A New Audio Coding Scheme Using a Forward Masking Model and Perceptually Weighted Vector Quantization

Yuan-Hao Huang, *Member, IEEE*, and Tzi-Dar Chiueh, *Member, IEEE*

***Abstract*—This paper presents a new audio coder that includes two techniques to improve the sound quality of the audio coding system. First, a forward masking model is proposed. 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. In the proposed audio coder, the forward masking is first modeled by a nonlinear analog circuit and then difference equations for finding the solution of this circuit are formulated. The parameters of the circuit are derived from several factors, including time difference between masker and maskee, masker level, masker frequency, and masker duration. Inclusion of this model in the coding process will remove more redundancy inaudible to humans and thus improves coding efficiency. Secondly, we propose a new vector quantization technique, whose codebooks are generated by a perceptually weighted binary-tree self-organizing feature maps (PW-BTSOFM) algorithm. This vector quantization technique adopts a perceptually weighted error criterion to train and select codewords so that the quantization error is kept below the just-noticed distortion (JND) while using the smallest possible codebook, again reducing the required coded bit rate. Experimental objective and subjective sound quality measurements show that the proposed audio coding scheme requires about 30% less bits than the MPEG layer III audio coding standard.**

***Index Terms*—Forward masking, perceptually weighted error criterion, vector quantization.**

I. Introduction

AUDIO SIGNAL compression has found application in many areas, such as multimedia signal coding (e.g., motion picture expert group (MPEG) systems [1]) high-fidelity audio for radio broadcasting (e.g., digital audio broadcasting (DAB) system [2]), audio transmission for HDTV, audio data transmission/sharing through internet, etc. High-fidelity audio signal coding demands a relatively high bit rate of 705.6 kbps per channel using the compact disc format with 44.1-kHz sampling and 16-bit resolution. With the proliferation of exchange and transmission of audio information through internet and wireless systems, efficient (i.e., low-bit-rate) audio coding algorithms need be devised.

Two major classes of techniques can be used in audio source coding to reduce coded bit rate. The first class employs some signal processing so that essential information and perceptually irrelevant signal components can be separated and the latter removed. This class include techniques such as subband coding [3], transform coding [9], critical band analysis [7], and masking effects [7]. The second class takes advantage of the statistical redundancy in audio signal and applies some form of digital encoding. Examples of this class include entropy coding in lossless compression [8] and scalar/vector quantization in lossy compression [5], [6].

Since the terminal receiver of audio coding are humans, audio coding algorithms that take into account psychoacoustic characteristics of the human auditory system seem better positioned for coding with better efficiency. In order to incorporate these characteristics in audio coding, the human auditory system needs be modeled to a certain degree of accuracy. Most psychoacoustic analysis aims to determine the maximum quantization noise not perceptible to even well-trained listeners. With this information, audio signals can be coded more efficiently while keeping the coding distortion below just-noticed distortion (JND). Among these characteristics, masking effects are some of the most important and they have been adopted in various audio and speech coders [4]–[7], [9].

Masking effects [10], [11] occur in the frequency domain as well as in the time domain. There are three types of masking effects: simultaneous masking, backward masking, and forward masking (see Fig. 1). Simultaneous masking is a frequency-domain phenomenon, where a lower-level signal component (maskee) is made inaudible by a simultaneously occurring stronger signal (masker). The masking threshold depends on the sound pressure level (SPL) of the masker and the frequency difference between the masker and the maskee. Backward masking, as its name suggests, masks signal components that occur before the masker. It can help to mask pre-echoes caused by the spreading of a large quantization error. This property has been utilized in the pre-echo control of the psychoacoustic model in MPEG Layer III standard [1]. Forward masking masks signal components that occur after the masker, and it has an effective duration ten times that of backward masking. Therefore, the forward masking effect can improve coding efficiency better than backward masking since more signal components are masked and need not be coded.

Manuscript received November 9, 1999; revised April 16, 2002. This work was supported by the National Science Council, Taiwan, R.O.C., under Grant NSC87-2213-E-002-019. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Bryan George.

Y.-H. Huang was with the Department of Electrical Engineering and Graduate Institute of Electronics Engineering, National Taiwan University, Taipei 10617, Taiwan, R.O.C. He is now with the VXIS Technology Corporation, Hsinchu 300, Taiwan, R.O.C.

T.-D. Chiueh is with the Department of Electrical Engineering and Graduate Institute of Electronics Engineering, National Taiwan University, Taipei 10617, Taiwan, R.O.C.

Publisher Item Identifier 10.1109/TSA.2002.800559.

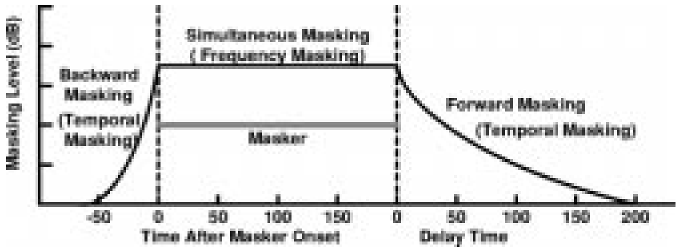


Fig. 1. Frequency and temporal masking effects.

Despite its potential, the forward masking effect receives little attention from audio coding researchers due to the difficulties in modeling its nonlinearity. In the first part of the paper, we will introduce a nonlinear circuit, which models the behavior of the forward masking effect. A discrete-time formulation of this circuit is integrated with the frequency-domain psychoacoustic model in the MPEG layer III standard. The resultant psychoacoustic analysis eliminates more signal components than the original MPEG layer III coding, making the proposed forward masking analysis even more efficient in eliminating imperceptible signal components and reducing coded bit rate.

Presently, most audio coding standards apply some sort of signal processing to obtain a frequency-domain representation of an audio frame. Then the coefficients are scalar quantized because scalar quantization with psychoacoustic modeling can achieve very good coding performance. On the other hand, only a few studies addressed vector quantization of the coefficients. Chan and Gersho [12] investigated using multi-stage tree-structured vector quantization (MSTVQ) technique in encoding discrete cosine transform (DCT) coefficients. In another of their study, with the constraint-storage VQ (CSVQ) [13], their proposed audio coding system can strike a balance between rate-distortion performance and codebook searching complexity. Recently, Iwakami [14] developed transform-domain weighted interleave vector quantization (TWIN-VQ), which is adopted in the MPEG4 standard. In this method, the modified DCT coefficients are first flattened by the signal spectral envelope. Then, a subvector, formed by sample interleaving, is quantized using a criterion weighted by the corresponding LPC envelope components. Subjective evaluation showed that the sound quality of the decoded audio of TWIN-VQ exceeds that of the MPEG1 Layer II coder at the same bit rate [14]. Several other reports also showed the advantages of vector quantization in audio coding [15]–[19]. However, none of these methods take psychoacoustic effects into account during codebook design and vector encoding.

In the second part of this paper, we will propose a neural-network-based vector quantization scheme that encodes the MDCT-polyphase-filter coefficients using a psychoacoustic feature related criterion. In this scheme, VQ codebooks are derived from perceptually weighted binary-tree structured self-organizing feature map (PW-BTSOFM), a modified version of binary-tree structured self-organizing feature map (BTSOFM) [20]. The distribution of the codewords in the proposed tree-structured codebook not only reflects underlying data statistics but also the signal-to-masking ratio (SMR) values. Furthermore, a perceptually weighted SMR-based error criterion is used to determine the best codeword during encoding. Simulation results show that, at all fixed bit rates, decoded sound quality of the PW-BTSOFM VQ audio coder with the forward masking model outperforms that of the MPEG1 audio layer III coder.

The rest of this paper is organized as follows. Section II describes the forward masking effect in psychoacoustics. In this section, we also introduce a circuit model for forward masking threshold estimation. The binary-tree structured self-organizing feature mapping for VQ codebook training is introduced in Section III. Section IV describes the algorithm of the perceptually weighted BTSOFM VQ and the complete architecture of the proposed audio coding scheme. Simulation and experimental results on the sound quality are then given in Section V. Finally, Section VI summarizes and concludes this paper.

II. Forward Masking Model

Forward masking effect, as shown in Fig. 1, occurs when a signal (maskee) follows a masker signal. Its effective duration, on the order of hundreds of milliseconds, is longer than that of backward masking. The basic principle of forward masking is still unclear. It is conjectured to be caused by ripple response in basilar membrane filtering, reduction in sensitivity of recently stimulated cells, or persistence in neural activity patterns evoked by the masker [11]. On the other hand, experimental phenomena of forward masking are well known. Forward masking lasts for about 200 ms after the end of a long masker [21]. The decay rate of the masking level depends nonlinearly on the masker level and masker duration [21]–[23]. The decay rate is high for short masker and masker with high energy level, while the decay rate is low for long masker and masker with low energy level. In addition, the decay rate also depends on the masker frequency [24].

A. Psychoacoustic Forward Masking Model

Research has been carried out in building psychological [25]–[28] and psychoacoustic models [21], [24], [29], [30] of forward masking using electronic circuits. Since the movement of electrons in a circuit is similar to the adaptation of the neural charges in the neural system, forward masking can be modeled fairly well by a nonlinear circuit. In psychoacoustics, forward masking has long been regarded as an indication of the decay of the hearing system's internal loudness [31]. So, it is often modeled using psychoacoustic specific loudness versus critical-band rate and time. For the ith critical band, we can first compute, as specified in [1], the excitation level $E(i)$ by convolving the signal with a spreading function

$$E(i) = s(i) * P(i) \tag{1}$$

where $P(i)$ is scaled signal energy in the ith critical band and $s(i)$ is the spreading function and

$$s(i) = 15.81 + 7.5 \cdot (i + 0.474) - 17.5\sqrt{1 + (i + 0.474)^2}. \tag{2}$$

The convolution actually spreads the signal among neighboring frequency components and is used to model the simultaneous frequency-domain masking.

A nonlinear circuit, as shown in Fig. 2, was proposed in [29] to estimate the output specific loudness, one circuit for each

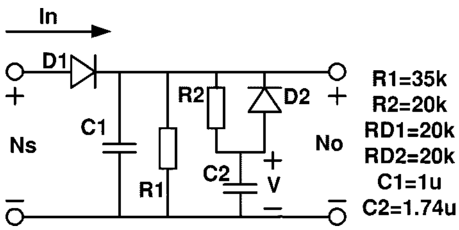


Fig. 2. RC circuit for loudness estimation.

critical band. In each circuit, input $N_s(i)$ is the ith critical-band specific loudness of the current frame, which is given by

$$N_s(i) = 0.08(E_T(i)/E_0)^{0.23} \cdot[(0.5 + 0.5 \cdot E(i)/E_T(i))^{0.23} - 1] \quad (3)$$

where $E_T(i)$ is excitation at absolute threshold, and E_0 is the excitation that corresponds to the reference intensity $I_0 = 10^{-12}$ W/m^2. The above equation transforms the external physical energy values $(E(i))$ to the internal loudness values $(N_s(i))$ [10]. Then, the low-pass circuit in Fig. 2 is used to emulate the nonlinear loudness processing in the neural system and generates the output specific loudness $(N_o(i))$. Two resistors and capacitors are used to model the two time constants in the specific loudness decay profiles [21], [24]. One constant is smaller (for short maskers), and the other is larger (for the maskers with duration longer than 100 ms).

Some simulation results of the aforementioned nonlinear circuit are shown in Fig. 3. The output specific loudness (N_o of the corresponding critical band) of input 2-kHz signals with different duration [shown in Fig. 3(a)] are depicted in Fig. 3(b). The total loudness (N), defined as the sum of the output specific loudness in all critical bands, is known to be directly related to the forward masking level [10]. As shown in Fig. 3(c), the total loudness (N) saturates and the forward masking level reaches its maximum if the masker duration is longer than 100 ms. After saturation, the total loudness decreases with a large time constant. On the other hand, if the masker duration is less than 100 ms, the total loudness starts to decrease, without even reaching saturation level, with a time constant that increases with increasing masker duration. The above simulated total loudness behavior in cases with different masker duration agree quite well with the observed forward masking level [10]. Therefore, in the proposed audio coding system we use the simulated total loudness as an estimate of the forward masking level.

B. Application of the Forward Masking Effect to Audio Coding

We now propose a model that integrates the frequency-domain simultaneous masking effect and the time-domain forward masking effect. The proposed model depends not only on the current frame but also on previous frames. To determine the total masking level of the ith critical band at time t, one computes the maximum of the current simultaneous masking level and the total masking level of the previous frame decayed by some constant

$$M(t, i) = \max\left\{M_s(t, i),\, M(t, i)^* \cdot \exp^{-\Delta t/(\tau(i)\cdot N)}\right\} \quad (4)$$

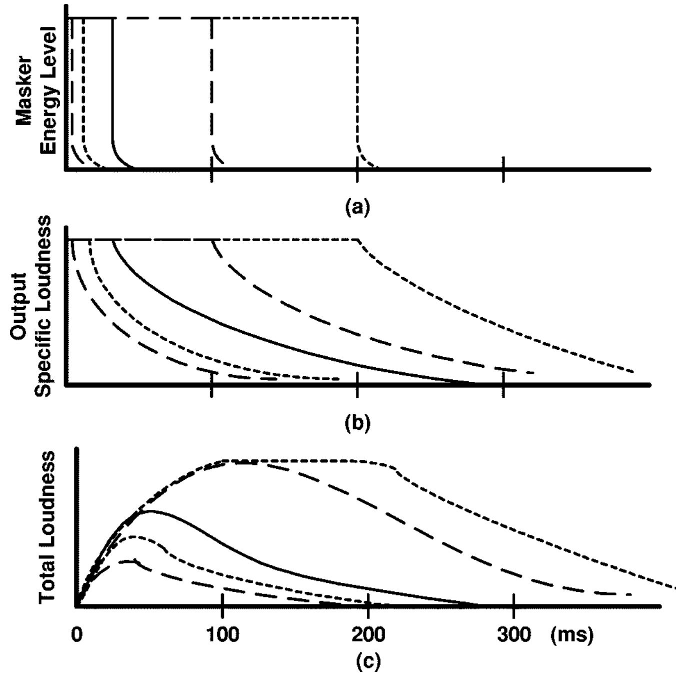


Fig. 3. Loudness profiles of the maskers with different durations (adapted from [8]).

where $M(t, i)$ and $M(t, i)^*$ are the total masking levels of the current frame and the previous frame, respectively; $M_s(t, i)$ is the masking level computed from the simultaneous masking model [1]; Δt is the time difference between two frames; $\tau(i)$ is the maximum decay time constant in each critical band [24]; and N is the total loudness level. Note that N is now normalized by the total loudness of a 60 dB uniform masking noise (UMN). If N is larger than one, N is set to one. So, N lies between zero and one. When N is one, the energy in the basilar membrane saturates and total masking level decays with a maximum time constant. When N is zero, there is no signal energy spilt over from previous frames and thus no forward masking.

The nonlinear RC circuit in Fig. 2 is used to find the output specific loudness N_o in each critical band. For numerical computation, we convert the differential equations into the following difference equations:

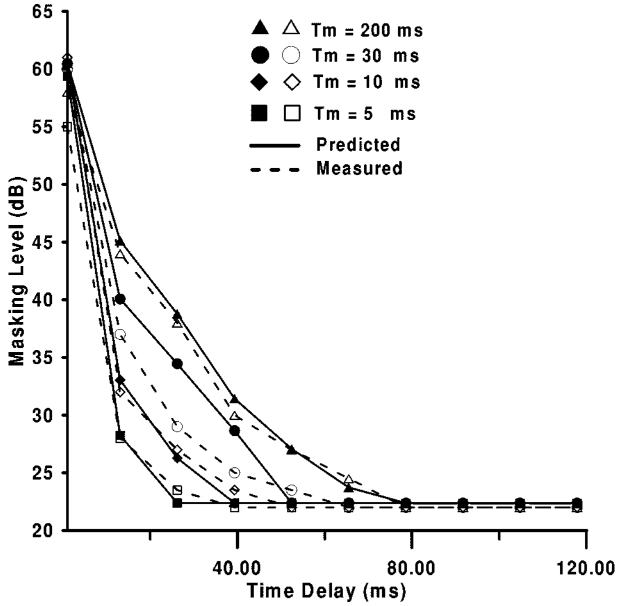
$$I_n = \frac{N_s - N_o^*}{R_{D1}} \quad (5)$$

where R_{D1} is on resistance of diode $D1$, and N_s and N_o^* are the specific loudness of the current frame and output specific loudness of the previous frame, respectively. If $I_n < 0$, it is set to zero since diode D1 is off. Let N_o^* and V^* be the voltages across capacitors $C1$ and $C2$ in the previous frame, respectively. If $N_o^* > V^*$ (meaning diode D2 is off), then

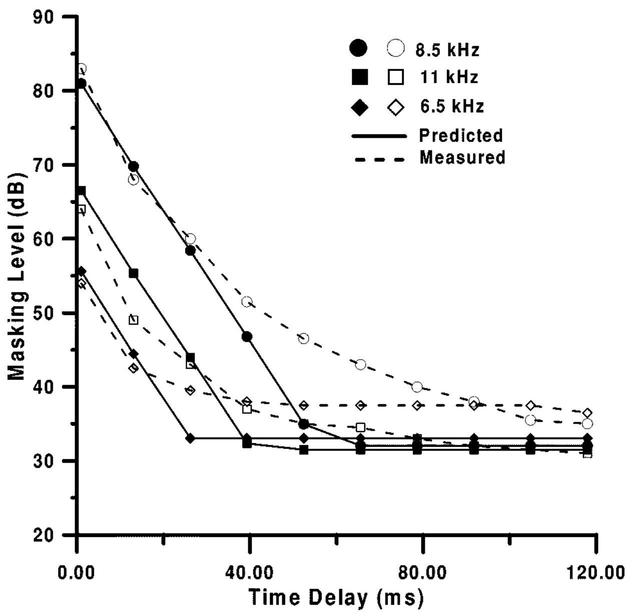
$$N_o = \frac{I_n + C2 \cdot V^*/(\Delta t + C2 \cdot R2) + C1 \cdot N_o^*/\Delta t}{1/R1 + C2/(\Delta t + C2 \cdot R2) + C1/\Delta t} \quad (6)$$

and

$$V = N_o^* \cdot \Delta t/(\Delta t + R2 \cdot C2) + V^* \cdot C2 \cdot R2/(\Delta t + C2 \cdot R2). \quad (7)$$



(a)



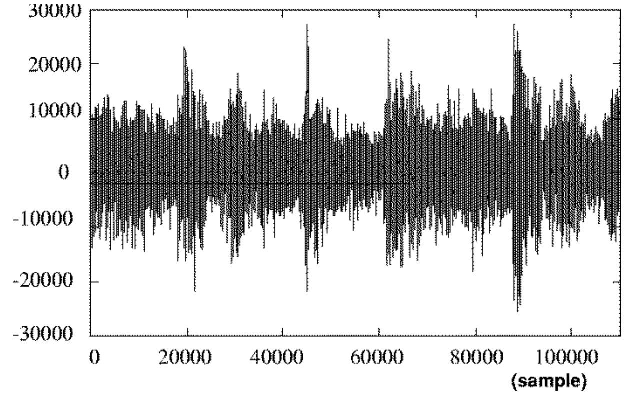
(b)

Fig. 4. (a) Forward masking patterns for a 5-ms tonal probe at 2 kHz. The masking at are drawn both for the estimated and measured cases with 5, 10, 30, and 200 ms UMN maskers at 60 dB. (b) Forward masking levels of 6.5, 8.5, and 11 kHz tonal probes by a 8.5 kHz CBN masker.

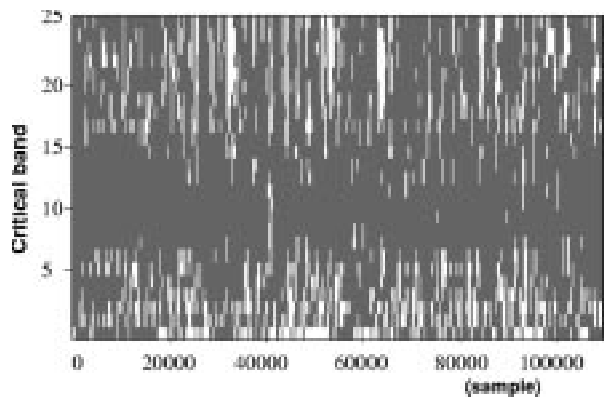
On the other hand, if diode D2 is on, then

$$V = N_o = \frac{I_n + (C1 + C2) \cdot V^*/\Delta t}{1/R1 + (C1 + C2)/\Delta t}. \tag{8}$$

In the RC circuit, the increase in masker duration corresponds to storing more charges into the capacitors. Therefore, if the capacitors are charged with a longer pulse, N_o will be larger. On the contrary, if the impulse is short, the output of the circuit will be smaller. The values of the resistors and capacitors are determined with a view to matching the phenomena measured using maskers longer than 200 ms [21], [22].



(a)



(b)

Fig. 5. (a) Piece of orchestra music and (b) the dominance pattern in the time-frequency space: simultaneous (frequency) masking effect (black) and forward masking effect (white).

C. Simulation of the Forward Masking Model

Three kinds of masker signals, uniform masking noise (UMN), critical band noise (CBN), and an orchestra signal are used in the following simulations about the forward masking model.

UMN [10] is filtered from white noise so as to produce constant masking level at all critical bands. Using a signal that has absolute masking level of 0 dB as the reference signal, we generate a 60-dB UMN as masker to find the threshold of a 5-ms tonal probe (maskee) at 2 kHz. In Fig. 4(a), both the masking level predicted by the simulation, denoted by solid line, and the measured data in [21], denoted by dashed line, are depicted for different masker duration T_m. The figure shows that the predicted total masking level always lies within 5 dB of the measured results. Fig. 4(b) shows both the simulated and the measured forward masking levels of a CBN masker centered at 8.5 kHz. The probes are 6.5, 8.5, and 11 kHz tonal signals, respectively. Again in most cases, the predicted masking levels agree quite well with the measured data.

Finally, we use a piece of orchestra music to examine the forward masking effect in a more realistic setting. Referring to (4), the total masking level is determined either by simultaneous masking or forward masking depending on which effect produces stronger masking level. An interesting issue is to examine which of the two masking effects is more dominant. The orchestra music segment as shown in Fig. 5(a) is encoded using the proposed masking model. In Fig. 5(b), the white

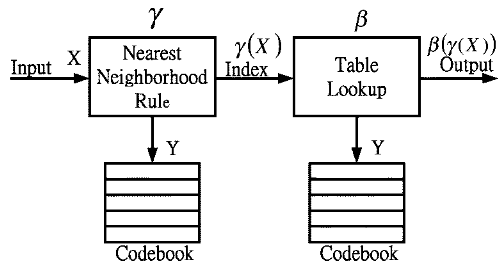


Fig. 6. Architecture of a vector quantization system.

regions denote time-frequency slots where the forward masking effect is dominant and the black regions mean otherwise. All white regions imply higher masking level than if only simultaneous masking is considered, as in the traditional audio coders [1]. Therefore, more quantization noise can be tolerated and fewer bits are needed for coding. Since the forward masking effect is dominant in quite a significant portion of the whole time-frequency space, we expect the proposed masking model will improve audio coding efficiency significantly.

III. Vector Quantization and Self-Organizing Feature Map

In this section, we introduce the basic formulation of vector quantization (VQ) and the algorithm of self-organizing feature map (SOFM). Furthermore, a binary-tree-structured SOFM (BTSOFM) is introduced as a flexible VQ coding scheme for perceptual audio coding.

A. Vector Quantization

In a vector quantization system shown in Fig. 6, there are two mappings: one in the encoder and the other in the decoder. For each input pattern X, an encoder assigns a symbol $\gamma(X)$ from the symbol set according to the nearest neighbor rule. The decoder then looks up the codeword corresponding to the symbol, $\beta(\gamma(X))$. So, a vector quantizer can be defined as a mapping Q from the L-dimensional Euclidean space R^L into a finite set $\boldsymbol{Y}$ consisting of M points in R^L. Thus, the vector quantizer, Q, is defined as

$$Q: R^L \rightarrow \boldsymbol{Y}$$

where $\boldsymbol{Y} = \{W^{(m)};\, m = 1, 2, \ldots, M\}$ is the set of codewords (codebook) and M is the size of $\boldsymbol{Y}$. Usually, a VQ system chooses the codeword whose Euclidean distance to the input pattern X is minimum. Thus, the reconstructed signal suffers minimal sum-of-squared-error distortion.

B. Self-Organizing Feature Map

A famous neural network model, self-organizing feature map (SOFM) [32], is often used for training vector quantization codebooks in various lossy signal compression systems because of its capability of clustering without supervision and the flexibility of the codebook structure it generates. SOFM has been shown to yield VQ codebooks that are better than those generated by the conventional generalized Lloyd algorithm [33].

Typically, SOFM network is composed of a discrete one-dimensional (1-D) or two-dimensional (2-D) lattice of

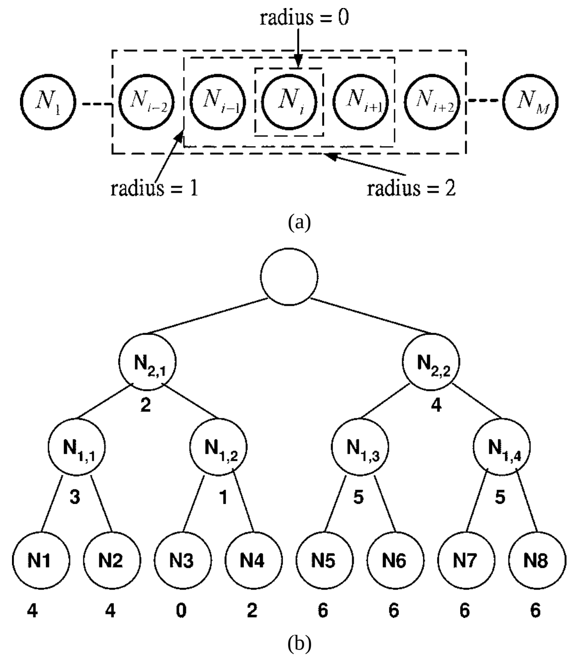


Fig. 7. (a) One-dimensional neuron structure in SOFM and (b) binary tree neuron structure in BTSOFM.

SOFM Algorithm

(1) Randomly initialize all L-demensional weight vectors of an M-neuron SOFM network

(2) Input a training pattern

(3) Exhaustively search all neurons to find the one with closest weight vectors

(4) Update weight vectors of all neurons within "current neighborhood".

(5) If not in equilibrium, go to step (1). If in equilibrium, then shrink neighborhood, decrease learning rate, and go to step (2). If in equilibrium and the neighborhood has only one neuron, then stop.

Fig. 8. Self-organizing feature map learning algorithm.

neurons. Each of the M neurons in the network has a weight vector W with dimension L. The weight vector of each neuron represents one codeword in the codebook. In the beginning, all weight vectors are initialized. During the training of the network, the neuron whose weight vector is closest to the current input training pattern is identified. Then, the weight vectors of the winning neuron and all neurons in its "current neighborhood" are updated in the direction toward the input pattern. In other words

$$W^{(m)} = W^{(m)*} + \alpha(T,\, d)\left(X - W^{(m)*}\right) \tag{9}$$

where $W^{(m)}$ and $W^{(m)*}$ are the updated and original weight vector of the mth neuron, respectively; T is the training iteration count; d is the distance between the neuron m and the winning neuron; and $\alpha(T,\, d)$ is the learning rate that controls the speed of the training process.

After the neurons are trained by all input patterns, an *epoch* is completed and the network is retrained by all input patterns for another epoch. For each epoch in the training process, the

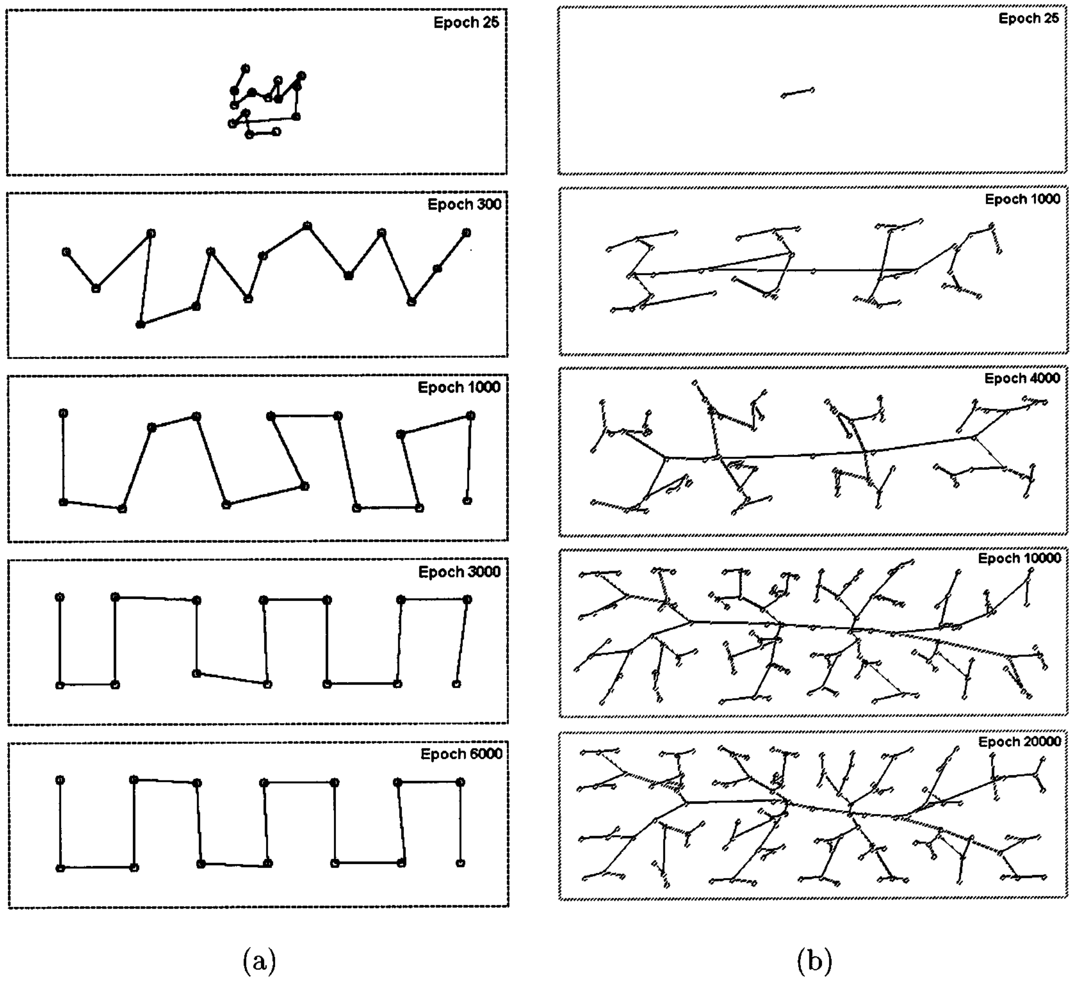


Fig. 9. Neuron distributions during the training process for (a) the 1-D SOFM network and (b) the BTSOFM network.

"current neighborhood" for updating weight vectors must be redefined. For a better clustering performance, the learning rate and the size of the current neighborhood should be gradually decreased. Fig. 7(a) depicts the generic 1-D SOFM network. The three neighborhoods contain all neurons inside the regions centered at the winning neuron N_i with radius two, one, and zero, respectively. Similarly, the current neighborhood of other neuron structures (e.g., 2-D lattice) can be defined as a set of neurons centered at the winning neuron. At the end of the training process, the weight vectors of the neurons are distributed to reflect the statistical nature of the training patterns and they are used as codewords in the codebook. Moreover, neighboring neurons will represent codewords that are alike since neighboring neurons are updated almost simultaneously. The SOFM learning algorithm is summarized in Fig. 8.

An example of the training process of a 1-D SOFM network is shown in Fig. 9(a). The learning rate is defined as

$$\alpha(T, d) = \frac{0.3}{1 + T/20} \cdot e^{-d^2/(r+1)^2} \tag{10}$$

where r is the radius of the current neighborhood and d is the distance between the winning neuron and the neuron being updated. Input patterns are randomly and uniformly distributed in the rectangular region. Early in the training process, the learning rate is high and the neighborhood is large, so the neurons are quickly pulled apart. Later in training, the neurons move slowly because the learning rate is low and the neighborhood is small. The learning rate and the neighborhood decrease gradually during the training process. Therefore, the SOFM algorithm is less likely to be stuck at local minima and it usually yields better codebooks than those designed by traditional descent-based methods with fixed learning rate.

The SOFM algorithm generates vector quantization codebooks with better quality. However, 1-D and 2-D SOFM networks are not suitable for designing high-dimensional codebooks. To remedy this, a binary-tree structure SOFM (BTSOFM) is proposed [20]. With binary-tree structure among the codewords, BTSOFM is suitable for progressive coding and variable bit-rate coding. In addition, the tree-structured codebook makes tree search of the codebook possible, thus reducing the encoding complexity.

The BTSOFM learning algorithm is similar to SOFM except for the neuron structure and the current neighborhood definition. In BTSOFM, tree search is used to locate the nearest neuron to a training pattern. Fig. 7(b) illustrates a three-level, eight-terminal-node BTSOFM structure. The numbers shown under the terminal nodes are the distance between the fourth node (N_4) and all respective nodes. The distance is defined by the number of hops between two nodes along the binary tree. Initially, all the nodes, including the terminal nodes and inter-level nodes, form a full binary-tree. As the training goes on, this high-dimensional tree is gradually stretched so that the terminal nodes reflect training-pattern distribution. At the same time, the inter-level nodes are also updated in such a way that a binary tree structure is always retained, so progressive coding using different bit rates (codebook sizes) can be adopted. Another example of using a six-level BTSOFM network is depicted in Fig. 9(b). Initially, neurons are all located in the center and input patterns are again randomly and uniformly distributed in the rectangular region. The neurons are finally stretched to reflect the distribution of the input patterns.

The BTSOFM possesses one characteristic that is crucial for codebooks for perceptual audio coding. Note that in addition to the codebook made up of all the terminal nodes, progressively smaller codebooks, each consisting of all nodes on a higher-level, are generated during BTSOFM training. These smaller codebooks can also be used for vector quantization, albeit with higher quantization noise. With these progressively smaller codebooks at hand, one can choose the smallest one that yields quantization noise just below the masking threshold so as to make the noise imperceptible. With smaller codebooks and fewer bits for codeword index, the coding efficiency can be enhanced while the sound quality is not compromised perceptually. To this end, a modified BTSOFM algorithm will be proposed to design better codebooks for perceptual audio coding in the next section.

IV. Audio Coding With Perceptually Weighted BTSOFM VQ

In this section, we propose a VQ-based perceptual audio coding scheme whose codebooks are designed by a perceptually weighted BTSOFM algorithm. This algorithm is based on a perceptually weighted error criterion. We will introduce the new criterion first and explain why it is better than the traditional error criterion.

A. Audio Coding Scheme

In the proposed audio coding scheme shown in Fig. 10, we use the same hybrid analysis as in the MPEG1 Layer III scheme, which has 576 MDCT coefficients in the frequency domain. The vector quantization block uses a gain-shape vector quantization codebook whose codewords have $2 \times n$ components, n MDCT coefficients from each of the two consecutive frames. The exact grouping of the MDCT coefficients is shown in Fig. 11. On the total, 22 bands cover 576 MDCT coefficients and 15 codebooks with different vector dimension ($2n$) are used to encode these MDCT coefficients. The bandwidth n increases approximately exponentially with frequency, somewhat consistent with the critical band scale or Mel scale [1].

To exploit perceptual characteristics in vector quantization, MDCT coefficient vectors are supposed to be surrounded by a masking region in the vector space as shown in Fig. 12(a). If the selected codeword is inside the masking region, the quantization distortion is imperceptible. Otherwise, perceptible noise will occur. The concept is similar to shaping scalar quantization error below the masking level except that the 1-D masking curve in the frequency domain is now extended to the general masking region in the L-dimensional space.

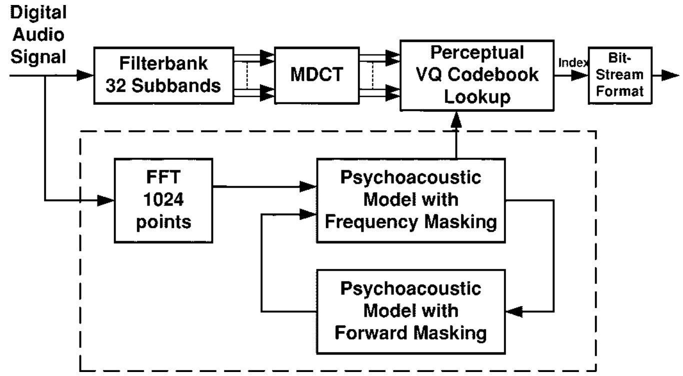


Fig. 10. Architecture of the proposed perceptual VQ-based audio coding system.

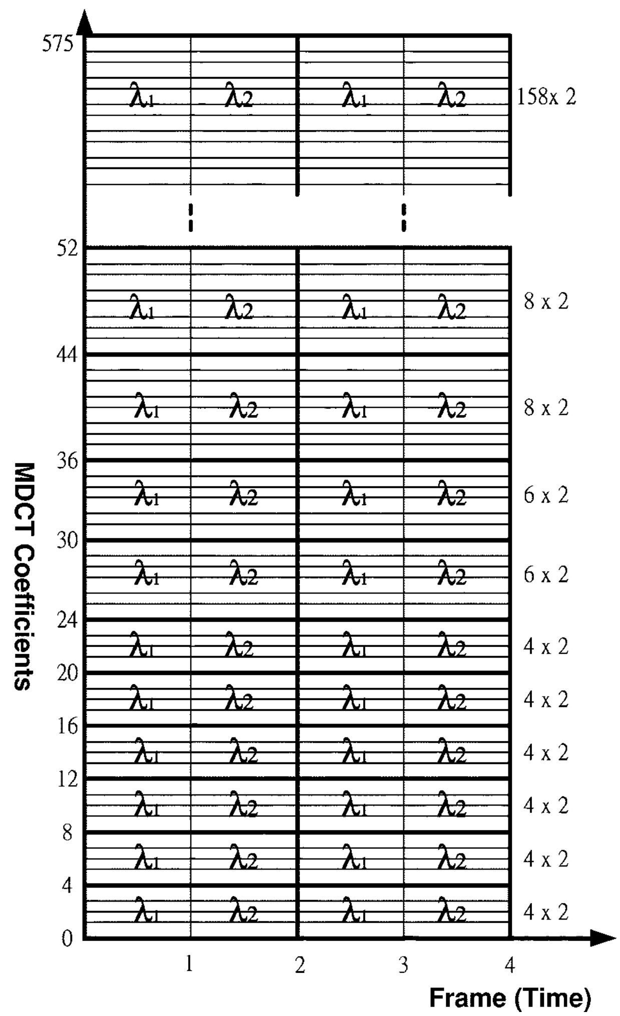


Fig. 11. Definition of critical band vectors used in the proposed audio coding scheme.

For each $2n$-dimensional MDCT coefficient vector, two signal-to-masking ratio (SMR) values (λ_1 and λ_2), as shown in Fig. 11, are calculated. The masking level is computed according to (4), thus includes both the frequency masking and the forward masking effects. The SMR values determine the masking region in the vector space, with which the codeword that causes least perceptible noise can be selected. In addition, the SMR values are also supervisory information used during

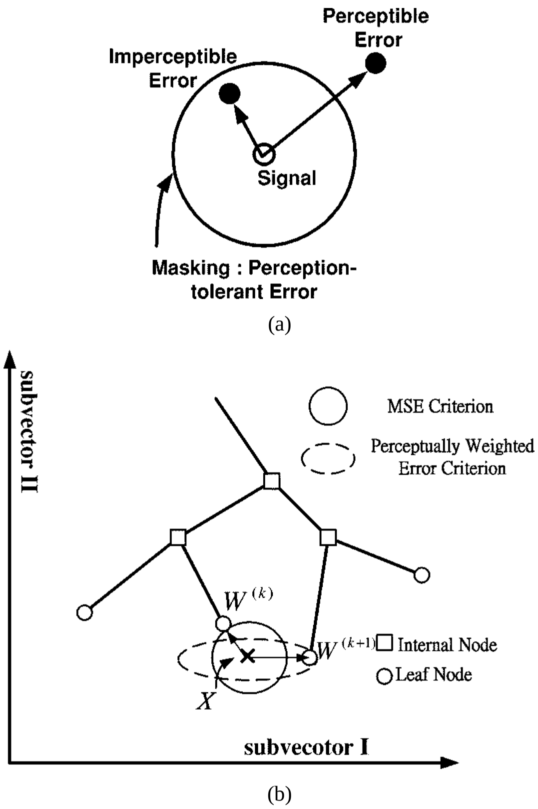


Fig. 12. (a) Concept of masking region in vector quantization and (b) perceptually weighted error criterion used in searching for the nearest codeword.

codebook training so as to generate a perceptually optimal codebook.

B. Perceptually Weighted Error Criterion

In traditional vector quantization, minimum sum-of-squared-difference is used to determine the nearest codeword. Geometrically, the winning neuron (codeword) has the smallest Euclidean distance to the input pattern among all codewords. In our case, however, the two SMR values can be quite different, implying different levels of noise can be tolerated. One reasonable approach to accommodate this fact is to use a weighted error (distance) criterion.

Let a $2n$-dimensional normalized (shape) input vector be denoted as

$$X = (X_1, X_2) = (x_{11}, \ldots, x_{1n}, x_{21}, \ldots, x_{2n}) \tag{11}$$

where x_{1i} and x_{2i} are computed from the MDCT coefficients from the first frame and the second frame, respectively. Let the mth codeword in a codebook be denoted as

$$W^{(m)} = \left(W_1^{(m)}, W_2^{(m)}\right) = (w_{11}, \ldots, w_{1n}, w_{21}, \ldots, w_{2n}). \tag{12}$$

According to the perceptual masking criterion, the quantization error should be smaller than the masking level. In other words, the signal-to-quantization-noise ratio should be larger than the signal-to-masking ratio (SMR) in each of the subvectors

$$\lambda_1 \leq \frac{\|X_1\|^2}{\|E_1\|^2} \quad \text{and} \quad \lambda_2 \leq \frac{\|X_2\|^2}{\|E_2\|^2} \tag{13}$$

where $E = X - W^{(m)} = (E_1, E_2)$. To facilitate codebook training and searching, we propose a new perceptually weighted error criterion E_{per}, where

$$E_{per} = \sum_{i=1}^{2} \lambda_i \cdot \frac{\|X_i - W_i^{(m)}\|^2}{\|X_i\|^2}. \tag{14}$$

If E_{per} is less than one, then the SNR is higher than the SMR is both subvectors, ensuring imperceptible quantization noise. Geometrically, the imperceptible-noise region (masking region) becomes an ellipsoid, thus the name perceptually weighted error criterion. In the example shown in Fig. 12(b), codewords of a tree-structured codebook and an input pattern are shown in the coordinates of the two subvectors. The SMR in the first subspace is smaller than that in the second subspace. Therefore, instead of the nearest codeword in Euclidean distance, $W^{(k)}$, the perceptually nearest codeword $W^{(k+1)}$ should be chosen.

C. Perceptually Weighted BTSOFM

The strategy of the perceptually weighted BTSOFM algorithm, a modified version of the BTSOFM algorithm, consists of the perceptually weighted error criterion and an updating process according to the SMR values. We use perceptually weighted BTSOFM to train 15 tree-structured codebooks needed in the proposed audio coding scheme. The BTSOFM training procedure is described in the following.

1) Each codebook is a binary tree of depth 12, containing 4094 inter-level neurons and 4096 terminal neurons. In the beginning, the weight vectors of all neurons are initialized and the training starts with a neighborhood of distance 24, covering all nodes in the network.
2) For each training pattern, a binary tree search is used to identify the perceptually nearest neuron to the input pattern according to the perceptually weighted error criterion formulated in (14).
3) All neurons inside the current neighborhood of the winning are updated according to the SOFM update rule in (9) except that now the distance is the BTSOFM tree distance and the learning rate depends on not only T and d, but also λ_i.
4) After ten runs of all patterns in the training set, "equilibrium" is assumed and the radius r is decreased by one. When the radius reaches zero and the neighborhood contains only the winning node itself, then the training stops after reaching equilibrium.

The learning rate α is given by

$$\alpha(T, d, \Lambda_i) = \frac{0.4}{1 + T/500} \cdot e^{-d^2/(r+1)^2} \cdot \frac{e^{\Lambda_i}}{e^{\Lambda_i} + e^{-\Lambda_i}} \tag{15}$$

where $\Lambda_i = 10 \log_{10} \lambda_i$ is the SMR value expressed in log scale and it is used to compute the learning rate for all components in the ith subvector. The term that depends on Λ_i is in the form of a sigmoidal function, which saturates to 1 if Λ_i is large and to zero is Λ_i is small. This has the effect of limiting the adjustment strength of high-SMR input patterns as well as completely ignores the input patterns overwhelmed by masking.

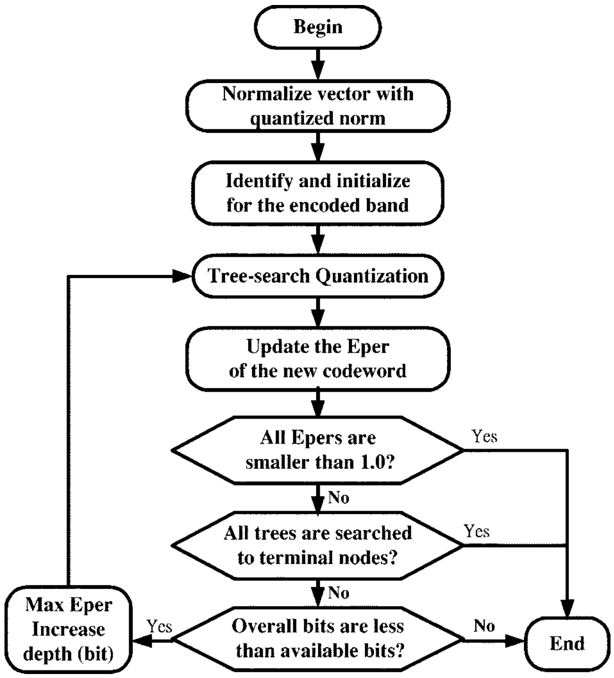


Fig. 13. Flow chart for bit allocation and tree-structured codebook searching.

D. Bit Allocation Using Tree-Structured VQ

In the proposed audio coding scheme, 576 MDCT coefficients in each frame are divided into 22 vectors (critical bands) and each quantized by one of the 15 codebooks of corresponding size. These 15 tree-structured codebooks are trained by the perceptually weighted BTSOFM. As mentioned in the previous section, a D-level binary-tree-structured codebook actually contains in it smaller tree codebooks of size 2^t for all t smaller than D. Therefore, bit allocation for each critical band needs to determine the size of the codebook used to encode the critical band vector. The available bits must be precisely assigned to control the bit rate and the distortion caused by the gain-shape vector quantizer. Thus, we employed an iteration loop in Fig. 13 for bit allocation and tree-search of the gain-shape tree-structured VQ. The algorithm is basically a greedy method with the perceptually weighted error criterion and is described in the following.

1) The magnitude (gain) of an input vector is quantized with four bits for the first 11 bands and three bits for the next 11 bands. The input vector is then normalized by the quantized gain to get the normalized input vector (shape). No bits will be assigned to encode the bands with both Λ_i lower than -3 dB due to strong masking.
2) In the tree-search quantization stage, the perceptually weighted error E_{per} is used as the criterion for selecting codewords for shape vectors in 22 bands. All 22 shape vectors search in their corresponding codebook tree in parallel. Bit allocation and tree searching proceed simultaneously, i.e., traversing down one level in the tree leads to one more bit assigned to encode the shape vector.
3) For each iteration, if at least one band has a E_{per} higher than one, one more bit is assigned to the band that has the highest E_{per} value. Then, the codebook tree is traversed one level down and E_{per} of this band is updated.
4) The iteration stops when E_{per} for all bands are less than one, or all the codebook trees are searched to terminal nodes, or all available bits have been allocated. Otherwise, the procedure goes back to the previous step.

TABLE I
TWENTY SOUND ITEMS USED IN AUDIO CODING QUALITY MEASUREMENT EXPERIMENTS

Sound Items	
Saxophone	Tenor (Jose Carreras)
Male voice	Female song
Female voice, drum and cello	Synthetic music
Electrical guitar and violin	Electrical guitar
Violin and piano	Rock drum
Orchestra	Folk guitar
Piano	Harmonica
Chorus song	Folk song (Male)
Flute	Folk song (Female)
Bass	Chinese violin and flute

When the procedure stops and if the bit rate is high enough, then quite possibly all E_{per} are less than one, which means that no quantization noise will be perceptible in any of the 22 bands. However, if the bit rate is not high enough, then the bits will be assigned so as to make E_{per}, and thus the perceptually weighted distortion, in 22 bands as small as possible.

V. EXPERIMENTAL RESULTS

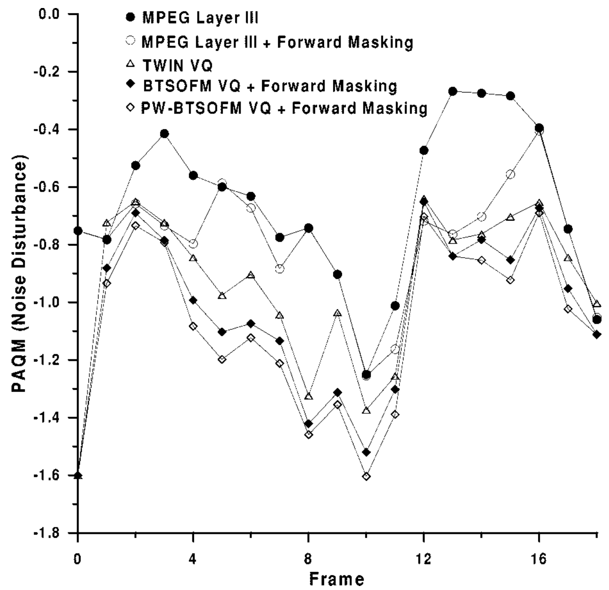
Twenty mono sound items with 44.1 kHz sampling rate, 16-bit resolution, and 30-s duration are selected for experiments (see Table I). To evaluate sound quality at different bit rates, seven different bit rates of 80, 64, 56, 48, 40, 32, and 24 kbps are used. The decoded sound items by MPEG layer III coding scheme are used as the baseline for comparison.

A. Objective Sound Quality Measurements

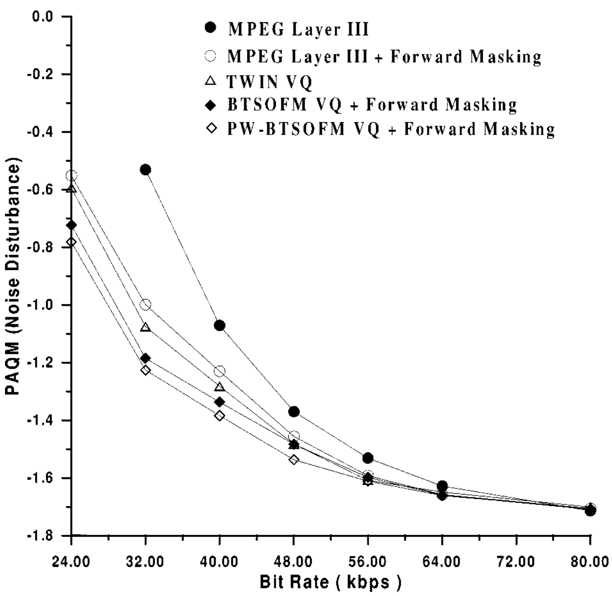
To evaluate the sound quality, we use the perceptual audio quality measure (PAQM) [34] for objective assessment of the sound quality. It is one way of measuring noise disturbance, which ranges from -1.7 to -0.3 corresponding to 5 to 1 in the mean opinion score (MOS). This method measures the quality of an audio coding scheme by mapping the input and output of the coding scheme from physical signal representation onto a psychoacoustic representation. This mapping enables quantification of perceptual degradation introduced by the audio coding scheme. With this mapping, subjective quality of the reconstructed audio signal can be estimated. Besides, this method can measure the sound quality at different time in a sound segment. Thus, we can see the variation of sound quality in a sound segment.

Fig. 14(a) shows the sound quality of a sound segment using five different audio coding schemes. By comparing the MPEG Layer III scheme and the MPEG Layer III scheme with the forward masking model at 48 kbps (both use scalar quantization), we can see that the sound quality is markedly improved since the forward masking model is adopted. In addition, the performance of PW-BTSOFM VQ is better than that of BTSOFM VQ and TWIN-VQ in [14] because of the additional sigmoidal function and perceptually weighted error criterion in PW-BTSOFM VQ. Of course the three vector quantization audio schemes are better than the two scalar quantization schemes, with or without forward masking.

The PAQM noise disturbance versus bit rate for all five audio coding schemes are shown in Fig. 14(b). The figure shows that



(a)



(b)

Fig. 14. (a) PAQM noise disturbance profiles in a sound segment and (b) PAQM noise disturbance versus different bit rated using five audio coding schemes.

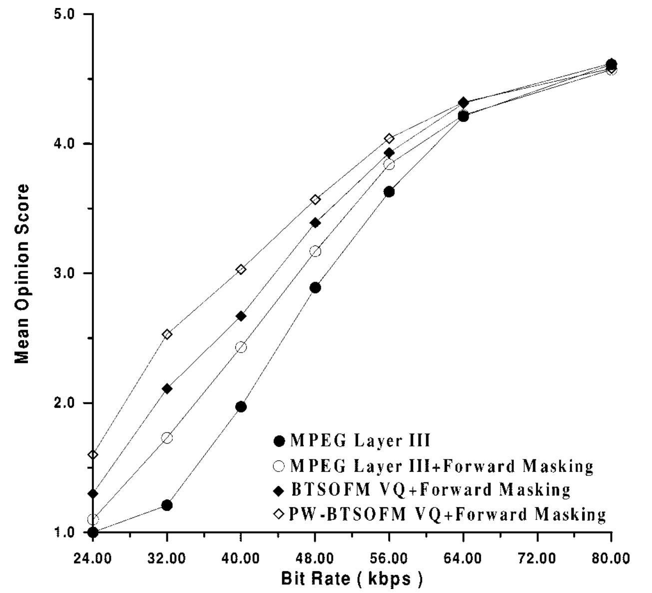


Fig. 15. MOS measurements for MPEG layer III standard with and without forward masking, BTSOFM, and PW-BTSOFM VQ coding schemes at different bit rates.

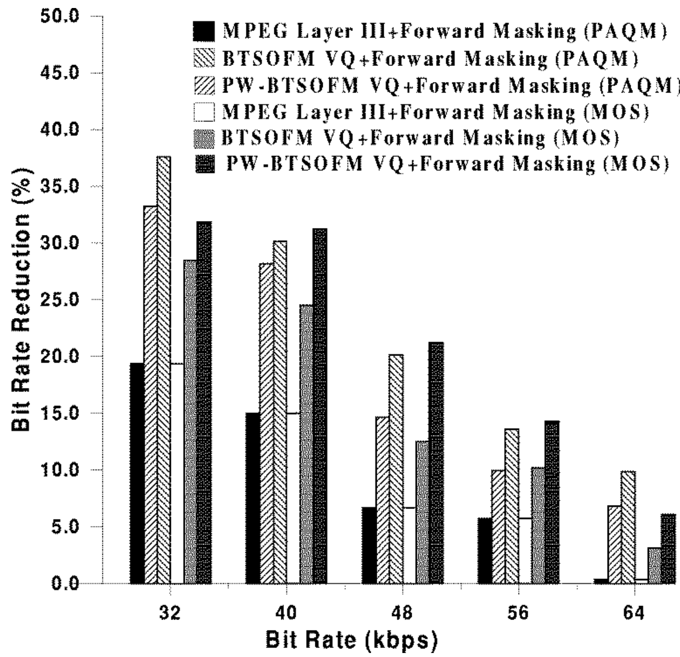


Fig. 16. Percentages of bit rate reduction using BTSOFM, PW-BTSOFM, and MPEG layer III scheme with forward masking.

PAQM noise disturbance of the MPEG layer III coding scheme with forward masking model increases with decreasing bit rate at a much slower pace than the original MPEG layer III scheme. This implies that the forward masking model is very crucial to sound quality performance, especially in low bit-rate audio coding. Moreover, the PW-BTSOFM VQ coding further improves sound quality and it outperforms the MPEG layer III scheme with the forward masking model and the TWIN-VQ scheme, especially in low bit rate settings.

B. Subjective Sound Quality Measurements

Subjective listening tests are carried out to evaluate the subjective quality of the proposed coding scheme by mean opinion score (MOS). This score goes from 1.0 (very annoying distortion) to 5.0 (inaudible distortion). Fourteen listeners evaluate the sound quality of the 20 sound items at 24- to 80-kbps bit rates. Their results are shown in Fig. 15. The average MOS curves also demonstrate the superiority of the forward masking model and the perceptually weighted BTSOFM VQ.

By simple interpolation, we can derive the percentages of bit reduction, shown in Fig. 16, for different audio coding schemes given that the same sound quality as the MPEG layer III scheme is required. The rate reduction computed from the PAQM experiments and that from the MOS experiments are similar and they both illustrate enhanced performance of the proposed scheme.

With the forward masking model, the bit-rate reduction of the three schemes is less than 10% above 56 kbps. On the other hand, the two VQ-based schemes require much less bits than MPEG layer III below 48 kbps. At the bit rate of 32 kbps, the MPEG layer III with the forward masking model require 25% less bits than the original MPEG layer III scheme. Moreover, BTSOFM VQ and PW-BTSOFM VQ schemes achieve 33% and 38% bit-rate reduction in the PAQM measurements, and 28% and 32% bit-rate reduction in the MOS listening tests, respectively.

VI. Conclusion

In this paper, the forward masking model using a RC analog circuit is exploited to estimate the forward masking effect. Since the proposed RC analog circuit is time-varying and nonlinear, it is more accurate in estimating the forward masking level. Due to the long-term (200 ms) forward masking effect on the ensuing signals, more imperceptible signal components can be eliminated and the coding efficiency improved with the forward masking effect taken into account.

Moreover, we proposed the perceptually weighted BTSOFM algorithm that considers a perceptually weighted error criterion in vector quantization codebook design. The codebook has a binary-tree structure and inherently contains several progressively smaller codebooks with similar codeword distribution. This feature makes feasible a new bit assignment algorithm proposed in this paper to minimize the perceptible quantization noise. With all the above properties, the proposed audio coding scheme can achieve up to about 40% bit rate reduction when compared to the standard MPEG Layer III audio coding scheme.

References

[1] ISO/IEC Std. 11 172-3, "Information technology—Coding of moving pictures and associated audio for digital storage media at up to about 1,5Mbits/s," Switzerland, 1993.

[2] ETS 300 401, *Radio Broadcasting Systems; Digital Audio Broadcasting (DAB) to Mobile, Portable and Fixed Receivers*, 2nd ed, France: Eur. Telecommun. Std. Inst., 1997.

[3] D. Sinha and J. D. Johnston, "Audio compression at low bit rates using a signal adpative switched filterbank," in *Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Processing*, 1996, pp. 1053–1056.

[4] N. Jayant, J. Johnston, and R. Safranek, "Signal compression based on models of human perception," *Proc. IEEE*, vol. 81, pp. 1385–1421, Oct. 1993.

[5] A. Gersho, "Advances in speech and audio compression," *Proc. IEEE*, vol. 82, pp. 900–918, June 1994.

[6] P. Noll, "Digital audio coding for visual communications," *Proc. IEEE*, vol. 83, pp. 925–943, June 1995.

[7] M. R. Schroeder, "Optimizing digital speech coders by exploiting masking properties of the human ear," *J. Acoust. Soc. Amer.*, vol. 66, pp. 1647–1651, Dec. 1979.

[8] S. D. Stearns, "Arithmetic coding in lossless waveform compression," *IEEE Tran. Signal Processing*, vol. 43, pp. 1874–1879, Aug. 1995.

[9] J. D. Johnston, "Transform coding of audio signals using perceptual noise criteria," *IEEE J. Select. Areas Commun.*, vol. 6, pp. 314–323, Feb. 1988.

[10] E. Zwicker and H. Fastl, *Psychoacoustics—Facts and Models*. Berlin, Germany: Springer-Verlag, 1990.

[11] B. C. J. Moore, *Introduction to the Psychology of Hearing*, 4th ed. New York: Academic, 1997.

[12] W. Y. Chan and A. Gersho, "High fidelity audio transform coding with vector quantization," in *Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Processing*, Apr. 1990, pp. 1109–1112.

[13] ——, "Constrained-storage vector quantization in high fidelity audio transform coding," in *Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Processing*, Apr. 1991, pp. 3597–3600.

[14] N. Iwakami, T. Moriya, and S. Miki, "High-quality audio-coding at less than 64 kbp/s by using transform-domain weighted interleave vector quantization," in *Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Processing*, May 1995, pp. 3095–3098.

[15] H. Najafzadeh-Azghandi and P. Kabal, "Perceptual coding of narrowband audio signals at 8 kbp/s," in *Proc. IEEE Workshop on Speech Coding for Telecommun.*, Sept. 1997, pp. 109–110.

[16] ——, "Improving perceptual coding of narrowband audio signals at low rate," in *Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Processing*, vol. 2, Mar. 1999, pp. 913–916.

[17] P. Philippe, F. M. de Saint-Martin, and M. Lever, "Wavelet packet filterbanks for low time delay audio coding," *IEEE Trans. Speech Audio Processing*, vol. 7, pp. 310–321, May 1999.

[18] P. Monta and S. Cheung, "Low rate audio coder with hierarchical filterbanks and lattice vector quantization," in *Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Processing*, vol. 2, Apr. 1994, pp. II/209–212.

[19] K. Ferens and W. Kinsner, "Adaptive wavelet subband coding for music compression," in *Proc. IEEE Data Compression Conf.*, Mar. 1995, p. 460.

[20] T. D. Chiueh, T. T. Tang, and L. G. Chen, "Vector quantization using tree-structured self-organizing feature maps," *IEEE J. Select. Areas Commun.*, vol. 12, no. 9, pp. 1594–1599, Dec. 1994.

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

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

[23] H. Fastl, "Temporal masking effects: II. Critical band noise masker," *ACUSTICA*, vol. 36, no. 5, pp. 317–330, 1977.

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

[25] T. Dau and D. Puschel, "A quantitative model of the 'effective' signal processing in the auditory system. I. Model structure," *J. Acoust. Soc. Amer.*, vol. 99, pp. 3615–3622, Jun. 1996.

[26] ——, "A quantitative model of the 'effective' signal processing in the auditory system. II. Simulations and measurements," *J. Acoust. Soc. Amer.*, vol. 99, pp. 3623–3631, June 1996.

[27] F. Baumgarte, "A physiological ear model for specific loudness and masking," in *Proc. IEEE ASSP Workshop on Application of Signal Processing to Audio, Acoustics*, Oct. 1997, p. 4.

[28] J. M. Kates, "A time-domain digital cochlear model," *IEEE Trans. Signal Processing*, vol. 39, pp. 2573–2591, Dec. 1991.

[29] E. Zwicker, "Procedure for calculating loudness of temporal variable sounds," *J. Acoust. Soc. Amer.*, vol. 62, pp. 675–682, Sept. 1977.

[30] B. C. J. Moore and B. R. Glasberg, "A revision of Zwicker's loudness model," *ACUSTICA*, vol. 82, pp. 335–345, 1996.

[31] H. Fletcher and W. A. Munson, "Relation between loudness and masking," *J. Acoust. Soc. Amer.*, vol. 9, pp. 1–10, 1937.

[32] T. Kohonen, *Self-Organization and Associative Memory*. New York: Springer-Verlag, 1988.

[33] N. M. Nasrabadi and Y. King, "Vector quanization of images based upon the Kohonen self-organizing feature maps," in *Proc. IEEE Int. Conf. Neural Networks*, July 1988, pp. I101–108.

[34] J. G. Beerends and J. A. Stemerdink, "A perceptual audio quality measure based on a psychoacoustic sound representation," *J. Audio Eng. Soc.*, vol. 40, no. 12, pp. 963–978, Dec. 1992.

Yuan-Hao Huang (S'98–M'02) was born in Taiwan, R.O.C., in 1973. He received the B.S. and Ph.D. degrees in electrical engineering from National Taiwan University, Taipei, in 1995 and 2001, respectively.

He is currently a Member of Technical Staff at VXIS Technology Corporation, Hsinchu, Taiwan. His primary research interests include speech and audio coding, digital signal processing system design and VLSI design of speech, audio, and telecommunication systems.

Tzi-Dar Chiueh (S'87–M'90) was born in Taipei, Taiwan, R.O.C., in 1960. In 1983, he received the B.S.E.E. degree from the National Taiwan University, Taipei. He also received the M.S. and Ph.D. degrees in electrical engineering from the California Institute of Technology, Pasadena, in 1986 and 1989, respectively.

Since 1989, he has been with the Department of Electrical Engineering, National Taiwan University, where he is presently a Professor. His research interests include IC design for digital communication systems and analog neuromorphic systems.