OPTIMIZING HIGH QUALITY AUDIO CODING: ADVANTAGES OF FULL SYSTEM OBSERVABILITY

Aníbal J. S. Ferreira

INESC, Largo Mompilher 22, 4000 Porto Portugal

ABSTRACT

Perceptual audio coders rely on the efficient reduction of perceptually irrelevant components of the audio signal as well as on the removal of statistical signal redundancies to achieve good coding gains. In order to reach high compression ratios without reducing the subjective quality of the encoded audio signal, it is necessary to identify critically interdependent functional units of the encoding algorithm and to jointly optimize their performance. A flexible and interactive simulation and analysis environment has been programmed to assist the development and optimization of a new perceptual audio coder. The main features of this environment will be explained and the most relevant aspects that were found to limit the encoding performance will be presented.

1. INTRODUCTION

Perceptual coding is an approach to the problem of signal compression based essentially on models of human perception. In the context of audio coding, the properties and tolerances of the human auditory system are represented by psychoacoustical models. These models are used to produce an estimate of the signal dependent noise profile that does not cause noticeable impairments when actually added to the original signal. It is know as the Threshold of Masking of the Just Noticeable Distortion (JND) [1]. A compression strategy that takes advantage of this fact involves an analysis of the audio signal so as to estimate the subjectively irrelevant components of the signal and replace them by quantization noise. This operation is called irrelevancy extraction and is based on the concept of masking. Masking refers to the total or relative inaudibility of one sound component due to the presence of another one, with particular relations in amplitude, frequency, time and space[2]. Among the usually considered aspects of masking (simultaneous masking, forward masking and backward masking), simultaneous masking is the most important source of irrelevancy extraction and is essentially characterized in the frequency domain. This is the main reason why most perceptual audio coders operate in the frequency domain by means of a time-frequency transformation of the audio signal.

Taking the Compact Disc (CD) quality as a reference, compression ratios of about 6:1 are reachable with transparent quality (*i.e* the decoded and reconstructed audio signal remains perceptually indistinguishable from the original signal), by proprietary algorithms [3][4], as well as by the recently developed standard MPEG-Audio algorithm [5].

In general, due to the algorithmic complexity of high quality audio coders, it is very difficult and time consuming to optimize their coding performance, particularly at very low bit rates. In fact, the best possible audio quality for a given bit rate, may not be achievable by optimizing each functional unit of the coder independently since the sum or product of local optimizations may not lead to a global optimization. Instead, critical interdependencies between functional units must be identified, understood and jointly optimized.

This paper presents a pragmatic approach to the problem of system design and optimization which considers that full system observability is necessary before an appropriate system controllability, aiming at a global performance optimization.

In section 2 the paradigm of perceptual signal compression will be expressed as a suitable compromise between information reduction and subjective coding transparency. In section 3 the concerns of coding efficiency are shown to face practical difficulties in the system design and system optimization phases. In order to identify and to overcome these difficulties section 4 presents a new simulation paradigm that was adopted to design and optimize a new perceptual audio coder by programming a highly flexible and interactive simulation environment. Section 5 presents a few aspects that were found to have a relevant influence on the coding performance and section 6 presents the main conclusions of the study.

2. INFORMATION REDUCTION *VERSUS* CODING TRANSPARENCY

The redundancy removal of an audio signal, which is an information lossless compression method, only provides a modest (average) compression gain in the order of 2:1. Removing the irrelevancy of an audio signal has the potential to provide compression gains higher than 4:1. The upper bound for the transparent compression gain can be estimated using the Perceptual Entropy (PE) measure. This measure was developed to estimate the information contents of an acoustical stimulus from a perceptual point of view [6] or, in other words, to estimate the minimum number of bits needed to code (*i.e.* to properly quantize), in a transparent way, an arbitrary audio signal. Several results indicate that

This work was supported by JNICT, the Portuguese Science and Technology Research Agency.

0-7803-2431-5/95 $4.00

2 bit/sample is a good estimate for wideband audio signals although it is expected 1.5 bit/sample could be reached in the near future. This figure is not objectively measurable, it is just an extrapolation based on existing and simplified psychoacoustical models. In fact, there are not perfect nor exact psychoacoustical models because if feasible, an exact psychoacoustical would be subject dependent. It is only possible to have approximated models derived statistically from several studies and experiments [2].

Transparent coding of audio signals is fairly reachable at $128Kbit/s$ per channel by most known perceptual audio coders such as the AT&T PAC [3], the Dolby AC-2 [4] and the standard ISO/IEC MPEG-Audio Layer II and Layer III [5]. In fact, since at compression ratios in the order of 6:1, the full margin of statistical redundancy and irrelevancy of audio signals is not completely exploited, there is no need for special concern about a global optimization of the encoding algorithm. However, at higher compression ratios, in the order of 12:1, the corresponding high encoding efficiency calls for the maximization of the global encoding performance of the algorithm which depends, obviously, on the efficiency of each functional unit of the coder, but depends essentially on the joint efficiency of critical coder functional units.

3. THE EFFICIENCY CHALLENGE

If exactly known, the PE limit for a given audio signal can not be reached because neither any practical analysis/synthesis structure will be able to mimic the complex time and frequency signal analysis as performed by the human hear, nor any practical quantization strategy will mirror the complex non-linear behavior of the human auditory system particularly when we admit audio signals with high dynamic range (about $100dB$) and high spectral bandwith ($e.g\ 20KHz$ or higher as in the case of very sharp attacks).

For practical reasons, perceptual audio coders use simplified psychoacoustical models, simple quantization schemes and computationally efficient signal analysis/synthesis structures. For an improved efficiency, all quantized data and side information can be further entropy encoded. It has been observed that very low bit rate coding conditions force the relevant coding variables and parameters to interact very intensely and their interaction was verified to become more and more dependent on the actual audio data. The search for an "average optimal" coding performance, becomes an even more complex task because quite frequently at least a brief listening test is necessary to assess and validate results. This means it may take a few years for a perceptual audio coder to be developed and finely tuned before achieving the best coding performance. A pragmatic approach to overcome the above problems is to consider that for a fast evaluation and efficient controllability of the encoding system performance, a necessary prerequisite is to have a simulation environment with full, immediate and interactive observability and analysis capabilities. These capabilities are also justified by considering that due to the dynamics of high quality audio signals, the performance of the encoding algorithm may be very poor on sporadic but very critical signal passages.

4. A NEW SIMULATION PARADIGM

A new frequency domain encoding/decoding algorithm (Audio Spectral Coding -ASC) has been developed and programmed according to the paradigms of an object oriented language (C++) and using the powerful graphics and audio capabilities of a workstation (Silicon Graphics, IrisIndigo). This approach made very easy the inclusion of system observability concerns in the programming structure of ASC. The audio capabilities were included in a independent software module that controls the input and output audio ports of the workstation, and can be used as an audio editor and analysis tool. Besides the capabilities of a conventional audio editor, it also has extensive spectral analysis features. Any new feature can be easily programmed and used interactively as well as the existing ones. This audio editor can be used simultaneously with the audio encoding algorithm thus keeping track of all the audio data previously encoded and reconstructed, or can be used after the complete encoding and decoding process. In the following, we will focus only in the encoding/decoding simulation environment.

4.1. Encoder/Decoder Structure

The simplified encoding section of ASC for a single channel is represented in Figure 1. The analysis/synthesis filterbank is based on a multiresolution 50% overlap TDAC filterbank and the frequency response of each channel is given by an optimized window [7]. The psychoacoustical model was derived from Zwicker results [2] and the quantization unit uses a linear quantizer. All main and most side information is entropy encoded and the encoding algorithm has the ability to generate statistics for the update of Huffman code books.

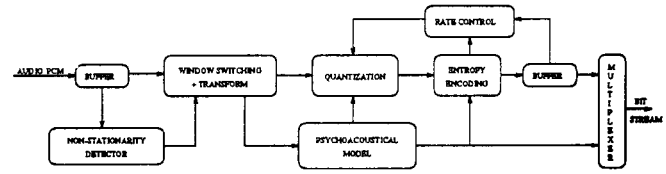

Figure 1: Simplified structure of ASC.

When constraints are placed on the bitstream, *e.g.* a constant bit rate then, use is made of an output buffer and a closed loop interaction between the (re)quantization unit and the entropy coding unit, in order to match coding quality with bit rate constraints.

The decoding section of ASC is simpler than the encoding counterpart and involves only the demultiplexing and entropy decoding operations, the inverse quantization and the frequency-time mapping.

4.2. Graphical Simulation Environment

The full system observability is achieved by representing graphically the main processed audio data and system variables associated to each relevant functional unit of the complete encoding *and* decoding process for each new processed segment of audio data. The whole simulation system may be inspected in a interactive (mouse driven) way. The graphical representation does not bring a significant additional delay to the simulation process because of the direct

call of high performance native graphic subroutines. Furthermore, it is possible to exit the graphics mode and let the simulation run at full speed.

The complete system observability is based on a multi-window representation. Each window is a particular object with multicolor overlapped data. For the simulation of single channel encoding, five windows are represented:

The Time Window represents the input audio data overlapped with reconstructed output audio data (only 50% since a TDAC filterbank with 50% overlap is used) together with the time window sequence used to filter the input data (typically one long window for stationary regions of the signal, or a sequence of short windows in non-stationary regions of the signal to prevent pre-echo effects).

The Linear Frequency Window represents the spectral power of the original signal as well as tonality information (which is an important factor of the psychoacoustical model efficiency) and the spectral power of the decoded audio signal.

The non-linear frequency window (Bark Window) represents the spectral power of the original signal, tonality information, the estimated spectral JND profile as well as the final masking profile that has to be used to quantize the spectral coefficients in order to comply with bit rate constraints. The non-linear frequency scale is related to the Bark rate.

A Bit Allocation Window shows the distribution of bit utilization among coder bands by main and side information as well as the assignment of different Huffman tables to clusters of bands.

Finally, an Information Window displays all relevant values about axis ranges of each window as well as other figures, *e.g.* number of encoded coefficients and scale factors, SNR of the fully reconstructed half-segment, local bit rate, ratio main/side information, etc.

An simplified example (there is data and axis information omited for clarity) of audio data and processing variables associated to the coding of a $11.6ms$ segment of female speech can be appreciated in Figures 2-5.

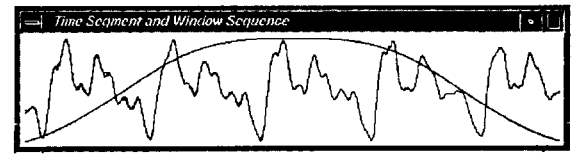

Figure 2: Segment of input audio data overlapped with the analysis window and the reconstructed sub-segment (first half) of decoded audio.

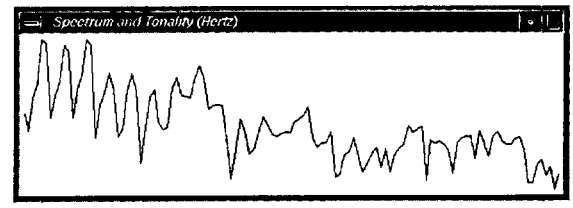

Figure 3: Power spectrum of the audio segment on a linear frequency scale.

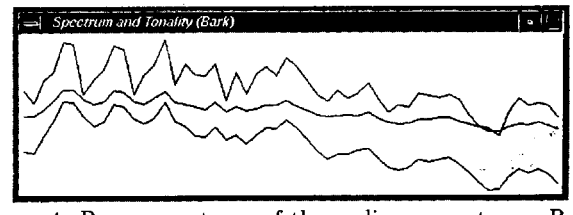

Figure 4: Power spectrum of the audio segment on a Bark scale (upper curve). The lower curve is the Threshold of Masking and the one in the middle is the final quantization noise that corresponds to a bit rate of $40Kbit/s$.

Figure 5: Bit utilization only for quantized coefficients.

4.3. Interface Functionality

Besides resizing and moving any graphical window which are intrinsic features of the native X window programming environment, the interactive functionality of the encoder/decoder is possible mainly by an extensive use of the mouse In fact, it is possible within each window to call a pup-up menu and select, for instance, the detailed analysis of the displayed data. In this case new range values must be typed within a small editor window that is created for this purpose. It is also possible to inspect interactively the exact coordinates of any point within any graphical window. Other existing interactive capabilities include skipping the visual inspection of a desired number of subsequent encoded segments, ignoring visual inspection after any segment and naturally, simulation abortion. Given the high flexibility of the simulation environment, it is possible to study, for any given bit rate constraint (variable or fixed), the influence of the transform size (frequency resolution) which may be any arbitrary number power of two, the sampling frequency of the audio signal, different window switching alternatives (it is possible to improve the time resolution of the encoding process to 1/2, 1/4 or 1/8 of the "normal" segment size), and different window types (satisfying the perfect reconstruction property), *etc.* The simultaneous coding of any number of channels till 5 is also an option although not yet completely implemented because the optimization of a single channel coding is still on-going work.

5. CRITICAL EFFICIENCY FACTORS

A number of critical aspects that influence the global optimization has been identified. For instance, if extensive spectral zero bit quantization (relative to half the sampling rate of the audio signal) is consistent over several consecutive encoded audio frames, the quantization noise may not be assumed anymore simply as additive as usually considered for simplicity, but rather multiplicative which implies a circular convolution of the sampled segment of audio data with the equivalent "quantization filter". The resulting circularly aliased noise terms must be attenuated. One possibility as reported in a study conducted recently [7] for the

case of the TDAC filterbank is to find an optimized window that minimizes the amount of circularly aliased terms for an arbitrary "quantization filter", while insuring the perfect reconstruction property. It was verified at the same time that a better frequency selectivity could be obtained relatively to a commonly used (sine taper) window. Other important factors that are likely to limit the coding efficiency were found:

- the suitable raise criterion of the JND noise profile for a "graceful" quality of the undercoded audio signal (*i.e.* coded with less bits than those required for the quality corresponding to the edge of transparency),
- the suitable spectral resolution (relative to the Bark rate) of the encoding process since an excessive (sub-Bark) resolution can expose too much the inaccuracies of the psychoacoustical model and imply a significant increase in side information, whereas a poor resolution can lead to a sub-optimal bit allocation but allow a better tradeoff between side information (mainly scale factors) and main information (quantized spectral coefficients),
- the adjustment, correction or compensation of the psychoacoustical model considering the idiosyncrasies of the analysis/synthesis filterbank,
- the existence of extensive spectral zero bit quantization and other concurrent factors that expose the "aliasing signature" of the analysis/synthesis filterbank,
- the efficient utilization of Huffman code books, *i.e* the minimization of the bit count of the audio data symbols by means of an optimal assignment of several Huffman code books tuned for different statistical distributions,
- the update of Huffman tables according to the bit rate, and the associated criterion to collect statistics which depends on the overcoding or undercoding expectation given the available bit rate,
- the type of quantizer (*e.g.* linear, logarithm) used to quantize scale factors and spectral coefficients and the width of the "zero" zone of the quantizer,
- the use of an output buffer to accommodate different bit rate requests among adjacent encoded audio segments,
- the judicious exploitation of inter segment correlations (*i.e.* of short windows when the time resolution of the encoding process is improved),
- the characterization and efficient encoding of all side information.

Quantitative results of ASC performance taking the MPEG-Audio Standard as a reference are being prepared and will be published in a future paper. As an example, Figure 6 illustrates the necessary bit rate to code transparently a representative mix of critical mono audio signals (SQUAM CD) as a function of the segment size, *i.e* of the processing delay.

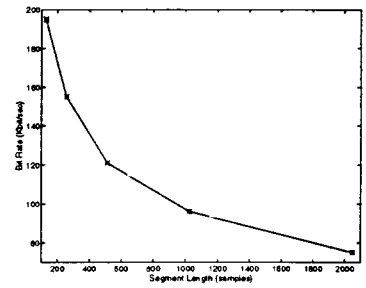

Figure 6: Bit rates for transparency as a function of the transform window size. The sampling rate is 44.1 KHz.

6. CONCLUSION

An efficient perceptual audio coder is a complex algorithm whose global performance depends essentially on the joint optimization of interacting critical functional units. An highly flexible and interactive simulation environment has been presented that helped significantly to develop, debug and optimize a new perceptual audio coder (ASC) in just a few months. This environment was programmed using the paradigms af an object oriented language and took advantage of the powerful graphic and audio features of a workstation. Preliminary results show that ASC codes transparently any monophonic audio signal with CD quality using a bit rate not exceeding 96 Kbit/s and is also able to code speech signals at 16 Kbit/s with a quality close to that of G.728. For signals sampled at 48 KHz/s the total encoding and decoding delay is $43ms$.

On-going work and next studies will concentrate on further optimization of single channel coding, particularly at very low bit rates and later, on multichannel coding. This latter extension is trivial since conceptually, it matches two of the most powerful paradigms of C++: object replication and class inheritance.

7. REFERENCES

[1] Nikil Jayant *et al.* "Signal Compression Based on Models of Human Perception". *Proc. IEEE*, 81(10):1385–1422, Oct 1993.

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

[3] J. D. Johnston and A. J. Ferreira. "Sum-Difference Stereo Transform Coding". In *IEEE ICASSP*, pages II–569 II–572, 1992. San Francisco, CA.

[4] L. Fieldier and M. Bosi. "AC-2: High Quality Digital Audio Coding for Broadcasting and Storage". In *Broadcast Eng. Conf.*, pages 98–105, April 1992.

[5] Karlheinz Brandenburg and Gerhard Stoll. "The ISO/MPEG-Audio Codec: A generic Standard for Coding of High Quality Digital Audio". *92nd Conv. of AES*, March 1992. Preprint n. 3336.

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

[7] Aníbal J. S. Ferreira. "Convolutional Effects in Transform Coding with TDAC: An Optimal Window". *Accepted for pub. in Trans. Speech and Audio Processing.*