PERCEPTUALLY HIDDEN DATA TRANSMISSION OVER AUDIO SIGNALS

Paolo Prandoni, Martin Vetterli

LCAV, Ecole Polytechnique Fédérale de Lausanne, Switzerland
email: [prandoni,vetterli]@de.epfl.ch

ABSTRACT

A data transmission framework is proposed to embed digital data into an audio signal in a perceptually undetectable or almost undetectable way. The resulting signal can be reproduced as is with no loss of acoustic quality; the embedded data can be exactly retrieved at the decoder. The transmission process exploits the perceptual redundancy of the audio signal to conceal the acoustic impact of the embedded data; encoding of side information is used to inform the receiver of the time-varying structure of the masking properties of the audio signal. A sample implementation is described with a throughput of the order of 30 kbit/sec over CD-quality audio.

1. INTRODUCTION

State of the art audio coding algorithms such as MPEG [1] or AC3 can provide acoustically transparent compression ratios of the order of 4:1 to 6:1. As an example, a PCM CD-audio stereo bitstream (whose raw bitrate is 1.41 Mbit/sec) can be encoded by the MPEG algorithm at 384 kbit/sec without any perceivable loss of quality. These results are the outcome of a clever exploitation of the *masking phenomena* inherent to human hearing [3]. Simply stated, maskings 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.

From a different perspective, the fact that the perceptual quality of an audio signal is not affected as long as the injected noise is shaped to fall below the masking curve could be exploited to embed digital modulated data onto the audio waveform in an acoustically imperceptible way. The main advantage of this technique lies in the *layered* structure of the processed waveform: it occupies the same storage medium as the unprocessed one, either permanent or volatile; users with access to more sophisticated decoding equipment can retrieve the embedded data, but the audio data per se can be played by standard equipment as before.

Perceptual hiding of embedded data has been proposed previously in the context of audio watermarking [2]; our approach, although directly applicable to the problem of digital watermarking, is however more general in that the goal is to arrive at a data concealment system which allows perfect extraction at the receiving end.

2. PSYCHOACOUSTIC MODELING

Acoustic masking is a consequence of the nonlinear processing mechanisms of the human ear; frequency selective areas in the cochlea, called *critical bands*, exhibit a saturation characteristic whereby loud frequency components render inaudible the weaker components in the same critical band which lie below a certain threshold. The value of the threshold follows the decaying pattern of an experimentally determined masking function in the vicinity of the masker, and its absolute value depends on the index of the critical band, on the power of the masker (the loud component), and on the type of the masker (whether an isolated spectral line or part of a noise-like spectral component). The masking function can be approximated as a linear function in the log-power, bark frequency domain, where a unit of one bark corresponds to the width of a critical band; since the width of successive critical bands increases by approximately a third of an octave, there is a logarithmic mapping between bark scale and linear frequency scale. 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*.

The algorithmic process of estimating the masking curve is called *psychoacoustic modeling*. The outline of the estimation procedure 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 [4]:

- *Computation of the power spectrum;* this is performed by a short time Fourier transform analysis.
- *Separation of tonal and non-tonal components;* since the masking power of isolated spectral lines is less that 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 to the linear frequency domain; the final result will be denoted by $T(t,\omega)$. If the spectrum is divided in subbands (as is the case for MPEG compression, or for the signaling schemes we will examine later), the masking curve is generally discretized to one single value per subband; this is accomplished by selecting for each subband n

0-7803-4428-6/98 $10.00

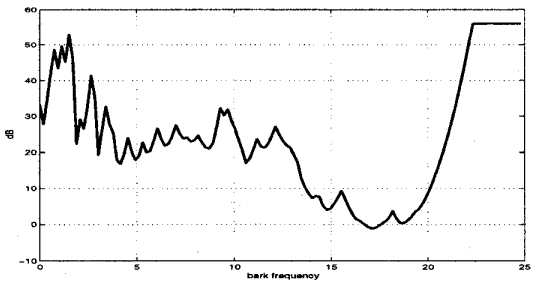

Figure 1: Masking curve for one frame.

the minimum of the masking curve in the interval straddled by the subband. The resulting set of masking levels will be denoted by $T_s(t,n)$.

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 refer to this analysis interval as a *frame* and denote its duration by t_f. Figure 1 displays a typical masking curve for a single frame of audio data. Here and in the following, the examples are obtained from track 2 of [5].

3. MULTICHANNEL SIGNALING

The time-varying masking curve indicates for each frame the portion of the signal which is in fact inaudible. From the perspective of embedding data into the audio waveform, the goal is to shape the power spectrum of the modulated data signal so that it falls below the masking threshold. In other words, over the bandwidth used for signaling (which can be the entire signal's bandwidth), the masking threshold represents the ideal power constraint of the channel. Since the instantaneous channel conditions are exactly known at the transmitter, the design of an embedded signal which fulfills the masking threshold requirements translates to a classic waterfilling problem.

Multichannel modulation ideas can be usefully employed to simplify the task at hand. By splitting the available bandwidth into adjacent non-overlapping subbands, the waterfilling problem is discretized into a finite number of independent signal design problems. Each subband n can be considered a time-varying fading channel where the instantaneous masking level $T_s(t,n)$ represents the power constraint for the embedded data signal and therefore the signal's power at the receiver. The peculiar feature of this transmission scheme (as opposed to [6], for instance) is that the channel conditions are exactly known at the encoder; the signaling scheme can therefore be adjusted on line to minimize audio distortion while maximizing throughput.

At the receiver, however, the situation is different; channel conditions are not known with precision, since the composite signal (original plus embedded data) possesses a different masking structure with respect to the original waveform. Informally stated, the psychoacoustic model at the transmitter singles out the perceptual "gaps" in the audio signal; the embedded data is shaped as to fill these gaps, so that psychoacoustic analysis of the original and composite waveforms can differ substantially. Relying on psychoacoustic modeling at the receiver to infer channel conditions would therefore result in signal-dependent errors which are very difficult to analyze and counteract. For this reason we choose to use side information, piggybacked onto the data signal, to inform the receiver of incoming switches in the transmission model[1].

Since side information reduces the net throughput of the signaling scheme, it is desirable to minimize the number of transmission model switches; on the other hand, in order to minimize distortion, there should be enough model switches to allow the embedded signal to closely follow the time-varying masking threshold; depending on the application, one might be willing to increase the throughput by keeping a signaling model constant over short periods in which its output power is above the masking threshold. This fundamental tradeoff can be looked at as a general rate-distortion problem, whereby we seek to maximize the throughput B while fulfilling a distortion constraint $D_{\max}$:

$$\begin{cases} \max\{B\} \\ D \leq D_{\max} \end{cases} \tag{1}$$

Efficient dynamic programming techniques exist to select the optimal number of model switches and their location [7, 8], and we will examine this in the next section.

4. OPTIMAL SIGNALING STRATEGY

The above constrained minimization problem can be reformulated in unconstrained form by considering the inverse cost functional

$$J(\lambda) = S + \lambda B \tag{2}$$

where S is the signal to noise ratio of the composite signal and B is the throughput. If D is the distortion introduced by the embedded data signal, $S = S_{\max} - D$, where $S_{\max}$ is the maximum dynamic range of the audio signal dependent on its physical storage format (for instance, $S_{\max} = 96$ dB for CD audio). Solution of (1) is obtained by finding

$$\hat{J}(\lambda^*) = \max\{J(\lambda^*)\} \tag{3}$$

where λ^* is the value which guarantees $S = S_{\max} - D_{\max}$ and which is found by a fast iteration of (3) over λ.

In the discussion which follows, a single subchannel is considered; the same principles equally apply to all the independent subchannels. For subchannel N, it is assumed that the transmitter can choose between M transmission

[1]Other reasons to use side information to notify the receiver of channel conditions are decoder simplicity and the possibility, much in the MPEG fashion, to use arbitrarily complex psychoacoustic schemes at the encoder as long as the resulting data stream complies with the mandated bitstream syntax. Evolution of the encoding process would not therefore require modifications at the decoder.

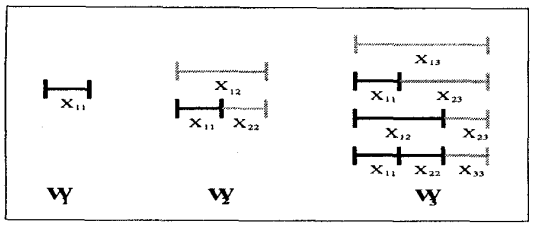

Figure 2: Incremental construction of the set of possible segmentations.

schemes, each with a given average power p_k and bitrate b_k, $1 \leq k \leq M$; using model k for a transmission interval from frame n to frame $n+m-1$ results in throughput and SNR values of

$$B_k(n,m) = \max\{-c_k + \sum_{i=n}^{n+m-1} b_k, 0\} \quad (4)$$

$$S_k(n,m) = S_{\max} - \sum_{i=n}^{n+m-1} \mu(i,p_k) \quad (5)$$

where c_k is the side information (in bits) used to indicate the model index and the duration of the frame interval. The distortion function is

$$\mu(i,p) = \begin{cases} 0 & \text{if } p \leq T_s(i,N) \\ p - T_s(i,N) & \text{if } p > T_s(i,N) \end{cases} \quad (6)$$

Since throughput and distortion are non negative, additive, and independent over disjoint signaling intervals we can make use of dynamic programming to find the optimal number of model switches and the optimal sequence of models for a given distortion budget. Under these conditions we can indeed rewrite (3) the following way: let for convenience be

$$J_{\max}(\lambda,n,m) = \max_{k=1,\ldots,M} \{S_k(n,m) + \lambda B_k(n,m)\};$$

then

$$\max\{J(\lambda)\} = \max_{\tau \in W}\{\sum_{(n,m)\in\tau} J_{\max}(\lambda,n,m)\} \quad (7)$$

where τ is a set of pairs (n,m) which define a segmentation of the data, and W is the set of all possible such segmentations. By the optimality principle the maximization over W can be carried out incrementally: let $W_0 = \emptyset$; at each step j, form W_j by extending all the segmentations in W_i by segment $(i+1,j)$, for all i from 0 to $j-1$ (see Figure 2, where the extending segments are drawn in gray). In a parallel way, the maximum cost functional for W_j is found amongst the sums of the cost of coding the extension $(i+1,j)$ plus previously computed max cost for subpartition W_i, for $0 \leq i < j$. Therefore we start by computing $\hat{J}_0(\lambda) = 0$ and then, at each step j,

$$\hat{J}_j(\lambda) = \max_{0\leq i<j} \{\hat{J}_i(\lambda) + J_{\max}(\lambda, i+1, j)\} \quad (8)$$

At each step we need only keep track of the newly computed value $\hat{J}_j(\lambda)$ and of the value of i yielding the maximum.

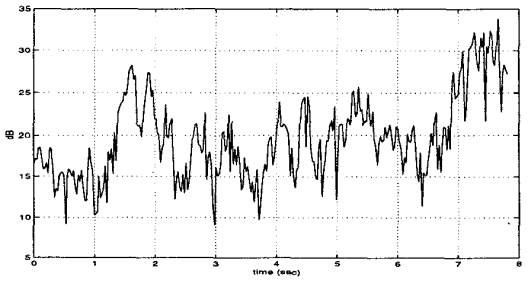

Figure 3: Masking threshold over time for a single subchannel.

This defines an incremental algorithm which, for a given operating point λ, yields the optimal segmentation and the optimal sequence of models with only quadratic computational complexity and linear storage requirements; as with all dynamic programming techniques, the complexity can be made linear by operating the maximization process of (8) over a sliding window, at the price of a slight suboptimality.

5. IMPLEMENTATION

To exemplify the signaling strategy described above we employed the basic building blocks of the MPEG-audio compression algorithm [1]. A 32-channel filterbank is used to decompose a CD audio signal (44.1 KHz sampling rate, stereo, 16 bits/sample) into subchannels 389 Hz wide; of these, the first 6 are used as independent subchannels for the embedded data. A psychoacoustic model provides the masking levels for each subchannel with a time resolution of 26 ms., corresponding to a transmission frame duration of 36 samples per subchannel.

In this simple example, the data is tagged onto the audio signal by modifying the least significant bits (LSB's) of the subband samples; the composite waveform is synthesized via the dual of the analysis filterbank and quantized to 16 bits. Due to this quantization, bit tagging starts from the second bit of each subband sample, since the LSB is affected by roundoff error. At the receiver, the analysis filterbank decomposes the signal and the tagged bits can be retrieved from the subband samples. By using fixed point arithmetic

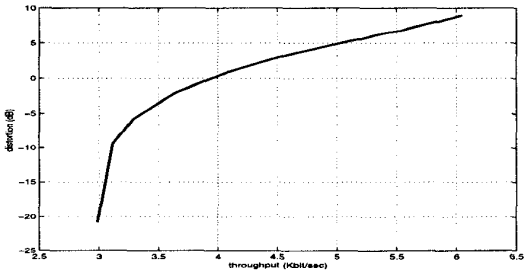

Figure 4: Throughput/Distortion curve for channel 5.

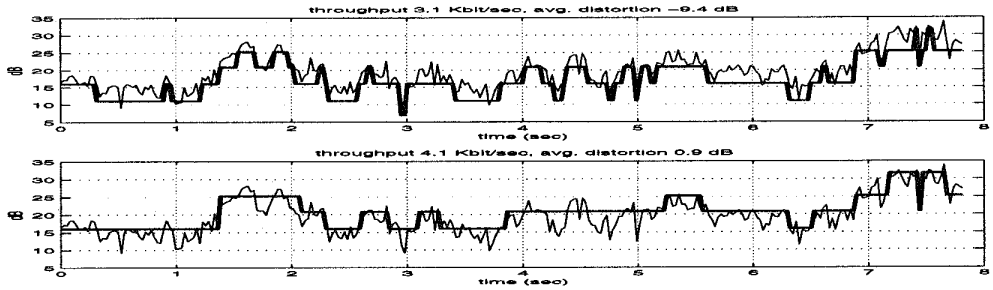

Figure 5: Two possible signaling model sequences.

with at least 27 bits of precision one can guarantee perfect recoverability of the tagged bits in the absence of channel errors.

With respect to the previous discussion, in this scheme we selected $M = 8$, corresponding to tagging zero to seven bits to each subband sample. The corresponding power levels p_k are given in table 1; these are the same values as introduced by a k-bit noise source, where the additional bit comes from the unusability of the subband LSBs. The side information comprises the model index k, which can be coded by three bits, and the signaling interval duration, 9 bits, for a maximum interval length of 512 frames or 13.3 seconds; since side information is vital to the decoding process, a rate 1/3 error control code is used, bringing the cost of side information to 36 bits per model switch.

Initialization of a subchannel takes place after detecting a frame-long syncword; the next frame is by convention a 1-frame segment containing solely the side information for the segment to follow encoded with model number two. Each segment starts with the side information for the next segment.

The test audio signal used in the experiments of Figures consists of 8 seconds of music for string quartet [5]. In the remainder of this section we will illustrate the results relative to one subband, the fifth of the right channel; qualitatively identical results hold for all the other subbands. Figure 3 displays $T_s(i, 5)$, the time-varying masking level for this sample subband. Iteration of the dynamic segmentation algorithm for several values of λ yields the throughput/distortion curve of Figure 4. The y-axis corresponds to the *average* distortion introduced by the embedded data signal; negative values for the distortion indicate that the noise introduced by the bit tagging is on average below the masking threshold. Figure 5 displays two different sequences of model switches; the thin line is the masking level of Fig. 3 while the thick line indicates the power of the chosen signaling model at each instant. The first segmentation corresponds to no perceptible distortion, the second one to an average distortion of 0.9 dB. It can be noted in this second case that the decreased distortion constraint determines a looser match between signaling power and masking threshold; the increase in throughput comes directly from the reduced number of signaling transitions.

For the global signaling scheme, at zero or less average distortion for each subband, the total throughput for the 6 left and right subchannels is 24.6 kbit/sec.

6. CONCLUSIONS

A data transmission system has been presented which allows to tag data to an audio signal in a perceptually transparent way. Audio files which demonstrate the algorithm as well as an implementation of the coding process can be retrieved at:
`http://lcavwww.epfl.ch/~prandoni/optimal.html`

7. REFERENCES

[1] ISO/IEC. *Internat. Standard IS 11172 (MPEG).* ISO, 1993.

[2] K.N. Hamdy L. Boney, A. H. Tewfik. Digital watermarks for audio signals. In *Proc. of MULTIMEDIA '96*, pages 473–480, 1996.

[3] J. D. Johnston. Transform coding of audio signals using perceptual noise criteria. *IEEE Journal Sel. Areas Commm.*, 6(2):314–323, Feb. 1988.

[4] D. Pan. A tutorial on mpeg audio compression. *IEEE Multimedia Journal*, Summer 1995.

[5] F. Schubert. String quartet no. 13 in a minor, op. 29. *The Guarneri String Quartet*, 1997.

[6] J. M. Cioffi P.O. Okrah. Multichannel modulation as a technique for transmission in radio channels. *IEEE id 0-7803-1266-x*, 1993.

[7] Z. Xiong, K. Ramchandran, C. Herley, and M. T. Orchard. Flexible tree-structured signal expansions using time-varying wavelet packets. *IEEE Trans. SP*, 45(2):333–345, Feb 1997.

[8] P. Prandoni, M. Goodwin, and M. Vetterli. Optimal time segmentation for signal modeling and compression. In *ICASSP Proc.*, volume 3, pages 2029–2032, Munich, Germany, April 1997.

k	p_k (dB)	k	$p_k(dB)$
1	$-\infty$	5	31.59
2	7.0	6	37.75
3	16.0	7	43.84
4	25.28	8	49.89

Table 1: Power associated with the signaling models.