A BS.1387 Based Psychoacoustic Model for Perceptual Audio Coder

Xiaopeng Hu, Guiming He, and Xiaoping Zhou

Abstract—This paper attempts to combine the basic version of psychoacoustic model for audio quality evaluation in BS.1387 with perceptual audio coder. First, the principle of this new psychoacoustic model is analyzed in theory, and then corresponding improvements are proposed to make it be effectively applied to actual audio coder. Finally, both this new model and MPEG-2 psychoacoustic model 2 are implemented in the latest AVS reference coder of China, comparison of output masking parameter and subjective hearing test between the two models are conducted. Experimental results show that the proposed psychoacoustic model is feasible.

I. INTRODUCTION

PERCEPTUAL audio coding exploit the properties of human auditory perception to make coding loss perceptually indistinguishable as far as possible. Since middle of 20th century, human beings have realized the importance of human auditory perception for audio compression[1][2][3], and the theory foundation of psychoacoustic model has been developed gradually. MPEG-1 ISO/IEC11172-3[4] was released in 1992, in which two psychoacoustic models were present. Afterwards both ISO/IEC 13818-3 (MPEG-2 Audio part)[5] and 13818-7 AAC[6] adopted the basic framework of psychoacoustic model 2 firstly introduced as a whole by J.Johnston in [9]. In stead of simulating physiological structure of human ear, psychoacoustic model 2 makes the output results approaching to the examination data.

ITU-R BS.1387 is the objective measurements Recommendation Standard of perceived audio quality established in 1998. It utilizes software to simulate perceptual properties of human ear, then integrates multi-indices to evaluate subjective quality of test audio. BS.1387 has been considered as the most effective objective measurement schemes of perceptual audio quality. Most of objective evaluating software such as EQUAL [7] conform to this standard. BS.1387 standard has two options: basic version and advanced version. Psychoacoustic model is its core module and design emphasis. Compared to MPEG, the psychoacoustic model in BS.1387 takes into account perceptual properties of human ear more comprehensively.

The goal of this paper is to combine objective measurement of audio subjective quality with audio coder, and design a more perfect psychoacoustic model in audio coder to improve coding efficiency and coding quality. This paper is organized as follows. In next section, the fundamental of psychoacoustic model adopted in the basic version is introduced. In section 3, algorithmic implementation of proposed psychoacoustic model in audio coder is presented. In section 4, comparison experiments between proposed psychoacoustic model and MPEG psychoacoustic model 2 are conducted. Finally, the major works of this paper are summarized and the further research and direction of improvement is presented.

II. ANALYSIS OF THE PSYCHOACOUSTIC MODELS

In MPEG psychoacoustic model, SMR (ratio between signal energy and masking threshold) is determined by experiential value of examination observation. This model can be easy to correct and adjust, moreover fit between model and examination result is satisfied. However, restricted in examination condition, so far most of examination results obtained were based on pure tone, only a few examination used dual tone[8]. In fact, a complex signal is combination of infinite sinusoids, which has much more complicated masking phenomena than pure tone. This kind of model only fit the examination data to simple signals and is hard to be extended to complex signals in theory. Therefore the maximum masking threshold could not be obtained in this model unless the examination measure is improved.

Basic version of BS.1387 adopted a design approach different from MPEG audio standard. It attempts to combine physiological structure of human ear with masking effect of simple signal represented from examination to find the inherent consequence, and then use mathematic model to emulate the structure of human ear. Fig. 1 illustrates the block diagram of this design[10], in which the function of outer/middle ear, inner ear, and audio perception related nerve cell and brain are emulated. This kind of psychoacoustic model could be extended conveniently to acoustic masking of complex signal.

BS.1387 was initially designed to measure audio quality, without special consideration for requirements of audio coder, such as window switching, unification of critical band scale, estimation of masking properties on transient signal, etc. Thus, we must take into account these requirements and correct the psychoacoustic model of BS.1387 basic version so as that it could be used in audio coder.
window function is
\[ w(n, N_F) = \begin{cases} \sqrt{\frac{\sin(\pi n/N_F)}{\pi n/N_F}} & 0 \leq n < N_F/2, \\ 0 & \text{otherwise} \end{cases} \]
(1)

Where \( N_F \) is frame length of input signal.

Because this window function changes the original signal energy, it must be normalized [12]. The actual normalized window function is
\[ w(n, N_F) = \begin{cases} \frac{\sqrt{\sin(\pi n/N_F)}}{\sin(\pi/N_F)} & 0 \leq n < N_F/2, \\ 0 & \text{otherwise} \end{cases} \]
(2)

Compared to the window function used in MPEG, here energy change along with windowing is calculated more precisely.

Then \( x_w(n) \) is converted into frequency domain signal \( X(k) \) using a Discrete Fourier Transform, as shown in (2).

\[ X(k) = \frac{1}{N_F} \sum_{n=0}^{N_F-1} x_w(n)e^{-j2\pi nk/N_F} \]

**B. Outer/Middle Ear Transfer Model**

Human ear’s sensitivity varies for different frequency. Signal is weighted by Emulating the outer/middle ear transfer model, so as to be consistent with human ear sensitivity curve. There is no such consideration in MPEG standards.

Since auditory perception of human ear is pertinent to loudness, loudness correction for \( x(k) \) must be done before weighting. Correction factor \( G_L \) is defined as [12]:

\[ G_L = \frac{L_p}{10^{\frac{8 A_{\text{max}} N_F-1}{3 4 N_F}}} \]

(3)

Where \( L_p \) is the sound pressure level (SPL) corresponding to sine wave with maximum amplitude. \( A_{\text{max}} \) is the max amplitude. \( \gamma(f_c) \) is center frequency related factor altered from 0.84 to 1.

Weighting transfer function is defined as

\[ A_{\text{tg}}(f_c) = 2.184 / 1000^{0.8} + 6.8e^{-0.05(f_c/1000-3.3)^2} - 0.005 / 1000^{3.6} \]

(4)

Thus spectrum energy after weighting is

\[ |X_w(k)|^2 = G_L^2 W^2(k)|X(k)|^2 \quad (0 \leq k \leq N_F/2) \]

**C. Frequency Grouping**

In order to be consistent with the critical band partition of MPEG audio coder, the frequency domain must be regrouped according to new critical band. Then Energy \( E_b \) is calculated for each band. Since critical band partition is defined by discrete DFT bins, and energy of each DFT bin represents continuous frequency energy distributed over bin width. Thus special processing for maximum and minimum frequency is necessary. The distribution scale of energy for bin \( k \) in band \( i \) is given by (5), and the energy of band \( i \) is given by (6), where \( F_s \) is sampling rate, \( f_{l}(i), f_{h}(i), k_{i}, k_{h}(i) \) are upper limit, lower limit and bin in band \( i \) respectively, and \( U_{l}(i), U_{h}(i) \) are energy distribution corresponding to upper limit and lower limit respectively.
and nonlinear adding was added to the input signal. This noise is simulated on the base of physiological structure of human ear, whereas it is represented as absolute hearing threshold in MPEG psychoacoustic model. The internal noise function is given by

$$E_{IN}(i) = 10^{0.1456(f_i/1000)^{-0.8}}$$

(Eq. 7)

D. Adding of Internal Noise

Internal noise caused by blood flow within human ear, is added to the input signal. This noise is simulated on the base of physiological structure of human ear, whereas it is represented as absolute hearing threshold in MPEG psychoacoustic model. The internal noise function is given by

$$E_{IN}(i) = 10^{0.1456(f_i/1000)^{-0.8}}$$

(Eq. 7)

E. Energy Spreading of Frequency Domain

Energy spreading of frequency domain is the important phenomenon of auditory properties. In MPEG psychoacoustic model 2, energy spreading curve is added linearly with independent of loudness. But it was discovered that energy spreading curve at a frequency point varied according to its loudness [13], and nonlinear adding was much coincidental to auditory properties of human ear [14]. Thereby, the loudness is taken effect into energy spreading curve, as shown in (8).

$$S_{bb}(i,l,E) = \begin{cases} \frac{27(i-l)}{A(l,E)} & i \leq l \\ \frac{1}{A(l,E)} & i > l \end{cases}$$

(Eq. 8)

Where \( i, l \) are the center frequencies respectively of the \( i_b \) excitation band and \( i_b \) spread band in unit of Bark. \( f_c(l) \) is center frequency of the \( i_b \) excitation band in unit of Hz. \( \log_{10}(E) \) relates loudness and spreading function. In addition, (8) is required to be normalized, as shown in (9).

$$S(l,i,E) = \begin{cases} \frac{1}{A(l,E)} & i \leq 1 \\ \frac{1}{A(l,E)} & i > 1 \end{cases}$$

(Eq. 9)

Where \( A(l,E) \) is normalization factor, \( A(l,E) = 10^{0.1456 \cdot \frac{27(i-l)}{A(l,E)}} \).

Then we make use of nonlinear adding to calculate distribution curve spread energy as follows.

$$E_{s}(l) = \frac{1}{B_{s}(l)} \left( \sum_{j=0}^{N_s-1} S(l,j,E) \right)^{0.4}$$

(Eq. 10)

$$B_{s}(l) = \left( \sum_{j=0}^{N_s-1} S(l,j,E) \right)^{0.4}$$

(Eq. 11)

Where \( N_s \) is the number of band groups. \( B_{s}(l) \) is normalizing factor, which is the nonlinear sum of spreading functions of the \( i_b \) band.

F. Time Domain Spreading

Time domain spreading concerns temporal masking effects. As for auditory properties, there are pre- and post- masking effects in time domain besides frequency domain masking [8], while such temporal masking effects are not involved in psychoacoustic model of MPEG. In this model, temporal masking effect is considered by means of first-order smooth filtering.

Time domain spreading is calculated in unit of band partitioned above as shown below:

$$E_{f}(i,n) = \max(0,E_{s}(i,n-1) + (1 - \alpha(i)E_{s}(i,n))$$

(Eq. 12)

Where \( E_{f}(i,n) \) is energy of time domain spreading of band \( n \) at frame \( i \). \( \alpha(i) \) is the decaying coefficient of band \( i \) controlled by time constant, as shown in (13):

$$\alpha(i) = \exp(-\frac{1}{F_s \cdot \tau(i)})$$

(Eq. 13)

Where \( F_s = F_c / 2 \) is the frame rate, \( F_c \) is the sampling rate. \( \tau_{100} \) and \( \tau_{min} \) are time decaying constants. After tuning for AVS audio coder, \( \tau_{100} = 0.03 \) and \( \tau_{min} = 0.008 \) for long window, while \( \tau_{100} = 0.003 \) and \( \tau_{min} = 0.001 \) for short window.

G. Calculation of Masking Parameters

After processing described above, signal energy curve concerned both temporal masking and simultaneous masking is obtained. The energy distribution of masking threshold can be calculated utilizing masking curve. Masking property and masking energy of each critical band are given by (14).

$$E_{mask}(dB)(i) = E_{f}(i,n) - m_{dB}(i)$$

(Eq. 14)

Where \( z \) is center frequency of the \( i_b \) band in unit of Bark, and \( z_l = 0.8594 \).

IV. EXPERIMENTAL RESULTS

This new psychoacoustic model has been implemented in reference coder based on AVS audio standard [11]. Fig. 3 shows distribution curve of masking threshold for tone at 1 kHz calculated by utilizing the two different psychoacoustic models respectively. From Fig. 3, we can observe that distribution of masking curve in new psychoacoustic model matches better with the real masking threshold. Whereas masking curve in MPEG psychoacoustic model 2 is more sensitive to distribution of spectrum energy, and the slope of pre- energy spreading at peak energy is nearly equal to that of post- energy spreading. This is different from observed acoustic phenomena. In contrast, the new psychoacoustic model represents preferably the hearing property that pre-energy spreading effect is bigger than post-energy spreading effect.

$$E_{s}(l) = \max(0,E_{s}(l))$$

(Eq. 5)

$$E_{s}(l) = \frac{\alpha(l) E_{s}(l) + \beta(l) E_{s}(l-1)}{\gamma(l)}$$

(Eq. 6)
In this paper, we discussed application of a new psychoacoustic model based on BS.1387 standard to perceptual audio coder. We first described the principles of this new psychoacoustic model, then presented the implementation of algorithms. In order to meet requirements of audio coder, we proposed improving methods as well as corresponding parameters setting. Compared to hearing examination and observation, this new model is proved to be practical. Due to consideration of more acoustic properties, the proposed psychoacoustic model can characterize the auditory properties of human ear more precisely than MPEG psychoacoustic model 2. Thus the work in this paper provides new approach to audio coder’s optimization. In addition, the results of subjective listening test show the proposed model is faced with some problems for short window and window transition, which is one of our further research focus.

V. CONCLUSION

In this paper, we discussed application of a new psychoacoustic model using noise masking to provide better quality than MPEG psychoacoustic model 2. For test sequences with few window switching, the new model can be improved further; On the other hand, it is necessary to adjust original rate control and bit allocation strategy due to changes of psychoacoustic model. Subjective quality has been remarkably improved after adjustment. For test sequences with few window switching, the new model can provide better quality than MPEG psychoacoustic model 2.

REFERENCES


![Fig. 3. distribution curve of masking threshold for tone at 1 kHz](image)

![Fig. 4. distribution curve of masking threshold for complex signal](image)

Fig. 3. distribution curve of masking threshold for tone at 1 kHz

Fig. 4. distribution curve of masking threshold for complex signal at 1 kHz