A Masking-Threshold-Adapted Weighting Filter for Excitation Search

Wen-Whei Chang, *Member, IEEE,* and Chin-Tun Wang

***Abstract*—Most LPC-based audio coders improve reproduction quality by using predictor coefficients to embody perceptual masking in noise spectral shaping. Since the predictor coefficients were originally derived to characterize sound production models, they cannot precisely describe the human ear's nonlinear responses to frequency and loudness. In this paper, we report on new approaches to exploiting the masking threshold in the design of a perceptual noise-weighting filter for excitation searches. To track the nonstationary evolution of a masking threshold, an autoregressive spectral analysis with finite order has been shown to be capable of providing sufficient accuracy. In seeking faster response, an artificial neural network was also trained to extract autoregressive modeling parameters of the masking threshold from typical audio signals via mapping. Furthermore, we propose the concept of sinusoidal excitation representation to better track the intrinsic characteristics of prediction error signals. Simulation results indicate that the combined use of a multisinusoid excitation model and a masking-threshold-adapted weighting filter allows the implementation of an LPC-based audio coder that delivers near transparent quality at the rate of 96 kb/s.**

I. Introduction

FOR MANY years there has been considerable interest in transparent reproduction of bit rate reduced audio signals, not just for using statistical correlation to remove redundancies but also to eliminate the perceptual irrelevancy by applying psychoacoustic measures [1]. In many coding applications, it is generally sufficient to assume that coded signals are contaminated by some additive reconstruction noise. Many perceptual coding schemes have been proposed to prevent the appearance of such audio artifacts. In essence, the system is not modeled with respect to the source waveform itself, but is designed to take into consideration the human perception of sound. Recent research has placed emphasis on either transform coding [2] or subband coding [3]. In both cases, the audio frequency range is subdivided into critical bands and then quantized in accordance with the estimated masking threshold.

An alternative approach to audio representation is based on the linear predictive coding (LPC) model, in which audio signals are decomposed into the product of excitation and system spectra. Unfortunately, however, most psychoacoustic experiment results are expressed in the frequency domain and are not directly applicable for use in conjunction with the LPC model. For this reason, conventional LPC-based coders employ relatively simple techniques for incorporating the perceptual masking properties either in postfiltering [4] or in noise feedback coding [5]. Even recently proposed analysis-by-synthesis predictive coders also utilize the predictor coefficients to implement perceptual noise-weighting filters for excitation searches [6]. The basic problem with this approach is that the predictor coefficients were originally determined to characterize sound production models and hence cannot precisely describe human perception of sound. Further improvement can only be realized through some intelligent exploitation of new findings in psychoacoustic studies. Since the audibility of noise depends heavily on its spectral shaping, we attempt to improve performance using a newly designed noise-weighting filter based on the properties of the peripheral auditory system, as opposed to those based on the properties of the sound production mechanism.

When dealing with the LPC model, it is also important to address the accuracy of representation for the prediction error signals after inverse filtering. The majority of the proposed coder candidates rely on either the multipulse or stochastic codebooks [7], [8]. Indeed, analysis of experimental data shows that real residual signals exhibit predominantly pulse-like trends in the frequency domain. To better reflect this, we propose to represent the excitation waveform in terms of a multiplicity of amplitude and frequency-modulated sinusoids. The concept of sinusoidal representation has been successfully applied in providing an approximation of speech waveforms [9], [10]. Because of the time-varying nature of the parameters, this straightforward approach leads to parameter discontinuities at the frame boundaries and causes audible artifacts in the steady-to-transient regions. As shortcomings appear, both the frequency-tracking and the parameter-smoothing techniques must be introduced to deal with rapid changes during the transients. As we shall see, the parameter continuity problem is not a serious obstacle provided that sinusoidal analysis is performed on the prediction error signal instead of on the incoming sound, as in [9] and [10].

In this paper, we attempt to capitalize more fully on psychoacoustic knowledge and then develop a new perceptual noise-weighting filter for use in analysis-by-synthesis predictive audio coders. The first part of this paper focuses on analyzing the masking thresholds evoked by incoming sounds. We use an autoregressive (AR) spectral estimator that allows tracking of the masking threshold's evolution by means of

Manuscript received July 18, 1994; revised October 14, 1995. This work was supported by the National Science Council, Taiwan, ROC, under Grant NSC82-0404-E009-175. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. James H. Snyder.

W.-W. Chang is with the Department of Communication Engineering, National Chiao Tung University, Hsinchu, Taiwan, R.O.C.

C.-T. Wang was with the Department of Communication Engineering, National Chiao Tung University, Taiwan. He is now with the Power Research Institute, Taiwan Power Company, Shu-Lin, Taiwan, R.O.C.

Publisher Item Identifier S 1063-6676(96)02450-9.

an all-pole model with finite order. Of more concern is the heavy computational load required for calculating the masking threshold functions. To reduce computational complexity, we trained an artificial neural network to extract the perceptually significant features from audio signals via mapping. The next step of the present investigation was concerned with the accuracy of excitation representation. Toward this end, we explored the benefits of sinusoidal approximation for its use in modeling the pulselike characteristics of residual spectra.

The paper is organized as follows. This section provides an overall view of the investigation. In Section II, we briefly review the basic aspects of auditory perception and then propose a perception-oriented objective measure for quality assessment. Algorithms for AR modeling of the masking threshold are presented in Section III. Comparative performance results for various weighting filter configurations are listed in Section IV in terms of signal-to-noise ratio (SNR), segmental SNR (SNRSEG), generalized Bark spectral distortion, and subjective listening tests. In Section V, we introduce the multisinusoid excitation model and develop an efficient algorithm for extracting the associated modeling parameters. Finally, Section VI presents a short summary and list of conclusions.

II. Auditory Perception

Since psychoacoustic interpretation is central to the design of perceptual coding systems, we first summarize the relevant aspects of auditory perception here; more comprehensive accounts can be found in [11]–[13]. In essence, information received by human ears can be described most conveniently as nonlinear auditory responses to frequency selectivity and perceived loudness. The general properties of frequency selectivity are related to the concept of critical band. Fletcher's band-widening experiment [14] laid the foundation for the critical-band concept by virtue of the assumption that incoming sounds are preprocessed by the peripheral auditory system through a bank of bandpass filters. Each auditory filter behaves like a frequency-weighting function, and corresponds closely to the ear's frequency selectivity across the critical bands. Since the critical bandwidth increases toward higher frequencies, we find the human ear has poorer discrimination in the higher frequency region than in lower ones. The process of auditory filtering involves two steps: critical-band analysis, which accounts for nonlinear perceptual resolution; and critical-band integration, which describes the spread of masking effect across the critical bands. In correspondance with the Hertz-to-Bark transformation [15], critical-band analysis is first carried out to derive the critical-band density $X(b)$ by substituting the frequency variable f in the magnitude spectrum $X(f)$ with the Bark scale b. Next, we perform critical-band integration to determine the excitation pattern $D(b)$ by taking the convolution of the critical-band density $X(b)$ with the basilar-membrane spreading function $B(b)$ [16].

Loudness is another important attribute of auditory perception in terms of which sounds can be ranked on a scale extending from quiet to loud. To measure the loudness quantitatively, two commonly used scaling criteria are the loudness level (in phons) and the subjective loudness scale (in sones)

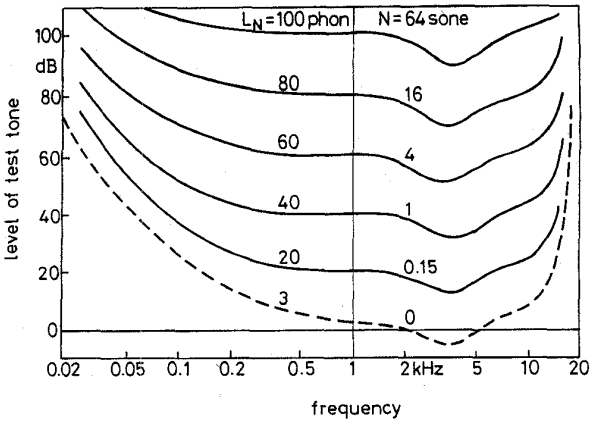

Fig. 1. Equal-loudness contours for pure tones (after [17]).

[11]. The loudness level of a test sound is defined as the intensity level in decibels (dB) of a 1 000 Hz reference tone to which it sounds equally loud. It is well known that the human ear is more sensitive in the frequency range of 1–5 kHz and less sensitive at higher and lower frequencies. The phenomenon that relates audibility to frequency may be demonstrated with the equal-loudness curve [17], which indicates how the intensity level of a tone must be varied with frequency in order to maintain a constant level of loudness. To illustrate this, some typical examples of equal-loudness curves for pure tones are shown in Fig. 1. For purposes of comparison, the absolute threshold (dashed line) is also included to indicate the minimum audible intensity level in quiet surroundings. As seen in the figure, the equal-loudness curves have a shape for low loudness levels that runs almost parallel to the absolute threshold. This is especially true for frequencies above about 200 Hz and, in that frequency range, also holds for higher loudness levels [11].

The shape of the equal-loudness curve provides a useful model for perceptual weighting of spectral energy in the design of objective measures of sound quality. Among the many measures to be considered for sound quality evaluation [18], the most frequently used one is represented in terms of the mean-squared error distortion between original and coded waveforms, as in the SNR. Such objective measures are derived to quantify signal waveform differences, and hence often result in a less perceptually relevant assessment. To compensate for this shortage, Wang *et al.* [19] proposed a Bark spectral distortion (BSD) measure, which does not simply sum up the waveform differences, but rather performs an equal-loudness preemphasis process on the excitation pattern. Consequently, the BSD measure has been shown to correlate more closely with the results of human preference tests than those obtained by other conventional objective measures. As outlined in Fig. 2, several steps are required to compute the BSD measure. They are: a fast Fourier transform (FFT), a critical-band analysis, a critical-band integration, an equal-loudness preemphasis, and a subjective-loudness conversion. For use in the telephone band (300–3 400 Hz, 40–80 dB intensity level), a bilinear preemphasis filter [19] has been proposed to approximate the equal-loudness response. However, the

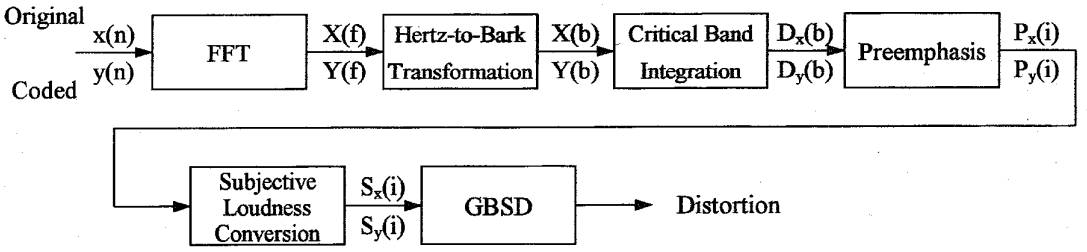

Fig. 2. Bark spectral distortion (BSD) calculation.

equal-loudness contour for wideband audio (20 to 20k Hz) should be emulated more precisely to cover the spectral range of interest. Recognizing this, we propose a generalized BSD (GBSD) measure in which the equal-loudness preemphasis process is carried out by weighting the excitation pattern with the sign-inverted absolute threshold dB values. Accordingly, we calculate the weighted excitation pattern within the ith critical band as follows:

$$P(i) = 10 \log D(i) - LT_q(i) \text{ dB} \quad 1 \le i \le 26 \qquad (1)$$

where $LT_q(i)$ denotes the absolute threshold in dB and $D(i)$ represents the down-sampled version of the excitation pattern $D(b)$ at one-Bark intervals. Note that the total number of critical bands required to cover the entire spectral range is 26. Finally, a subjective-loudness conversion is needed to compensate for the difference between the loudness level and the truly perceived loudness. This difference is due to the perceptual nonlinearity principle, which states that the increase in phons needed to double subjective loudness varies with the loudness level. A 1 000 Hz reference tone of 40 dB intensity level was used to give a perceived loudness of one sone. Since an increase of ten phons tends to double the subjective loudness, we can calculate the Bark spectrum $S(i)$ from the weighted excitation pattern $P(i)$ by the conversion [20] that follows:

$$S(i) = 2^{(P(i)-40)/10}. \qquad (2)$$

The resulting Bark spectrum, which reflects the ear's nonlinear transformation of frequency and loudness, yields a measure in terms of which subjective quality can be evaluated. Let $S_x^k(i)$ and $S_y^k(i)$ denote the Bark spectra of the original signal $x(n)$ and coded signal $y(n)$, respectively, at the ith critical band within the kth frame. The overall distortion—the GBSD—is then calculated by taking the average squared difference over the entire utterance of K frames, as follows:

$$\text{GBSD} = \frac{1}{K} \sum_{k=1}^{K} \left\{ \sum_{i=1}^{26} \left[S_x^k(i) - S_y^k(i) \right]^2 \right\}. \qquad (3)$$

III. AR Modeling of the Masking Threshold

Most perceptual coding systems rely, at least to some extent, on the auditory masking effect to reduce the subjective impairments of reconstruction noise [6–8], [13]. The phenomenon of masking lies in the observation that the ear's perceptual resolution is insufficient to perceive the signal in the presence of another masking signal. In the context of audio coding, the signal to be masked is undesired reconstruction noise, and the masking signal is typically the incoming sound. The audio source can generate a perceptual concealment function below which simultaneously existing artifacts become inaudible. In the present work, we calculate the masking threshold according to the psychoacoustic model of layer I as specified in the ISO/MPEG Audio Standard 11 172-3 [21]. The calculation starts with a precise spectral analysis on 512 windowed audio source samples to generate the magnitude spectrum. The spectral lines are then examined to discriminate between tonal and nontonal maskers by taking the local maximum of audio spectrum as an indicator of tonality. Among all the labeled maskers, only those above the absolute threshold are retained for further calculation. Using rules known from psychoacoustics, the individual masking thresholds for the relevant maskers are then calculated dependent on frequency position, loudness level, and the nature of tonality. Finally, we obtain the global masking threshold from the upper and lower slopes of the individual masking thresholds of tonal and nontonal maskers and from the absolute threshold in quiet. A more detailed description can be found in the informative part of the ISO/MPEG standard [21].

Once the masking threshold has been estimated, we can determine the amount and spectral shape of noise that might be inaudibly inserted into the audio signal. In this respect, the frequency-dependent masking threshold can be regarded as the desired reshaping of the noise spectrum, denoted as $S_n(e^{jw})$. Its highly nonstationary evolution has been investigated by using an autoregressive (AR) spectral estimator with the autocorrelation method [22]. This choice is motivated in part by the success of AR modeling in the discipline of spectral estimation, and partly because accurate estimates of AR parameters can easily be found by solving a set of linear equations. With a pth order AR spectral estimator, the masking threshold is modeled by a linear filter for which the transfer function is all-pole of the form $1/A_m(z)$, where $A_m(z) = 1 - \sum_{i=1}^{p} c_i z^{-i}$. Using such a parametric modeling, the original spectral estimation problem can be formulated as one of the optimal identification of AR parameters $\{c_i,\ 1 \le i \le p\}$. In order to solve for the optimal AR parameters, we first compute the autocorrelation functions $\{r(i), 0 \le i \le p\}$ by taking an inverse Fourier transform of the desired noise spectrum $S_n(e^{jw})$. When a pth order all-pole model is fitted to the noise spectrum, their normalized autocorrelation functions should match exactly for the first $(p+1)$ time lags. Given the autocorrelation functions, we can then determine the optimal values of $\{c_i\}$ by solving the least-squares Yule–Walker

equations [23], as follows:

$$\sum_{k=1}^{p} c_k r(|i-k|) = r(i) \quad 1 \le i \le p. \tag{4}$$

The process above can be considerably simplified if Levinson's recursive algorithm [24] is applied by exploiting the Toeplitz nature of the autocorrelation matrix.

While the masking threshold is conceptually useful in noise spectral shaping, it has some limitations as far as its time-consuming estimation process is concerned. To overcome this problem, we attempt to examine whether the perceptual attributes of the masking threshold can be extracted from typical audio signals via neural network mapping. As illustrated in Fig. 3, a multilayer perceptron consists of feedforward connections with parallel layers of computational elements (neurons). The neurons between successive layers are fully interconnected through the weighting coefficients. The input layer receives data from the audio source, the hidden layer models the ear's physiological mechanism, and the output layer provides relevant aspects of the masking threshold. A series of experiments were performed to optimize the neural network design based on various input-output mapping pairs. It was concluded that optimum mapping is achieved by presenting the input with predictor coefficients $\{a_i\}$ and using AR parameters $\{c_i\}$ of the masking threshold as the desired response at the output layer. To describe the binary activity of neural firing, a neuron's output is generally obtained by applying a sigmoid function to the weighted sum of its inputs [25], [26]. A sigmoid, or S-shaped, function is defined by $f(y) = 1/(1+e^{-y})$, i.e., $0 \le f(y) \le 1$. To understand this, consider the j-th neuron at the hidden layer, whose output is given by

$$O'_j = 1 \Big/ \left(1 + \exp\left(-\sum_l w_{jl}^{(1)} O''_l\right)\right) \tag{5}$$

where $w_{jl}^{(1)}$ denotes the weight connecting the input-layer neuron O''_l to the hidden-layer neuron O'_j. Unfortunately, this nonlinear sigmoid description fails to apply for the activation of neurons at the output layer because the present output-layer neurons are associated with AR parameters of the masking threshold and, hence, may have output values with magnitudes greater than one. Recognizing this, we propose using linear nodes at the output layer of the neural net

$$O_i = \sum_j w_{ij}^{(2)} O'_j \tag{6}$$

where $w_{ij}^{(2)}$ denotes the interconnecting weight between the hidden-layer neuron O'_j and the output-layer neuron O_i.

Before a neural network can perform any specific mapping function, it must be trained by presenting a set of input patterns and adjusting the weights until the desired response occurs within a small error margin. Two updating schemes can be differentiated: one for the linear units at the output layer, the other for the sigmoid units at the hidden layer. We begin by defining an efficient learning procedure that adjusts the

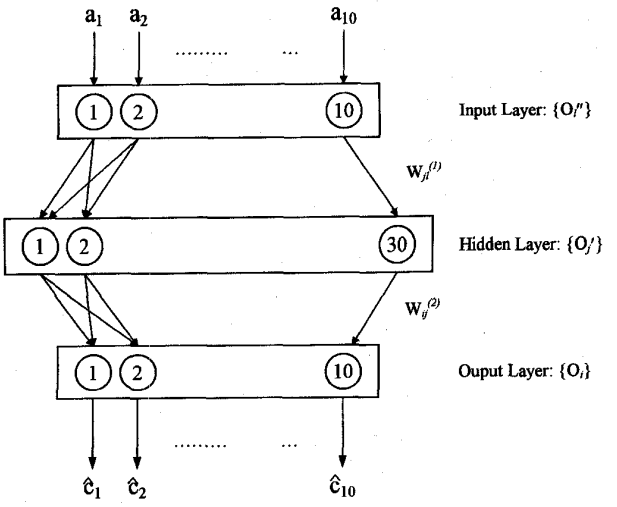

Fig. 3. Multilayer neural network used for extracting AR modeling parameters of masking threshold.

weights $\{w_{ij}^{(2)}\}$ between the hidden layer and the output layer. For the ith output-layer neuron, we calculate the squared-error distortion between the desired response T_i and the actual output O_i, as follows:

$$\delta_{2i}^2 = (T_i - O_i)^2 = \left(T_i - \sum_j w_{ij}^{(2)} O'_j\right)^2. \tag{7}$$

By differentiating the above with respect to the weight, we get the following value for the gradient:

$$\frac{\partial \delta_{2i}^2}{\partial w_{ij}^{(2)}} = \frac{\partial \delta_{2i}^2}{\partial O_i}\frac{\partial O_i}{\partial w_{ij}^{(2)}} = 2(O_i - T_i)O'_j = -2\delta_{2i}O'_j. \tag{8}$$

According to the least-mean-square algorithm [25], successive corrections needed to drive the weight toward the optimum value should be in a direction opposite to its gradient. In view of this, we proceed with updating as follows:

$$w_{ij}^{(2)} = w_{ij}^{(2)} + \eta \delta_{2i} O'_j \tag{9}$$

where the learning step η is chosen empirically to be 0.05 here. The next problem to be addressed is updating the weights $\{w_{jl}^{(1)}\}$ between the input layer and the hidden layer. We stress that these weights must be trained through backpropagation for lack of a desired response as an update reference. Specifically, the output error δ_{2i} is propagated back through the interconnecting weights $w_{ij}^{(2)}$ to the hidden layer. The output error of the jth hidden-layer neuron is given [25] by

$$\delta_{1j} = O'_j(1 - O'_j)\sum_i \delta_{2i} w_{ij}^{(2)} \tag{10}$$

which is used, in turn, to adjust all the weights feeding into the jth hidden-layer neuron

$$w_{jl}^{(1)} = w_{jl}^{(1)} + \eta \delta_{1j} O''_l. \tag{11}$$

A preliminary experiment was conducted to examine whether AR parameters of the masking threshold can be efficiently identified through neural network mapping. Toward this end, we implemented a multilayer perceptron network with 30 neurons in the hidden layer and ten neurons in the input and

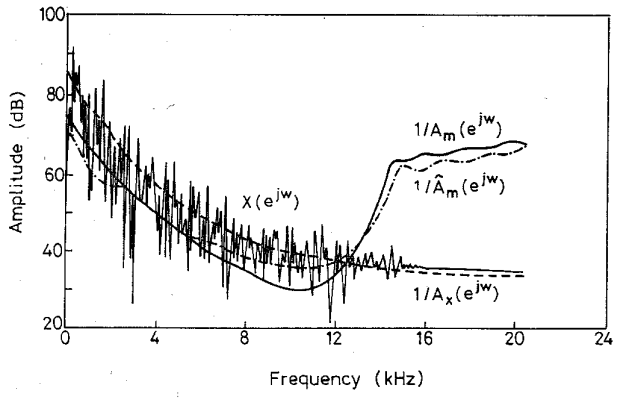

Fig. 4. Fitting of AR models to the audio spectrum and masking threshold.

output layers. Before starting the training process, all of the weights were initialized to small random numbers. The outputs of the well-trained neural network, designated by $\hat{c}_i$, were then used to implement an all-pole filter $1/\hat{A}_m(z)$, where $\hat{A}_m(z) = 1 - \sum_{i=1}^{10} \hat{c}_i z^{-i}$. Although this straightforward approach could lead to the filter's instability, representing the $\hat{c}_i$ in terms of equivalent partial correlation (PARCOR) coefficients $\hat{\pi}_i$ [27] and using them to implement the all-pole filter in a lattice form [28] guarantees stability. Such an alternative provides a convenient test and, if necessary, compensation for stability control in view of the bounded condition , i.e., $-1 < \hat{\pi}_i < 1$.

To test the validity of the neural network, an all-pole fit to the masking threshold obtained by mapping, e.g., $1/\hat{A}_m(z)$, was first compared with that obtained by means of direct estimation, e.g., $1/A_m(z)$. Extensive experiments were conducted using various audio sources with different characteristics. Our general conclusion is that the neural network can learn to provide an approximation of the estimated masking threshold to a reasonable degree of accuracy. This is illustrated, for a typical audio segment, in Fig. 4 where the true spectrum $X(e^{jw})$ and the corresponding AR modeled version $1/A_x(e^{jw})$ are also included for purposes of comparison. A few comments must be made concerning these results. First, the masking threshold tends to become flattened at high frequencies (above 16 kHz). As stated in ISO/MPEG standard 11 172-3 [21], for layer I at the sampling rate of 44.1 kHz, it is generally sufficient to assume a constant value of 68 dB for the absolute threshold above 16 kHz. Second, a comparison between the $1/A_x(e^{jw})$ and $1/A_m(e^{jw})$ indicates that in the former, the spectral envelope decays monotonically, whereas in the latter, the spectral envelope rises abruptly above 12 kHz. The reason for this is that at higher frequencies the sound pressure level of audio input is far below the absolute threshold in quiet. Hence, it appears that absolute threshold dominates the spectral evolution of the masking threshold above 12 kHz. It is thus reasonable to expect that new work based on the masking threshold will yield the noise spectral shaping that differs noticeably from those based on the spectral envelopes of incoming sounds.

IV. A Masking-Threshold Adapted Weighting Filter

Many perceptual coding schemes have been proposed to alleviate the adverse effects of reconstruction noise with varying degrees of success. This is made possible by exploiting the new findings in psychoacoustics, which suggest that better sound reproduction quality can be achieved by perceptual reshaping of noise spectra than by reducing noise power. Hiding noise under the masking threshold is a particularly desirable feature in view of the ear's noise-masking properties [13]. In analysis-by-synthesis predictive coders, this task is generally accomplished by using a perceptual noise-weighting filter for excitation searches [6]–[8]. A weighting filter is considered to be perceptually optimum if the inverse of its magnitude response matches the ear's sensitivity to reconstruction noises in different frequency ranges. In other words, a small value of magnitude response indicates that a high value of reconstruction noise variance is acceptable at that frequency, and vice versa. Depending upon the choice of parameters, a number of different weighting filter configurations can be realized. In order to compare them on the same basis, all the weighting filters considered here are implemented by means of the bandwidth expansion of a 10th th order denominator polynomial using a weighting factor $\gamma = 0.8$.

One approach consists of starting with a conventional LPC-based weighting filter, whereby the noise power is distributed in accordance with the input spectrum $X(e^{jw})$. In this case, linear prediction analysis provides a simplistic approach to implementing the linear predictor $A_x(z)$ as well as the perceptual weighting filter $W_1(z)$. For further discussion, this type of LPC-based weighting filter configuration is refered to as PWF1. Its transfer function is given by

$$W_1(z) = \frac{A_x(z)}{A_x(z/\gamma)} = \frac{1 - \sum_{i=1}^{10} a_i z^{-i}}{1 - \sum_{i=1}^{10} a_i \gamma^i z^{-i}} \tag{12}$$

where $\{a_i\}$ denotes the linear predictor coefficients and the weighting factor γ controls the energy of the error embedded in the formant regions. There has been considerable experience with such LPC-based weighting filters; they are currently the basis for many practical coders. The basic problem with this approach is that the predictor coefficients were originally derived to provide an all-pole fit to the audio spectrum. We thus attempt to improve on this approach by redistributing the noise power in relation to the masking threshold produced by the audio signal, instead of the audio spectrum itself. In the proposed system, we implement the weighting filter using masking threshold AR parameters, which may be obtained either through direct processing or through neural network mapping. The corresponding transfer functions, designated by PWF2 and PWF3, are given, respectively, by

$$W_2(z) = \frac{A_m(z)}{A_m(z/\gamma)} = \frac{1 - \sum_{i=1}^{10} c_i z^{-i}}{1 - \sum_{i=1}^{10} c_i \gamma^i z^{-i}} \tag{13}$$

$$W_3(z) = \frac{\hat{A}_m(z)}{\hat{A}_m(z/\gamma)} = \frac{1 - \sum_{i=1}^{10} \hat{c}_i z^{-i}}{1 - \sum_{i=1}^{10} \hat{c}_i \gamma^i z^{-i}}. \tag{14}$$

The suitability of each of the noise-weighting filters introduced above has been evaluated for use in noise spectral shaping. The experimental arrangement of a code-excited LPC (CELP) encoder (after [29]) is shown in Fig. 5, and we denote CELP systems with weighting filters PWF1, PWF2, and PWF3 as CELP1, CELP2, and CELP3, respectively.

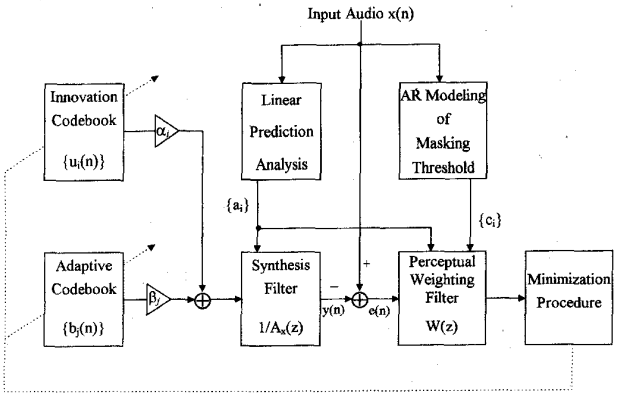

Fig. 5. Block diagram for the CELP encoder. The perceptual weighting filter can be characterized using either predictor coefficients $\{a_i\}$ or masking threshold AR parameters $\{c_i\}$.

A CELP system produces its excitation by summing the gain-scaled codevectors from an innovation codebook and an adaptive codebook. The adaptive codebook, which accounts for long-term pitch periodicity, is updated using the past history of excitation sequences. The innovation codebook, on the other hand, is usually populated by Gaussian random samples whose statistics resemble noiselike residual signals. In this experiment, audio source samples were segmented into frames of 160 samples long. Each frame was further divided into eight subframes. The filter's coefficients were updated once per frame. The excitation parameters were transmitted once per subframe. The optimal values of system parameters were determined in two steps. First, the predictor coefficients $\{a_i,\ 1 \leq i \leq 10\}$ were calculated using linear predictive analysis with the autocorrelation method. Prior to transmission, these predictor coefficients were transformed into line spectrum pair (LSP) frequencies and linearly quantized. As mentioned in [30], LSP representation has the advantages of allowing more efficient quantization and conserving the synthesis filter's stability after quantization provided that its natural ordering relationship is preserved. Next, the system finds the excitation parameters with the least squared-error distortion by sequentially feeding all possible codewords to the synthesis filter and using a weighted distortion measure to evaluate the reconstructed signal until the best fit is found.

Table I presents the comparative performance results for CELP coding of audio in conjuction with different noise-weighting filter structures. The transmission rate is 92.61 kb/s with bits allocated to parameters as listed in Table II. The monophonic audio database for these studies consisted of piano, drum, and horn signals of four seconds duration and sampled at a rate of 44.1 kHz. SNR, SNRSEG, and GBSD were measured on the reconstructed signals. We also conducted informal listening tests. GBSD results were consistent with the results of the listening tests. Informal listening tests indicated that the CELP1 output is subjectively inferior to the CELP2 and CELP3 in spite of its higher values of SNR and SNRSEG. In other words, the nominal advantage in waveform difference measures should not be interpreted as an indication of subjective preference. Since GBSD is more indicative of perceptual cues than SNR and SNRSEG, the GBSD comparison performed on three noise-weighting filter configurations should be instructive. As the table shows, CELP2 and CELP3 yielded substantial improvement over CELP1 for all test samples. The results also indicated that CELP2 and CELP3 produced comparable performance, with perhaps a slight advantage going to CELP2. Among the reasons for success, we find that an all-pole system with finite order is sufficient to model the masking threshold because its spectral evolution has few deep valleys that require a large number of poles for adequate spectral approximation. It is also important to note that audibility of noise depends heavily on the relative distribution between noise power and frequency response of the noise-weighting filter. In CELP2 and CELP3, the noise spectrum much more closely duplicates the masking threshold, and thus, even though the noise power is greater than that of CELP1, it is better masked and so less audible.

TABLE I
SNR/SNRSEG/GBSD PERFORMANCE OF CELP CODER WITH VARIOUS WEIGHTING-FILTER CONFIGURATIONS

Coder \ Music	Piano	Horn	Drum
CELP1	27.4/30.3/219.8	31.8/32.3/149.5	32.2/32.2/149.2
CELP2	26.5/29.1/144.4	26.8/28.5/83.1	27.6/28.1/74.6
CELP3	25.3/28.1/131.0	25.7/27.8/88.3	27.9/28.8/78.5

TABLE II
BIT ALLOCATION FOR CELP CODERS AT 92.61 KB/S

Update Rate	Item	Bits
Subframe	Adaptive Codebook Gain	9
	Adaptive Codebook Index	9
	Innovation Codebook Gain	9
	Innovation Codebook Index	10
Frame	LPC Coefficients	40
Total Bits Per Frame (Frame Length = 160 Samples)		336

We next compare the proposed system with the well-established ISO/MPEG 96 kb/s audio-coding system [21]. In layer I, the MPEG system employs a filterbank to create 32 critically sampled representations of the input signal, which are then quantized using adaptive block companding under the control of the estimated masking threshold. As listed in Table III, none of the CELP-based audio coders can outperform the audio coder in conforming to the ISO/MPEG standard. The reason for this seems to be that real residual samples are highly correlated and hence fail to meet the white noise assumption used in populating the stochastic innovation codebook. Therein lies the motivation for further work on the incorporation of sinusoidal excitation representation, which is detailed in the following section.

V. THE MULTISINUSOID EXCITATION MODEL

A code-excited LPC coder decomposes the signal into the product of excitation and system spectra, and then represents

TABLE III
SNR/SNRSEG/GBSD PERFORMANCE OF VARIOUS AUDIO CODERS

Coder \ Music	Piano	Horn	Drum
CELP2	26.5/29.1/144.4	26.8/28.5/83.1	27.6/28.1/74.6
MPEG	27.8/32.3/61.6	28.7/34.7/39.0	26.6/37.0/54.4
MSLPC	28.9/32.3/67.1	35.5/39.8/30.8	27.5/42.8/24.2

the excitation by using a stochastic codebook. When dealing with periodic sounds, its near-white spectrum fails to provide a good approximation of the pulselike envelope of a real residual spectrum. As an alternative to the above models, we propose to represent the excitation waveform by a sum of sine waves with arbitrary amplitudes, frequencies, and phases [9]. Fig. 6 illustrates the functional block diagram of a MultiSinusoid LPC (MSLPC) encoder. The general form of a multisinusoid excitation model is given by

$$e(n) = \sum_{i=1}^{M} r_i \cos(w_i nT + \phi_i), \qquad 1 \le n \le N \tag{15}$$

where N is the subframe length, M is the number of sinusoids, and the r_i, w_i, and ϕ_i represent the associated amplitude, frequency, and phase, respectively, of the ith sinusoid. Letting $h(n)$ denote the impulse response of the weighted-synthesis filter, we produce the output signal $y(n)$ by taking the convolutional sum

$$y(n) = e(n) * h(n) \tag{16}$$

$$= \sum_{i=1}^{M} [\alpha_i h_{ci}(n) + \beta_i h_{si}(n)], \quad 1 \le n \le N \tag{17}$$

where $\alpha_i = r_i \cos\phi_i, \beta_i = -r_i \sin\phi_i, h_{ci}(n) = \cos(w_i nT) * h(n)$, and $h_{si}(n) = \sin(w_i nT) * h(n)$. It is more convenient to rewrite the above set of linear equations in vector form as follows:

$$\vec{y} = \sum_{i=1}^{M} [\alpha_i \vec{h}_{ci} + \beta_i \vec{h}_{si}]. \tag{18}$$

An accurate identification of excitation parameters should be the basis for the success of MSLPC. Assuming that the frequencies $\{w_j,\ 1 \le j \le M\}$ have been determined, the optimal values of α_j and β_j can then be found by minimizing the mean-squared error E defined as

$$E = \sum_{n=1}^{N} (x(n) - y(n))^2 = \vec{x} \cdot \vec{x}^t - 2\vec{x} \cdot \vec{y}^t + \vec{y} \cdot \vec{y}^t. \tag{19}$$

This minimization results in the set of linear equations for $j = 1, 2, \ldots, M$ as follows:

$$\frac{\partial E}{\partial \alpha_j} = 2\left(\vec{x} - \sum_{i=1}^{M} \alpha_i \vec{h}_{ci} - \sum_{i=1}^{M} \beta_i \vec{h}_{si} \right) \cdot \vec{h}_{cj}^{\,t} = 0 \tag{20}$$

$$\frac{\partial E}{\partial \beta_j} = 2\left(\vec{x} - \sum_{i=1}^{M} \alpha_i \vec{h}_{ci} - \sum_{i=1}^{M} \beta_i \vec{h}_{si} \right) \cdot \vec{h}_{sj}^{\,t} = 0 \tag{21}$$

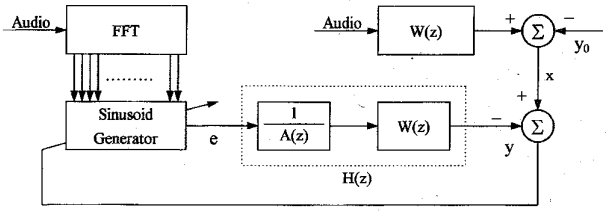

Fig. 6. Block diagram for the multisinusoid LPC (MSLPC) encoder.

or, rewritten in matrix form as

$$\vec{S} \cdot \vec{g} = \vec{c} \tag{22}$$

where the entries in $\vec{g}, \vec{c}$ and $\vec{S}$ are given, respectively, for $1 \le j \le 2M$ and $1 \le k \le 2M$, as follows:

$$g_j = \begin{cases} \alpha_{(j+1)/2}, & j\text{: odd} \\ \beta_{j/2}, & j\text{: even} \end{cases} \tag{23}$$

$$c_j = \begin{cases} \vec{x} \cdot \vec{h}_{c(j+1)/2}^{\,t}, & j\text{: odd} \\ \vec{x} \cdot \vec{h}_{s(j/2)}^{\,t}, & j\text{: even} \end{cases} \tag{24}$$

$$S_{jk} = \begin{cases} \vec{h}_{c(j+1)/2} \cdot \vec{h}_{c(k+1)/2}^{\,t}, & j\text{: odd}, k\text{: odd} \\ \vec{h}_{s(j/2)} \cdot \vec{h}_{s(k/2)}^{\,t}, & j\text{: even}, k\text{: even} \\ \vec{h}_{c(j+1)/2} \cdot \vec{h}_{s(k/2)}^{\,t}, & j\text{: odd}, k\text{: even} \\ \vec{h}_{s(j/2)} \cdot \vec{h}_{c(k+1)/2}^{\,t}, & j\text{: even}, k\text{: odd}. \end{cases} \tag{25}$$

Indeed, the above equation can be solved more efficiently by taking advantage of the symmetric nature of the matrix $\vec{S}$. According to the Cholesky factorization theorem [31], a symmetric matrix $\vec{S}$ can be decomposed into the form of $\vec{G}\vec{G}^t$ where $\vec{G}$ is a lower triangular matrix whose nonzero entries are given by the following expressions:

$$G_{jj} = \sqrt{S_{jj} - \sum_{k=1}^{j-1} G_{jk}^2}, \quad 1 \le j \le 2M \tag{26}$$

$$G_{jk} = \left(S_{jk} - \sum_{l=1}^{k-1} G_{jl} G_{kl} \right) \Big/ G_{kk}, \quad 1 \le k \le j-1. \tag{27}$$

In correspondence with this factorization, we can rewrite (22) in terms of $\vec{G}$ as follows:

$$\vec{G}\vec{q} = \vec{c} \tag{28}$$

$$\vec{G}^t \vec{g} = \vec{q} \tag{29}$$

where the entries in $\vec{q}$ are given by

$$q_j = \left(c_j - \sum_{k=1}^{j-1} G_{jk} q_k \right) \Big/ G_{jj}, \quad 1 \le j \le 2M. \tag{30}$$

Substituting these optimum parameters into (19) leads to the least-squared error expression

$$E_{\min}^{(M)} = \vec{x} \cdot \vec{x}^t - \vec{q} \cdot \vec{q}^t = E_{\min}^{(M-1)} - \left(q_{2M-1}^2 + q_{2M}^2 \right) \tag{31}$$

where the superscript "M" denotes the number of sinusoids used in approximating the excitation waveform. From inspection of (31), it follows that the squared error distortion is guaranteed to converge by increasing the number of sinusoids.

From the perspective of computational efficiency, the Cholesky factorization technique also provides an ideal

TABLE IV
BIT ALLOCATION FOR THE MSLPC CODER AT 96 KB/S

Update Rate	Item	Bits
Subframe	Frequency Patterns	12
	Amplitudes	7x8
	Phases	6x8
Frame	Frequency Candidates	77
	LPC Coefficients	40
Total Bits Per Frame (Frame Length = 480 Samples)		1045

framework for independently estimating the parameters $\{w_i\}$ and $\{r_i, \phi_i\}$ in a two-step procedure. Consider the frequencies$\{w_i,\ 1 \leq i \leq M\}$, which are exclusively embedded in the entries of $\vec{q}$. Under such conditions, the frequency of the ith component sine wave can be determined as the location of the spectral peak, which maximizes the term $(q_{2i-1}^2 + q_{2i}^2)$. Though the sine wave frequencies could be tracked using the procedure above, the necessity of an exhaustive search makes algorithm implementation impossible. Fortunately, computational complexity can be reduced because sinusoidal components should correspond to the occurence of spectral peaks. It is therefore sufficient to search only the spectral region around the spectral peaks to find the best fit for the underlying sine waves.

In this experiment, a set of L frequencies were chosen once per frame by locating the predominant peaks of the associated audio spectrum. Only these L frequency candidates were examined to find the M best frequencies needed within each constituent subframe. We empirically chose $L = 12$ and $M = 8$ as the best compromise between coding gain and implementational complexity. Together with the weighting filter PWF2, the proposed multisinusoid excitation model was evaluated to assess its suitability for use in developing an LPC-based audio coder. The performance of MSLPC at a transmission rate of 96 kb/s is listed in Table III. With an analysis frame length of 480 samples, the total number of bits allocated per frame is 1 045, with the breakdown according to parameters as shown in Table IV. For transmission to receiver, the phases and amplitudes of the sinusoidal components were linearly quantized. In addition, the frequencies were quantized in a two-step procedure. First, we employed a differential coding strategy to quantize twelve frequency candidates once per frame. Second, each constituent subframe was associated with a 12-b pattern in which the absence or presence of a sinusoid is indicated by a "0" or a "1." The results indicate that the sinusoidal excitation model is preferred to the codebook-excited model for use in audio representation, because the former better fits the pulselike natures of residual spectra. We also stress that it is the combined use of a multisinusoid excitation model and a masking-threshold adapted weighting filter that allows the implementation of an LPC-based coder to outperform its ISO/MPEG counterpart.

VI. CONCLUSION

This paper presents and discusses technical options that allow an LPC-based audio coder to deliver near- transparent quality at 96 kb/s. We first emphasized the importance of matching the weighting filter's response to the noise-masking threshold. This task was done by using AR parameters of the masking threshold to implement the noise-weighting filter rather than using linear predictor coefficients, as do conventional LPC-based coders. Furthermore, an efficient neural network was trained to extract most of the perceptual information concerning the masking threshold from the audio sources via mapping. One enhancement that further increases performance is the use of a sinusoidal excitation representation that more closely matches the intrinsic natures of residual spectra.

ACKNOWLEDGMENT

The authors are very grateful to the unknown reviewers and the associate editor, James H. Snyder, for their careful readings of this paper and their constructive suggestions. They also acknowledge Li-Wei Wang for carrying out the multisinusoid excitation model experiments.

REFERENCES

[1] N. S. Jayant, "Signal compression: Technology targets and research directions," *IEEE J. Select. Areas Commun.*, pp. 796–818, June 1992.
[2] J. D. Johnson, "Transform coding of audio signals using perceptual noise criteria," *IEEE J. Select. Areas Commun.*, pp. 314–323, Feb. 1988.
[3] G. Theile, G. Stoll, and M. Link, "Low bit-rate coding of high-quality audio signals," *EBU Tech. Rev.*, no. 230, pp. 71–94, Aug. 1988.
[4] J. H. Chen and A. Gersho, "Real-time vector APC speech coding at 4 800 b/s with adaptive postfiltering," in *Proc. ICASSP*, Apr. 1987, pp. 2185–2188.
[5] B. S. Atal and M. R. Schroeder, "Predictive coding of speech signals and subjective error criteria," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-27, pp. 247–254, June 1979.
[6] P. Kroon and E. F. Deprettere, "A class of analysis-by-synthesis predictive coders for high quality speech coding at rates between 4.8 and 16 kb/s," *IEEE J. Select. Areas Commun.*, vol. 6, no. 2, pp. 353–363, Feb. 1988.
[7] S. Singhal, "High quality audio coding using multipulse LPC," in *Proc. ICASSP*, 1990, pp. 1101–1104.
[8] X. Lin, R. A. Salami, and R. Steele, "High quality audio coding using analysis-by-synthesis technique," in *Proc. ICASSP*, 1991, pp. 3617–3620.
[9] R. J. McAulay and T. F. Quatieri, "Speech analysis and synthesis based on a sinusoidal model," *IEEE Trans. Signal Processing*, vol. SP-34, pp. 744–754, Aug. 1986.
[10] E. B. George and M. J. T. Smith, "A new speech coding model based on a least-squares sinusoidal representation," in *Proc. ICASSP*, 1987, pp. 1641–1644.
[11] E. Zwicker and H. Fastl, *Psychoacoustics.* Berlin: Springer-Verlag, 1990.
[12] B. C. J. Moore, *An Introduction to the Psychology of Hearing.* London, U.K: Academic, 1989.
[13] M. R. Schroeder, B. S. Atal, and J. L. Hall, "Optimizing digital speech coders by exploiting masking properties of the human ear," *J. Acoust. Soc. Amer.*, vol. 66, pp. 1647–1651, Dec. 1979.
[14] H. Fletcher, "Auditory patterns," *Rev. Mod. Phys.*, pp. 47–65, 1940.
[15] A. Fourcin, "Speech processing by man and machine—Group report," in *Recognition of Complex Acoustic Signals*, T. Bullock, Ed., Life Sciences Res. rep., Dahlem Workshops, Berlin, Germany, 1977.
[16] A. Sekey and B. Hanson, "Improved one-Bark bandwidth auditory filter," *J. Acoust. Soc. Amer.*, vol. 75, pp. 1902–1904, June 1984.
[17] D. Robinson and R. Dadson, "A redetermination of the equal-loudness relations for pure tones," *Brit. J. Appl. Phys.*, pp. 166–181, 1956.
[18] S. Quackenbush, T. Barnwell, and M. Clements, *Objective Measures of Speech Quality.* Englewood Cliffs, NJ: Prentice-Hall, 1988.
[19] S. Wang, A. Sekey, and A. Gersho, "An objective measure for predicting subjective quality of speech coders," *IEEE J. Select. Areas Commun.*, vol. 10, no. 5, pp. 819–829, June 1992.

[20] R. Bladon, "Modeling the judgement of vowel quality differences," *J. Acoust. Soc. Amer.*, vol. 69, pp. 1414–1422, May 1981.
[21] ISO/IEC Int. Std. IS 11172-3, "Information technology-coding of moving pictures and associated audio for digital storage media up to about 1.5 mbits/s," Part 3: Audio.
[22] S. M. Kay, *Modern Spectral Estimation.* Englewood Cliffs, NJ: Prentice-Hall, 1988.
[23] N. S. Jayant and P. Noll, *Digital Coding of Waveforms.* Englewood Cliffs, NJ: Prentice-Hall, 1984.
[24] N. Levinson, "The Wiener RMS (root mean square) error criterion in filter design and prediction," *J. Math. Phys.*, vol. 25, pp. 261–278, 1947.
[25] P. D. Wasserman, *Neural Computing: Theory and Practice.* New York: Van Nostrand Reinhold, 1989.
[26] R. P. Lippmann, "An introduction to computing with neural nets," *Acoust., Speech, Signal Processing*, pp. 4–22, Apr. 1987.
[27] F. Itakura and S. Saito, "Speech analysis-synthesis system based on the partial autocorrelation coefficient," *Acoust. Soc. Japan*, 1969.
[28] J. D. Markel and A. H. Gray Jr., *Linear Prediction of Speech.* New York: Springer-Verlag, 1976.
[29] M. R. Schroeder and B. S. Atal, "Code-excited linear prediction (CELP): High quality speech at very low bit rates," in *Proc. ICASSP*, 1985, pp. 937–940.
[30] F. Itakura, "Line spectrum representation of linear predictor coefficients of speech signal," *Acoust. Soc. Japan*, 1979.
[31] G. H. Golub and C. F. Van Loan, *Matrix Computations.* Baltimore, MD: Johns Hopkins Univ. Press, 1989.

Wen-Whei Chang (S'86–M'89) received the B.S. degree in communication engineering from National Chiao Tung University, Hsinchu, Taiwan, R.O.C, in 1980 and the M.Eng. and Ph.D. degrees in electrical engineering from Texas A&M University, College Station, TX, in 1985 and 1989, respectively.

Since August 1989, he has been an associate professor with the Department of Communication Engineering, National Chiao Tung University, Hsinchu, Taiwan, R.O.C. His current research interests include speech processing, language identification, and secure communication.

Chin-Tun Wang received the B.S. degree in electrical engineering from National Taiwan Institute of Technology, Taiwan, R.O.C, in 1989, and the M.S. degree in communication engineering from National Chiao Tung University, Hsinchu, Taiwan, ROC, in 1993.

Since August 1993, he has been an electrical engineer at the Power Research Institute, Taiwan Power Company, R.O.C, where his work has focused on power-frequency electromagnetic fields and power electronic applications.