discussion of future research implications of the work.

Keywords: Subband coding, critical bands, perceptual entropy, masking threshold, bit allocation

1. Introduction

In the context of evolving digital audio and Internet multimedia applications a need to transmit/store maximal amount of information consuming minimal relevant resources persists. Audio coding thus remains a field of continuous research interest. Perceptual coding is a class of lossy audio compression algorithms (e.g. [1]) that exploit the perceptual properties [2] of the human auditory mechanism to achieve coding gains. The signal spectrum is split into subbands of equal or unequal bandwidths, and bits are allocated according to distribution of energy within these bands. Compression is achieved by inducing noise (through coarse quantization) into the signal which is imperceptible by the human ear. If the reconstructed signal is indistinguishable from the original, the coding is termed as transparent. The current widely famous perceptual audio coder—the MPEG-1 Layer 3 (MP3) [3] achieves transparent compression at around 64 kbit/s for monophonic signals. The performance of a codec however varies with different types of signals, and the choice of an appropriate filter bank heavily contributes towards the performance of a perceptual audio codec [4]. Several efforts have been made to adapt the filterbank design to nature of the incoming signal. However these works have focused on shaping the quantization noise according to the masked threshold in order to reduce perceptible coding artifacts due to transient signals[5] [6] [7]. In this work we present a new strategy for the adaptation of subband configurations to changes in audio signal, with an aim of improving the coding gain/compression ratio. The approach results in an improved quality of reconstruction compared to MPEG-1 Layer-3 at similar bit rates or alternately a reduced bit rate at similar quality. The paper is organized as follows. Section II describes the encoder design. Section III explains the novel idea of dynamic subbands selection strategy based on a minimization of PE estimates. Section IV briefly explains the bit allocation and Section V presents the results achieved by testing the codec on files from the SQAM database. Section VI concludes the paper with a discussion of future research implications of the work.

2. Encoder Structure

Figure 2 shows the basic encoder structure. The resolution selection block checks the presence of transient signal portions, information regarding which is embedded in the side information.

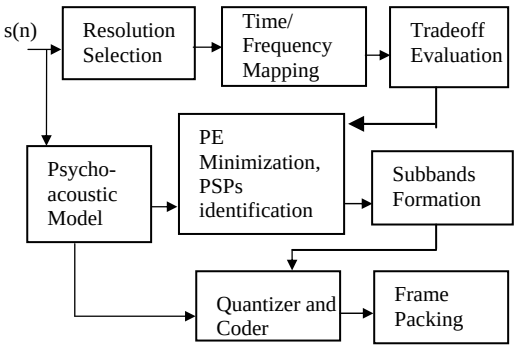

Fig 1. Encoder Structure

The time frequency mapping block employs a modified discrete cosine transform (MDCT) [8] to obtain a spectral representation of the audio signal. The tradeoff evaluation block evaluates the best tradeoff between empty quantizer slots and amount of side information. The following PSP identification block extracts frequency and phase information from the MDCT coefficients [9] and identifies prominent spectral peaks (PSPs) with an aim of utilizing them to suppress maximum inject able coding noise with bit allocation procedures. It also estimated perceptual entropy (PE) estimates for the decision of selection of an appropriate configuration by the subbands formulation block. The PSPs are identified using the following methodology:

2.1 PSPs Identification

PSPs are identified from the power spectral density (PSD) estimate by looking for local maxima that exceed their neighboring samples by some threshold ***T***. As a starting point, ***T*** is set at 7 dB. However, this threshold is adjusted to satisfy the following condition:

$$31<N_{PSP}<129 \quad (1)$$

where N_{PSP} is the number of PSPs in one block of data. The range in (1) is selected to match the number of PSPs with that of possible subbands in a block, which are dynamically formulated, by a method explained in section 3. The PSPs are identified as:

$$PSP_n = \{ P(k) \mid > P(k \pm 1), P(k) > P(k \pm \Delta) + T \} \quad (2)$$

where Δ is adjusted dynamically to satisfy (1). The PSPs differs from "tonal-maskers" in that, energy from three adjacent spectral components centered at the peak are combined to form a single "tonal-masker", where as PSPs are the stand alone peaks. An estimation of the noise and tonal maskers in the signal spectrum follows and a global masking threshold is computed by using the established rules of psychoacoustics.

3. Dynamic Subbands Selection Strategy

Irrelevance reduction in perceptual subband coders is achieved by splitting the spectrum into subbands, and obtaining a lower bit rate by inducing the allowable amount of noise as suggested by the psychoacoustic model. This is done by quantization with a step size dynamically selected for each subband. The choice of the number of spectral samples per subband is made by considering the tradeoff between bits used for side information versus bits used for empty quantizer slots [10]. Most coding schemes including the MPEG-1 Layer-3 use a fixed configuration for the distribution of spectral samples in subbands. The novelty of our work is that every block of data is analyzed in terms of its masking potential, and a different configuration of subbands is selected in each frame which is signaled to the decoder through side information. Although an infinitely large number of configurations are possible, only a finite set is established and used in this work. Specifically, a 6-bit code is used to signal the chosen configuration which can be among any of the following:

1. Critical Bands (1 configuration)
2. Uniform Bands ranging from 32 to 128 (13 configurations)
3. NUNC Bands (40 configurations)

The actual decision on selection of the appropriate configuration is done according the following steps:

1. Formulate a critical bands distribution.

2. Formulate all uniform bands distributions.
3. Formulate all NUNC distributions.
4. Compute the PE for all configurations formulated in serial 1 to 3 above.
5. Select the configuration with minimum PE

The outlined procedure may seem to be computationally expensive at the first reading, but the actual measurements have shown it to be just a quarter of the expense incurred by the FFT, which is quite acceptable. Further, it should be noted, that the number of subbands in serial 2 and 3 above is not only powers—of—two, which is a novelty of this work. Most previous works in perceptual audio coding have formulated the number of uniform bands only as powers—of—two, probably for equal number of samples in each subband and subsequently an easy identification of subbands data from the bitstream. However our coding scheme averts this limitation by embedding a code representing the number of samples in each subband in case of any of the 64 distributions indicated at an expense of a 6-bit code.

The coding efficiency enhancement in our scheme starts with an optimization of the empty-quantizer-slots vis-à-vis side information tradeoff. The dynamic selection of different subband patterns minimizes the number of empty quantizer slots as compared to fixed number of subbands and greatly enhances the compression ratio. An example of this capacity wastage is illustrated in figure 2 with a uniform configuration having 8 samples per subband. The dotted boundary above each subband represents the amount of information which can be conveyed given the number of bits for each subband.

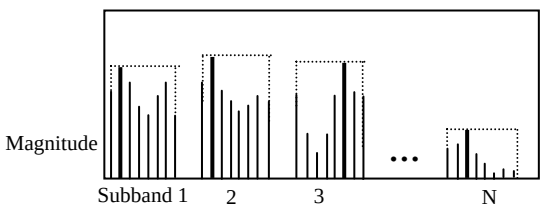

Fig 2. Quantizer Slots for Uniform Subbands with 8 samples each

As it can be seen, this capacity has been utilized quite optimally in subband 1, however in subbands 2 and 3 a lot of space is empty/wasted as more number of bits have been allotted to these subbands because of the PSP magnitudes represented by a dark line. In the proposed coding scheme, we minimize this wastage by changing the subband configuration in every block depending upon the occurrence of such PSPs. For example if instead of the subband configuration shown in figure 2, a uniform configuration with 4 samples per subband is applied, the situation will improve as depicted by figure 3.

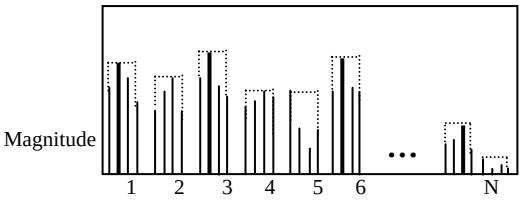

Fig 3. Quantizer Slots for Uniform Subbands with 4 Samples each

Now, the number of empty quantizer slots has been minimized thus giving a boost in the compression ratio. A final step to further lower the bit rate is to reduce upon the amount of side information where the major portion of it remains the bits allotted to every subband. Consequently, an increased number of subbands will warrant a bigger amount of side information. To counter this, capitalizing on the fact that a 6-bit code is already being sent per block of data to indicate the subband configuration, we have about 40 combinations which could be used for NUNC subbands with varying number of samples per subband. Using these un-conventional configurations we don't have the restriction of sending huge side information in case of a large number of small subbands, while still having small subbands where required. Table 1 shows the NUNC distribution for one particular code. Here, another consideration is made to not cross any critical band boundaries coming in between. Although the global masking thresholds have been calculated and the spread of masking accounted for,

when the rate distortion based bit allocation procedure allocates lesser bits to a particular subband compared to its actual demand, the corresponding coding noise is spread in the whole subband. The rate distortion loop in the bit allocation algorithm of our coding scheme allots lesser bits to PSP containing bands first, and if that coding noise is restricted to the critical band where the PSP exists, the distortion posed to the human listener will be significantly less than the case where it is spread to neighboring auditory filters. However, mostly at higher frequencies, more than one NUNC subbands may exist within a single critical band (such cases in bold font in column D of table 1). The NUNC distribution patterns are selected through extensive experimentation and identification of most commonly required configurations for PE minimization. Enlisting of all these patterns for the remaining code words comprising of 6-bits requires a lot of space, and only an example is presented in table 1.

A	B	C	D
Code	**NUNC Subband Number**	**No of Samples**	**Corresponding Critical Band Number(s)**
1110	1-3	2 each	1-3
	4	3	4
	5	2	5
	6,7	3 each	6,7
	8-10	4 each	8-10
	11,12	5 each	11,12
	13,14	7 each	13,14
	15	4	**15**
	16	5	
	17	6	**16**
	18	5	
	19,20	6 each	**17**
	21	5	**18**
	22,23	6 each	
	24-26	7 each	**19**
	27,28	8 each	**20**
	29	9	
	30,31	8 each	**21**
	32,33	7 each	
	34-39	7 each	**22**
	40-45	8 each	**23**
	46	10	
	47-126	8 each	**24**
	127	9	
	128	153	**25**

Table 1. Distribution of samples for a NUNC configuration corresponding to the code 001110

Table 2 shows the 6-bit codes and interpretations for the critical and the uniform subbands configurations. The number of subbands is computed by dividing total number of spectral coefficients available (according to the window size in use) by the number of samples in each subband (Column B):

$$N_{usb} = \text{int}(N_c / S_i),\ i = \{1, 2, 3 ... N_{usb}\} \quad (3)$$

Where, int stands for "integer",
N_{usb}= Number of uniform subbands
N_c = Number of spectral coefficients in the current window
S_i = Number of samples in uniform subbands

There is an additional subband at $i= N_{usb}+1$. The convention for number of samples in it is as follows:
If Number of samples in last additional subband $< S_i/2$ *for* $i=1..(N_{usb}$ -1) then include them in S_i *,for* $i=N_{sub}$

A	B	C	D
Code	**Samples per subband**	**Subband Configuration**	**Samples in Last Subband**
000000		Critical Bands	N.A
000001	16	**32** Uniform with **16** samples each	Nil
000010	15	**33** Uniform Subbands with **15** samples each	17
000011	14	**36** Uniform Subbands with **14** samples each	8
000100	13	**38** Uniform Subbands with **13** samples each	18
000101	12	**42** Uniform Subbands with **12** samples each	8
000110	11	**46** Uniform Subbands with **11** samples each	6

000111	10	**50** Uniform Subbands with **10** samples each	12
001000	9	**56** Uniform Subbands with **9** samples each	8
001001	8	**64** Uniform Subbands with **8** samples each	Nil
001010	7	**72** Uniform Subbands with **7** samples each	8
001011	6	**84** Uniform Subbands with **6** samples each	8
001100	5	**101** Uniform Subbands with **5** samples each	7
001101	4	**128** Uniform Subbands with **4** samples each	Nil

Table 2. Embedded codes and meanings for uniform subbands

The number of samples computed in this way for the additional subband are as shown in Column D. It should be noted however, that these computations are not required to be made in every frame by the decoder, as they are incorporated in the source code according the bitcode provided in column A. Column D in table 1 is presented for the understanding of the reader only.

4. Bit Allocation

Traditional perceptual coders perform a bit allocation to subbands based on the distribution of the signal power in the frequency domain to minimize the total noise power [11]. To this end, either a signal to mask ratio (SMR), or and energy based bit allocation is performed. We take the energy based approach where a total audible distortion minimization scheme is adopted. We use an inverse greedy algorithm to achieve the objective in (3). First all bands are allotted the bits according to their actual demand. Then a rate-distortion loop operates in an iterative manner to cut-down on the bit allocation while maintaining the quality. This is done by exploiting the masking potential of the PSPs first. The masking thresholds around the spectral vicinity of PSPs allow more coding noise to be injected, and already the subband configurations have been formulated in such a manner that allocating fewer bits to the bands with PSPs will have a minimum damage to the audio quality. In a single iteration, one bit each is removed from bands with PSPs, leaving others bands as such. The process is continued till either the target bit rate is met, or the bit cut induces enough noise in PSP bands as allowed. Any further iterations if required remove bits from all bands based on an energy criteria where the first bit is removed from the band with maximum total energy.

5. Results

The set of signals from the European Broadcasting Union (EBU) Sound Quality Assessment Material (SQAM) [12] was used for testing the codec performance. Subjective listening tests were conducted separately with a listening set up based on ITU-R recommendation BS.1116 [13] for our coding scheme and MPEG-1 Layer 3 at 56 Kbps (monophonic).

File	SDGs for proposed scheme	SDGs for MPEG-1 Layer 3
Electronic tune (Frère Jacques)	-0.2	-0.2
Harpsichord	-0.5	-0.7
Trumpet	-0.2	-0.5
Soprano	-0.3	-0.3
Glokenspiel	-0.4	-0.5
Female speech (English)	-0.1	-0.3
Male speech (German)	-0.3	-0.5

Table 3. Subjective Difference Grades for PSP based Codec and MPEG-1 Layer 3 at 56Kbps

The subjects were a mix of expert and non-expert listeners. They were presented the audio files in an A,B,C sequence, with A always being the reference un-encoded signal, and the coded and hidden reference signals randomly as B and C. The averaged subjective difference grades (SDGs) for selected SQAM test files are listed in table 3. It can be seen that our coding scheme performs better as

compared to MPEG-1 Layer 3 for most of the signal types. Our coding scheme performed slightly better possibly because the all the frequency regions are always matched strictly to the critical bands distribution by the NUNC configurations.

File	Proposed codec bitrate for good quality	% Improvement in Bitrate compared to MP3 (64kbps)
Electronic tune (Frère Jacques)	56	5
Harpsichord	60	2.5
Trumpet	52	8
Soprano	54	6.5
Glokenspiel	60	2.5
Female speech (English)	54	6.5
Male speech (German)	56	5

Table 4. Percentage of improvement in bitrate for good quality compared to MP3.

Another analysis was carried out shown by table 4, where the percent bitrate savings (at 'good' quality) for each class of signals by the proposed coding scheme as compared to MPEG-1 Layer 3 were observed. The proposed codec was operated at varying bitrates (controlled by the bit allocation loop), and the bit rate offering 'good' quality was marked for computation of the bit rate savings. It can be seen that on the average a 5-6% of bit rate is saved by employing the proposed coding scheme compared to MP3 codec.

6. Conclusion

A novel high quality perceptual audio coding scheme is proposed in this paper. The idea of prominent spectral peaks PSPs is presented, which are utilized to dynamically switch the filter bank design in order to minimize the total over all bitrate. A number of unconventional NUNC subbands configurations are formulated in addition to a critical bands distribution, and a uniform distribution with 32-128 subbands. A final decision on the selected filterbank is made depending upon minimization of PE estimates from all these configurations. An associated bit allocation strategy uses an inverse greedy algorithm, and starts with scaling bits from bands with PSPs. The idea is, that the global masking curve is higher around the PSPs thereby introducing minimum distortion by bit removal. Finally, the codec is tested for performance using ITU-R recommendations on subjective listening tests, and found to perform slightly better than MPEG-1 Layer 3. The work can be extended to include more accurate measures other that the PE estimates for the filter bank switching decisions. Moreover, a 7 or 8 bit code can be used to further enhance the NUNC distribution optimizations.

References:

[1] K. Brandenburg and J. D. Johnston, "Second generation perceptual audio coding: The hybrid coder," in *Proc. 88th Conv. Aud. Eng. Soc.*,Mar. 1990, preprint 2937..

[2] E. Zwicker and H. Fastl, *Psychoacoustics Facts and Models*. Berlin, Germany: Springer-Verlag, 1990.

[3] K. Brandenburg, G. Stoll, Y.F. Deh6ry, J.D. Johnston, L.v.d. Kerkhof, E.F. Schroeder: "The ISO/MPEG-Audio Codec: A Generic Standard for Coding of High Quality Digital Audio", 92nd AES Convention, Vienna 1992, preprint 3336

[4] J. Johnston, S. Quackenbush, G. Davidson, K. Brandenburg, and J. Herre, "MPEG audio coding," in *Wavelet, Subband and Block Transforms in Communications and Multimedia*, A. Akansu and M. Medley, Eds, Boston, MA:Kluwer Academic, 1999.

[5] M. Iwadare, A. Sugiyama, F. Hazu, A. Hirano, and T. Nishitiani, "A128 kb/s hi-fi audio CODEC based on adaptive transform codingwith adaptive block size MDCT," *IEEE J. Select. Areas Commun.*,pp. 138–144, Jan. 1992.

[6] Princen, J. and Johnston, J. D., ``Audio coding with signal adaptive

filterbanks," *IEEE ICASSP*, 1995, pp. 3071-3074.

[7] Herre, J. and Johnston, J. D., ``A continuously signal-adaptive filter bank for high quality perceptual audio coding," *IEEE ASSP Workshop on Applications of Signal Processing to Audio and Acoustics*, Mohonk, NY, 1997.

[8] Princen and A. Bradley, "Analysis/synthesis filter bank design based on time domain aliasing cancellation," *IEEE Trans. Acoust Speech, Sig Processing*, vol. ASSP-34, pp. 1153–1161, Oct.1986

[9] Corey I. Cheng. Method for Estimating Magnitude and Phase in the MDCT Domain, 116th AES Convention 2004 May 8–11 Berlin, Germany

[10] J. D. Johnston, "Transform Coding of Audio Signals Using Perceptual Noise Criteria," *IEEE J. Select Areas in Communications*, vol. 6, pp. 314 - 323 (1988 Feb.).

[11] R. Zelinski and P. Noll, "Approaches to adaptive transform speech coding at low bit rates," *IEEE Trans.Acoustics, Speech, Signal Processing*, vol. 27, pp. 89–95, Feb. 1979.

[12] SQAM -Sound Quality Assessment Material. Available at www.tnt.uni-hannover.de/project /mpeg/audio/sqam/

[13] "Methods for subjective assessment of small impairments in audio systems including multichannel sound systems,", ITU-R BS 1116, 1994.