Proc. 1999 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics, New Paltz, New York, Oct. 17-20, 1999

A SYSTEMATIC HYBRID ANALOG/DIGITAL AUDIO CODER

Richard Barron and Alan Oppenheim

Research Laboratory of Electronics
Massachusetts Institute of Technology
Cambridge, MA 02139

ABSTRACT

This paper describes a signal coding solution for a hybrid channel that is the composition of two channels: a noisy analog channel through which a signal source is sent unprocessed and a secondary rate-constrained digital channel. The source is processed prior to transmission through the digital channel. Signal coding solutions for this hybrid channel are clearly applicable to the in-band on-channel (IBOC) digital audio broadcast (DAB) problem. We present the design of a perceptually-based subband audio coder, with complexity comparable to conventional coders, that exploits a signal at the receiver of the form $y[n] = g[n]*x[n]+u[n]$, where $x[n]$, $g[n]$, and $u[n]$ denote respectively the source, the impulse response of convolutional distortion, and additive Gaussian noise. Concepts from conventional subband coding, *e.g.* subband decomposition, quantization, bit allocation, and lossless signal coding, are tailored to exploit the analog signal at the receiver such that frequency-weighted mean-squared error is minimized.

1. INTRODUCTION

In some source coding scenarios, there exist observations of signals at the decoder that are correlated with the source which may be used jointly with a digital representation to reconstruct the source. For example, in the case of in-band on-channel (IBOC) digital audio broadcast (DAB), an existing noisy analog communications infrastructure may be augmented by a low-bandwidth digital side channel for improved fidelity. As another example, in a two-sensor scenario, one sensor may observe a distorted full-bandwidth form of the source signal, while the other observes the source undistorted but can only record or transmit a low-bandwidth representation of the signal. A final example is a source coding scheme which devotes a fraction of available bandwidth to the analog source and the rest of the bandwidth to a digital representation. This scheme is applicable in a wireless communications environment, where analog transmission has the advantage of a gentle "roll-off" of fidelity with SNR.

This paper describes a method for subband signal coding, using algorithms of comparable complexity to conventional coders, that exploits a noisy analog signal at the decoder. We assume the analog signal is the output of a channel through which the source is sent uncoded. Clearly by using the analog signal at the receiver, we should be able to reduce the required digital bit rate while offering comparable fidelity to conventional coding systems that ignore the analog signal. In the DAB scenario broadcasters can use the bits saved on audio source coding either for improved error-correction or transmission of non-audio data.

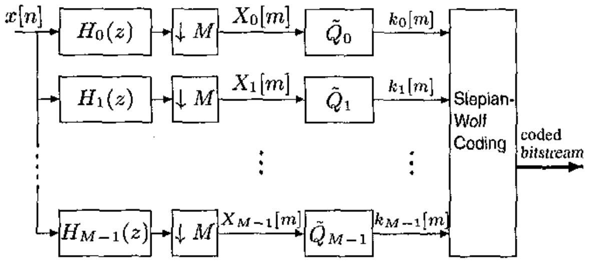

Figure 1: The digital encoder. The $H_i(z)$s are analysis filters. The $\tilde{Q}_i$s are modulo-uniform scalar quantizers.

Shamai *et. al.* in [1] use the term "systematic" to describe source coding with analog information at the receiver as an extension of a concept from error-correcting channel codes. A *systematic* error-correcting code is one whose codewords are the concatenation of the uncoded information source string and a string of parity-check bits. Similarly, in the systematic hybrid source coding scenario, we have an uncoded analog transmission and a source-coded digital transmission.

In our design, concepts from conventional subband coding, *e.g.* subband decomposition, quantization, bit allocation, and bitstream coding, are tailored to exploit the analog signal at the receiver such that frequency-weighted mean-squared error (MSE) is minimized. Because we code subband coefficients, all results pertaining to perceptual masking are easily applied to this method of coding. In addition, the methods proposed here require very little additional overhead as far as source side information. Although our results are applicable to coding of all signals, we emphasize the application of these digital coding techniques to the perceptual coding of audio as a solution to the IBOC DAB problem. In a series of experiments we augment, with a digital bitstream, a 30 dB analog signal corrupted by additive white Gaussian noise at the receiver. Bit rates for the digital stream as low as 10 to 20 kbits/sec yield perceptually near-transparent coding of mono audio sampled at 44.1 kHz.

This paper is organized as follows. Section 2 will describe the

This research was supported in part by the U.S. Air Force Office of Scientific Research under Grant AFOSR-F49620-96-1-0072 and in part through collaborative participation in the Advanced Sensors Consortium sponsored by the U.S. Army Research Laboratory under Cooperative Agreement DAAL01-96-2-0001. Richard Barron is supported by the National Defense Science and Engineering Graduate Fellowship. The views expressed are those of the authors and do not reflect the official policy or position of the U.S. Government.

structure of the encoder and decoder. Section 3 describes the implementation of the coding method for audio and presents the results of informal listening experiments on a variety of audio tracks.

2. THE ENCODER AND DECODER

In this section we describe the encoder and decoder as diagrammed in Figs. 1 and 2. We will assume that the source is some colored Gaussian sequence $x[n]$ and the analog observation is $y[n] = g[n]*x[n]+u[n]$, where $g[n]$ is the impulse response of some convolutional distortion and $u[n]$ is additive Gaussian noise, which is independent from the source and may be colored. These assumptions will assist in analysis but the system design can be applied to general sources and a broad class of additive noise. The assumption that audio is approximately Gaussian has been successfully applied to a number of problems in audio processing. The Gaussian channel model very accurately represents the AM channel and closely approximates the FM channel in the high SNR case [2]. We refer the reader to [3] for an explanation of the optimality of the encoder/decoder structures with respect to mean-squared error. The encoder is very similar to a conventional encoder, while the decoder has some additional complexity induced by the incorporation of the analog information into the estimate. As with conventional coding, systematic hybrid coding is composed of three essential elements: an analysis/synthesis filter bank, quantization, and bitstream coding. We describe the encoder and decoder structures that compose each of the first two elements. The optional bitstream coding stage involves the use of Slepian-Wolf codes, as shown in Fig.1. With the added complexity of bitstream coding, we can lower the digital bit rate while maintaining the same performance. As the construction of Slepian-Wolf codes is beyond the scope of this paper, we assume the indices output from the quantizers are transmitted uncoded to the the decoder.

2.1. The Filter Bank

The hybrid coder operates on the basic premise of subband coding. The source signal $x[n]$ is decomposed by a filter bank into a set of M subband signals $\{X_i[m]\}_{i=0}^{M-1}$, which are subsequently coded (quantized) for the particular bit rate allowed by the digital channel. A particular filter bank is described by its analysis filters, denoted by $\{H_i(z)\}_{i=0}^{M-1}$ in Fig. 1, and corresponding synthesis filters at the decoder, denoted by $\{F_i(z)\}_{i=0}^{M-1}$ in Fig. 2. The synthesis bank takes the subband signal estimates $\{\hat{X}_i[m]\}_{i=0}^{M-1}$, derived from the analog and digital data, and creates a time domain signal estimate, $\hat{x}[n]$.

We now establish some facts about filter banks used in this paper. We assume that all filter banks are critically sampled, *i.e.* the number of samples into the filter bank is equal to the number of samples out. Critical sampling is achieved by downsampling by M after the analysis filters. We use n to denote the index for the original source and the index m to denote the time index for the (decimated) subband signals. If the filter bank is implemented by a fast transform like the modified discrete cosine transform (MDCT), each index m corresponds to a windowed frame of signal data. We therefore use the nomenclature "frame" to refer to a particular index m.

There exists a wealth of results on the design of filter banks for a variety of signal processing tasks. For use in conventional signal coding, a filter bank usually satisfies several criteria. First, the filter bank is perfect reconstruction, so that in the absence of any quantization of subband signals the source can be reconstructed exactly using the matching synthesis filter bank. Secondly, we desire strong stopband rejection for each synthesis filter so that any noise injected into the system by quantization will not affect neighboring subbands significantly. Finally, the filter bank should be implementable by fast algorithms, usually involving the FFT, to minimize algorithmic complexity of the coder. The MDCT satisfies these criteria nicely, and is used in many state of the art transform audio coders. We show in [3] that, in the sense of maximum coding gain, a good filter bank for conventional source coding is also a good filter bank for coding with analog information at the decoder. Therefore we can enjoy all of the advantages of standard filter bank constructions for our problem. The audio coder that we implement using analog information at the decoder will use the MDCT for subband decomposition. In order to alleviate time-domain artifacts such as pre-echo, many state of the art audio coders use signal-dependent switched filter banks. These filter banks may also be used for systematic hybrid audio coding, but our initial implementation uses a fixed filter bank.

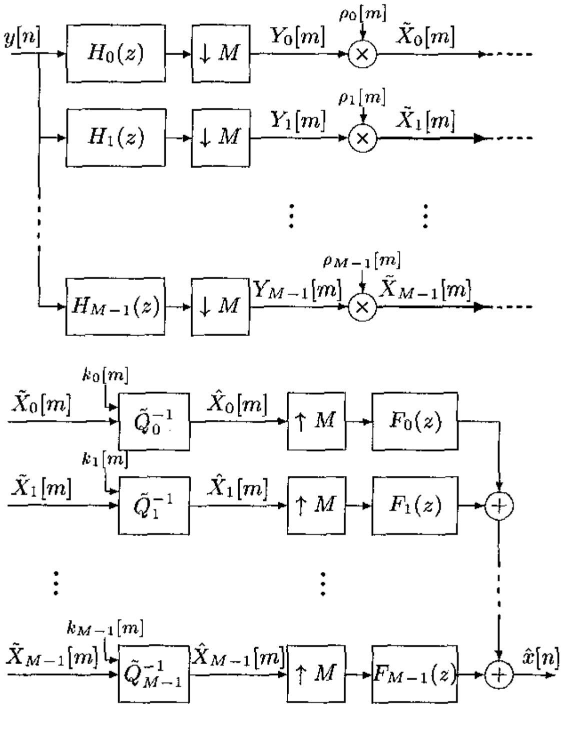

Figure 2: The hybrid decoder. The $\tilde{Q}_i^{-1}$s are hybrid reconstruction functions. The $F_i(z)$s are synthesis filters.

2.2. Quantization

The encoder must quantize the subband coefficients under a bit rate constraint, anticipating that the decoder will have access to analog information correlated to the source. As shown in Fig. 1, we let $\{\tilde{Q}_i(\cdot)\}_{i=0}^{M-1}$ denote the bank of hybrid quantizers that encode the subbands. For most of the remainder of Sec. 2, we omit the indices i and m as they will be considered implicit.

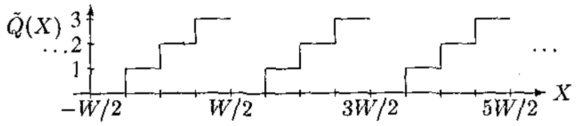

Figure 3: Hybrid quantization with modulo uniform quantizers, a 2 bit example.

Lattice	Symbol
$\mathcal{L}_0$	○
$\mathcal{L}_1$	△
$\mathcal{L}_2$	□
$\mathcal{L}_3$	×

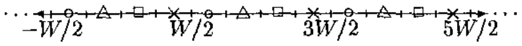

Figure 4: Lattice interpretation of hybrid quantization.

For systematic hybrid coding, we use quantizer structures that have complexity comparable to conventional scalar quantizers. Simply the composition of a modulo operation and a conventional uniform quantizer, the quantizer $\tilde{Q}(X)$ for the hybrid scenario is given by:

$$\begin{aligned} \tilde{Q}(X) &= Q(X \operatorname{mod} W) & (1) \\ Q(\nu) &= k, \quad \frac{kW}{K} \leq \nu + \frac{W}{2} < \frac{(k+1)W}{K} & (2) \\ & \quad k = 0, 1, ..., K-1, \end{aligned}$$

where K is the number of levels allocated to the quantizer. We show in [3] that these modulo-uniform quantizers very closely approximate the optimal scalar hybrid quantizers with respect to mean-squared error. We discuss the determination of appropriate values for K for each of the subbands when we address bit allocation in Sec.2.4. Note that the graph of $\tilde{Q}(X)$, shown in Fig. 3 for a 2 bit quantizer, is a cascade of staircases, where W is the width of each staircase. A *cell* is the interval described by a step of the staircase, and its width is given by $\Delta = W/K$. Note that each quantizer level $k \in 0, 1, ..., K-1$ is the image of the union of several disjoint cells, rather than just one cell.

The quantizer $\tilde{Q}(X)$ may also be interpreted in terms of a collection of interleaved lattices, $\{\mathcal{L}_i\}_{i=0}^{K-1}$. To each quantizer output $k = \tilde{Q}(X)$ we assign a lattice $\mathcal{L}_k$, as shown in Fig 4, with lattice points uniformly separated by length W. Each lattice point is the center of a cell region defined by the function $\tilde{Q}$. Each successive lattice is the previous lattice shifted by Δ units. An alternative description of $Q(X)$ in terms of lattices is as follows. The function $Q(X)$ is the index of the lattice that contains the lattice point closest in Euclidean distance to X. The lattice interpretation is useful to describe the reconstruction of subband coefficients from $k = Q(X)$ and the analog signal, a procedure we describe in Section 2.4.

In order to determine a good value for the staircase width W we focus attention on the operation of the decoder. Note that at the decoder the analog signal $y[n]$ is also decomposed into subbands $Y_i[m]$ by the same analysis bank as at the encoder. Given that $y[n] = g[n] * x[n] + u[n]$ we argue that a given subband signal Y is closely approximated by $Y = hX + U$ where h is a known gain, and U is an additive Gaussian noise variable. Note that the signal $U_i[m]$, as a function of frame m, is a highly correlated sequence, because it is bandlimited by the i^{th} subband filter. In this implementation we do not exploit this correlation, but a future implementation may use prediction across frames to do so. We assume that convolutional distortion is approximated closely by constant gain for each subband. Several authors have shown that this is a valid approximation [5, 6]. If more accuracy is desired we can use the results in [7] indicating that for appropriate choice of analysis and synthesis filters, convolution may be implemented by low order filters on each subband signal.

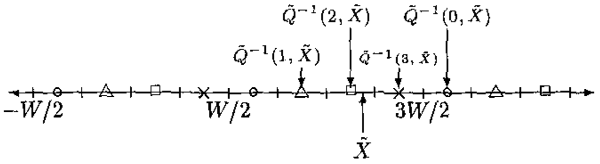
$\tilde{Q}^{-1}(2, \tilde{X})$
$\tilde{Q}^{-1}(0, \tilde{X})$
$\tilde{Q}^{-1}(1, \tilde{X})$
$\tilde{Q}^{-1}(3, \tilde{X})$
$-W/2$
$W/2$
$3W/2$
$\tilde{X}$

Figure 5: Example of signal reconstruction

Let σ_X^2 and σ_U^2 be the variances of X and U respectively. In [3] we argue that the subband decomposition approximately orthogonalizes the source samples and the noise samples in a given frame. Due to this fact, the minimum mean-squared error (MMSE) estimate $\tilde{X}$ of subband coefficient X from a frame of $y[n]$ is ρY, simply a gain times the analog subband coefficient. The value ρ, shown in Fig. 2 at the decoder, is the correlation coefficient between X and Y:

$$\rho = \frac{h\sigma_X^2}{h^2\sigma_X^2 + \sigma_U^2}.$$

The error in the estimate is given by $e = X - \tilde{X}$, and the error variance is given by:

$$\sigma_e^2 = \frac{\sigma_X^2 \sigma_U^2}{h^2\sigma_X^2 + \sigma_U^2}. \quad (3)$$

Given the analog observation, the minimum mean-squared error variance σ_e^2 is constant and is always less than the source variance, σ_X^2. Therefore we let $W = C\sigma_e$, for some constant $C > 1$, so that a single staircase will contain the region of support for the MMSE error. In [8], we show that $C \approx 7$ is the optimal value for C for the typical range of operation in an audio application. Note that the source variance, noise variance, and gain h in a given subband must be known in order to calculate C. Typically, the values σ_U^2 and h are given by some known channel model. The variance of the source, however, must be communicated as side information in the digital bit stream, perhaps in some low-bandwidth parametric form. As we shall see in Section 2.4, this information is sent as side information to specify bit allocation across subbands, so the analog estimation stage requires no additional overhead.

We have established that the subband coefficients will be coded by modulo-uniform quantizers, and the modulo factor W equals $7\sigma_e$, where σ_e is the standard deviation of the MMSE estimation error based on the analog observation.

2.3. Signal Reconstruction

The index k that is output from the quantizer $\tilde{Q}$ is sent to the decoder, where it is used jointly with the analog signal $y[n]$ to reconstruct the subband coefficient X. As shown in Fig. 2, the reconstruction function for each subband is denoted $\tilde{Q}^{-1}(k, \tilde{X})$, and it requires the MMSE analog estimate $\tilde{X}$ in addition to the index k from the encoder as input. The reconstruction function provides an improved estimate $\hat{X}$ of the subband coefficient X. Illustrated in Fig. 5 for a 2 bit quantizer, the function is implemented as follows. The index $k = \tilde{Q}(X)$ from the encoder defines a particular uniform lattice $\mathcal{L}_k$. The reconstructed subband signal $\hat{X} = \tilde{Q}^{-1}(k, \tilde{X})$ is the lattice point of $\mathcal{L}_k$ that is the minimum Euclidean distance from $\tilde{X}$. We show in [3] that this minimum-distance reconstruction rule closely approximates the rule for MMSE reconstruction.

2.4. Bit Allocation

Due to digital bandwidth constraints, a particular number of bits B are allocated for a frame of audio data. The bit allocation problem addresses the allocation of b_i bits to each quantizer $\tilde{Q}_i$ such that a perceptually-weighted error is minimized and $\sum_i b_i = B$. Variable rate coders vary bit rates from frame to frame. We do not discuss the procedure to determine the allocation of bits across frames, as methods from conventional coding extend obviously to the hybrid coder.

The considerable coding gains attained by most state of the art audio coders may be attributed to bit allocation based on a signal-dependent masking threshold we refer to as the just-noticeable-distortion (JND) level. In the hybrid coding scenario, use of perceptual masking is as straightforward as in conventional coding. The JND, as a function of critical band frequency, may be calculated by one of several methods outlined in the research literature. Let $M_{\rm CB}$ be the number of critical bands, and let J_i, $i = 0, 1, ..., M_{\rm CB} - 1$, denote the JND function. The JND is most often calculated as a function of two variables for each critical band: source variance and level of tonality or noise-like character. Note that since we are sending the source variance to facilitate the analog estimation stage, this information is already provided.

Bits are allocated according to the following greedy algorithm. The frame starts with a reservoir of B bits. Initially, each critical band has an associated weighted analog estimation error $(\sigma^2_{e_i})_{CB}/J[i]$, where $(\sigma^2_{e_i})_{CB}$ is simply the sum of the mean-squared estimation errors in the subbands contained in critical band i. Bit allocation is performed by inverse waterpouring on the weighted analog estimation error. If the number of bits B in the reservoir is large enough for every frame of audio, perceptual transparency (CD-quality audio) is achieved when the mean-squared error in every critical band is less than the JND.

3. IMPLEMENTATION AND EVALUATION OF PERFORMANCE

This section describes the implementation of the signal coder for the coding of audio at 44.1 kHz sampling rate with observations of the source corrupted by additive white Gaussian noise at the receiver. It is clear from the design that in a broadcast situation, coding for a worst case SNR will enable proper decoding for all SNRs greater than the worst case value.

The filter bank is implemented by a 2048 sample MDCT/IMDCT operating on data windowed by an integrated Kaiser window at 50% overlap. Each subband coefficient is quantized as in Sec. 2.2. Reconstruction from the quantized coefficients requires that the subband energy envelope be communicated to the decoder as side information. We use a frequency-warped all-pole model proposed in [10] to describe the spectral envelope with between 20 and 40 poles depending on the source. The frequency warping gives equal emphasis to the spectral components on a Bark frequency scale. The spectral envelope is encoded as log-area ratios that are quantized at 5 bits per coefficient. Thus the side information uses 4.3-8.6 kb/sec of bandwidth. Reusing the side information, we calculate the JND level according to [9] using the parametric representation of the spectral envelope. In this current implementation we use no tonal/noise-like properties to calculate the JND, so the masking thresholds are in general more conservative than necessary.

Analog Channel SNR (dB)	Bit Rate (kb/sec)
10	25-55
20	15-40
30	10-20

Table 1: Required bit rate for transparent audio given analog channel output at certain SNR

As an evaluation of performance, we coded audio for transparency assuming 10, 20, and 30 dB SNR observations at the receiver. We coded several different types of audio and show the ranges of required bit rates for each SNR in Table 1. Systematic hybrid audio coders clearly have significant coding gain over coders that ignore the analog signal at the receiver. Preliminary results also suggest that there are similar coding gains for the FM channel.

4. REFERENCES

[1] Shamai, S., Verdu, S., Zamir, R., "Systematic Lossy Source/Channel Coding," *IEEE Trans. Inform. Theory*, vol. IT-44, pp. 564-579, 1998.

[2] Vitterbi, A.J., "Phase-Locked Loop Dynamics in the Presence of Noise by Fokker-Planck Techniques", *Proc. IEEE*, vol. 51, pp. 1737-1753, December 1963.

[3] Barron, R.J., "Systematic Hybrid Analog/Digital Signal Coding," *PhD Thesis*. Massachusetts Institute of Technology, Cambridge, 1999 (unpublished at time of submission).

[4] Gersho, A., Gray, R.M., "Vector Quantization and Signal Compression," Kluwer Academic Publishers, Boston, 1992.

[5] Ramstad, T.J., "IIR Filter Bank for Subband Coding of Images," *IEEE Intl. Symp. Circuits Syst.*, Espoo, Finland, pp. 827-830, June 1988.

[6] Smith M.J.T., and Barnwell, T.P. III, "A new Filter Bank Theory for Time-Frequency Representation," *IEEE Trans. Acoust., Speech, Signal Processing*, ASSP-35, March 1987, pp. 314-327.

[7] Vetterli, M., "Running FIR and IIR Filtering Using Multirate Filter Banks," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 36, May 1988, pp. 730-738.

[8] Barron, R.J., Oppenheim, A.V., "Signal Processing for Hybrid Channels," *Proceedings of 3rd Annual ARL Fedlabs Symposium* pp. 481-484., Feb. 1999.

[9] International Standards Organization, "ISO/IEC 11172-3 MPEG 1 Audio Coding Standard," pp. 129, January 1996.

[10] Smith, J.O., Abel J.O., "The Bark Bilinear Transform," 1995 IEEE ASSP Workshop on Applications of Signal Processing to Audio and Acoustics. pp. 202-205, October 1995.