A BACKWARD-ADAPTIVE PERCEPTUAL AUDIO CODER

João Manuel Rodrigues[*] Ana Maria Tomé
Departamento de Electrónica e Telecomunicações / INESC
Universidade de Aveiro
3810 AVEIRO, PORTUGAL
Tel: +351-34-370500; Fax: +351-34-370545
e-mail: jmr@inesca.pt

ABSTRACT

This paper presents a new audio compression algorithm that includes a nonuniform filter bank, gain-adaptive logarithmic quantizers, arithmetic entropy coding and an explicit psychoacoustic model to adapt the quantization according to perceptual considerations. Unlike existing perceptual coders, the new system is backward-adaptive, i.e., adaptation depends exclusively on already quantized samples, not on the original signal. We discuss the advantages of backward adaptiveness and show that it can be successfully applied to perceptual coding.

1 INTRODUCTION

Compression of digital signals can be achieved in two ways: *redundancy removal*—exploitation of known properties of the signal source; and *reduction of irrelevance*—exploitation of the limited sensibility of the final receiver. Traditionally, speech and audio waveform coders have focused on techniques to remove redundancy such as predictive, transform, and entropy coding. In these systems, the relevance or irrelevance of the introduced quantization noise is equated to somewhat arbitrary distortion measures such as the signal-to-noise ratio (SNR) or the mean square error (MSE). It is now widely recognized that these traditional coders cannot generally perform well in audio applications because: 1) audio signals, due to their vast diversity, wide bandwidth and large dynamic range, are hard to model and that limits the amount of redundancy that can be removed; 2) SNR, MSE and similar distortion measures do not reflect real human auditory perception so that irrelevancy is not fully exploited [6]. To achieve high quality coding at very low bit rates, then, it is necessary to employ knowledge of perceptual phenomena in order to minimize the audibility of the introduced distortion. This is the essence of *perceptual audio coding.*

In recent years, several perceptual audio coding systems have been proposed which claim “perceptually transparent” or “high quality” coding with rates as low as 2 bits per sample per audio channel. Most of these systems share a common generic structure: they are transform or sub-band coders with forward-adaptive quantizers controlled by some form of perceptual adaptation algorithm that evaluates the signal-dependent masking threshold and shapes the quantization noise accordingly.

We propose a new perceptual coding algorithm for audio signals that differs from other existing coders in one important respect: it is a *backward-adaptive* system. The coder structure is described in Section 2. In Section 3, we discuss the advantages of using backward adaptation, and show that this approach is not incompatible with perceptual coding. Some preliminary results are presented in Section 4.

2 CODER STRUCTURE

Figure 1 shows the block diagram of the proposed coding system. The input is a single-channel, CD-quality signal: 44100 samples per second, 16 bit PCM. The analysis filter bank (T) splits the input signal into 62 maximally decimated frequency bands of varying widths. Sub-band samples are then discretized by a set of gain-adaptive logarithmic quantizers, and entropy-coded for transmission. The adaptation algorithm uses a psychoacoustic model to continuously evaluate the signal-dependent masking threshold, and sets the gains of the quantizers accordingly, in order to avoid or reduce audibility of the quantization noise. At the receiver, the bit stream is decoded, and the quantized sub-band samples are recovered. Finally, the synthesis filter bank combines the sub-bands to form a replica of the original signal. The dequantization gains are adapted using the same algorithm and the same quantized samples as the encoder. Therefore, there is no need for any side information to keep synchronism between transmitter and receiver.

2.1 The Filter Bank

Signal decomposition is performed by a two-stage nonuniform analysis structure designed to approach the demanding time and frequency resolution properties of the human ear. The first stage splits the signal using a

[*] This work was supported in part by JNICT (Portugal) under grant number BM/2979/92-IA.

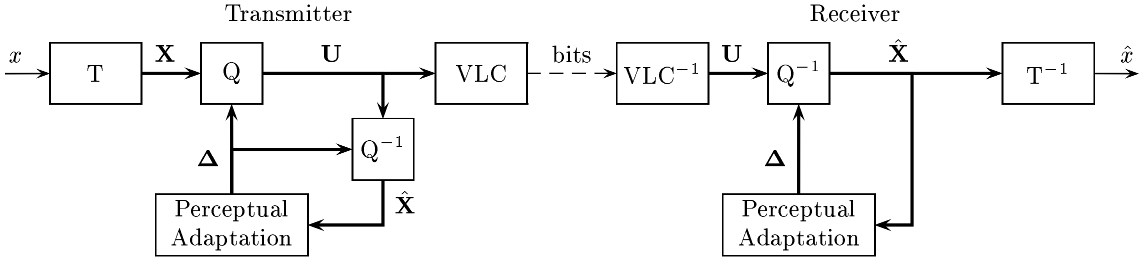

Figure 1: Backward-adaptive perceptual audio coding system.

256-band extended lapped transform (ELT) [8]. In the second stage, the first 32 sub-bands are left untouched while the others are merged together in groups of two, four, eight or sixteen with inverse ELTs in order to form wider bands with improved time resolution. Note that this structure differs from the more commonly used tree structure in which each stage further splits the outputs from the preceding stage. This type of structure was studied in [3].

The prototypes for the ELTs were all designed for 75% overlapping, and were optimized for minimization of the energy in the band $[\pi/M, \pi]$, where M is the number of bands of the transform. This resulted in similarly shaped prototype responses which, as noted by Cox [3], is a desirable property to achieve partial aliasing cancellation across the splitting and merging stages of the two-stage nonuniform structure. Delay units were added to each output line to normalize the overall delay and allow perfect reconstruction to take place on the dual synthesis structure. Table 1 shows the time and frequency resolution of each of the resulting 62 channels. The bandwidth is tabled both in Hertz and in the perceptually more significant Bark scale. The complete filter bank can be implemented with 21 additions and 12 multiplications per sample. The analysis-synthesis combination introduces a delay of 40.6 ms.

Bands	Δt (ms)	Δf (Hz)	Δz (Bark)
1–32	5.80	86	0.85–0.19
33–40	2.90	172	0.35–0.22
41–48	1.45	345	0.42–0.26
49–54	0.73	689	0.47–0.31
55–62	0.36	1378	0.54–0.23

Table 1: Sampling period (Δt) and bandwidth (Δf and Δz) of each band of the filter bank.

2.2 Quantization and Coding

The sub-band quantizers are mid-tread with 127 output levels distributed according to a logarithmic companding rule. The gain or step size Δ of the quantizers is allowed to vary from band to band and from sample to sample. For samples in the range -16Δ through 16Δ, the quantizer behaves like a uniform quantizer with step Δ. Larger amplitudes suffer proportionally coarser quantization. This provides a wide dynamic range and some intrinsic noise shaping.

Outputs from the quantizers are entropy-coded using an arithmetic code [9]. Since each quantizer presents a different probability distribution, the arithmetic coder maintains 62 distinct source models and dynamically switches between them as it processes each quantized sample. This effectively multiplexes the sub-band data without any overhead. The source models, implemented as lookup tables of cumulative counts for the 127 possible levels, are updated after processing each sample by increasing the appropriate counters. Regular rescaling of the tables avoids overflows and provides a means to "forget" distant past events and adapt to changing input statistics. During startup the tables are initialized with distributions measured from real music signals.

2.3 Perceptual Adaptation Algorithm

The adaptation algorithm, as in other perceptual coders, consists of a psychoacoustic model to estimate the amount of noise that can be added to the signal without causing perceptible distortion—the so-called *masking threshold* of the signal—and uses that information to set the quantization steps Δ. The estimation of the masking threshold follows a scheme similar to that of [5] and comprises four steps:

1. Time-domain smearing to take a running average of the energy in each band while modelling forward-masking at the same time. This is achieved simply by squaring each sample and passing it through a first-order recursive low-pass filter. The impulse responses of these filters are exponential decay sequences, and their time constants vary from band to band in order to reflect the dependence of forward masking on frequency. This step was inspired by the model described in [1].

2. Convolution with a frequency spreading function models masking phenomena in the frequency domain. The computation is somewhat involved because of the multiresolution nature of the spectral data and the use of backward adaptation.

3. The “tone-masking-noise” index is subtracted (in a logarithmic scale) from each band to produce the masking threshold estimate. In this first version, the “noise-masking-tone” index was not taken into account.

4. A final correction is introduced so that no masking level falls under the absolute threshold of hearing.

The quantization steps are determined so that the injected noise power falls just under the estimated masking threshold. The steps are then multiplied by a global parameter ϕ—known as the *quality factor*—to allow several quality/rate operating points: setting $\phi > 1$ enlarges the quantization steps, resulting in lower quality and reduced bit rate; setting $\phi < 1$ reduces the steps, providing a “safety margin” under the masking threshold at a cost of increased bit rate. This parameter can be transmitted just once for an entire track or it may be transmitted more regularly to provide a simple form of rate control with a very low overhead. The complete adaptation algorithm involves about 70 operations per sample.

A few remarks should be made about this algorithm. First, the spectral data used in the estimation are taken from the output of the filter bank described above, not from a different filter bank. This saves computation time. Also, the time-frequency structure of the transformed signal is kept through all computations thus avoiding potentially lossy conversions between domains. Finally, the masking threshold is estimated from previously quantized samples ($\hat{\mathbf{X}}$). This contrasts with existing perceptual coders which use the original samples ($\mathbf{X}$). The consequences of this will be discussed next.

3 BACKWARD ADAPTATION

A fundamental feature of the proposed coder is that it is backward adaptive. Although backward adaptation has been used in traditional coders such as CCITT G.722 [7], it is not common in *perceptual* audio coders. A comparison of these strategies seems in order.

The obvious implication of backward adaptiveness is the elimination of side information because adaptation parameters (the quantization steps, in this case) are locally generated in both the transmitter and the receiver. In a forward adaptive coder, on the contrary, side information can consume a considerable fraction of the full bit rate.[1] Since a backward adaptive coder does not have to quantize, encode, and multiplex this extra information, it has a simpler algorithm and the design procedure is straightforward—there is no need to find the best coding compromise between the main and side information channels. Another consequence is that adaptation can proceed on a sample-by-sample basis, unlike forward-adaptive systems which adapt block-by-block to minimize side information.

On the other hand, the receiver in a backward-adaptive system must implement the adaptation algorithm. This increases receiver complexity and can hamper upgradeability, especially in broadcast applications where the number of receivers is very large.

A major question that can be raised is whether the quantization noise introduced at the input of the adaptation process will disturb the masking threshold computation to such a point as to make it useless. Reflection on the behavior of the system under both normal and extreme input conditions convinced us that this would not be the case. To confirm it, we performed a simple simulation: using a representative test signal, we computed the “exact” threshold Ψ_0 from unquantized samples; we also computed the threshold Ψ_ϕ from samples quantized with quality factor ϕ; finally, we plotted histograms of the ratios Ψ_ϕ/Ψ_0. The results are shown in Figure 2 for $\phi = 4$. Even in this case, which represents very coarse quantization, almost 60% of all samples deviate no more than 0.25 dB from the “exact” threshold; only about 1% deviate more than 2 dB. If we consider that in a forward-adaptive system such as MPEG Layer I the adaptation parameters (scalefactors) are quantized with a resolution of 2 dB and that only one scalefactor is transmitted for each block of 12 sub-band samples [2], it seems that backward estimation might in fact result in closer tracking of the real masking threshold.

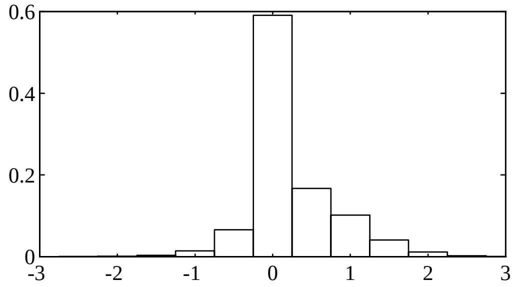

Figure 2: Histogram of differences (in dB) between thresholds computed from quantized (with $\phi = 4$) and unquantized samples.

4 PRELIMINARY RESULTS

The Backward-Adaptive Perceptual Audio Coding (BAPAC) system was simulated in software using a combination of MATLAB and C functions. The encoding and decoding algorithms can be implemented with about 110 to 120 arithmetic operations per sample each.

The performance of the coder was assessed over a set of seven music segments taken mainly from the EBU SQAM compact disc [4]. Four different coded versions of each segment were tested in a random order. Three of these were obtained with BAPAC at three decreasing

[1] In MPEG Layer I at 128 kb/s, for instance, this fraction can reach 20–25%.

quality levels: $\phi = 1$, $\phi = 2$ and $\phi = 3$. The other was produced with the Fraunhofer IIS shareware implementation of the MPEG Layer III algorithm at 64 kb/s and was included to permit comparisons of performance in identical testing conditions. For each version, the bit rate was measured and the coding quality was evaluated by listening tests.

The listening tests were based on the triple stimulus, hidden reference, double blind testing methodology. In each test, three signals were presented to a subject: R, X and Y. Signal R was always the original segment to be used as a reference. One of X and Y, chosen randomly, was the encoded/decoded version while the other was a repetition of the reference. The subject could play the signals repeatedly and in any order he or she pleased. The subject completed each test by grading each of X and Y with a score taken from the CCIR 5-point impairment scale.

Ten subjects participated in the listening tests. Each subject performed a pair of tests for each of the four coded versions of each music segment. The series of 56 tests took around one hour to complete.

Table 2 shows the mean score given to each of the four coded versions of the various music segments. (The mean score given to the hidden reference was very close to 5 in all cases.) The average of the scores (MOS) and the collective mean bit rate are also shown.

	BAPAC	BAPAC	BAPAC	Layer III
	$\phi = 1$	$\phi = 2$	$\phi = 3$	64 kb/s
Castanets	4.20	3.85	3.70	4.20
Harpsichord	4.30	3.45	2.55	4.30
Sarasate	4.60	3.75	2.40	4.75
Sting	4.75	4.65	4.30	4.70
Stravinsky	4.85	4.50	3.90	4.40
Suzanne	3.00	1.85	1.40	3.25
Violin	3.00	1.65	1.20	3.40
MOS	4.10	3.39	2.78	4.14
Rate	2.35	1.78	1.46	1.42

Table 2: Results of the listening tests: mean scores given to each coded version, Mean Opinion Score (MOS) and mean bit rate (in bits per sample).

At 2.35 bits per sample, BAPAC produced high quality output similar to that of the Layer III system. A comparable bit rate is only achieved at the cost of a lower quality. Notice that some music segments consistently got low scores across all coding schemes. This suggests that the psychoacoustic models need some improvements.

5 CONCLUSIONS

We have presented a new perceptual audio coder that uses backward-adaptive quantization. The coder is very simple and has a low computational complexity. The bit rate shows a predictable monotonic dependence on ϕ which should facilitate the implementation of a rate control mechanism. Informal subjective testing shows high-quality coding at a mean bit rate of around 2.35 bits per sample.

The effect of quantization noise on the backward estimation of the masking threshold was assessed, and it was found to be very small even when coarse quantization is employed. In fact, we argue that samplewise backward adaptation may track the real masking threshold closer than blockwise forward adaptation.

This and other interesting properties justify further work on backward-adaptive perceptual coding.

REFERENCES

[1] John G. Beerends and Jan A. Stemerdink. A perceptual audio quality measure based on a psychoacoustic sound representation. *Journal of the Audio Engineering Society*, 40(12):963–978, December 1992.

[2] Karlheinz Brandenburg, Gerhard Stoll, and al. The ISO/MPEG-Audio codec: A generic standard for coding of high quality digital audio. In *92nd AES-Convention*, Vienna, March 1992. Audio Engineering Society. preprint 3336.

[3] Richard V. Cox. The design of uniformly and nonuniformly spaced pseudoquadrature mirror filters. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, ASSP-34:1090–1096, October 1986.

[4] European Broadcasting Union, Brussels. *Sound Quality Assessment Material: Recordings for Subjective Tests*, April 1988.

[5] James D. Johnston. Transform coding of audio signals using perceptual noise criteria. *IEEE Journal on Selected Areas in Communications*, 6(2):314–323, February 1988.

[6] James D. Johnston and Karlheinz Brandenburg. Wideband coding—perceptual considerations for speech and music. In Sadaoki Furui and M. Mohan Sondhi, editors, *Advances in Speech Signal Processing*, chapter 4. Marcel Dekker, Inc., New York, 1991.

[7] Xavier Maitre. 7 kHz audio coding within 64 kbit/s. *IEEE Journal on Selected Areas in Communications*, 6(2):283–298, February 1988.

[8] Henrique S. Malvar. *Signal Processing with Lapped Transforms*. Artech House, Norwood, MA, 1992.

[9] Ian H. Witten, Radford M. Neal, and John G. Cleary. Arithmetic coding for data compression. *Communications of the Association for Computing Machinery*, 30(6):520–540, June 1987.