Rate-Distortion Allocation for Time-Frequency Dependent Audio Coding

Ricky Der[1] *Peter Kabal*[1] *Wai-Yip Chan*[2]

[1] Electrical & Computer Engineering
McGill University
Montreal, Quebec H3A 2A7

[2] Electrical & Computer Engineering
Queen's University
Kingston, Ontario K7L 3N6

Abstract

A stream coding framework is presented for solving the distortion-constrained time-frequency dependent quantization problem that naturally arises when overlapped time-frequency decompositions are used. The main contributions of this paper are (1) an efficient rate-distortion allocation algorithm for dependent quantization when the neighborhood of dependency is large; and (2) demonstration that a perceptual Excitation Distortion measure produces better coded audio quality than the conventional Noise-to-Mask Ratio measure.

1 Introduction

Consider a standard transform audio coder, operating in the constrained distortion mode. An input file is segmented into overlapping frames, and a linear transform is applied to give a time-frequency representation of the signal. For each frame, a set of quantizers $\mathbf{q} = \{q_i\}$ acting upon transform coefficients $\mathbf{x} = \{x_i\}$ is selected by an allocation algorithm to minimize some rate function R, subject to the constraint that the distortion $D(\mathbf{x}, \mathbf{q}(\mathbf{x}))$ is below a target threshold K. Finally, the time-domain version of the signal is reconstructed by inverting the transform and applying an overlap-add synthesis.

Figure 2 gives a block-diagram overview of the coding process just described, illustrated for the case of frame-length M and 50% overlap. Our key observation is that, given any distortion function D, there actually exist *two* distortion estimates that may be sensibly defined. Estimate 1 corresponds with the standard one, and is derived from a comparison between the original spectral coefficients $\mathbf{x}$ and the quantized spectral coefficients $\mathbf{q}(\mathbf{x})$. Estimate 2 is computed from the reference transform coefficients and a spectral analysis of the final time-domain reconstruction — *after* the overlap-add operation.

These two estimates are not, in general, the same, since the reconstructed frequency coefficients after synthesis are different from the quantized frequency coefficients. When D is perceptually motivated, then it is also clear that the "correct" estimate is the one obtained after time-domain addition, corresponding as it does to the actual signal upon which a listener performs perceptual processing. For instance, if D is the Noise-to-Mask Ratio (NMR), the noise factor should be computed as an end-to-end difference between reference and reconstructed spectra; similarly, if D is the dB-distance between two excitation patterns, as introduced in [2], then comparison must occur between the excitation pattern of the original signal, and the excitation pattern of the final, coded signal as presented to the ear. The intermediary quantized frequency spectrum $\mathbf{q}(\mathbf{x})$, alone, has little perceptual value, being merely a by-product of the particular signal decomposition.

If we agree to use the end-to-end distortion estimate 2 over the standard estimate 1, a number of important ramifications result. The most significant one is that the quantization problem becomes time-frequency dependent. Indeed, from Figure 2, it is readily seen that *each* of the M reconstructed frequency coefficients of frame n are functionally dependent not only on the M quantized transform coefficients in frame n, but the M quantized coefficients in both frame $n-1$ and $n+1$. Moreover, the range of dependency does not decrease with smaller overlap; independence occurs only in the case of zero overlap.

Other problems arise in relation to the constrained-distortion coding mode. While the end-to-end distortion estimate D' is correlated with the standard estimate $D(\mathbf{x}, \mathbf{q}(\mathbf{x}))$, and indeed $D = 0$ if and only if $D' = 0$, the two measures will not generally agree. For example, a bit allocation satisfying $D < K$ will not in general satisfy $D' < K$. Figure 1 provides one such example. Here, a coder using a 50% overlapped DFT decomposition obtains quantization parameters for a speech file such that the NMR distortion D before overlapped synthesis is below 1, for all frames — i.e. noise is below masking threshold. This threshold is violated for a number of frames, however, upon calculation of the end-to-end NMR D'.

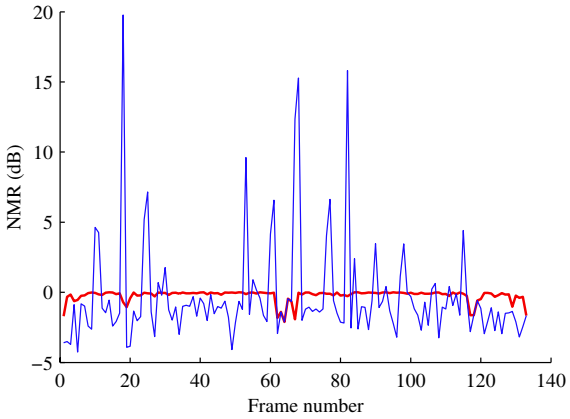

Fig. 1 Maximal noise-to-mask ratio as a function of frame number. Thick line: distortion estimate 1 (NMR before overlapped synthesis); thin line: distortion estimate 2 (NMR after synthesis). Masking threshold corresponds to the line 0 dB.

The conventional speech or audio coder is a "frame-by-frame" coder; namely, each frame is processed sequentially, and in as much that there is any dependence between frames, only past frames $n-1, n-2, \ldots$ can influence the coding decisions of the n-th frame. An implication of the time-dependent structure imposed by an end-to-end distortion measure is that constrained distortion coding, in the sense that $D_n < K$ for all frames n, is not generally

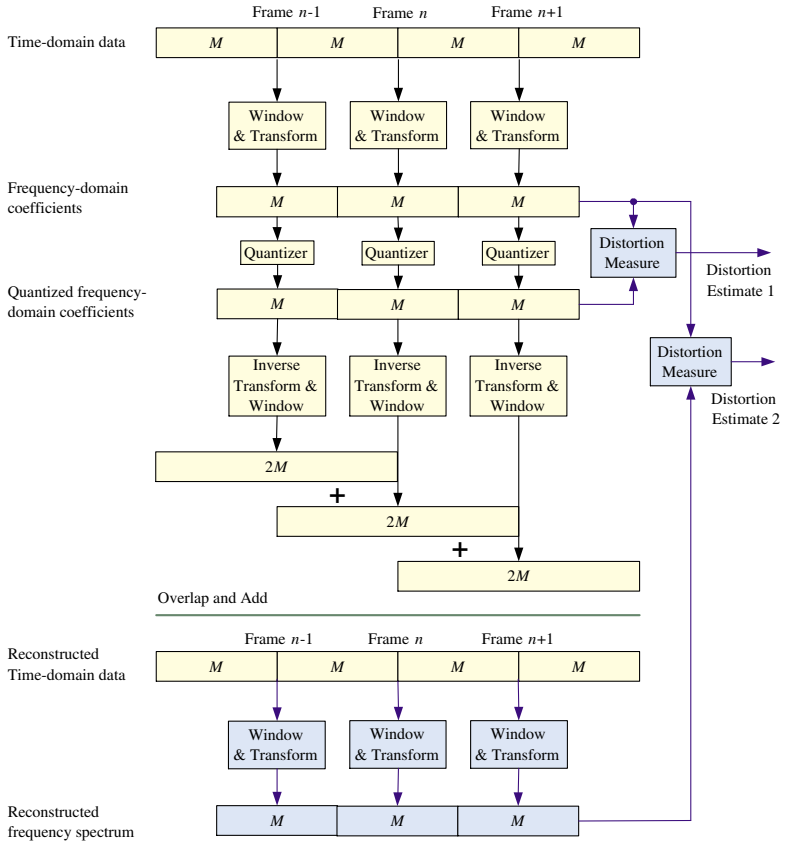

Fig. 2 Transform coding process with 50% frame overlap. Because of overlap-add reconstruction, the quantized frequency coefficients are not the same as the reconstructed frequency coefficients after synthesis.

possible with frame-by-frame coding. This is because there is no guarantee that, even if the first n frames are quantized to satisfy $D_k < K$, $k \leq n$, it is possible to achieve the distortion constraint for frame $n+1$, i.e. $D_{n+1} < K$ — even with lossless coding of the $(n+1)$-th frame.

We shall approach the problem by renouncing frame-by-frame coding and adopting a stream coding paradigm. That is, instead of processing time-frames in sequence, a rate-distortion optimization is performed over the entire audio file (or at least over segments of a file each containing some number of frames, depending upon delay constraints), in the time-frequency plane. Coding decisions are made simultaneously for a group of frames, instead of individually. The procedure will allow us to produce true, end-to-end D-constrained files.

2 Stream Coding: Formulation

Given an input signal segment S, we assume the existence of some (invertible) time-frequency transform producing the discrete time-frequency representation $x(t_i, \omega_j) \equiv x_{ij}$ for S, with $1 \leq i \leq n$ and $1 \leq j \leq m$. For every (i, j) transform coefficient, we associate a set of quantizers q_{ijb}, parameterized by bits b, and ordered such that $\lim_{b\to\infty} q_{ijb}(x) = x$. Moreover, we will write $b = b_{ij} \geq 0$ as a way of indicating the quantizer used in location (i, j), so that b_{ij} has the interpretation of being a bit-allocation. We define the rate R as the average bit-allocation over the entire segment:

$$R = \frac{1}{nm} \sum_{i=1}^{n} \sum_{j=1}^{m} b_{ij} \tag{1}$$

Let the vector $\tilde{\mathbf{x}}_i = \{\tilde{x}_{ij}\}_{j=1}^{m}$ denote the set of quantized transform coefficients, $\hat{\mathbf{x}}_i = \{\hat{x}_{ij}\}_{j=1}^{m}$ the set of reconstructed frequency coefficients *after* overlapped synthesis, both at frame index i, and let $d(\cdot,\cdot)$ be a distortion measure between two spectra. The *end-to-end distortion* d is

$$d(\mathbf{x}_i, \hat{\mathbf{x}}_i) = d(\mathbf{x}_i, \tilde{\mathbf{x}}_{i-1}, \tilde{\mathbf{x}}_i, \tilde{\mathbf{x}}_{i+1}) \tag{2}$$

where we have made clear the dependency of d on *three* frames of quantized coefficients. In line with the goal of distortion-constrained coding, we define D over S as the maximum end-to-end distortion over all time frames:

$$D = \max_{1\leq i\leq n} d(\mathbf{x}_i, \hat{\mathbf{x}}_i) \tag{3}$$

Now the rate-distortion optimization problem can be stated thusly: *Given a target threshold K, find a bit allocation $b_{ij} \geq 0$ such that $D < K$ with minimal rate R.*

2.1 Optimal and Incremental Approaches to Bit Allocation

If B is the maximum number of bits b that may be allotted to any quantizer, then an optimal, brute-force evaluation of D for all possible combinations of bit allocation is of complexity $O(B^{nm})$, which is of course unfeasible (nm will range in the thousands).

Less demanding yet still optimal/near-optimal search procedures typically involve dynamic programming. Most such methods assume that the distortion function D is additively separable in the sense that $D = \sum_i d_i(x_i, \tilde{x}_i)$, for some functions d_i; the requirement allows a Lagrangian relaxation, and has resulted in some very efficient algorithms, for instance that of [5]. Unfortunately, the function (3) is neither additive nor separable, the former because the requirement of constrained distortion induces a max-type definition, and the latter because the overlapped representation induces frame-dependencies.

A very general trellis-based approach to the min-max dependent quantization problem (in both rate and distortion constrained variants) has been formulated in [4]. While, in principle, that algorithm can be applied to the non-separable distortion (3); the large neighborhood of dependence in the time-frequency plane introduced by frame overlap makes such an implementation prohibitively complex. Indeed, the complexity of the algorithm in [4] is of order $O(nmB^N)$, where N is the cardinality of the region of dependence associated with each distortion point. For the problem at hand, one might typically have 20 bit-allocation bands, with a 3-frame time-dependency, so that $N = 3 \cdot 20 = 60$, which, while a significant improvement on exhaustive search, is still entirely unfeasible.

2.1.1 Sub-Optimal Incremental Algorithms

Given the enormous complexity of the methods presented in the previous section, we must forego rate-distortion optimality and make use of more heuristic procedures. One class of such methods are so-called "greedy" algorithms.

The standard greedy search approaches the rate-distortion problem in the following way: beginning with an initial bit-allocation $b_{ij} = 0$ for all i, j, the algorithm finds the (i, j) location for which the bit increase $b_{ij} = b_{ij} + 1$ results in a maximal decrease in distortion D. Computational complexity for this algorithm is upper-bounded by $O(B(nm)^2)$, which is large but feasible.

Unfortunately, while performing well for independent quantization problems, the greedy algorithm can fail to halt for dependent quantization problems [2]. In particular, allocation distributions can arise such that for *no* time-frequency location does the bit increase $b_{ij} = b_{ij} + v$ result in a decrease in distortion, for *any* positive integer v. This phenomenon is more or less the consequence of trying to minimize an irreducibly multivariate, non-separable distortion function by checking changes in the objective function only along the axial directions (1,0,0,...), (0,1,0,0,...), (0,0,1,0,0,...), etc. Indeed the very notion of separability implies a function that is "naturally" oriented along axial directions, and hence the greedy search tends to work well for separable distortion functions. The converse is that greedy search tends to perform poorly or not at all for non-separable functions, and in particular cannot be applied to the distortion function (3).

2.1.2 Forward-Backward Allocation

The idea of a "reverse" greedy algorithm may have first been proposed in [2], though the procedure was there formulated only for a distortion measure of specific type. However, the experimental results therein showed that reverse allocation could meet distortion targets at approximately 50% the rate of a suitably defined multi-coefficient (forward) greedy algorithm. We now provide a general formulation for a reverse-type incremental algorithm, applicable for the bit allocation of a wide range of distortion measures.

The basic idea is extremely simple and consists of the following: we first obtain, by any method, a bit allocation satisfying the distortion constraint without necessarily worrying about rate-optimality. This bit allocation is used as an initialization to a de-allocation procedure, which successively *removes* bits until the distortion constraint K is breached. More specifically, at each iteration a bit is removed from the location (i, j) which results in the smallest updated distortion. The process continues until the condition $D < K$ is first violated; the last allocation for which the target is achieved is retained.

There are a variety of ways to obtain the initializing bit distribution. In [2] an initialization was found by utilising the fact that, under certain conditions, there exists a simpler *separable* function D' that can over-bound the non-separable distortion measure D. A standard forward greedy algorithm applied to the resulting *independent* quantization problem induced by D' obtained the desired initialization. With general distortion measures, one does not always have the luxury of such a structure. There always exists one crude estimate, however: simply set $b_{ij} = B, \forall i, j$ with B sufficiently large. A more intelligent design is to use some type of forward multi-allocation procedure:

Algorithm 3.1 (Forward Search)
Given a distortion target K,

1. Initialise $b_{ij} = 0, \quad \forall i, j$
2. Compute D using bit allocation b_{ij} for all i, j.
3. If $D < K$, exit. If not, locate $i^* = \arg\max_i d_i$.
4. Find the set $\mathcal{P}$ of all indices (i, j) such that d_{i^*} is a function of q_{ij}.
5. Increment $b_{ij} = b_{ij} + 1, \quad \forall (i, j) \in \mathcal{P}$.
6. Go to Step 2.

Step 4 of the algorithm involves finding the dependency neighborhood for d_i; when the coding process involves an overlapped representation with overlap O such that $0\% < O \leq 50\%$, this neighborhood is always given by $\mathcal{P} = \{(i, j) : i^* - 1 \leq i \leq i^* + 1\}$, i.e. a 3-frame dependency. The above initialization algorithm is guaranteed to converge as long as the distortion function d of (3) satisfies the continuity requirement $\lim_{x \to y} d(x, y) = 0$ and the upper bound B on the number of bits is chosen sufficiently large.

We now give a formal description of the de-allocation phase, in the process introducing a complexity-scaling partition that allows the user to trade-off time-complexity for rate-distortion optimality.

Algorithm 3.2 (Backward Search)
We assume that some initialization process has already arrived at a bit allocation b_{ij} satisfying $D < K$. Begin by partitioning the set of time-frequency locations into disjoint sets $A_k, 1 \leq k \leq L$. Define the function $T_{b_{ij}}$ as the updated distortion D when the bit allocation b_{ij} is replaced with $b_{ij} - 1$ (only in the (i, j) location) if $b_{ij} \geq 1$ *and* when the updated distortion satisfies $D \leq K$; define $T_{b_{ij}} = \infty$ otherwise.

1. FOR $k = 1..L$
   Compute $T_{b_{ij}}$ for all $(i, j) \in A_k$.
   Find $(i^*, j^*) = \arg\min_{i,j} T_{b_{ij}}$.

Set $b_{i^*j^*} = b_{i^*j^*} - 1$ if $T_{b_{i^*j^*}} < \infty$.

END

2. If b_{ij} changed for at least one location (i, j), go to Step 1. If not, exit.

We shall call the concatenation of Algorithm 3.1 and 3.2 by the name "Forward-Backward Greedy Algorithm". It is a rather general procedure which can be applied to any constrained-distortion coding problem, without any separability requirements on the distortion function. The algorithm is guaranteed to halt in a finite number of iterations, for any distortion target K, unlike the forward greedy search.

The partitioning of the time-frequency locations into sets A_k provides a way of trading off computational effort and rate-distortion optimality. The main point to the partitioning is in reducing the number of distortion evaluations required before removing one bit. For instance, if $L = 1$ and one takes A_1 the entire set of indices over the file segment S, the algorithm must test every location in the file before de-allocating a single bit. On the other hand, by choosing a fine partition so that $A = \max_k |A_k|$ is small, the algorithm is required at most to evaluate the distortion A times before attempting to remove a bit. Thus the latter algorithm is potentially nm/A times faster than the former. However, because the size of the search-field is reduced, it will tend to remove relatively fewer bits before exceeding the distortion constraint.

The complexity of the forward phase is no more than $O(Bnm)$, while the complexity for de-allocation is no more than $O(B(nm)^2)$. The total complexity of the forward-backward greedy algorithm is therefore no more than that $O(B(nm)^2)$ — the same as the standard greedy algorithm. Depending on how well localised the neighborhood of dependence is, a careful implementation of the partitioning $\{A_k\}$ can obtain $O(ABnm)$, so that it is even possible to have linear complexity $O(Bnm)$ in some cases by using the finest partition A_k possible: each A_k containing exactly one time-frequency location.

3 Experimental Results

The stream coding framework can be applied to gauge the relative performance of any two distortion functions in the audio coding context. We shall compare coding results for two different distortion measures (d in (2)): (1) the standard NMR measure, and (2) a distortion measure posed entirely in the perceptual domain, following [1], [2]. For the latter, define the Excitation Distortion (ED) to be the maximum dB-difference between reference and coded excitation patterns; in symbols:

$$d_{\mathrm{ED}} = \max_j \left| 10 \log_{10} \left(E_j / \hat{E}_j \right) \right| \tag{4}$$

where E_j and $\hat{E}_j$ are the excitation powers respectively of the reference signal $\mathbf{x}$ and the reconstructed spectrum $\hat{\mathbf{x}}$, at frequency j.

We use a transform coder with uniform scalar quantizers in the discrete Fourier domain, frames overlapped at 50%, and 18 sub-bands of unit critical bandwidth partitioning the frequency interval $[0, 4000]$ Hz. Each sub-band is parameterized by a step-size δ, and a single bit quantum is associated with an increase or decrease of quantizer step-size by the factor 0.9. The Glasberg-Moore excitation model [3], with small variations, is used to compute high-resolution excitation and masking patterns (in conjunction with an appropriate masking offset) required in the calculation of D_{ED} and D_{NMR}, respectively.

Since the excitation pattern is a function only of the power spectrum of a signal, the distortion measure (4) is phase-blind. As a consequence, we restrict ourselves in this experiment by examining how each distortion measure evaluates magnitude-distortion only, placing the quantizers in the magnitude Fourier domain and assuming perfect phase coding. In each case, the Forward-Backward Greedy algorithm is applied to find, given a target K, the step-size parameters necessary to drive $D < K$. Given the definition of the global file distortion in (3) as the maximum distortion over all time frames, the coded files will be such that the end-to-end NMR and ED patterns are no greater than K, in both time and frequency.

To compare files coded to different distortion targets, one may use the average bit allocation measure of (1). More realistically, given the coded file, we compute the empirical entropy of the quantized levels for each frequency bin; the resulting average entropy over all bins will be used as our measure of rate, which emulates the rate of a coder in which the levels are Huffman coded.

For each audio segment, three ED targets were fixed at ED=3.83, ED=2.85, and ED=2.23, the forward-backward rate-distortion allocation performed, and the resulting empirical entropies recorded. These three target levels roughly correspond to coding levels ranging from low to high quality. The NMR distortion target K was then tuned so that the resulting empirical entropy of the NMR-coded files matched those of the ED-coded files. The following table gives an example of the distortion targets necessary to produce matched-entropy pairs in the case of a trumpet file.

Table 1 Distortion Targets for Trumpet File

Entropy (bits/coeff.)	ED Target	NMR Target (dB)
0.33	3.83	6.90
0.42	2.85	5.56
0.56	2.23	3.80

Five different audio selections were tested: (1) male speech, (2) vocal quartet, (3) trumpet, (4) orchestra, (5) organ, for a total of fifteen constrained-distortion, matched-rate pairs. An informal subjective listening test revealed that, for all files, and at all the prescribed rates, the quality of ED coding was higher than that of NMR. The quality improvement of ED over NMR increased as the rate dropped, corroborating the observations contained in [1] concerning the inadequacies of NMR for non-transparent-quality audio coding.

References

[1] R. Der, P. Kabal, W.-Y. Chan, "Towards a New Perceptual Coding Paradigm for Audio Signals", *Proc. ICASSP*, 2003.

[2] R. Der, P. Kabal, W.-Y. Chan "Bit Allocation Algorithms for Frequency and Time Spread Perceptual Coding", *Proc. ICASSP*, 2004.

[3] B. Glasberg, B. Moore. "Derivation of auditory filter shapes from notched-noise data", *Hearing Research*, 47, pp. 103–138, 1990.

[4] G. M. Schuster, G. Melnikov, A. K. Katsaggelos. " A review of the minimum maximum criterion for optimal bit allocation among dependent quantizers". *IEEE Trans. on Multimedia*, vol. 1, pp. 3–17, 1999.

[5] Y. Shoham, A. Gersho. "Efficient Bit Allocation for an Arbitrary Set of Quantizers," *IEEE Trans. Acoustics, Speech, Signal Processing*, vol. 36, pp. 1445–1453, 1988.