A NEW FORWARD MASKING MODEL AND ITS APPLICATION TO PERCEPTUAL AUDIO CODING

Yuan-Hao Huang and Tzi-Dar Chiueh

Room 511, Department of Electrical Engineering
National Taiwan University, Taipei, Taiwan R.O.C
chiueh@cc.ee.ntu.edu.tw

ABSTRACT

This paper presents a new forward masking model for perceptual audio coding. This model exploits adaptation of the peripheral sensory and neural elements in the auditory system, which is often deemed as the cause of forward masking. Nonlinearity of the ear is modeled by a nonlinear analog circuit with difference equations. We incorporate this model in the MPEG Layer III audio coding scheme and construct a masking plane in the frequency-time space. With some extra computations, the new audio coding scheme can improve the sound quality of the decoded audio signals. In our experiments, subjective and objective sound quality measurements show that, to achieve the same reconstructed sound quality, the new scheme requires 12% to 23% less bits than the original MPEG Layer III scheme.

1. INTRODUCTION

Audio signal compression has found application in many systems, such as multimedia signal coding and high quality audio transmission and storage. Because of the limited bandwidth of transmission or storage, encoding the audio signal with the minimum perceived loss of sound quality at a limited and fixed bit rate becomes an important issue. As a result, several perceptual audio coding algorithms exploit the masking effects of the auditory system to reduce the bit rate by keeping the quantization error below the just-notice distortion (JND) [1]. The AC-3 audio coding scheme further includes forward masking effect by adaptively adjusting the filter banks [2]. Besides, many RC circuit models are proposed to simulate adaptation of the auditory system [4] and spiking phenomenon of the neural elements in the ears [5]. Because the RC circuits are time-dependent, they are often used to model the temporal effects in psychoacoustics. As a result, forward masking effect can also be modeled by an electronic circuit.

In this paper, we first adopt a psychoacoustic model to formulate the forward masking of the human auditory system. We use an analog circuit model proposed in [7] to predict the loudness and the amount of forward masking. Next, we combine the forward masking model and the frequency masking model in the MPEG layer III audio coding scheme [3] by constructing a masking plane in the frequency-time space. Therefore, more accurate masking effect can be modeled and the optimal bit allocation can be achieved under fixed bit rates. Both subjective and objective measurements of sound quality are performed in our experiments. Under fixed bit rates, the sound quality of the decoded audio signal is better than that of the original MPEG-Layer III scheme.

2. FORWARD MASKING MODEL

The amount of forward masking is strongly influenced by the signals in the previous frames. Energy of the previous frames may remain in the basilar membrane of the cochlea. This is often regarded as the cause of forward masking in psychoacoustics. Furthermore, the amount of the remaining energy in the basilar membrane is related to the forward masking level. As a result, the forward masking level changes with masker duration in such a way that masking level decreases more rapidly for shorter masker impulses and more slowly for longer masker impulses. The masker-dependent forward masking effect reveals the fact that the auditory system is not a linear system with simple time constants.

We now propose a perceptual audio scheme that includes the forward masking effect, as shown in Figure 1. This masking model depends not only on the present signal frame but also on the previous frames. To determine the final masking level for encoding, one computes the maximum of the frequency masking and the forward masking as follows

$$M(t,f) = max\{M_f(t,f), M(t-\Delta t, f)\cdot exp^{-\Delta t/(\tau(f)\cdot N)}\},$$

where $M(t,f)$ is the final masking level for coding constraints. $M_f(t,f)$ is the masking level computed from the frequency masking model. Δt is the time difference between frames. $\tau(f)$ are the maximum decaying time con-

stants of different critical bands [6]. N is the total loudness level in psychoacoustics, which has close relation to the masker duration and decaying time constant in forward masking [7]. N is limited between 1 and 0. When N is 1, the energy in the basilar membrane saturates and each band has maximum decaying constant. When N is 0, there is no signal energy from the previous frames and thus no forward masking effect.

To model the nonlinearity of the auditory system, we use the analog circuit in Figure 2 to estimate the amount of forward masking. N is the sum of the output specific loudness N_o in each critical band, which is the output of the analog circuit in Figure 2. The input N_s of the circuit is the specific loudness of the present frame, which is calculated from the following equation [4]

$$N_s = 0.08(E_T/E_0)^{0.23}[(0.5 + 0.5 * E/E_T)^{0.23} - 1].$$

We can get excitation level E by convolving the spectrum energy with the spreading function. This equation transforms the external physical energy values to the internal loudness values [4]. The non-linear RC circuit in Figure 2 is used to find the output specific loudness N_o in each critical band. To use the model in a digital system, we convert the differential equations to the difference equations as follows.

$$I_n = \frac{N_s - {N_o}^*}{R_{on}},$$

where R_{on} is on resistance of diode $D1$.
If $I_n < 0.0$,

$$I_n = 0.0.$$

If $N_o^* > V^*$

$$N_o = \frac{I_n + C2 \cdot V^*/(\Delta t + C2 \cdot R2) + C1 \cdot {N_o}^*/\Delta t}{1/R1 + C2/(\Delta t + C2 \cdot R2) + C1/\Delta t},$$

and

$$V = {N_o}^* \cdot \Delta t/(\Delta t + R2 \cdot C2) + V^* \cdot C2 \cdot R2/(\Delta t + C2 \cdot R2),$$

else

$$V = N_o = \frac{I_n + (C1 + C2) \cdot V^*/\Delta t}{1/R1 + (C1 + C2)/\Delta t},$$

where V^* and ${N_o}^*$ are the respective voltages of the previous frame with frame time difference Δt. In the RC circuit, the increase in masker duration corresponds to storing more charges into the capacitors through a diode and a resistor. Therefore, if the capacitors are charged with a longer impulse, the output specific loudness N_o will be larger. Then, the decaying time constant will become larger. On the contrary, if the impulse is short, the output of the circuit will be smaller. Then, the decaying constant will be smaller. The diode here is used to limit the charges into the capacitors when the remaining energy saturates in the basilar membrane. Consequently, the time constant in each critical band is maximum. The values of the resistors and capacitors are designed to match the phenomena that the forward masking effect saturates when the masker signal is longer than 200 ms [7][8].

3. COMPUTATION COMPLEXITY

The computation complexity of the forward masking model is lower than that of the frequency masking model. The computation complexity of each function in the psychoacoustic model is listed in Table 1. It is clear that the computation overhead of the forward masking process is quite limited and almost negligible. This table does not include the computations of the MDCT, bit allocation algorithm, scalar quantization and Huffman coding. If they are included, the computation overhead of the forward masking effect will be even smaller.

4. SUBJECTIVE AND OBJECTIVE SOUND QUALITY MEASUREMENTS

We collected ten mono audio items, which are listed in Table 2, from the compact discs with 44.1 kHz sampling rate and 16-bit resolution. The ten audio items are coded and decoded by the MPEG Layer III coding scheme with and without the forward masking model. To rate the quality in different bit rates, we used 80, 64, 56, 48, 40, and 32 kbps as defined in MPEG layer III standard.

We conducted the mean opinion score (MOS) listening test as subjective assessment of sound quality. Ten subjects listened to random listed sound segments using headphones in a quiet office environment and gave a 5-point MOS rating for each segment.

Moreover, we used the perceptual audio quality measurement (PAQM) in [9] as the objective assessment of the sound quality. This method measures the quality of an audio codec by mapping the input and output of the audio codec from the physical signal representation onto a psychoacoustic representation. This mapping enables quantification of perceptual degradation introduced by audio codecs. From this mapping, the subject quality can be predicted to a certain extent. This method can measure the sound quality in different time points of each sound segment. More negative value of noise disturbance in PAQM means better sound quality. Figure 3 shows the noise disturbance variation of the signals encoded in 48 kbps bit rate. The result shows that the signal encoded with the forward masking model has better sound quality than that encoded without the forward masking model.

The maximum of the frequency and the forward masking level is utilized as the masking threshold for encoding.

An interesting issue is to examine which of these two is the more dominant effect. Supposed a sound segment as shown in Figure 4 (a) is encoded using the proposed masking scheme. In Figure 4 (b), the white areas correspond to dominance of the forward masking and the black areas correspond to dominance of the frequency masking. One sees that the forward masking makes up a significant portion of the whole frequency-time space, which explains why the proposed scheme is effective in reducing bit rate and improving sound quality. The final results of the MOS scale and the mean PAQM measurements are depicted in Figure 5 (a) and (b). The same audio quality can be maintained with 12% and 23% less bits in subjective and objective measurements respectively when the forward masking model is included in the MPEG Layer III audio coding scheme.

5. CONCLUSION

In this paper, a forward masking model was proposed and incorporated into the standard perceptual audio coding scheme. This model exploits the forward masking effect with dynamic adaptation of the auditory system. The sound quality measurements of subjective MOS and objective PAQM show that the decoded audio segments have better sound quality than the original audio coding scheme. Furthermore, the overhead in computation complexity is almost negligible. Therefore, we believe that the new audio coding scheme forms a solid foundation for future audio coding technology.

6. REFERENCES

[1] Johnston, "Transform Coding of Audio Signals Using Perceptual Noise Criteria," *IEEE J. of Selected Area In Comm.* , vol. 6,No. 2, Feb. 1988, pp. 314–323.

[2] M. Bosi and G. Davidson, "High Quality, Low-Rate Audio Transform Coding for Transmission and Multimedia Applications" *http://www.dolby.com/tech/* presented at the 93rd AES Convention, Oct. 1993.

[3] K. Brandenburg, "ISO-MPEG-1 Audio: A Generic Standard for Coding of High-Quality Digital Audio" *J. of Audio Eng. Soc.* , vol. 42, no. 10, Oct. 1994, pp. 780–792.

[4] E. Zwicker and H. Fastl,"Psychoacoustics-Facts and Models," Springer-Verlag, 1990.

[5] James M. Kates, "A Time-Domain Digital Cochlea Model," *IEEE Trans. on Signal Processing*, vol. 39, No. 12, Dec. 1991, pp. 2573–2592.

[6] W. Jesteadt, S. P. Bacon and J. R. Lehman, "Forward masking as function of frequency, masker level and signal delay" *J. Acoust. Soc. of Amer.* , vol. 71(4), Apr. 1982, pp. 950–962.

[7] E. Zwicker "Dependence of post-masking on masker duration and its relation to temporal effects in loudness" *J. Acoust. Soc. of Amer.* , vol. 75(1), Jan. 1984, pp. 219–223.

[8] G. Kidd Jr. and L. L. Feth, "Effects of masker duration in pure-tone forward masking" *J. Acoust. Soc. of Amer.* , vol. 72(5), Nov. 1982, pp. 1384–1386.

[9] J. G. Beerends and J. A. Stemerdink, "A Perceptual Audio Quality Measure Based on a Psychoacoustic Sound Representation" *J. of Audio Eng. Soc.* , vol. 40, no. 12, Dec. 1992, pp. 963–978.

Table 1: Computation complexity of the psychoacoustic model: First eight functions are frequency masking computations. Last three functions are forward masking computations.

Function	Complexity	operation	N
FFT	$N\log(N)$	* + cos sin	1024
unpredictive	N	* + sqrt cos sin	512
grouped energy	N	* +	512
spreading	N^2	* +	63
tonality cal.	N	* + log comp.	63
pre-echo	N	comp.	63
percep. entropy	N	* + log comp.	63
threshold cal.	N	* + exp	63
Trans. to loudness	N	* + / power	63
diff. equation	N	* + /	63
compare masking	N	comp. + exp	63

Table 2: The recorded music items for sound quality measurements.

no.	music
1	saxophone
2	male voice
3	female voice, drum and cello
4	electrical guitar and violin
5	violin and piano
6	orchestra
7	piano solo
8	chorus song
9	flute
10	bass

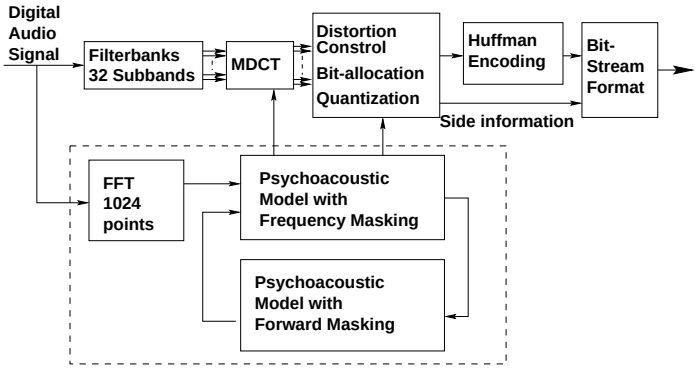

Figure 1: Block diagram of the MPEG layer III audio coder with the forward masking model.

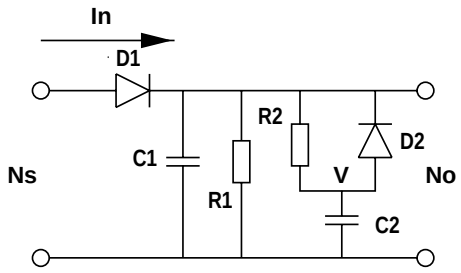

Figure 2: Circuit model of the forward masking effect.

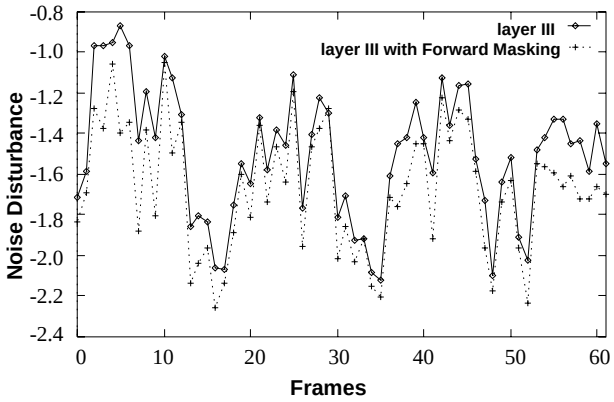

Figure 3: Noise disturbance of decoded audio signals using the MPEG layer III standard with and without the forward masking model.

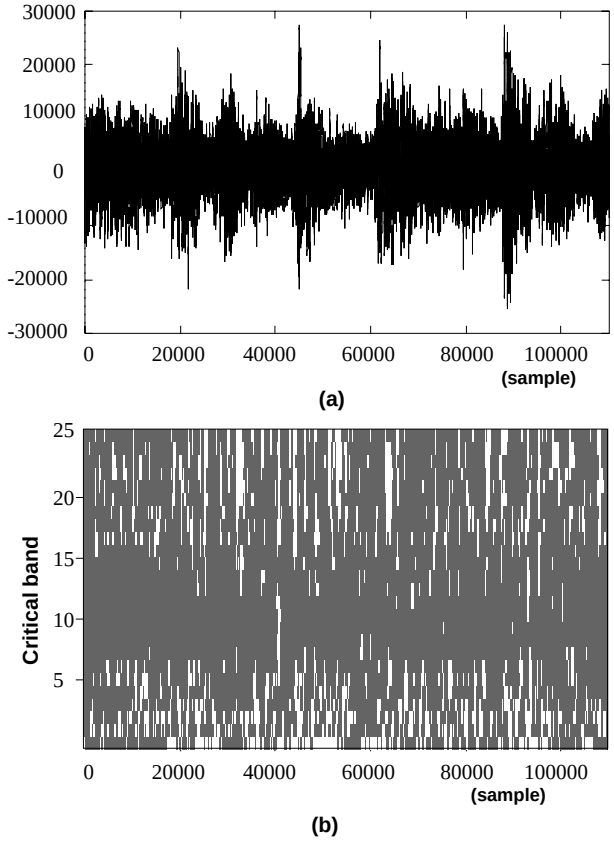

Figure 4: (a) the original sound, and (b) the dominance pattern in the frequency-time space: frequency masking effect (black) and forward masking effect (white).

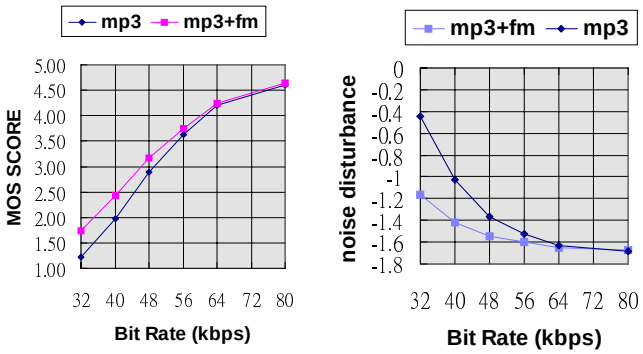

Figure 5: (a) MOS score vs. bit rate, and (b) noise disturbance vs. bit rate.