R/D Optimal Data Hiding

Paolo Prandoni[a] and Martin Vetterli[b]

[a]École Polytechnique Fédérale de Lausanne,
CH - 1015 Lausanne, Switzerland.

[b]Electrical Engineering and Computer Science Department,
University of California at Berkeley, Berkeley CA 94720
and École Polytechnique Fédérale de Lausanne,
CH - 1015 Lausanne, Switzerland.

ABSTRACT

State of the art audio coders exploit the redundancy in audio signals by shaping their quantization noise below the signal's masking curve, which is a signal-dependent threshold of audibility. This framework can be extended to the context of data hiding, where the data play the role of noise. To minimize audio distortion the data power should be closely adapted to the time-varying masking curve; each power switch, however, reduces the net throughput via its associated side information. This tradeoff can be cast in a rate/distortion framework: the optimal sequence of power levels and the optimal sequence of power switchpoints is found by minimizing a Lagrange cost functional relating perceptual audio distortion to throughput, and is implemented as a linear-time trellis search. For 16-bit, 44.1 KHz PCM stereo signals, a net throughput of the order of 30 kbits/sec can usually be achieved at no perceptual cost in an algorithmically efficient way.

Keywords: Steganography, data hiding, rate-distortion optimality, psychoacoustic modeling, MPEG audio

1. INTRODUCTION

Data hiding, or steganography, is concerned with embedding data into a "host" message (the *cover message*) in a way which is undetectable to an external observer; in this sense, it differs from cryptography (although the hidden data can itself be encrypted) since the cover message remains unaltered and entirely meaningful from a perceptual point of view. In recent years, there has been a lot of interest in steganographic techniques in connection with watermarking for copyright protection of audio and video material[1]; in these cases the goal is to embed a marker or a digital signature in proprietary material in order to track potential copyright infringements. Clearly, this requires that the data hiding method be extremely resilient to common data processing techniques such as compression, filtering, or cropping, which are viewed by the copyright owner as "attacks" on the watermark. Intuitively, it is easy to see that the capacity of the steganographic channel and its robustness to attacks are inversely related to each other; in watermarking applications this should not be a fundamental limitation since the amount of data to be hidden is relatively modest, yet it is still an active area of research to find a universally robust watermarking technique.[2]

There is however another situation which can make good use of steganographic techniques, and that is the case when different data sources share the same, fixed physical channel. Multiple access channels are extremely well studied in the case of mobile communications, for instance, and we will see that some techniques that are commonly employed in dealing with fading or nonflat multiple access links can be of help in designing a data hiding system. However, the fundamental difference between such communication protocols and a data hiding scenario is that, in the latter, one particular data source takes priority with respect to the others; furthermore, some of the receivers might not have the ability to extract the hidden message, yet the composite data stream should be entirely equivalent (in some perceptual way) to the prioritary data stream alone. Everyday examples of such a scenario are for instance a color TV signal, which can be reproduced by older black and white sets; or teletext messages, inserted at the blanking intervals between frames in a way that does not affect the standard TV decoding process. Both these methods work by exploiting gaps in a decoding protocol; more sophisticated techniques can exploit the *perceptual* gaps in the human visual or auditory system instead, much in the same fashion as standard compression algorithms

Further author information: send correspondence to P. Prandoni:
email: `prandoni@de.epfl.ch`, tel: +41 21 693 5629, fax: +41 21 693 4312.

do. The common strategy, in all cases, is to maximize the amount of information that can be hidden in a given signal while preserving its perceptual properties. This maximization of throughput necessarily implies a very low resilience to signal manipulation, which sets us apart from watermarking techniques; at any rate, the goal is not secrecy or robustness towards attacks but the expansion (or better, the splitting) of the communication channel throughput in a "backward dependent" way.

In this paper we will present a data hiding technique for digital PCM audio, based on the masking properties of the human auditory system (HAS). The setting might be that of a compact disc (CD), for instance, to which we want to tag additional data such as lyrics or pictures without altering either the format of the support (the "Red Book" protocol) or its audio capacity (74 minutes approximately); the resulting CD will therefore be entirely compatible with standard CD players, will play to the maximum allowable duration, and yet will allow more sophisticated equipment to retrieve the hidden data. Some commercial examples which implement such an idea (albeit in minimal form) to improve audio quality are already available[3]; we will show that, by casting the data hiding technique in a rate/distortion framework, the optimal compromise between throughput and perceptual distortion can be achieved for all types of audio signal and for any data hiding protocol.

2. PERCEPTUAL AUDIO CODING

State of the art audio coding algorithms such as MPEG or Dolby's AC3 can provide acoustically transparent compression ratios of the order of 4:1 to 6:1 by cleverly exploiting the *masking phenomena* inherent in human hearing.[4,5] Simply stated, masking occurs when weaker signal components are made inaudible by the presence of louder components; such weaker components are said to lie below the *masking curve* of the signal. Compression algorithms quantize the signal so that the bulk of the overall quantization noise is hidden below the masking curve, and is therefore inaudible. In many audio steganography techniques, the hidden data source can be modeled by an equivalent additive noise generator; by shaping the power spectrum of such noise so that it lies below the masking curve, one could in theory achieve a perfectly transparent embedding. Perceptual hiding according to this general line has been indeed proposed in the context of audio watermarking[6]; in this paper we are however concerned with a tagging system for which maximum throughput and perfect extraction at the receiving end are the fundamental targets, which requires a finer exploitation of the masking properties of a signal and a more sophisticated hiding strategy.

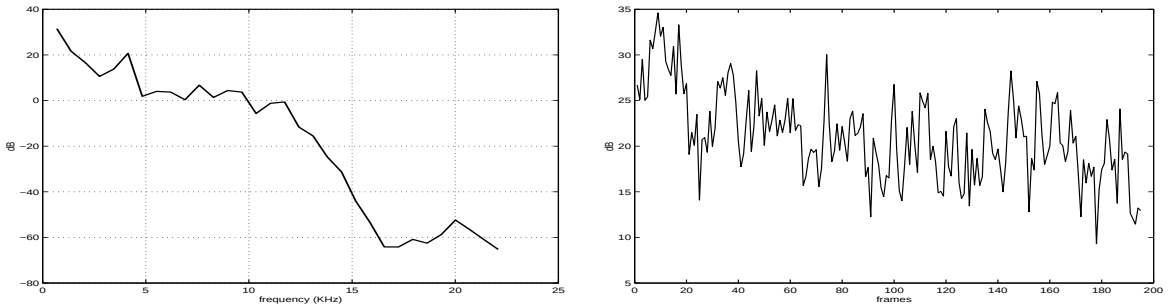
2.1. Psychoacoustic Modeling

In this section we will briefly review the fundamentals of psychoacoustic modeling from an operative, computational point of view; this simplified approach neglects some finer aspects of the hearing mechanism such as temporal masking but proves entirely adequate for the task at hand. From a physiological perspective, acoustic masking is a consequence of nonlinear processing in the inner ear; frequency-selective areas in the cochlea, called *critical bands*, exhibit a saturation characteristic whereby a loud sound component (the *masker*) renders inaudible all the weaker components in the same critical band which lie below a certain power threshold. The threshold is in fact a function of frequency and it decays rather rapidly in an interval centered around the masker; its magnitude depends on the critical band number, on the power of the masker, and on the type of masker (whether an isolated spectral line or a noise-like component).

A psychoacoustic model tries to algorithmically reproduce these hearing mechanisms in order to obtain an estimate of the effectively inaudible portion of a given audio signal. Much study has been devoted to the characterization of masking functions,[4] and different psychoacoustic models differ essentially in the parametrization of their shapes and decay rates. In the following we will rely on a simple model in which masking functions are approximated by piecewise linear functions in the log-power, bark frequency domain[7]; a unit of one bark corresponds to the width of a critical band and, since the width of successive critical bands increases by approximately a third of an octave, there is essentially a logarithmic mapping between bark scale and linear frequency scale. For complex tones the masking phenomenon is distributed across the entire audible spectrum, since each spectral component originates a local masking function; the sum of all masking thresholds for all components across the signal's bandwidth yields the overall *masking curve*.

An algorithmic process estimating the masking curve for a given signal can be illustrated with reference to the MPEG standard psychoacoustic model 1, which will also be used in section 5. It comprises the following steps[8]:

- *Computation of the power spectrum;* this is performed by a short time Fourier transform analysis.



(a) (b)

Figure 1. (a) Masking curve for one frame; (b) time varying masking threshold for a frequency around 1 KHz.

- *Separation of tonal and non-tonal components;* since the masking power of isolated spectral lines is less than that of noise-like spectral components, the former are separated from the latter.
- *Computation of the individual masking thresholds;* this step is accomplished by convolving each spectral component by the appropriate (tonal or non-tonal) masking function.
- *Computation of the global masking curve;* the masking curve is obtained as the sum of the individual masking thresholds.

An additional step is required to map the masking curve thus obtained back to the linear frequency domain; the final result will be denoted by the time-frequency function $\Theta(t, f)$. For a fixed time instant t_0, this function is one realization of the masking curve; for a fixed frequency f_0, it represents the evolution of the masking threshold over time for the given frequency. Figures 1-(a) and (b) display typical instances of a masking curve and of a time-varying masking threshold (at low frequency) respectively; here and in the following, the audio data are obtained from a string quartet CD.[9] The time-frequency resolution of the psychoacoustic model depends on the underlying short time spectral analysis used to compute the power spectrum, and it is a tradeoff between accuracy of the tonal and non-tonal representation and responsiveness to fast signal transients. In MPEG layer II, for an input sampling rate of 44.1 KHz, the psychoacoustic model produces a masking curve every 26 ms.; in the following, we will call such an analysis interval as a *frame*.

2.2. Compression

As stated previously, the fundamental coding step of a perceptual audio coder is to quantize the signal separately over different subbands so that the quantization noise level for each subband is less than the minimum value of the masking curve over the subband. This minimum value is generally referred to as the masking threshold for the subband. Without going into further details, the most important point is to remark that, due to this signal dependent quantization, part of the total bit budget for the coder must be spent to inform the decoder of the different subband levels.

3. PERCEPTUAL DATA HIDING

Consider an audio signal $s(t)$, $0 \leq t \leq t_0$ bandlimited to f_N Hz: we have seen that the masking curve computation yields a function $\Theta(t, f)$, of which two typical time and frequency "slices" are displayed in Figure 1. This function represents a two dimensional power constraint for the transmission power which, in principle, should lie below the threshold. In order to fulfill this constraint, we choose to adopt a separable approach by discretizing both the time and frequency axis; while this is admittedly not the most sophisticated way to achieve a time-frequency power shaping, it is in line with the current techniques which implement the psychoacoustic analysis model and allows for an easy integration of our transmission scheme in compression engines such as MPEG. In particular, spectral power shaping in frequency for a given time interval is accomplished by multicarrier modulation, which implements a discretized approximation of the reverse waterfilling algorithm.[10] The key point is that, if the entire frequency interval is split into M bands by a M-ary filterbank with perfect or almost perfect reconstruction capabilities, then *power shaping along the time axis can proceed independently for each filterbank subchannel.*

3.1. Discretization

Assume the (positive) frequency axis $[0, f_N]$ is subdivided into M disjoint bands $[f_k, f_{k+1})$, $f_0 = 0, f_{M-1} = f_N$: at any time instant t we can obtain M corresponding masking thresholds values by selecting the minimum value of $\Theta(t, f)$ in each subband. Also, the psychoacoustic model computes a full masking curve via short-time windowing over frames K samples apart; for a sampling frequency of f_s Hz, there are a total of $N = f_s t_0 / K$ frames, with $\Theta(t, f)$ constant for $hK/f_s \leq t < (h+1)K/f_s$, $h = 0, \ldots, N-1$. The completely discretized set of masking thresholds can therefore be denoted by the (discrete) function $\theta_k(h)$ for $k = 1, \ldots, M-1$ and $h = 0, \ldots, N-1$:

$$\theta_k(h) = \min_{f_k \leq f < f_{k+1}} \{\Theta(hK/f_s, f)\}. \tag{1}$$

Using the same discretization grid, we would like to obtain a sequence of signal energy levels relative to the time-frequency tiles over which the masking thresholds are computed; these can be determined as:

$$\sigma_k(h) = \int_{hK/f_s}^{(h+1)K/f_s} |s(t) * g_k(t)|^2 dt \tag{2}$$

where the Fourier transform of $g_k(t)$ is the indicator function for the interval $[f_k, f_{k+1}]$; in practice, these energy levels can be computed as the average energies over a time interval of one frame at the outputs of a M-band filterbank.

As for the data tagging scheme, we assume we have M independent "transmitters" which operate over the $[f_k, f_{k+1})$ bands; we also assume the power of the transmitters can be arbitrarily varied across the K-point intervals from a minimum level l_1 to a maximum level l_Q. In the rest of the analysis, the actual transmission method is irrelevant as long as it can be modeled by a modulated, bandlimited additive noise source of equivalent power; we will also assume that the number of bits which can be successfully transmitted to the decoder each frame is a given function $r(l)$ of the transmission level l only.

3.2. Data Hiding and Side Information

At first, one might think to use a sequence of transmission power levels which lies just below the sequence of masking thresholds for each subchannel; the receiver would re-determine the sequence of thresholds via a psychoacoustic model identical to the transmitter and decode the data. Unfortunately, this approach is not viable for two reasons. First, the masking curve computation for the audio signal and for the signal after the data has been embedded generally yields different results; this means that the decoder cannot recover the correct sequence of transmission levels, and therefore the data themselves. In addition, it might happen that the throughput requirements are stringent but cannot be fulfilled by a sequence of power levels strictly below the thresholds; one might then decide to forgo the absolute undetectability constraint and use a sequence of levels which sometimes exceeds the masking thresholds. For both reasons, it is necessary to inform the receiver of the actual sequence of power levels, and this can only be done via side information. The situation is akin to perceptual compression where, in order to exploit the psychoacoustic gaps in the signal, side information about these gaps had to be supplied.

In our transmission scheme we inform the receiver only when the transmission power level changes in time; the amount of side information is therefore proportional to the number of power level switches. Since side information necessarily uses up part of the overall throughput, the objective is to devise a transmission sequence that maximizes throughput with a low description cost. At the same time, fewer power switches mean that we will overshoot the quickly varying masking threshold requirements more often. A tradeoff between throughput and perceptual distortion is now apparent: the goal is to determine the optimal transmission sequence with respect to these two parameters. The problem can be effectively cast in a rate/distortion framework, where the maximization of the "rate" (the throughput) must be weighed against the corresponding increase in perceptual distortion. Efficient techniques for unequal resource allocation have been developed in the context of quantization and coding[11] and in the following we will make full use of their fundamental concepts.

4. RATE/DISTORTION FORMULATION

For a single discretized subchannel the sequence of masking thresholds and signal levels for the time interval $[0, N-1]$ can be expressed in vector form as $\boldsymbol{\theta} = \{\theta(0), \ldots, \theta(N-1)\}$ and $\boldsymbol{\sigma} = \{\sigma(0), \ldots, \sigma(N-1)\}$ (the channel subscript is inessential and is dropped throughout this section). Call a *power allocation* a sequence of N transmission power

levels $\boldsymbol{p} = \{p(0), \ldots, p(N-1)\}$, $l_1 \leq p(i) \leq l_Q$ and let P be the set of all possible allocations over $[0, N-1]$; it is clearly $|P| = Q^N$. Call $D(\boldsymbol{\theta}, \boldsymbol{\sigma}, \boldsymbol{p})$ the distortion we incur in if we use the sequence of power levels $\boldsymbol{p}$ to transmit data over the subchannel; the corresponding net throughput will be denoted by $R(\boldsymbol{p})$. The goal is to maximize throughput while keeping the associated distortion below an acceptable minimum level D_0; this can be expressed as a constrained maximization problem:

$$\begin{cases} \max\limits_{\boldsymbol{p} \in P} \{R(\boldsymbol{p})\} \\ \\ D(\boldsymbol{\theta}, \boldsymbol{\sigma}, \boldsymbol{p}) \leq D_0 \end{cases} \tag{3}$$

A direct maximization of (3) is obviously prohibitive computationally due to the cardinality of P for even moderate values of N. The remainder of this section will show how to reformulate (3) as an equivalent unconstrained problem and will discuss an efficient implementation of the resulting algorithm based on dynamic programming.

4.1. Efficient formulation

First of all, let us consider in detail the structure of rate and distortion. The measure we choose for the distortion is the sum of the segmental perceptual noise to signal ratios (NSR), since the hidden data in each subband can be represented as an equivalent narrowband noise source. A fundamental requirement is that the power of the noise be less than the power of the signal; then, if the noise level is below the masking threshold, the perceptual distortion is zero, otherwise its power is equivalent to the power of the transmitted data offset by the level of the threshold itself. For a single frame we can therefore write:

$$d(\theta(n), \sigma(n), p(n)) = \begin{cases} +\infty & \text{if } p(n) \geq \sigma(n) \\ \\ 0 & \text{if } p(n) \leq \theta(n) \leq \sigma(n) \\ \\ \dfrac{p(n) - \theta(n)}{\sigma(n)} & \text{if } \theta(n) < p(n) < \sigma(n) \end{cases} \tag{4}$$

and $D(\boldsymbol{\theta}, \boldsymbol{\sigma}, \boldsymbol{p}) = \sum_n d(\theta(n), \sigma(n), p(n))$. As for the rate, we can initially write

$$R(\boldsymbol{p}) = \sum_{n=0}^{N-1} [r(p(n)) - c(p(n), p(n-1))] \tag{5}$$

with $p(-1) = l_1$ by convention and where $c(p(n), p(n-1))$ is the cost (in bits) of encoding the side information relative to a power level switch in transmission between one frame and the next; note that, according to our signaling scheme, $c(p, p) = 0$ since only transitions between different levels need be notified. We will also assume that the cost of side information is always less than or equal to the minimum achievable frame rate for all frames and all models, so that all the terms in (5) are non negative. Assume now that the original (digital) audio channel allows for a maximum capacity of r_M bits per frame*; our goal is to split this capacity between an audio data channel proper and a tagged data channel so as to maximize the throughput of the latter. We can therefore introduce the following auxiliary quantity

$$\tilde{R}(\boldsymbol{p}) = N r_M - R(\boldsymbol{p}) \tag{6}$$

which represents the portion of the total capacity *not* devoted to the hidden data. It is easy to see that using $\tilde{R}$, the problem in (3) is equivalent to:

$$\begin{cases} \min\limits_{\boldsymbol{p} \in P} \{\tilde{R}(\boldsymbol{p})\} \\ \\ D(\boldsymbol{\theta}, \boldsymbol{\sigma}, \boldsymbol{p}) \leq D_0 \end{cases} \tag{7}$$

The next step towards an efficient solution involves reformulating (7) as an equivalent unconstrained problem. Note that each allocation $\boldsymbol{p}$ indexes a point on the $\{D, \tilde{R}\}$ plane as in Figure 2-(a); if we restrict the search for the

*For instance: for a frame length of 36 samples at 16 bits per sample, $r_M = 576$ bits

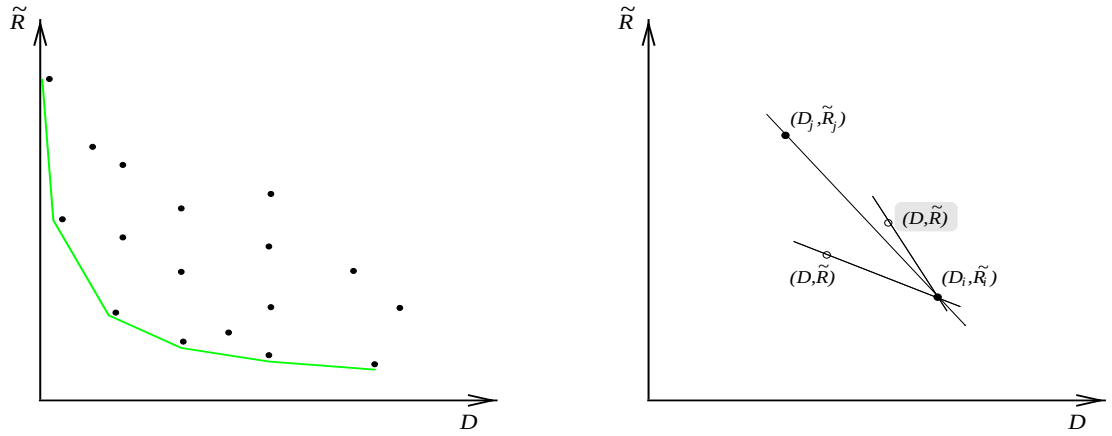


Figure 2. (a) $(D, \tilde{R})$ solution points and convex hull; (b) convexity of the hull.

solutions to (7) to the convex hull of the set of all $(D, \tilde{R})$ pairs, we can make use of Lagrange multipliers. Define a functional $J(\lambda) = \tilde{R}(\boldsymbol{p}) + \lambda D(\boldsymbol{\theta}, \boldsymbol{\sigma}, \boldsymbol{p})$; if, for a given λ,

$$\boldsymbol{p}^* = \arg\min_{\boldsymbol{p}\in P}\{J(\lambda)\} \tag{8}$$

then $\boldsymbol{p}^*$ (star superscripts denote optimality) indexes a point on the convex hull which solves the problem:

$$\begin{cases} \min\limits_{\boldsymbol{p}\in P}\{\tilde{R}(\boldsymbol{p})\} \\ \\ D(\boldsymbol{\theta}, \boldsymbol{\sigma}, \boldsymbol{p}) \leq D(\boldsymbol{\theta}, \boldsymbol{\sigma}, \boldsymbol{p}^*). \end{cases} \tag{9}$$

In other words, for each value of λ we also have the solution to an equivalent constrained problem where the distortion constraint is given by $D(\boldsymbol{\theta}, \boldsymbol{\sigma}, \boldsymbol{p}^*)$. An iteration over λ must be performed until this constraint coincides with D_0; luckily the overall distortion is a monotonically nonincreasing function of λ,[11] so that the optimal value can be found with a fast bisection search.[12]

It may help the intuition to show why the set U of $(D, \tilde{R})$ pairs solving (8) indeed defines a convex line on the "R/D" plane. Given any two solutions $(D_i, \tilde{R}_i)$ and $(D_j, \tilde{R}_j)$, $D_i > D_j$, the line in the R/D plane connecting them has an (absolute) slope $\gamma = (\tilde{R}_j - \tilde{R}_i)/(D_i - D_j)$. Convexity requires that all solutions $(D, \tilde{R})$ such that $D_i \leq D \leq D_j$ lie below this line. Suppose this was not the case for a solution $(D, \tilde{R}) \in U$; in terms of slopes connecting $(D, \tilde{R})$ to $(D_i, \tilde{R}_i)$ and to $(D_j, \tilde{R}_j)$ this implies

$$\frac{\tilde{R}_j - R}{D - D_j} < \gamma < \frac{\tilde{R} - \tilde{R}_i}{D_i - D}. \tag{10}$$

(see Figure 2-(b)). The $(D, \tilde{R})$ pair is by hypothesis a solution to (8) for a given λ; however, if $\lambda < \gamma$, (10) implies $\tilde{R}_i + \lambda D_i < \tilde{R} + \lambda D$ and we have a contradiction; otherwise, if $\lambda \geq \gamma$, we have again $\tilde{R}_j + \lambda D_j \leq \tilde{R} + \lambda D$; therefore $(D, \tilde{R}) \notin U$. This holds for all elements of U and, for a dense solution set, it provides an intuitive meaning to the value of λ in (8) as the derivative of the convex hull at the relative solution point.

Even in its unconstrained form, the minimization in (8) would still require $O(Q^N)$ comparisons; the fundamental simplification stems from rewriting (8) as the search for:

$$J^*(\lambda) = \min_{\boldsymbol{p}\in P}\{\sum_{n=0}^{N-1} j(n; p(n), p(n-1); \lambda)\} \tag{11}$$

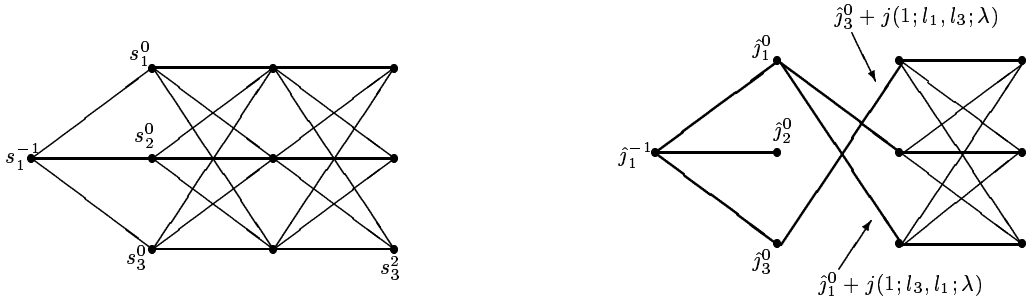


Figure 3. (a) Trellis structure; (b) metric computation and path population.

with

$$j(n; p(n), p(n-1); \lambda) = r_M - r(p(n)) + c(p(n), p(n-1)) + \lambda d(\theta(n), \sigma(n), p(n)). \tag{12}$$

Although all the terms in the summation are non negative, they are not independent because side information; the minimization, therefore, cannot be carried out term by term as usual.[11] However, suppose we know that the k-th element of the optimal allocation $\boldsymbol{p}^*$ is $p^*(k)$; since the dependence amongst terms extends only one step backward we can write:

$$\begin{aligned} J^*_{[0,N-1]}(\lambda) \quad = \quad & \min_{\{p(0),\ldots,p(k-1)\}} \{J_{[0,k-1]}(\lambda) + c(p(k-1), p^*(k))\} + \\ & + r_M - r(p^*(k)) + \lambda d(\theta(k), \sigma(k), p^*(k)) + \\ & + \min_{\{p(k+1),\ldots,p(N-1)\}} \{J_{[k+1,N-1]}(\lambda)\} \end{aligned} \tag{13}$$

(where subscripts for $J(\lambda)$ indicate the signal range for the minimization). The three terms on the right hand side of the equation are now independent; this means that, if $p^*(k)$ is an optimal allocation choice, the partial allocation up to k is independent of further data and that, at each step k, the optimal allocation can be built incrementally by formulating an optimality hypothesis over all transmission levels $p(k)$. Equation (13) is a particular case of the optimality principle in dynamic programming[13] and suggests an efficient way to implement the allocation process as a trellis search.

4.2. Trellis Algorithm

The minimization in (13) can be deployed on a Q-state, $(N+1)$-stage trellis as in Figure 3; a state will be denoted as s_k^n, where the superscript identifies the time index from 0 to $N-1$ and the subscript is a transmission power level index from 1 to Q; the same notation is used for the path metrics $\hat{\jmath}_k^n$.

A) Initialization: Define the initial state s_1^{-1} and set $\hat{\jmath}_1^{-1} = 0$.

Then, for $0 \le n \le N-1$:

B) Metric computation: for $1 \le k, h \le Q$ compute the cumulative Lagrangian cost $\hat{\jmath}_k^{n-1} + j(n; l_h, l_k; \lambda)$; if this is less than $+\infty$, connect s_k^{n-1} to s_h^n and associate the cost to the connecting path.

C) Pruning: for $1 \le k \le Q$, determine amongst all the paths converging to s_k^n the one with the minimum cumulative Lagrangian cost and prune all the others; call this minimum cost $\hat{\jmath}_k^n$ and associate it to the state.

In the end, select the state $s_{k^*}^{N-1}$ with the minimum cumulative cost and backtrack to s_1^{-1} along the surviving paths in the trellis: the sequence of traversed states provides the optimal allocation $\boldsymbol{p}^*$ for the chosen λ. From $\hat{\jmath}_{k^*}^{N-1} = J^*(\lambda) = \tilde{R}(\boldsymbol{p}^*) + \lambda D(\boldsymbol{\theta}, \boldsymbol{\sigma}, \boldsymbol{p}^*)$ the total throughput $R^*(\lambda) = N r_M - \tilde{R}$ and distortion $D^*(\lambda)$ can be obtained.

4.3. Iteration over λ

In some cases, especially when the distortion constraint is specified to within a tolerance interval, an educated guess for λ can avoid the need for an explicit search. In most cases, however, the value for λ must be determined iteratively until the constraint D_0 is met as closely as possible. The search algorithm exploits the convexity of the solution set and proceeds as follows[12]: first determine $\lambda_{\min}$ and $\lambda_{\max}$ so that $D^*(\lambda_{\max}) \leq D_0 \leq D^*(\lambda_{\min})$ (see below for details) and choose a starting value for λ between $\lambda_{\min}$ and $\lambda_{\max}$. Then, run the metric computation and pruning routines (Steps B and C above)[†]; if $D^*(\lambda) > D_0$ then replace $\lambda_{\max}$ by λ, else replace $\lambda_{\min}$ by λ; determine the new value as

$$\lambda = \frac{\tilde{R}^*(\lambda_{\min}) - \tilde{R}^*(\lambda_{\max})}{D^*(\lambda_{\max}) - D^*(\lambda_{\min})} \tag{14}$$

and repeat until the distortion constraint is met.

4.4. Initial values for the iteration

Given the monotonic relationship between λ and $D^*(\lambda)$, an obvious choice for the initial minimum value is $\lambda_{\min} = 0$, for which the maximum allowable real throughput is achieved at the expense of a very large distortion. A good estimate for $\lambda_{\max}$ can be inferred from the following argument. Assume we have obtained (for some large value of λ) the solution pair $(0, \tilde{R}_{\max})$, at zero distortion. As we now sweep λ from $+\infty$ to 0, consider the first value λ_1 for which $D^*(\lambda_1) > 0$; a nonzero distortion means that the power of the data is larger than the masking threshold in at least one frame, and therefore

$$D^*(\lambda_1) \geq \min_{0 \leq n \leq N-1} \{ \min_{i | l_i > \theta(n)} \{ d(\theta(n), \sigma(n), l_i) \} \} = D_{\min}; \tag{15}$$

$D_{\min}$ can be easily computed from a fast pass over $\boldsymbol{\theta}$. Because of (8), for all $(D, \tilde{R})$ pairs it is $\tilde{R} + \lambda_1 D \geq \tilde{R}^*(\lambda_1) + \lambda_1 D^*(\lambda_1)$ and, in particular, for $(0, \tilde{R}_{\max})$ we can write

$$\lambda_1 \leq \frac{\tilde{R}_{\max} - \tilde{R}^*(\lambda_1)}{D^*(\lambda_1)} \leq \frac{N r_M}{D_{\min}} \tag{16}$$

We can therefore set $\lambda_{\max} = N r_M / D_{\min}$.

4.5. Multiple subchannels

The previous analysis easily carries over to the case of data hiding over several subchannels simultaneously. Optimality of the global rate/distortion tradeoff means that *all subchannels must operate at the same value for* λ since the overall rate and distortion are the (unweighed) sums of the rates and of the perceptual distortions for all channels. For i different subchannels, by building

$$\boldsymbol{\theta} = \{\theta_{k_1}(0), \ldots, \theta_{k_1}(N-1), \theta_{k_2}(0), \ldots, \theta_{k_2}(N-1), \ldots, \theta_{k_i}(0), \ldots, \theta_{k_i}(N-1)\} \tag{17}$$

(and similarly for $\boldsymbol{\sigma}$) the trellis algorithm can be used to obtained an Ni-element allocation vector $\boldsymbol{p}$ from which the single subchannel allocations can be extracted in orderly fashion. A constant λ across subchannels implies an optimal distribution of the total amount of hidden data across channels, so that little or no data is tagged to frequency regions with poor masking capabilities such as in the high end of the audio spectrum.

5. A PRACTICAL IMPLEMENTATION

In the following section we will describe a practical implementation of the data hiding scheme for stereo, 44.1 KHz 16-bit PCM audio.

[†]Obviously, at each iteration only the new cumulative costs need be recomputed if both rate and distortion values for all transitions have been stored rather than just the cumulative metric. This increases the storage requirements by a factor of two but can significantly speed up the iteration process.

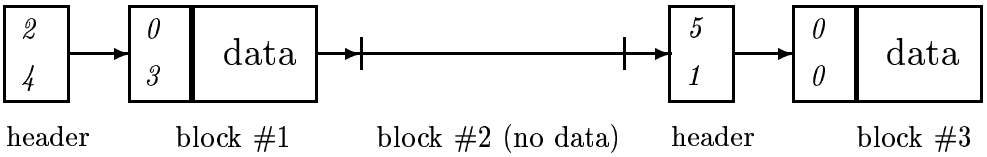


Figure 4. Signaling protocol for the allocation sequence $\boldsymbol{p}^* = \{l(2), l(2), l(2), l(2), l(0), l(0), l(0), l(5), l(5)\}$.

5.1. Signaling scheme

Discretization of the frequency axis is achieved by means of a 32-channel, uniform cosine-modulated filterbank as described in the MPEG Layer II specifications[8]; the width of a single subchannel is 690 Hz. The psychoacoustic model produces a masking curve every 1152 raw audio samples, which is then discretized to produce a masking threshold value for each subband; for a single subchannel, therefore, one threshold level is active over a frame of 36 subband samples.

Data tagging is achieved by replacing the least significant bits of the *subband* samples with data bits. Assuming the data has been scrambled, this is equivalent to an additive narrowband noise source with an approximate power of 6 dB per tagged bit; due to roundoff error in the analysis/synthesis filterbank computations, however, the least significant bit (LSB) of the subband samples is essentially random and therefore tagging can proceed safely only from the second LSB onwards. The transmission power (in linear units) corresponding to k tagged bits per sample is then:

$$l(k) = \begin{cases} 0 & \text{if } k = 0 \\ 4^{k+1} & \text{if } k > 0; \end{cases} \tag{18}$$

a signaling power of zero is needed when the finite distortion constraints cannot be met. Since there are 36 16-bit subband samples in a frame, $r_M = 576$ and $r(l) = 36k$; the minimum number of bits which can be tagged to a frame is therefore 36.

Side information is strictly related to the signaling protocol described below; its cost for a transmission power switch from k to h bits/sample is

$$c(l(k), l(h)) = \begin{cases} 0 \text{ bits} & \text{if } k = h \\ 36 \text{ bits} & \text{if } k \neq h \quad \text{and } k \neq 0 \\ 36h \text{ bits} & \text{if } k \neq h \quad \text{and } k = 0 \end{cases} \tag{19}$$

With these models, the optimal signaling sequence $\boldsymbol{p}^*$ is determined by the trellis algorithm, and tagging proceeds by passing the signal through the filterbank, replacing the LSB's of the subband samples according to $\boldsymbol{p}^*$, and reconstructing the PCM audio signal with the complementary synthesis filterbank. At the decoder, after synchronization, an identical analysis filterbank recovers the subband samples and the data can be retrieved following the structure conveyed by the side information.

5.2. Signaling protocol

Synchronization between transmitter and receiver is established at the audio PCM sample level by inserting a sync sequence in the LSBs of the audio PCM samples; this has no noticeable perceptual effect on the audio material. The hidden data transmission protocol is displayed in Figure 4 for an allocation vector

$$\boldsymbol{p}^* = \{l(2), l(2), l(2), l(2), l(0), l(0), l(0), l(5), l(5)\}.$$

First of all, the allocation vector is subdivided in blocks for which the number of bits per subband sample is constant; in this example there are three such blocks. Each tagged data block begins with the side information describing the number of bits/samples and the length (in frames) of the *next* block; for blocks which follow a zero bit allocation (and for the initial block) a separate one-frame side information header is necessary in which data is always tagged at one bit/sample; in all other blocks side information is encoded at the current bit/sample allocation level.

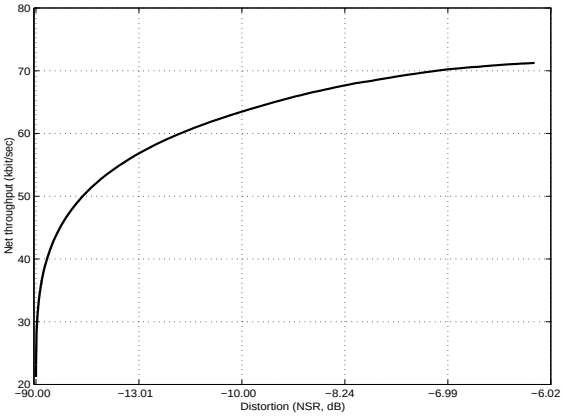


Figure 5. Throughput/distortion curve.

In the side information slot, 4 bits completely specify the power level and 8 bits are used for the block length; due to the catastrophic propagation of possible errors in the decoding of the allocation structure, a 3:1 repetition code is employed for a total of 36 bits. With this bit distribution the maximum block length is 6.6 seconds, which has proven largely sufficient in all practical cases; in the unlikely event of a longer block, the price to pay is an additional 36-bit side information slot after the 256th frame, which introduces a negligible suboptimality in the global allocation.

5.3. Results

In Figure 5 the net throughput is plotted against the perceptual NSR for a 1 minute audio excerpt.[9] At zero distortion (noise always below the masking threshold) the net rate is about 21 kbit/sec; at twice this rate the average NSR grows to -25 dB, the limit after which, in most cases, distortion becomes disruptive. For the intermediate case of a 30 kbit/sec throughput and -35 dB NSR, Figure 6 displays how the R/D optimal algorithm allocates the transmission power across bands and in time for the first four subbands; the dotted line represents the signal power while the thin continuous line represents the masking threshold; the power of the data is plotted with a thick line.

6. CONCLUSIONS

We have presented a data hiding algorithm to embed digital data into PCM audio signals. The data hiding strategy determines the optimal tradeoff between the capacity of the steganographic channel and the perceptual distortion added to the original signal for all possible rates by means of a Lagrangian minimization; for stereo CD audio, capacities on the order of 30 kbit/sec can easily be achieved at no perceptual cost. The rate/distortion framework is completely general and can be adapted to many different signaling schemes; furthermore, since the structure of the tagged data is conveyed to the receiver by means of side information, the system is independent of the particular psychoacoustic model employed in the analysis. Audio samples, together with the decoding software, can be retrieved at `http://lcavwww.epfl.ch/~prandoni/optimal/doa.html`.

REFERENCES

1. R. J. Anderson and F. A. P. Petitcolas, "On the limits of steganography," *IEEE Journal on Selceted Areas in Communications* **16**, pp. 474–481, May 1998.
2. I. J. Cox and J.-P. M. G. Linnartz, "Some general methods for tampering with watermarks," *IEEE Journal on Selceted Areas in Communications* **16**, pp. 587–591, May 1998.
3. Pacific Microsonics, "Hdcd: High definition compatible digital," `http://www.hdcd.com`.
4. B. C. J. Moore, *Psychology of Hearing*, Academic Press, San Diego, CA, USA, 1997.
5. J. D. Johnston, "Transform coding of audio signals using perceptual noise criteria," *IEEE Journal Sel. Areas Commm.* **6**, pp. 314–323, Feb. 1988.
6. K. H. L. Boney, A. H. Tewfik, "Digital watermarks for audio signals," in *Proc. of MULTIMEDIA '96*, pp. 473–480, 1996.
7. D. Pan, "A tutorial on mpeg audio compression," *IEEE Multimedia Journal* , Summer 1995.

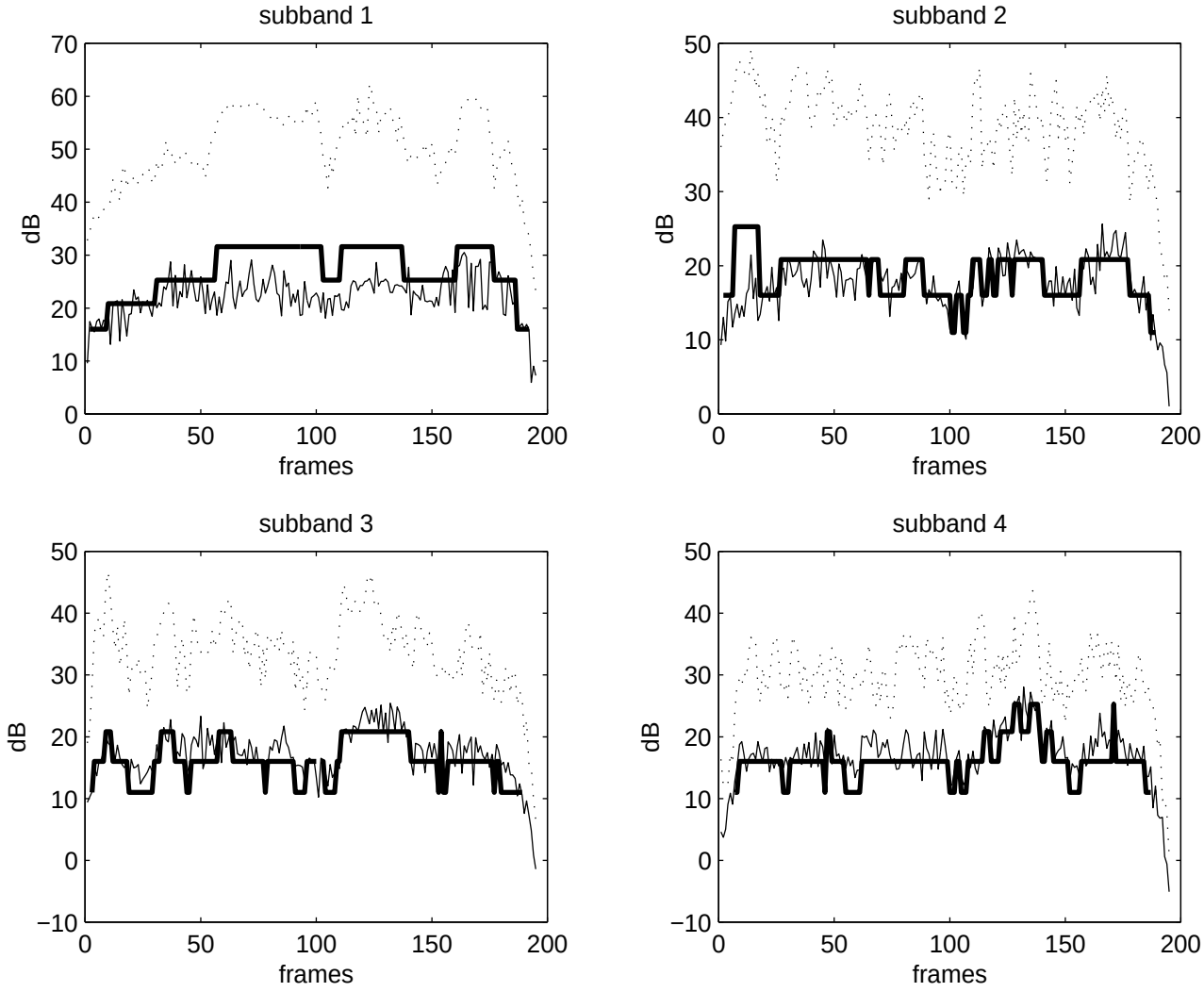


Figure 6. Transmission power allocations for the first four subbands.

8. ISO/IEC, *Internat. Standard IS 11172 (MPEG)*, ISO, 1993.
9. F. Schubert, "String quartet no. 13 in a minor, op. 29," *The Guarneri String Quartet* , 1997.
10. J. M. C. P.O. Okrah, "Multichannel modulation as a technique for transmission in radio channels," in *Proc. Vehicular Technology Conference*, pp. 29–33, (Secaucus, NJ, USA), May 1993.
11. Y. Shoham and A. Gersho, "Efficient bit allocation for an arbitrary set of quantizers," *IEEE Trans. Acoust., Speech, and Signal Proc.* **36**, pp. 1445–1453, September 1988.
12. K. Ramchandran and M. Vetterli, "Best wavelet packet bases in a rate-distortion sense," *IEEE Tran. on IP* **2**, pp. 160–175, April 1993.
13. R. Bellman, *Dynamic Programming*, Princeton University Press, 1957.