Proc. IEEE Int. Conf. Acoustics, Speech, Signal Processing (Istanbul), pp. 893–896, June 2000

PERCEPTUAL BIT ALLOCATION FOR LOW RATE CODING OF NARROWBAND AUDIO

Hossein Najafzadeh and Peter Kabal

Department of Electrical & Computer Engineering
McGill University, Montreal, Canada

ABSTRACT

In this work we consider adaptive bit allocation for perceptual coding of narrowband audio signals at low rates (down to 8 kb/s). Two different strategies are used to shape the audible noise spectrum. In one approach, the quantization noise spectrum is shaped in parallel with the masking threshold curve. This way the noise is equally audible in different frequency bands. The other approach generates a flat noise spectrum above the masking threshold. The noise power is not equally distributed over the frequency range, hence it is audible to various extents at different frequencies.

1 INTRODUCTION

In transform audio coding, the input signal is decomposed into the frequency components. The spectral components are adaptively coded using quantizers with different accuracies based on the short-term spectrum.

In traditional transform coders, bit assignment is performed based on the distribution of the signal power in the frequency domain to minimize the total noise power [1, 2]. Most audio signals are lowpass. For a limited total number of bits, relatively few bits are assigned to high frequency components which may result in a loss of high frequencies. Moreover, the masking phenomena are not fully exploited which often results in allocating bits to the spectral components which are inaudible.

Since the final goal of audio coding is to achieve the best possible quality for a fixed bit rate, the spectrum of the quantization error should be shaped based on perceptual principles. In adaptive bit assignment the coding noise can be shaped to be less audible than a noise with the same energy without noise shaping. In low rate coding of audio signals, due to the scarcity of bits, audible quantization noise is often inevitable. The final goal is to deliver acceptable quality with no annoying artifacts. This contrasts with the requirement for transparent coding in high rate wideband audio coding.

Two different strategies are considered to shape the audible noise spectrum [3]. In one approach, the quantization noise spectrum is shaped in parallel with the masking threshold curve. This way the noise is equally audible in different frequency bands. The second approach is to generate a flat noise spectrum above the masking threshold. According to [3, pp. 427–428], these two approaches are different in terms of auditory object formation. In the first approach, the quantization noise has a temporal modulation similar to that of the input signal. Therefore, the input signal and the noise will be perceptually fused to form one auditory object. In the second approach, the noise power is not equally distributed over the frequency range, hence it is audible to various extents at different frequencies. This way, the noise remains perceptually distinct from the input signal.

2 BIT ALLOCATION ALGORITHMS

In perceptual coding, a frequency-dependent masking level, determined from the excitation level minus an offset (in dB) [4], is calculated. Only quantization noise above the masking level is audible.

In the narrowband perceptual audio coder proposed in [5] the bit assignment is done both at the transmitter and the receiver. The masking thresholds are calculated using the quantized gain factors. Note that for each band we need to specify the offset value which is subtracted from the excitation level (in dB) in order to obtain the masking threshold. The offset value depends on whether the spectrum in each band is tone-like or noise-like. One way to identify the nature of the signal is to use the tonality factor defined as the ratio of geometric mean to the arithmetic mean of the spectral components [4][1]. The scheme proposed by Johnston [4] assigns a single tonality factor to a frame of audio without discriminating between different parts of the spectrum. We wish to have different tonalities in different frequency bands allowing for frequency-dependent offsets. Such frequency-dependent offsets based on the local spectral structure are used in audio coders such as the MPEG standard [6]. At low bit-rates we can not afford to code the offset value for each band. However in order to take into consideration the fact that noise-like signals offers higher masking levels, equivalently to a smaller offset value, compared to those of tone-like signals, we distinguish between two cases. In one case the input block of data has a harmonic structure which implies that th spectrum is more tone-like. In the

Web page: `http://WWW.TSP.ECE.McGill.CA`.

[1] The offset value is related to the tonality factor by

$$O_i = (14.5 + i)a + 5.5(1 - a)$$

where i is the critical band index, O_i is the offset value for critical band i and a is the tonality factor.

other case the input has a more noise-like spectrum.

In order to switch between the two cases, we measure the similarity between the spectral shapes for successive frames of data. If the average mean squared difference between the successive spectra (in the log domain) is less than 6 dB, we consider the current frame of data as tone-like. For tone-like frames of data, since the signal is more tone-like in the low frequency bands than the high frequency bands, we assume higher offset values for the low frequency bands. The offset value varies from 18 dB to 10 dB which is equivalent to a tonality factor that varies from 1 to 0.17 (for low to high frequency bands). By doing so, we assign more bits to the low frequency bands to maintain the pitch of highly structured signals such as voiced speech. For noise-like frames of data, we set the masking threshold for all bands 8 dB below the excitation level. Fig. 1 shows the offset values for different frequency bands.

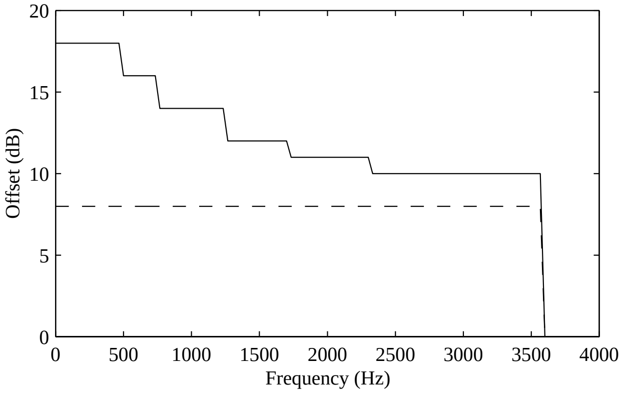

Fig. 1 Offset values for calculating the masking threshold for tone-like frames (solid lines) and noise-like frames (dashed lines).

Critical Band Rate-Distortion Relationship

In the narrowband audio coder proposed in [5], 120 transform coefficients are grouped into the critical bands and then vector quantized using an embedded codebook. Since the numbers of transform coefficients in the critical bands are not equal, we need the rate-distortion relationship for each band. A large set of vectors is used to measure the average distortion as different numbers of bits are assigned to each band. The rate-distortion data can be well represented by a line fitted to the experimental data. The correlation between the experimental data and the fitted lines for all the bands is more than 0.99. Note that all shape vectors in the test set are normalized and distortion is defined as the average energy of the quantization noise in decibels. As an example Fig. 2 shows the rate-distortion data for the codebook corresponding to critical band 2 which contains 3 coefficients. The slope of the line which has been fitted to the curve is -2.7 dB/bit.

In each band the distance between the energy and the masking threshold is upper bounded by the offset value (in dB). Hence the maximum number of bits allocated to each band is determined by dividing the corresponding offset value (in dB) by the distortion reduction rate. Fig. 3 shows the maximum number of bits allocated to each transform coefficient for tone-like and noise-like frames.

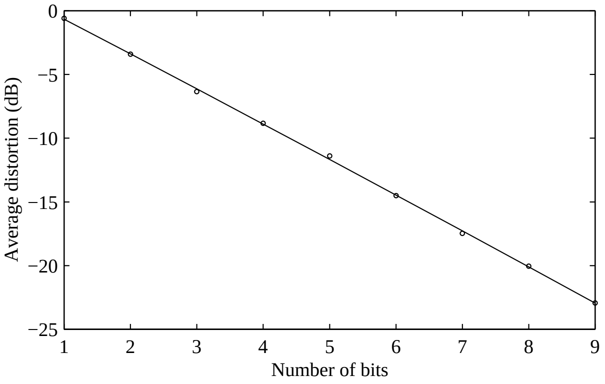

Fig. 2 Rate-distortion data for the embedded codebook corresponding to critical band 2 which contains 3 coefficients and its linear approximation.

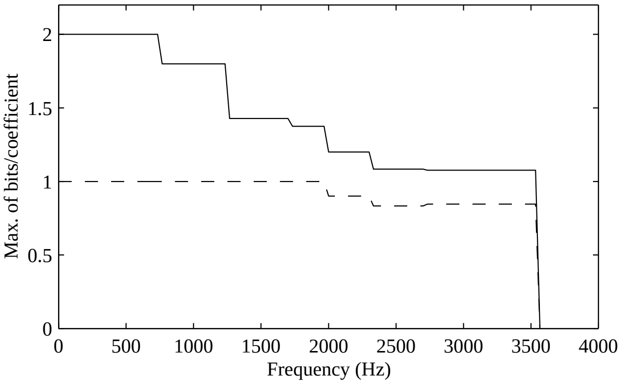

Fig. 3 Maximum number of bits per coefficient for tone-like frames (solid lines) and noise-like frames (dashed lines).

2.1 Signal-to-Mask Ratio (SMR)-based Bit Allocation

In this approach bit allocation is performed based on the Signal-to-Mask Ratio (SMR). This way, the resulting noise spectrum will be parallel to the masking threshold curve. Each critical band is considered as a single entity with its corresponding SMR. The SMR equals to the SNR when the quantization noise is at the threshold of audibility, i.e., when the noise level is at the masking threshold. The SMR for each band is calculated in the following manner

$$\text{SMR}_j = \hat{E}_j - T_j \qquad (1)$$

where $\hat{E}_j$ is the quantized log energy in band j, and T_j is the log masking threshold in that band. We assume that the initial distortion (in the log domain) for each band is equal to the corresponding SMR. A "greedy algorithm"[2] using the rate-distortion data can be employed to assign one bit at a time to the band with the largest (updated)*Noise-to-Mask Ratio (NMR)*. After assigning one bit to that band, its NMR on the average decreases by the amount given by the corresponding rate-distortion curve.

As a shortcut, a linear approximation of the rate-distortion data along with the values of SMR_j's can be used to allocate bits to each band according to the following formula

$$b_j = \max(\frac{\mathrm{SMR}_j b_T}{\lambda_j \sum_{i \in \Omega} (\mathrm{SMR}_i / \lambda_i)}, 0) \tag{2}$$

where Ω contains the indices of the bands with positive SMR and b_T is the total number of bits available to quantize the shape of the frequency spectrum within the critical bands. The slope of the rate-distortion line, λ_j, indicates the approximate reduction in the Noise-to-Mask Ratio (NMR) for one bit assigned to band j. Note that no bits are assigned to those bands whose SMR is negative. After the first round of bit allocation, the fractional parts of b_j's will be discarded to leave the integer parts. Therefore the total number of bits allocated in the first step will be less than b_T. To allocate the remaining bits, the Noise-to-Mask Ratio (NMR) is approximated for each band taking into account the bits already allocated in the first step,

$$\mathrm{NMR}_j = \hat{E}_j - T_j - \lambda_j b_j \tag{3}$$

After calculating the value of NMR's, one bit at a time is allocated to the band with the largest value of the updated NMR. This process will continue until all remaining bits are allocated.

2.2 Energy-based Bit Allocation

In this approach, bit assignment is performed based on the energy above the masking threshold. The distortion is considered as the audible part of the quantization noise, i.e. the noise above the masking threshold.

The level of audible noise will be relatively higher in spectrum valleys due to the fact that there is less energy above the masking threshold compared to unmasked spectral peaks. We consider two schemes to minimize the audible noise. In the first scheme the maximum of the distortion in the critical bands is minimized. In the second scheme the total audible noise is minimized.

Mini-Max Scheme

The mini-max bit assignment is done through the following optimization procedure

$$\arg\min_{b_j}(\max(D_j(bj))) \quad \text{s.t.} \quad \sum_{j=1}^{N_b} b_j = b_T \tag{4}$$

where N_b is the number of bands, b_T is the total number of bits available for each frame and D_j is the noise above the masking threshold.

We use a "greedy algorithm" to do the bit assignment. This way, one bit at a time is assigned to the band with the *largest updated distortion*.

Total Audible Distortion Minimization Scheme

The scheme minimizes the total audible distortion. Therefore the optimization objective function changes to

$$\arg\min_{b_i} \sum_{i=1}^{N_b} D_i \quad \text{s.t.} \quad \sum_{i=1}^{N_b} b_i = b_T \tag{5}$$

According to this approach, one bit at a time goes to the band which renders the *largest reduction* in distortion. This algorithm can be performed using either a greedy or an analytical approach. In the analytical algorithm, the energy above the masking threshold is related to the audible distortion through the following empirical formula

$$D_i = c_i \, \mathcal{E}_i \, 2^{-b_i/\beta_i} \tag{6}$$

where D_i is the energy of the audible noise in band i, $\mathcal{E}_i$ is the energy above the masking threshold, c_i and β_i are constants found from the corresponding rate-distortion curve for the codebook of band i.

The solution to the above is given by

$$b_i = \max(\frac{\beta_i b_T}{\sum_{j=1}^{N_b} \beta_j} + \log_2(\frac{\mathcal{E}_i c_i}{\mathcal{E}_{gm}}), 0) \tag{7}$$

where

$$\mathcal{E}_{gm} = \left(\prod_{i=1}^{N_b} (c_i \mathcal{E}_i)^{\beta_i} \right)^{\left(1/ \sum_{i=1}^{N_b} \beta_i \right)} \tag{8}$$

The integer parts of the b_i's are kept and the remaining bits will be distributed one at a time to the band which reduces the total distortion the most.

3 SUBJECTIVE EVALUATION OF THE BIT ASSIGNMENT ALGORITHMS

We ran an informal listening test to evaluate subjective preferences. We used the perceptual bit assignment schemes, i.e. energy-based approach (the mini-max scheme and the minimization of the total distortion scheme) and the SMR-based algorithm to code two speech files (male and female) and two pieces of music (soprano and guitar). We examined the impact of the masking effects on the quality of the decoded signals. In that test, we ignored the masking effects and performed the bit assignment based on the distribution of the signal power.

In the experiments the narrowband audio coder proposed in [5] was used. Operating at 8 kb/s (120 bits per frame),

[2] Note that the total distortion is a convex function of the bi's.

the coder assigns 80 bits to 17 bands to quantize the spectral shapes. The unquantized adjusted gains[3] were used to denormalize the quantized shape vectors.

Using the power-based bit allocation without involving the masking threshold, for all audio segments, the resulting outputs have a higher SNR compared to the outputs using perceptual bit allocation. However, because a power-based scheme allocates many bits to the low frequency bands and relatively few bits to the high frequency bands, the outputs suffer from incomplete coding of the high frequencies. This result verifies the importance of incorporating the masking effects into any bit assignment algorithm.

The energy-based algorithm which minimizes the total audible distortion resulted in output quality similar to that for the power-based bit allocation. Due to different dimensionality of different critical bands, the distortion reduction rate is higher for the narrower low frequency bands. Moreover for many audio signals, the power is concentrated in the low frequency bands. Therefore, more bits compared to other perceptual schemes are assigned to the low frequency bands. This results in finer quantization of low frequency bands and coarser quantization of the high frequency bands.

The other schemes (the SMR-based and the mini-max) deliver better quality with less high frequency distortion. The results show that both algorithms produce decoded signals which can be distinguished from the original. The SMR-based algorithm causes less high frequency distortion at the expense of a little degradation in the pitch structure. Due to this degradation, the speech segments which are coded using the SMR-based algorithm sound more harsh. On the other hand, the decoded audio signals using the energy-based algorithm carry higher levels of high frequency noise which sounds like an echo along with the original signal. Listeners showed a slight preference for the SMR-based allocation scheme over the mini-max scheme.

4 CONCLUSIONS

We have considered different bit allocation algorithms suitable for low rate audio coding. The SMR-based coder tends to maintain equal Noise-to-Mask Ratio (NMR) for different bands. This gives equal sensation of the audible distortion in different critical bands. The energy-based algorithm allocates bits based on the energy above the masking threshold. This is in contrast with the SMR-based approach in which the *ratio* of the energy to the masking threshold is used to perform bit assignment. Due to the lowpass nature of most audio signals, the energy-based algorithm assigns more bits to low frequency components. As a result of the fewer bits assigned to high frequency bands, the decoded signal sounds like a relatively clean lowpass signal along with high frequency noise.

Comparing these algorithms, the SMR-based algorithm is more suitable for low rate audio coding. However, we believe that the perceptually optimal bit allocation algorithm for low rate coding should be based on both the distribution of the audible noise and the SMR. This is a compromise between the schemes that may be better than either approach alone.

References

[1] R. Zelinski and P. Noll, “Adaptive transform coding of speech signals,” *IEEE Trans. Acoustics, Speech, Signal Processing*, vol. 25, pp. 299–309, Aug. 1977.

[2] R. Zelinski and P. Noll, “Approaches to adaptive transform speech coding at low bit rates,” *IEEE Trans. Acoustics, Speech, Signal Processing*, vol. 27, pp. 89–95, Feb. 1979.

[3] R. Veldhuis and A. Kohlrausch, “Waveform Coding and Auditory Masking,” in *Speech Coding and Synthesis* (W. B. Kleijn and K. K. Paliwal, eds.), pp. 427–428, Elsevier, 1995.

[4] J. D. Johnston, “Transform coding of audio signals using the perceptual noise criteria,” *IEEE J. Selected Areas in Comm.*, vol. 6, pp. 314–323, Feb. 1988.

[5] H. Najafzadeh-Azghandi and P. Kabal, “Perceptual Coding of Narrowband Audio Signals at 8 kb/s,” *Proc. IEEE Workshop on Speech Coding* (Pocono Manor, Penn.), pp. 109–110, 1997.

[6] ISO/IEC JTC1/SC29/WG 11, *Coding of Moving Pictures and Associated Audio*, 1993.

[3] The quantization error is reduced by adjusting the gain in each critical band through the following optimization

$$\rho_{\mathrm{opt}} \approx \arg\min_{\rho} \sum_{k=1}^{K} (X(k) - \rho\hat{X}(k))^2$$

where ρ is the gain adjustment factor, $X, \hat{X}$ are the original and quantized vectors of transform coefficients and K is the dimension of the subvector.