OPTIMUM BIT ALLOCATION AND DECOMPOSITION FOR HIGH QUALITY AUDIO CODING

Xiang Wei, Martyn J. Shaw and Martin R. Varley

Department of Electrical and Electronic Engineering,
University of Central Lancashire, Preston PR1 2HE, UK
x.wei1@uclan.ac.uk, m.j.shaw@uclan.ac.uk, m.r.varley@uclan.ac.uk
http://www.uclan.ac.uk/facs/destech/eleceng/staff.htm

ABSTRACT

Current audio compression schemes are capable of reducing the per channel bit rate of high quality audio signals from 16 bits per sample to around 2-4 bits per sample. In these schemes, knowledge of psychoacoustics is utilised and a uniform or nonuniform frequency decomposition method is used. In this paper we derive the optimum bit allocation to achieve the highest perceptual quality under a fixed bit rate, for an arbitrarily decomposed, critically sampled, filter bank. The resultant optimum bit allocation gives rise to a shaped reconstruction noise floor approximately parallel to the masking threshold level. Perceptual coding gain is defined and should be maximized for an optimum decomposition performed by the filter bank. Optimum band splitting is discussed and it is pointed out that decomposition in the manner of critical band splitting does not lead to optimal performance.

1. INTRODUCTION

Some of the current audio compression schemes, e.g. MPEG Layer I, II and III [1], utilise filter banks and knowledge of psychoacoustics to reduce the bit rate of high quality audio from 16 bits/sample to around 2-4 bits/sample. In these schemes, filter banks are used to split the signals into a number of frequency channels, enabling each channel to be quantized independently so that the quantization noise will be confined to this channel and be perceptually masked. The filter banks commonly in use are either uniform or nonuniform, but with fixed frequency resolution. Recently, it has been pointed out by Princen and Johnston [2] that resolution of the filter bank must vary with time if the system is to achieve optimum performance on a wide variety of signal types.

In an audio coding scheme, the performance is closely related to the bit allocation process, which assigns a certain number of bits to every channel according to a certain criterion. Usually, the bit allocation scheme is based on one of the following two optimization criteria.

- The reconstruction error is the least in terms of mean square error (MSE), or the signal to reconstruction error ratio (SNR) is the largest. This criterion was used in early systems for speech coding [3]. It is found that bit allocation under this criterion results in a flat reconstruction noise floor in the frequency domain.
- The reconstruction error is the least detectable by the human receiver. This criterion is important in digital audio coding because of the ultimate requirement of perceptual quality. It will be explained that this criterion is interpreted as having the least reconstruction error to masking threshold ratio (or noise to mask ratio, NMR), and bit allocation under this criterion results in a shaped reconstruction noise floor approximately parallel to the masking threshold level.

Though optimum bit allocation is not a new subject [4], and the optimum bit allocation under the least MSE (LMSE) criterion was studied in [5] for paraunitary analysis/synthesis systems (which give rise to perfect reconstruction), there has not been any literature about the optimum bit allocation specially for the least NMR (LNMR) criterion. In this paper, we'll derive the optimum bit allocation for the LNMR criterion and show that the practical bit allocation procedure used in MPEG/audio [1] can be modified to suit general systems.

Similar to the conventional concept of coding gain [3] defined under the LMSE criterion, a novel term 'perceptual coding gain' will be defined. We suggest that the optimum decomposition performed by the filter bank is to achieve the maximum perceptual coding gain, which results from the optimum bit allocation under LNMR criterion. Results obtained from calculating maximum perceptual coding gains for various audio signals show that the decomposition into auditory critical bands does not lead to the optimum performance.

Before we move on, some assumptions are given as follows. Assume the M-channel filter bank used in the subband coding is critically sampled and the decimation factor for channel i is κ_i.

Assume the decomposed samples in channel i are encoded to R_i bits/sample. Consequently, the average bit rate of the encoded signal is

$$R = \sum_{i=0}^{M-1} \frac{R_i}{\kappa_i} \quad \text{bits/sample} \qquad (1)$$

where κ_i satisfies $\sum_{i=0}^{M-1}\frac{1}{\kappa_i}=1$ to ensure critical sampling.

Assume the input audio signal variance is σ_x^2. In an R bits/sample PCM coder, the quantization error variance is given by [3]

$$\sigma_r^2=\varepsilon^2 2^{-2R}\sigma_x^2$$

where the constant ε^2 is the quantizer performance factor.

Assume the filter bank renders the decomposed signals uncorrelated with each other. As a result, the variance σ_x^2 of the input signal will be the summation of the variances σ_{xi}^2 of all its decomposed channels [3], i.e.

$$\sigma_x^2=\sum_{i=0}^{M-1}\sigma_{xi}^2$$

For channel i, assume the samples are quantized to R_i bits/sample. The reconstruction error variance is,

$$\sigma_{ri}^2=\varepsilon^2 2^{-2R_i}\sigma_{xi}^2$$

2. UNIFORM SUBBAND CODING FOR LNMR

Based on psychoacoustics, in audio coding, as long as the quantization noise is lower than the masking threshold at any frequency, the noise will be inaudible. Since the masking threshold varies with frequency, a noise floor that changes with the masking threshold would result in a lower bit rate for given perceptual effect.

For a given input audio power spectral density (psd) $S(e^{j\Omega})$, using a psychoacoustic model [1], we are able to obtain the masking threshold density $T(e^{j\Omega})$. We define the signal to mask ratio density (*SMRD*) as the ratio of signal psd and masking threshold density, i.e.

$$SMRD(e^{j\Omega})=S(e^{j\Omega})\,/\,T(e^{j\Omega})$$

In subband coding, the i^{th} channel covers frequency range $[\Omega_i, \Omega_{i+1}]$ and the masking threshold is given by [1]

$$\tau_i=\frac{\Omega_{i+1}-\Omega_i}{\pi}\min_{\Omega_i\le\Omega<\Omega_{i+1}}T\left(e^{j\Omega}\right)$$

and the signal to mask ratio (SMR) is given by

$$\text{SMR}_i=\sigma_{xi}^2/\tau_i=w_i\sigma_{xi}^2$$

where the weightings $w_i=1/\tau_i$.

In order to make the reconstruction error least detectable by the human ear, the NMR in each channel must be minimised. The purpose of coding is to minimise the total NMR; taking account of the bandwidth weightings on the individual NMRs, the problem is given as

$$\text{Minimize NMR}=\sum_{i=0}^{M-1}\frac{w_i}{\kappa_i}\sigma_{ri}^2=\sum_{i=0}^{M-1}\frac{w_i}{\kappa_i}\varepsilon^2 2^{-2R_i}\sigma_{xi}^2$$

Under the condition of equation (1).

Using the method of Lagrange multiplies [3], we yield

$$R_{i,opt}=\frac{\left(w_i\sigma_{xi}^2\right)(dB)}{6.02}-c=\frac{\text{SMR}_i}{6.02}-c \quad \text{bits/sample} \qquad (2)$$

where $c=-R+\frac{1}{2}\sum_{l=0}^{M-1}\frac{1}{\kappa_l}\log_2 w_l\sigma_{xl}^2$ is a constant.

Obviously, the bit allocation for uniform subband coding is a special case by replacing κ_i with M.

The error variance in channel i under the optimum bit allocation can be calculated as

$$\sigma_{ri,opt}^2=\frac{1}{w_i}\varepsilon^2 2^{2c}$$

It can be found that the weighted noise variance $w_i\sigma_{ri}^2$, or the NMR value in each channel, is a constant. This reveals that the noise floor is shaped in parallel to the masking threshold. In coding with transparent quality, the noise floor must not be higher than the masking threshold, or equivalently, $w_i\sigma_{ri}^2$ (dB) must not be greater than 0.

Because the purpose of the noise shaping is to achieve higher perceptual quality, the coding gain defined for LMSE criterion does not match this purpose. Now, we define the perceptual coding gain G as the ratio of total NMR in PCM and total NMR in subband coding (SB), taking bandwidth weightings into consideration,

$$G=\frac{\sum_{i=0}^{M-1}\frac{w_i}{\kappa_i}\sigma_{ri,PCM}^2}{\sum_{i=0}^{M-1}\frac{w_i}{\kappa_i}\sigma_{ri,SB}^2}=\frac{\sum_{i=0}^{M-1}\frac{w_i}{\kappa_i^2}\sigma_{r,PCM}^2}{\sum_{i=0}^{M-1}\frac{w_i}{\kappa_i}\sigma_{ri,SB}^2}$$

Substituting the optimum bit allocation results, we obtain the maximum perceptual coding gain as

$$G=\frac{\sigma_x^2\sum_{i=0}^{M-1}\frac{w_i}{\kappa_i^2}}{\prod_{i=0}^{M-1}\left[w_i\sigma_{xi}^2\right]^{1/\kappa_i}}=\frac{\sum_{i=0}^{M-1}\frac{w_i}{\kappa_i^2}}{\prod_{i=0}^{M-1}\left[\frac{w_i}{\kappa_i}\right]^{1/\kappa_i}}\;\frac{\sum_{i=0}^{M-1}\sigma_{xi}^2}{\prod_{i=0}^{M-1}\left[\kappa_i\sigma_{xi}^2\right]^{1/\kappa_i}}$$

Since the ratio of weighted arithmetic mean is larger than or equal to the weighted geometric mean for positive values [6], this gain is always larger than 1 for a non-flat noise shaping.

Note that the optimum bit allocations calculated from (2) may be less than zero. In practice the bit allocation in each channel is required to be non-negative with an upper limit.

Because of (2), interpretation to a practical bit allocation for a 4-band system can be made through Fig. 1, which is a similar illustration of bit allocation for LMSE made in [7]. The vertical axis for the thick curve is SMR, or the weighted signal variance $w_i\sigma_{xi}^2$. The vertical axis for the thin curve is the SMR density (SMRD). The dashed lines represent bit allocation decision thresholds λ_k, which are 6.02 dB apart from each other. From (2), the decision threshold λ_0 is initially assigned

$$\lambda_0=6.02c=10\log_{10}(w_i\sigma_{ri,opt}^2)-10\log_{10}\varepsilon^2 \text{ dB}$$

Because non-negative bit allocation is required, the decision threshold λ_0 may need to be adjusted. In the figure, NMR floor is shown to be parallel to the level of λ_0.

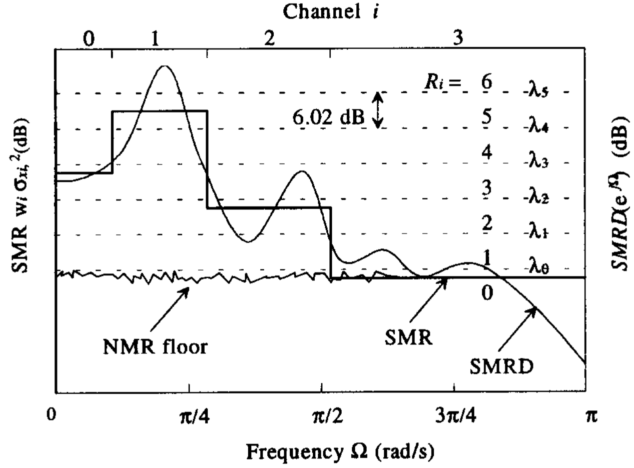

Fig. 1 Illustration of bit allocation for subband coding

In practice, integer bit allocation can be carried out using iterative procedures similar to the one used in NMR bit allocation for uniform subband coding, as conducted in [1]. In each iteration stage, the channel which has the highest NMR value will be given one more bit allocation. Note that different channels contain different numbers of decomposed samples.

3. OPTIMUM BAND SPLITTING FOR LNMR

In the previous section, we have discussed the optimum bit allocation and maximum coding gain for uniform and nonuniform subband coding to achieve LNMR. A question is raised on how to split the frequency band to achieve LNMR at a given average bit rate. The following is a discussion regarding this question.

As implied in equation (1), the number of bits used by channel i is proportional to R_i/κ_i. Since $\kappa_i = \pi/\Delta\Omega_i$ and because of (2), the number of bits used by channel i is proportional to $\Delta\Omega_i \cdot (\mathrm{SMR}_i - \lambda_0)$, which is proportional to the area below the thick line and above the lowest decision threshold λ_0 in channel i in Fig. 1. We call this area 'bits reserve area'.

To obtain the lowest bit rate for a given NMR criterion, we need to have the smallest bits reserve area given by

$$A = \sum_{i=0}^{M-1} \max\ \{0,\ \Delta\Omega_i \cdot (\mathrm{SMR}_i - \lambda_0)\}$$

Consider an arbitrary band of SMR density spectrum $SMRD(e^{j\Omega})$, $\Omega_a \le \Omega \le \Omega_b$, which in numerical calculation, is rewritten as $SMRD(k)$, where k is the spectral line index and $k_a \le k \le k_b$. Assume the SMR density spectrum is above λ_0.

If this band is regarded as a single channel, the bit reserve area A within this channel is yielded as

$$A_s \ge (k_b - k_a) \cdot 10\log_{10}\left[\frac{1}{k_b - k_a}\sum_{k=k_a}^{k_b} SMRD(k)\right] - (k_b - k_a)\lambda_0$$

If this band is regarded as k_b - k_a uniform channels, the bit reserve area A is obtained as

$$A_m = 10\log_{10}\left[\prod_{k=k_a}^{k_b} SMRD(k)\right] - (k_b - k_a)\,\lambda_0$$

Because the arithmetic mean is equal to or larger than the geometric mean for positive values, i.e.

$$\frac{1}{k_b - k_a}\sum_{k=k_a}^{k_b} SMRD(k) \ge \left[\prod_{k=k_a}^{k_b} SMRD(k)\right]^{\frac{1}{k_b - k_a}}$$

we have

$$A_m \le A_s$$

The equal occurs only if this band of spectrum is flat. This implies that to achieve the least NMR or the highest perceptual coding gain, we have the following band splitting rules.

- In a channel, if the SMR is not flat, split the channel.
- If the SMR over two adjacent channels is flat, we can either combine these two channels together to make a single channel or leave them as separate channels.

Note that the above rules only apply to cases where the SMR density spectrum is higher than λ_0. Otherwise, the continuous parts of the spectrum that are lower than λ_0 can be formed into a single channel.

Consequently, in terms of NMR, the highest perceptual coding gain is achieved when the band splitting scheme is one where the SMR is flat in each channel whose SMR density spectrum is above λ_0. Obviously, for general audio signals, the perceptual coding gain improves with larger M, namely with a better frequency resolution.

In certain applications a fixed M is required [8], [9]. The optimum band splitting under these circumstances is left as further work.

4. DISCUSSIONS ON BAND SPLITTING

In psychoacoustics, the critical band rate reflects the auditory frequency resolution which is about 100 Hz at low frequencies and about 3500 Hz at high frequencies [10]. However, in terms of perceptual coding gain, on many occasions a nonuniform band splitting according to the human auditory critical band rate is not the best choice, because such a band-splitting results in very coarse frequency resolution in high frequencies where the SMR curve is far from flat. This can be seen from the example shown in Fig. 2, where the SMRD spectrum of a violin signal is shown, and the critical band division is also depicted.

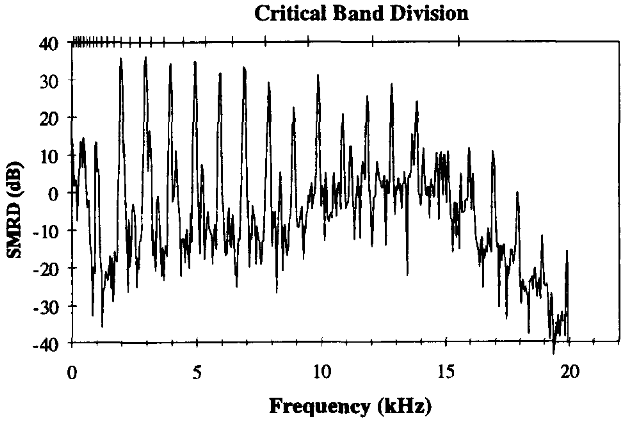

Fig. 2 Critical band division of the SMRD spectrum

In most cases, critical band splitting is worse than a uniform splitting with the same number of channels. To confirm this, a program in the C language was written to calculate the perceptual coding gains for 25-channel uniform band splitting and for 25-channel critical band splitting. The program adopts the MPEG psychoacoustic model 2 [1] to calculate the masking thresholds. The perceptual coding gain for each section of 384 input samples is calculated. In Fig. 3, the perceptual coding gains obtained for the piece of violin signal is shown. It shows that the uniform band splitting is superior to the critical band splitting. Results obtained for some of other audio signals reveal the same fact.

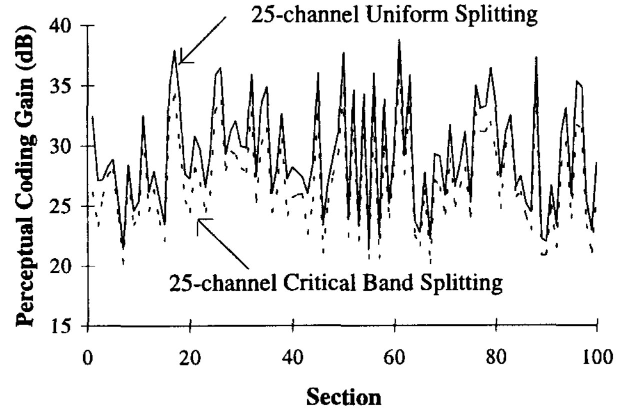

Fig. 3 Perceptual coding gains for a piece of violin signal

In the previous section, it was pointed out that band splitting into a larger number of channels gives better perceptual coding gain than a smaller number of channels. Since it is difficult to design a filter bank with a large number of channels, transforms are usually used for providing better frequency resolution. It is confirmed in [11] that transform coders tend to achieve better quality than uniform subband coders. However, the drawback of transform coders is that they lack time resolution and may cause pre-echo effect for signals with a period of silence followed by an attack [11].

Since audio signals are non-stationary and their SMR density changes with time, an adaptive nonuniform subband coding is desirable. However, this is a formidable task since an arbitrary nonuniform filter bank is difficult to design [9].

A feasible approach is to design an adaptive filter bank which enables a number of choices of band splitting, and the selection of a choice is decided by the psychoacoustic analysis of time-frequency energy distribution of the signal. For example, an implementation by Princen and Johnston [2] on an audio coder with adaptive filter banks gives promising performance which is better than MPEG Layer III at very low bit rates of 64 and 48 kbps for mono audio signals.

Acknowledgement: Xiang Wei is supported by the University of Central Lancashire.

REFERENCES

[1] Brandenburg K and Stoll G, "ISO-MPEG-1 audio: a generic standard for coding of high-quality digital audio", *J. Audio Eng. Soc.*, vol 42, no 10, pp 780-792, Oct. 1994.

[2] Princen J and Johnston JD, "Audio coding with signal adaptive filterbanks", *ICASSP-95*, pp 3071-3074, 1995.

[3] Jayant NS and Noll P, *Digital Coding of Waveforms, Principles and Applications to Speech and Video*, Prentice Hall, Englewood Cliffs, New Jersey, 1984.

[4] Huang JJ and Schultheiss PM, "Block quantization of correlated Gaussian random variables," *IEEE Trans. Communi. Syst.*, vol 11, pp 289-296, Sept. 1963.

[5] Soman AK and Vaidyanathan PP, "Coding gain in paraunitary analysis/synthesis systems", *IEEE Trans. Signal Processing*, vol 41, no 5, pp 1824-1835, May 1993.

[6] Taylor AE and Mann WR, *Advanced Calculus*, Xerox College Publication, 1972.

[7] Tribolet JM and Crochiere RE, "Frequency domain coding of speech," *IEEE Trans. Acoust., Speech, Signal Processing*, vol 27, no 8, pp 512-530, Oct. 1979.

[8] Shaw MJ, Wei X and Varley MR, "Nonuniform subband coding of audio signals employing frequency warping", to be published in *Applied Signal Processing*, 1996.

[9] Wei X, *Nonuniform Subband Coding of High Quality Audio Signals Employing Frequency Warping*, PhD thesis, University of Central Lancashire, 1996.

[10]Scharf B, "Critical bands", in vol 1 of *Foundations of Modern Auditory Theory*, edited by Tobias JV, Academic Press, pp 159-202, 1970.

[11]Musmann HG, "The ISO audio coding standard", *Globecom-90*, pp 511-517, 1990.