Proceedings of the 1997 IEEE ASSP Workshop on Applications of Signal Processing to Audio and Acoustics, October 19-22, 1997, New Paltz, NY

PERCEPTION-BASED BIT-ALLOCATION ALGORITHMS FOR AUDIO CODING

Stephen Voran

Institute for Telecommunication Sciences, U.S. Department of Commerce
325 Broadway, Boulder, Colorado 80303, sv@bldrdoc.gov

ABSTRACT

We describe six algorithms for bit allocation in audio coding. Each algorithm stems from the minimization of a different perceptually-motivated objective function. Three of these objective functions are extensions of existing ones, and three are new. Closed-form bit-allocation equations result in five cases, and an iterative approach is required in the sixth.

1. BACKGROUND

Sub-band and transform audio coders generate and encode frequency domain decompositions of audio signals. When combined with an understanding of human hearing, this approach offers the opportunity to encode signal components in a way that minimizes the audibility of coding distortions. In particular, when signal components are to be quantized, different quantizer resolutions can be selected for different signal components to minimize the audibility of the quantization process. The resolution of a quantizer is increased by assigning more bits to it. The total number of bits available for quantizing all signal components is usually fixed by the design of the audio coder and its bit rate. A bit-allocation algorithm dynamically distributes this fixed pool of bits over a number of signal component quantizers so that the audibility of the quantization process is minimized. This results in the highest possible audio quality for a given number of bits.

One important aspect of a bit-allocation algorithm is the objective function that represents the audibility of the quantization process. The bit-allocation algorithm minimizes this objective function, so higher quality coding is obtained with objective functions that more closely track the actual audibility of the quantization process. Existing objective functions are most often based on the noise-to-mask ratio (NMR)[1-5]. This approach involves decomposing an encoded signal into an original signal plus quantization noise. The NMR predicts the extent to which this theoretical noise signal would be masked by the original audio signal if both were presented to a listener at the same time. The NMR is based on well-documented masking effects, and has been shown to be extremely useful in audio coding and audio-quality assessment.

It is easy to argue that the NMR is a relevant and effective model for listeners judging the audibility of a quantization process. However, we argue that it may not be the *most* relevant or effective of all possible models: Listeners hear the encoded signal and must judge its purity against some internal reference. They do not hear an original signal that attempts to mask a quantization noise. The decomposition of a single signal into a masker and a maskee is generally an ad-hoc process. Finally, if quantization noise is to be treated as an actual signal, why not process it with a spreading function, just as the original signal is processed?

We propose three new perception-based objective functions to represent the audibility of the quantization process. Like the NMR, these objective functions also contain approximations and compromises. In addition, we extend three established objective functions to include additional perceptual weighting functions.

A second important aspect of a bit-allocation algorithm is the procedure for minimizing the objective function. Several procedures with varying restrictions on quantization functions and admissible bit-allocation values have been reported[6-8]. Such procedures could be used with the objective functions presented here. In addition, five of our objective functions yield closed-form expressions for the minimizing bit allocation under a fixed bit-rate constraint. These closed-form expressions allow for single step bit allocation, but the resulting bit-allocation values must be rounded to integers in any practical implementation. Our sixth objective function yields an iterative algorithm that results in integer bit allocations.

2. DEFINITIONS

The following symbols will be used:

N = number of frequency bands in bit-allocation problem,

$x_i = i$ th frequency domain sample of signal to be coded,

$p_i^2 = \mathrm{E}(x_i^2)$, E is the expectation operator,

$y_i = i$ th frequency domain sample of coded signal,

$$\underline{x} = [x_1, x_2, \ldots, x_N]^{\mathrm{T}},\ \underline{y} = [y_1, y_2, \ldots, y_N]^{\mathrm{T}},$$
$$\underline{p}^2 = [p_1^2, p_2^2, \ldots, p_N^2]^{\mathrm{T}},$$

$[\,\underline{c}\,]_i$ denotes the i th element of the vector $\underline{c}$,

b_i = number of bits allocated to i th band (bits/sample) ,

r_i = sample rate in i th band (samples/second) ,

$$R = \sum_{i=1}^{N} r_i \;=\; \text{total sample rate (samples / second)}\,.$$

A spreading function emulates the way in which a spectral component of an audio signal excites a neighborhood on the basilar membrane[9]. Thus, spreading functions can be used to convert spectral representations into excitation patterns. When a spectral representation is uniformly spaced on a critical band or Bark scale, a spreading function can be efficiently implemented as a matrix-vector multiply:

$$[\tilde{p}_1^2, \tilde{p}_2^2, \ldots, \tilde{p}_N^2]^{\mathrm{T}} = \underline{\tilde{p}}^2 = T_p \cdot \underline{p}^2 \,,$$

where $\underline{\tilde{p}}^2$ is the excitation pattern due to $\underline{p}^2$ and T_p is an N by N Toeplitz power spreading matrix. The first row of T_p is [1, α, α^2,…,α^{N-1}] and the first column of T_p is [1, β, β^2,…,β^{N-1} $]^{\mathrm{T}}$, where $10\cdot\log_{10}(\alpha) = -d\cdot\Delta$, $10\cdot\log_{10}(\beta) = -u\cdot\Delta$, d and u are the magnitudes of the downward and upward spreading slopes respectively, in dB/Bark, and Δ is the sample spacing in Bark. (We use d = 25 dB/Bark, u = 10 dB/Bark, Δ = 1 Bark, and N = 25 bands.) We also define the amplitude spreading matrix T_a, which is simply the element-by-element square root of T_p. We will exploit the invertibility of T_p:

$$T_p^{-1} = \left(\frac{1}{1-\alpha\beta}\right)\cdot\begin{bmatrix} 1 & -\alpha & & & & \\ -\beta & 1+\alpha\beta & -\alpha & 0 & & \\ & -\beta & 1+\alpha\beta & -\alpha & & \\ & & & \cdots & & \\ & 0 & & -\beta & 1+\alpha\beta & -\alpha \\ & & & & -\beta & 1 \end{bmatrix}.$$

We also define t_{ij} to represent the elements of T_p, and $\tilde{t}_{ij}$ to represent the elements of T_p^{-1}.

In many audio coding algorithms, groups of m samples from adjacent frequencies and/or times are divided by a single scale factor (chosen from a fixed set of n possible scale factors) before quantization. This scaling brings all samples in the group into the appropriate range for quantization. The scale factor multiplies the same group of samples at the decoder. We use s_i to represent the scale factor used in the i th band. The scaling operation can also be viewed as coarse quantization at a fixed (independent of bit allocation) bit rate of $\log_2(n)/m$ bits/sample.

We assume that quantization errors are zero mean, and are independent from each other and from the signal $\underline{x}$. The mean-squared quantization error in the i th band is assumed to be

$$E((x_i - y_i)^2) = k_i^2 \cdot 2^{-2b_i} \,,$$

where k_i^2 is a distribution-dependent scale factor. For example, when quantization errors are uniformly distributed

$$k_i^2 = \frac{s_i^2}{12}\,.$$

The bit allocations that follow are made under the total bit-rate constraint of B^* bits/second:

$$\sum_{i=1}^{N} b_i \cdot r_i \;=\; B \;\le\; B^* \text{ bits / second}\,. \qquad (1)$$

3. BIT-ALLOCATION ALGORITHMS

We present six objective functions that attempt to model the audibility of the quantization process. Sections 3.1-3.3 contain extensions of existing objective functions, and Sections 3.4-3.6 contain new objective functions. When combined with the bit-rate constraint in (1), each objective function leads to a bit-allocation algorithm. Additional observations on these results are offered in Section 4.

3.1. Average Weighted NMR

Average NMR has been used as an objective function. Examples can be found in [2,3]. We add a set of frequency-dependent perceptual weights to form Average Weighted NMR (AWNMR):

$$\mathrm{AWNMR} = \frac{1}{N}\sum_{i=1}^{N} \frac{\mathrm{E}((x_i - y_i)^2)}{[T_p \cdot \mathrm{E}(\underline{x}^2)]_i} \cdot 10^{(v_i + w_i)/10}\,. \qquad (2)$$

The weights are represented by w_i. They are on a dB scale and larger values represent greater listener sensitivity. The v_i are samples of the masking index. At each frequency i, v_i indicates how many dB must be subtracted from the excitation pattern $\underline{\tilde{p}}^2 = T_p \cdot \mathrm{E}(\underline{x}^2)$ to obtain the masking pattern[9]. Masking indices are based on prior knowledge of human hearing. Experiments to determine most effective values for the w_i have yet to be done. Once the w_i are determined, they can be combined with the v_i to form a single modified masking index.

Necessary conditions for the minimization of AWNMR under the bit-rate constraint in (1) can be found by invoking the Lagrange multiplier λ and solving

$$\frac{\partial}{\partial b_i}\big(\mathrm{AWNMR} + \lambda\cdot(B - B^*)\big) = 0\,,\; i = 1 \text{ to } N\,. \qquad (3)$$

This results in the closed-form bit allocations

$$b_i = \frac{B^*}{R} - \frac{1}{2R}\sum_{j=1}^{N} f_j \cdot r_j + \frac{1}{2} f_i \,,\quad i = 1 \text{ to } N, \qquad (4)$$

where $$f_i = \log_2\left(\frac{k_i^2}{\tilde{p}_i^2\, r_i}\right) + \frac{v_i + w_i}{10\cdot\log_{10}(2)}\,. \qquad (5)$$

As described in Section 4, our experiments have shown that, in practice, these necessary conditions are also sufficient. The resulting minimized value of AWNMR is

$$\mathrm{AWNMR}_{\min} = \tfrac{R}{N} \cdot 2^{\frac{1}{R}\left(\sum_{i=1}^{N} f_i \cdot r_i - 2B^*\right)} . \qquad (6)$$

3.2. Maximal Log Weighted NMR

Maximal Log NMR has been widely used as an objective function. Examples can be found in [4,5]. Again, we add a set of frequency-dependent perceptual weights represented by w_i. The weights are on a dB scale and larger values represent greater listener sensitivity. The Log Weighted NMR in the i th band (LWNMR$_i$) is given by

$$\mathrm{LWNMR}_i = 10 \cdot \log_{10}\left(\frac{\mathrm{E}((x_i - y_i)^2)}{[T_p \cdot \mathrm{E}(\underline{x}^2)]_i} \right) + (v_i + w_i) . \qquad (7)$$

Minimization of average LWNMR cannot result in meaningful bit allocations. The Maximal Log Weighted NMR (MLWNMR) is

$$\mathrm{MLWNMR} = \max_i(\mathrm{LWNMR}_i) . \qquad (8)$$

MLWNMR is minimized under the bit-rate constraint in (1) by forcing LWNMR$_\mathrm{i}$ = constant, for i = 1 to N. The resulting bit allocations are again given by (4), with

$$f_i = \log_2\left(\frac{k_i^2}{\tilde{p}_i^2} \right) + \frac{v_i + w_i}{10 \cdot \log_{10}(2)} . \qquad (9)$$

The resulting minimized value of MLWNMR is

$$\mathrm{MLWNMR}_{\min} = \frac{10}{\log_2(10)} \frac{1}{R} \left(\sum_{i=1}^{N} f_i \cdot r_i - 2B^*\right), \qquad (10)$$

using f_i as given in (9).

3.3. Maximal Log Weighted Noise-to-Signal Ratio

When the spreading of signal power and the masking index are eliminated, MLWNMR simplifies to Maximal Log Weighted Noise-to-Signal Ratio (MLWNSR):

$$\mathrm{MLWNSR} = \max_i\left(10 \cdot \log_{10}\left(\frac{\mathrm{E}((x_i - y_i)^2)}{\mathrm{E}(x_i^2)} \right) + w_i \right) . \qquad (11)$$

MLWNSR is minimized under the bit-rate constraint in (1) by the bit allocations given in (4), with

$$f_i = \log_2\left(\frac{k_i^2}{p_i^2} \right) + \frac{w_i}{10 \cdot \log_{10}(2)} . \qquad (12)$$

The resulting minimized value of MLWNSR is

$$\mathrm{MLWNSR}_{\min} = \frac{10}{\log_2(10)} \frac{1}{R} \left(\sum_{i=1}^{N} f_i \cdot r_i - 2B^*\right), \qquad (13)$$

using f_i as given in (12).

3.4. Maximal Normalized Excitation Error

Next we present an objective function based on the excitation patterns generated by the coded and uncoded signals. These patterns represent the auditory stimulation a listener would receive from the coder, and from a transparent coder. The difference between these excitation patterns is then normalized by the excitation pattern due to the uncoded signal, converted to a dB scale, and weighted. The weights are on a dB scale and larger values represent greater listener sensitivity. The Normalized Excitation Error in the i th band (NEE$_i$) is given by

$$\mathrm{NEE}_i = 10 \cdot \log_{10}\left(\frac{\left[\mathrm{E}((T_a \underline{x} - T_a \underline{y})^2)\right]_i}{[T_p \cdot \mathrm{E}(\underline{x}^2)]_i} \right) + w_i , \qquad (14)$$

and the Maximal Normalized Excitation Error (MNEE) is

$$\mathrm{MNEE} = \max_i(\mathrm{NEE}_i) . \qquad (15)$$

MNEE is minimized under the bit-rate constraint in (1) by forcing NEE$_\mathrm{i}$ = constant, for i = 1 to N. The resulting bit allocations are again given by (4), with

$$f_i = \log_2\left(k_i^2 \cdot \left[\sum_{j=1}^{N} \tilde{t}_{ij} \cdot \tilde{p}_j^2 \cdot 10^{-w_j / 10} \right]^{-1} \right) . \qquad (16)$$

The resulting minimized value of MNEE is

$$\mathrm{MNEE}_{\min} = \frac{10}{\log_2(10)} \frac{1}{R} \left(\sum_{i=1}^{N} f_i \cdot r_i - 2B^*\right), \qquad (17)$$

using f_i as given in (16). Note that if $w_i = v_i$, the linearity of T_a allows one to interpret MNEE as a modification of MLWNMR, where the quantization noise has been replaced with the excitation pattern created by the quantization noise. Note also that if w_i = constant, i = 1 to N, then the MNEE bit allocations given by (4) and (16) reduce to the MLWNSR bit allocations given by (4) and (12).

3.5. Probability of Detection

Given the excitation patterns generated by coded and uncoded signals, we can also model the probability that a difference can be detected, using generalizations of the work in [10,11]. The resulting objective function is called Probability of Detection (PDET):

$$\mathrm{PDET} = 1 - \prod_{i=1}^{N} 10^{w_i (1 - 10^{0.1 \cdot \Delta L_i})} ,$$

$$\text{where } \Delta L_i = \left| 10 \cdot \log_{10}\left(\frac{\left[T_p \cdot \mathrm{E}(\underline{y}^2)\right]_i}{\left[T_p \cdot \mathrm{E}(\underline{x}^2)\right]_i} \right) \right| \qquad (18)$$

$$= 10 \cdot \log_{10}\left(1 + \frac{1}{\tilde{p}_i^2} \sum_{j=1}^{N} t_{ij} \cdot k_j^2 \cdot 2^{-2b_j} \right) .$$

Each weight w_i is determined by the dB difference ΔL^*_{i} that is required for 50% probability of detection in the i^{th} band. Thus, larger values of w_i represent greater listener sensitivity:

$$w_i = \log_{10}(2) \Big/ (10^{0.1\cdot\Delta L_i^*} - 1)\ . \tag{19}$$

Necessary conditions for the minimization of PDET under the bit-rate constraint in (1) are derived as in Section 3.1. The resulting bit allocations are given by (4), with

$$f_i = \log_2\left(\frac{k_i^2}{r_i}\sum_{j=1}^{N}\frac{w_j\cdot t_{ji}}{\tilde{p}_j^2}\right). \tag{20}$$

As described in Section 4, our experiments have shown that in practice, these necessary conditions are also sufficient. The minimized value of PDET is found by inserting this bit allocation into (18).

3.6. Relative Excitation Sensitivity

Our final bit-allocation algorithm is an iterative one. The algorithm is initialized with an integer bit allocation that exceeds the constraint in (1) so that the quantization process is not audible. Bits are then removed one-by-one in a way that causes minimal disruption to the excitation pattern and maximal reduction of the total bit rate B.

We define the Relative Excitation Sensitivity in the i^{th} band to the j^{th} bit allocation as

$$\mathrm{RES}_{\mathrm{i}}(j) = w_i\cdot\frac{\left|\frac{\partial}{\partial b_j}10\cdot\log_{10}\left(\left[T_p\,\mathrm{E}(\underline{y}^2)\right]_i\right)\right|}{\frac{\partial}{\partial b_j}B} =$$
$$\frac{20}{\log_2(10)}\cdot\frac{w_i\,t_{ij}\,k_j^2\,2^{-2b_j}}{r_j\cdot\left(\tilde{p}_i^2+\sum_{q=1}^{N}t_{iq}\cdot k_q^2\cdot 2^{-2b_q}\right)}\ . \tag{21}$$

If decreasing the j^{th} bit allocation causes a relatively small change in the excitation pattern and/or a relatively large reduction in the total bit rate, then RES(j) will be relatively small, and b_j should be reduced. The relative impact of these two factors can be adjusted in each band using the weights w_i. If j^* minimizes $\min_j[\max_i[\mathrm{RES}_i(j)]]$, then b_{j^*} is decremented by one. Bits are selectively removed in this fashion until the bit-rate constraint in (1) is satisfied.

4. OBSERVATIONS

Equation (4) describes the bit allocations for five of these six algorithms. This equation agrees with our intuitions about bit allocations: Each allocation starts with B^*/R bits/sample, the global average. Next, each allocation is decremented by a factor that depends on a weighted average of the f_i for all bands. Finally, the i^{th} allocation is increased according to f_i. Because the f_i are all in some way inversely related to signal power, fewer bits are allocated to more powerful signal components. This is what we would expect based on the formulation of the objective functions, and our understanding of human hearing. Note that these bit allocations must be rounded to integer values in practical implementations. Equations (10), (13), and (17) show that the minimized distortion vs. rate curves for the objective functions that use decibel units will have slopes of -20/$\log_2$(10) = -6.02 dB/bit, as expected.

We have implemented all six algorithms, and have used them to allocate bits in N = 25 bands (each is one Bark wide) for 6860 audio frames (each frame covers 11.6 ms) taken from 20 musical and spoken selections. We verified the sufficiency of the necessary conditions given in Sections 3.1 and 3.5 by perturbing those bit allocations, and noting that those objective functions always increase in response to perturbations.

We plan to continue this work and quantize these 20 musical and spoken selections according to these six bit allocations. This will be followed by formal listening experiments to select values for the perceptual weights w_i , and to determine the relative encoded audio quality provided by each of the bit-allocation algorithms.

REFERENCES

1. Brandenburg, K. "Evaluation of quality of audio encoding at low bit rates," presented at the 82nd Convention of the Audio Engineering Society, London, 1987.
2. Mahieux, Y. & Petit, J.P. "Transform coding of audio signals at 64 kbit/s," in *Proc. IEEE Globcom `90*, 1990, pp. 518-521.
3. Sinha, D. & Tewfik, A.H. "Low bit rate transparent audio compression using adapted wavelets," *IEEE Trans. SP, vol. 41*, pp. 3463-3479, 1993.
4. Van der Waal, R.G. & Veldhuis, R.N.J. "Subband coding of stereophonic digital audio signals," *Proc. IEEE ICASSP `91*, 1991, pp. 3601-3604.
5. ISO/IEC 11172-3, "Coding of moving pictures and associated audio…, Part 3: Audio," 1993.
6. Shoham, Y. & Gersho, A. "Efficient bit allocation for an arbitrary set of quantizers," *IEEE Trans. ASSP*, vol. 36, pp. 1445-1453, 1988.
7. Westerink, P.H., et al.. "An optimal bit allocation algorithm for sub-band coding," in *Proc. IEEE ICASSP `88*, 1988, pp. 757-760.
8. Riskin, E.A. "Optimal bit allocation via the generalized BFOS algorithm," *IEEE Trans. IT*, vol. 37, pp. 400-402, 1991.
9. Voran, S. "Observations on auditory excitation and masking patterns," in *Proc. 1995 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, 1995.
10. Buss, S., et al. "Decision rules in detection of simple and complex tones," *J. Acoust. Soc. Am.*, vol. 80, pp. 1646-1657, 1986.
11. Colomes, C., et al. "A perceptual model applied to audio bit-rate reduction," *J. Audio Eng. Soc.*, vol. 43, pp. 233-239, 1995.