Perceptual Coding of Narrow-Band Audio Signals at Low Rates

Hossein Najaf-Zadeh and Peter Kabal, *Member, IEEE*

***Abstract*—This paper describes a coding paradigm using coding tools based on the characteristics of the human hearing system so as to accommodate a wide range of narrow-band audio inputs without annoying artifacts at low rates (down to 8 kb/s). The narrow-band perceptual audio coder (NPAC) employs a variety of algorithms to account for the perceptually irrelevant parts of the input signal in addition to statistical redundancies. The new algorithms used in the NPAC coder include a perceptual error measure in training the codebooks and selecting the best codewords which takes into account the audible parts of the quantization noise, a perception-based bit-allocation algorithm and a new predictive scheme to vector quantize the scale factors. The NPAC coder delivers acceptable quality without annoying artifacts for most narrow-band audio signals at around 1 bit/sample. Informal subjective tests have shown that the NPAC coder outperforms a commercial low-rate music coder operating at 8 kb/s.**

***Index Terms*—Adaptive bit allocation, masking model, modified discrete cosine transform filter bank, narrow-band audio coding, perceptual audio coding, perceptual distortion, perceptual vector quantization, predictive vector quantization.**

I. Introduction

AUDIO compression is concerned with the efficient transmission or storage of audio data with good perceptual quality. Audio files require a lot of bandwidth (or memory) for transmission (or storage). For instance, an audio signal sampled at 8 kHz and using 16 bits for each sample gives a data rate of 128 kb/s. In this work we show that for *narrow-band* audio signals, it is possible to reduce the data rate to less than 10 kb/s while maintaining acceptable quality (i.e., without annoying artifacts).

The increasing traffic in wireline (e.g., telephony or Internet) and wireless (e.g., cell phones) networks calls for high compression efficiency to better utilize the capacity of existing resources. As such, there is a need for bandwidth-efficient coding of a variety of sounds including speech, music and multiple simultaneous speakers.

Traditional speech coders designed specifically for speech signals achieve compression by utilizing models of speech production based on the human vocal tract. However, these speech coders are not effective when the signal to be coded is not human speech but some other signal such as music. These other signals do not have the same characteristics as human speech and are not well modeled with a voiced/unvoiced signal exciting a vocal tract filter. As a result, traditional speech coders often have uneven results for nonspeech signals. In contrast, perception-based coders can accommodate diverse signals by using human auditory masking phenomena [1] to identify the inaudible parts of audio signals. When the perceptually irrelevant information is not coded, the audio coder can operate at much lower bit rates and still provide good sound quality.

A. Motivation for Low Rate Coding of Narrow-Band Audio Signals

Although a lot of research has been done on high-quality coding of wide-band audio signals over the past decade [2]–[6], new applications such as Internet broadcasting, consumer multimedia products, narrow-band digital AM broadcasting[1] and satellite networks are emerging. For those applications moderate audio quality without annoying artifacts at low bit rates below 16 kb/s is adequate [7]–[9]. In some applications either the number of users is huge (e.g., Internet) or the available bandwidth is limited (e.g., satellite and radio communications) necessitating extreme bandwidth efficiency.

Coders suitable for narrow-band audio generally operate at bit rates above 16 kb/s (e.g., ITU G.726 ADPCM standard) to deliver moderate audio quality. On the other hand speech coders operating at bit rates lower than 16 kb/s are not suitable for encoding audio signals. There is a gap between the operating bit rates of state-of-the-art narrow-band speech coders (8 kb/s and below) and low bit rate audio coders operating at around 16 kb/s.

In this work we take on the challenge of designing a coding structure to accommodate narrow-band audio inputs (band-limited from 50 Hz to 3.6 kHz, sampled at 8 kHz, and represented with 16 bits per sample) at bit rates comparable to existing narrow-band speech coders. To accomplish this goal, we have developed an audio coding structure based on the characteristics of the human hearing system. The proposed coder, which is referred to as the *narrow-band perceptual audio coder (NPAC)*, provides moderate quality (i.e., without annoying artifacts) for narrow-band audio inputs at bit rates down to 8 kb/s [10]–[12]. Although in recent years the TwinVQ coder operating at 6–8 kb/s has been adopted as part of MPEG-4 Audio [13], the NPAC coder employs a variety of novel perception-based algorithms to take into account the perceptually irrelevant parts of the input signal in addition to statistical redundancies.

The NPAC coder has moderate complexity and a software implementation of the coder written in the C language runs in

Manuscript received February 28, 2001; revised December 17, 2004. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Walter Kellermann.

H. Najaf-Zadeh is with the Advanced Audio Systems, Communications Research Centre, Ottawa, ON K2H 8S2, Canada (e-mail: hossein.najafzadeh@crc.ca).

P. Kabal is with the Electrical and Computer Engineering, McGill University, Montreal, QC H3A 2A7, Canada (e-mail: kabal@ece.mcgill.ca).

Digital Object Identifier 10.1109/TSA.2005.855827

[1]Digital Radio Mondiale Consortium. See http://www.drm.org.

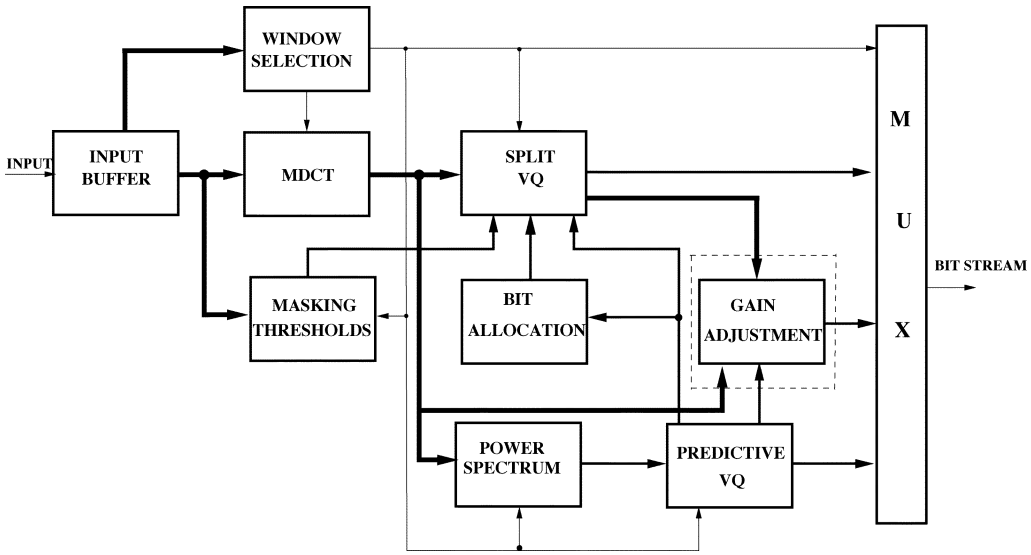


Fig. 1. Block diagram of the NPAC coder. The lines (in the order of their thickness) indicate the processed input data, the intermediate data and the window switching flag. The gain adjustment module can be added to the NPAC coder at the cost of around 2 kb/s.

real time on a computer using a 450 MHz Pentium processor. The algorithmic delay of this coder is 30 ms, which is reasonable for most applications.

Although the NPAC coder belongs to the family of perceptual audio coders, it is significantly different in concept and implementation from high-rate wide-band audio coders. While the goal of high-rate audio coding is to achieve transparent or near transparent quality of wide-band audio with a 7–20 kHz bandwidth [2]–[4], [14], our goal is to achieve moderate audio quality. State-of-the-art high-rate audio coders spend around 1.5 bits per sample (in the order of 64 kb/s) to reproduce high quality audio. Since important spectral features of natural audio signals are located between 300–5000 Hz [1], high-rate audio coders spend considerably more than 1.5 bits per sample on the low frequency spectral components. In the NPAC coder, we spend only 1 bit per sample for the frequency band 50–3600 Hz. In high-quality perceptual audio coders, the goal is to have the coding distortion totally masked by the input signal. However, at low bit rates the distortion may not be completely masked and the emphasis shifts to minimizing the audible artifacts.

II. Overview of the NPAC Coder

A block diagram of the NPAC coder is shown in Fig. 1. The blocks are described in this and the following sections. In this paper we consider monaural audio signals bandpass filtered to limit the spectrum between 50 Hz and 3.6 kHz and sampled at 8 kHz. A filterbank is used to decompose the input signal into spectral components. The masking threshold is estimated and used for both adaptive bit allocation and quantization of the transform coefficients.

III. Time to Frequency Mapping

A modified discrete cosine transform (MDCT) [15] is used to transform the audio data. The MDCT provides critical sampling, perfect reconstruction and due to the overlapping blocks, reduced block edge effects. There is a direct relationship between the MDCT and DFT [11] which implies that the MDCT coefficients represent the frequency content of the input signal. Moreover, FFT-like algorithms can be used to compute the MDCT.

A. Choice of MDCT Window

For the MDCT, windowing is used to select a portion of the input signal for analysis. The length of the window is a compromise between long windows (high coding gain[2]) and short windows (better transient coding by keeping coding noise localized in time). Since the characteristics of audio signals vary with time, and since the NPAC coder is also intended for speech use, we choose a compromise analysis frame length of 30 ms, a period over which speech signals can be considered to be pseudo-stationary. The coder takes in 240 samples (with 50% overlap) and uses an MDCT to decompose the block of data. However, sharp transient sounds require a higher temporal resolution (shorter time window).

The shape of the time window used for the MDCT affects the frequency selectivity of the equivalent filterbank. We need to trade off resolution in the main lobe versus high attenuation of the side lobes. A narrow main lobe keeps energy local to the MDCT coefficients and prevents loss of coding gain. The main lobe width should be less than the width of the narrowest critical band (100 Hz). This choice makes it easier to control the perception of the quantization noise and to compute the simultaneous masking thresholds more accurately. On the other hand, the stopband attenuation should be high to reduce spectral leakage.

In [6] a Kaiser-Bessel-Derived (KBD) window with high stop band attenuation is used. Although this window performs well for many audio signals, it has a poor frequency selectivity that makes it unsuitable for low-pitch harmonic signals. In [4], in order to accommodate a wider range of audio signals, the coder allows for switching between a KBD window and a sine window.

[2]Coding gain measures the ability of a transform to compact the energy into a few coefficients [16].

In the NPAC coder, we have designed the time (lowpass prototype) window with a 50-Hz bandwidth. The modulated response has a bandwidth of 100 Hz. The prototype window is designed by the following optimization procedure [12]

$$H_d(\cdot) = \arg\min_{H} \sum_{k=0}^{\frac{N_F}{2}+1} W(k)\left(H_{\text{ideal}}(k) - H(k)\right)^2$$

subject to

$$h(2M-1-n) = h(n), \quad \text{(symmetry)}$$
$$h^2(n) + h^2(n+M) = 1, \quad \text{(perfect reconstruction)} \quad (1)$$

where N_F is the Fourier transform length, H_d is the normalized DFT of the window $h(n)$, and M is the number of transform coefficients of the MDCT (half the length of the window). H_{ideal} is the DFT of the ideal lowpass filter defined as follows:

$$H_{\text{ideal}}(k) = \begin{cases} 1, & 0 \le k < k_p \\ 0, & k_p \le k \end{cases} \quad (2)$$

where k_p is the edge of the transition band. For an N_F-point DFT and M MDCT coefficients, each ideal bandpass filter is represented by $N_F/2M$ points of the DFT. Since the prototype lowpass filter generates the filterbank, its bandwidth is half the bandwidth of each bandpass filter. Therefore, the passband of the ideal lowpass filter is represented approximately by $(N_F/4M)+1$ points, meaning that $k_p \approx (N_F/4M)+1$.

The weighting function W gives different weights to the passband, transition band and the stop band. Note that we give more weight to the stopband to reduce the leakage between bands. W is defined by

$$W(k) = \begin{cases} 1, & 0 \le k < k_p \\ 0, & k_p \le k < k_s \\ 100, & k_s \le k \end{cases} \quad (3)$$

where k_s is the edge of the stop band. We assume that the width of the transition band can be larger than a critical band. Since the critical bandwidths vary with frequency, we take a value of 200 Hz for the transition band. For a sampling rate of 8000 Hz and an N_F-point DFT, the transition width becomes $N_F/40$ and therefore in (3), k_s is set to $k_p + (N_F/40)$. Although, for a window of 240 samples, the MDCT coefficients represent steps of 33.3 Hz, the choice of 100 Hz allows us to enhance the stopband rejection of the window response.

B. Handling Transients

For high-energy transient parts of the input signal, it is desired to localize short burst of quantization noise to prevent it spreading over a long period of time. We handle this problem by switching to a shorter window when a strong jump in energy is encountered [17]. As an alternative to switching to short windows at onsets, Herre and Johnston [18] use temporal noise shaping (TNS) to continuously adapt the temporal and frequency resolution of the filterbank.

Short windows reduce the coding gain and should be avoided when they do not improve the coded signal quality. Since backward temporal masking lasts for about 5 ms (for narrow-band stimuli) while forward temporal masking lasts for about 200 ms

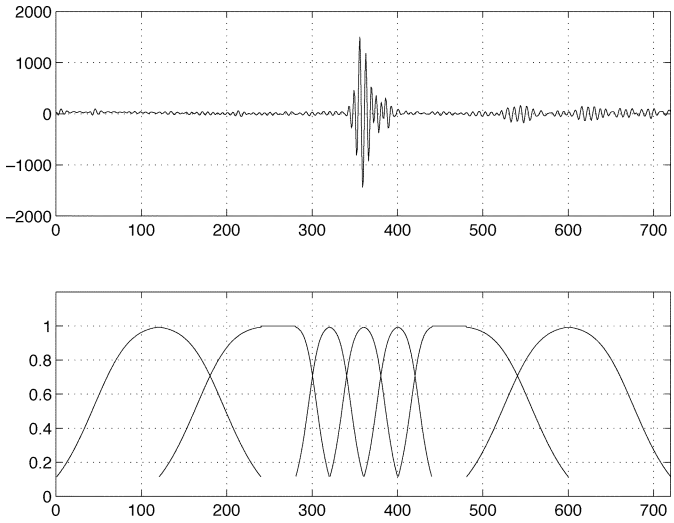


Fig. 2. Window switching for a piece of music containing a transient sound.

[19], a distinction should be made between rises and falls in the energy of the signal. A simple criterion based on the relative positive change in the energy of the input signal is used. In the time domain, a local estimate is made of the change in signal energy. This is done by splitting the input frame into intervals of three time samples and calculating the energy of the samples in each interval. The short integration interval allows the algorithm to respond to rapid fluctuations of the energy contour while being relatively unaffected by high-frequency noise. The maximum positive change will be found as follows:

$$r = \max\left(\frac{e_{j+1} - e_j}{e_j}\right) \quad (4)$$

where e_j is the energy of interval j. If r exceeds 10, we switch to a shorter window.

In order to maintain perfect reconstruction of the combined analysis and synthesis stages, a start window is used to switch from long to short windows, and a stop window switches back [14]. For the short windows we use a frame length of 10 ms (80 samples). Fig. 2 shows the switching of the longer window to a series of shorter windows for a piece of music containing a transient sound.

IV. Masking

Masking is a property of the hearing system by which a weaker audio signal becomes inaudible in the presence of a louder signal [1]. The masking depends both on the spectral composition of the masker and the signal to be masked as well as their variation with time [20], [21]. In audio coding, the masker is the original input signal and the signal to be masked is the quantization error.

The masking phenomena can be exploited to determine the optimal assignment of available bits. Bits need only be assigned to the audible spectral components. The step size and the number of bits for the quantizers can be chosen to give a quantization noise for the audible components that is below the masking threshold.

A. Simultaneous Masking

Simultaneous masking occurs when the masker and the maskee are presented at the same time to the hearing system. There are many models for computing the simultaneous-masking threshold [1], [14], [20], [22]. Since the MDCT is employed to decompose the input signal, we use a modified version of the model proposed by Johnston [22] which is based on the work by Zwicker [1] and Schroeder *et al.* [20] to calculate the masking threshold corresponding to the MDCT coefficients.

The masking calculation consists of the following steps.

- Calculate the Bark energy spectrum.
- Convolve the Bark energy spectrum with the following spreading function proposed by Schroeder *et al.* [20] to give the excitation curve

$$\mathrm{SpFn}(z)=15.81+7.5(z+0.474)-17.5\left(1+(z+0.474)^2\right)^{0.5} \tag{5}$$

  where z is the critical band number in Bark.
- Subtract an offset (in decibels) depending on a tonality factor from the excitation level to give the masking threshold.
- Compare the masking threshold with the hearing threshold to select the greater value.

Excitation Level: The Bark spectrum is derived from the DFT frequency spectrum with a nonlinear transformation of the frequency variable [22]. This gives a measure of the distribution of energies with respect to the critical band numbers. The Bark spectrum is convolved with the spreading function to give an excitation level [22].

Masking Level Calculation: The masking threshold is derived from the excitation level by subtracting an offset (in decibels) to give the masking level. The offset value depends on whether the signal is tone-like or noise-like.

In contrast to [22] in which a spectral flatness measure is used to identify the nature of the whole frame, we take another approach based on the predictability of the transform coefficients in each critical band. Note that, most audio signals have a noise-like structure at high frequencies despite the fact that they may have a strong harmonic structure at low frequencies. Considering this fact, it would be more accurate to identify the nature of the spectrum locally. The tonality factor will be calculated for each critical band using the predicted value of the current subvector [14]

$$\tilde{X}^{(i)} = 2X^{(i-1)} - X^{(i-2)} \tag{6}$$

where $\tilde{X}^{(i)}$ is the predicted vector at time index i, $X^{(i-1)}$ and $X^{(i-2)}$ are the vectors containing the transform coefficients in the same critical band in two previous frames (i indicates the newest vector). The relative prediction error is calculated as

$$\delta = \frac{\left\| X^{(i)} - \tilde{X}^{(i)} \right\|}{\left\| X^{(i)} \right\| + \left\| \tilde{X}^{(i)} \right\|}. \tag{7}$$

The relative prediction error will be converted to a tonality factor according to [14]

$$a = \min\left(1, \max\left(-0.3 - 0.43\log(\delta), 0\right)\right). \tag{8}$$

The offset value is determined by the tonality factor [22]

$$L_{\mathrm{offs}}(j,a) = a(14.5+j) + 5.5(1-a) \tag{9}$$

where j is the index of the critical band.

The masking level can then be calculated. Finally, the calculated masking threshold is compared to the threshold of hearing in the corresponding critical band to select the greater value.

MDCT Masking Threshold: Since the masking threshold is calculated based on the DFT, this masking threshold must be modified for use with the MDCT coefficients. Consider the following relationship between the DFT and MDCT coefficients [11]:

$$C(k)=\sqrt{\frac{2}{M}}\,|S(k)|\cos\left(\frac{\pi(M+1)(2k+1)}{4M}-\angle S(k)\right) \tag{10}$$

where $C(k)$ is the MDCT coefficient and $S(k)$, the corresponding DFT coefficient of the windowed input signal modulated by $\exp(-j\pi n/N)$, is calculated as follows:

$$S(k) = \mathcal{F}\left\{\exp\left(\frac{-j\pi n}{N}\right)x(n)h(n)\right\} \tag{11}$$

where $\mathcal{F}$ denotes the Fourier transform, $x(n)$ and $h(n)$ are the input block of data and the window function respectively. If $\mathcal{M}_{\mathrm{DFT}}(k)$ is the masking threshold corresponding to the kth DFT coefficient, then in order to have the same signal-to-mask ratio (SMR) at any coefficient in the DFT and MDCT domain, the following relation should hold:

$$\frac{C^2(k)}{\mathcal{M}_{\mathrm{MDCT}}(k)} = \frac{|S(k)|^2}{\mathcal{M}_{\mathrm{DFT}}(k)} \tag{12}$$

where $\mathcal{M}_{\mathrm{MDCT}}(k)$ is the masking threshold corresponding to the kth MDCT coefficient. The masking thresholds are then related as follows:

$$\mathcal{M}_{\mathrm{MDCT}}(k) = \frac{2}{M}\mathcal{M}_{\mathrm{DFT}}(k) \times \cos^2\left(\frac{\pi(M+1)(2k+1)}{4M} - \angle S(k)\right). \tag{13}$$

B. Temporal Masking

Temporal masking occurs when signals occur close in time, but not simultaneously. A signal can be masked by another signal that occurs later (premasking), or a signal can be masked by another signal that ends before the signal begins (postmasking). The duration of premasking is less than 5 ms (for narrow-band stimuli), whereas that of the postmasking is in the range of 50 to 200 ms [19]. Since incorporating the backward masking of the hearing system into the coder introduces extra

delay with little gain in compression, we neglect that effect and just exploit forward masking.

There are few analytical expressions which model post-masking (but see [23] and [24]). We have adopted the following model proposed in [23] as it takes both the effect of the frequency and the level of the masker into account:

$$m_{\mathrm{t\ dB}}(f, L) = \alpha + \beta \exp\left(-\frac{f}{\gamma}\right) \tag{14}$$

where $m_{\mathrm{t\ dB}}$ is the temporal masking in decibels, L is the sound (masker) level in the previous frame in decibels SPL, f is the frequency in hertz and α, β, and γ are parameters to be determined from experimental data. In [23], three expressions have been fitted to the experimental data for α, β, γ. In this work, we consider the temporal masking only if the masker level is more than 30 dB SPL. We assume that an audible sound usually has a level of more than 30 dB SPL in an ordinary (nonquiet) environment. Based on this assumption and the data given in [23] we have found the following expressions for the above-mentioned parameters:

$$\begin{aligned}\alpha &= 0.001L^2 + 0.2267L + 17.7142\\ \beta &= -0.0047L^2 + 1.2256L - 24.32548\\ \gamma &= -0.0002L^4 + 0.0546L^3 - 5.4685L^2\\ &\quad + 234.7411L - 3325.0350.\end{aligned} \tag{15}$$

Note that the data reported in [23] give the level of the temporal masking at 20 ms after the masker. The time interval between successive frames in the NPAC coder is 15 ms, and hence the masking level will be under-estimated using this formula.

In the NPAC coder, we calculate the temporal masking for each critical band. In doing so, we assume that all the energy in each critical band is concentrated in the center frequency (except the first band for which we set f to 100 Hz) and that the sound level in any critical band is due to the contribution of all the transform coefficients in that critical band. This way, for each frame we calculate 17 temporal-masking thresholds. If the temporal-masking threshold in any critical band is greater than the sound level in that band, we assume that all the coefficients in that critical band are masked. Otherwise, the temporal-masking threshold is equally divided among the coefficients in that critical band. Note that since the masking level is due to the contribution of all the transform coefficients in a critical band (from the previous frame), the equal division of the masking power among the transform coefficients in an unmasked critical band is an appropriate way to take into consideration the effect of the temporal-masking phenomenon.

C. Combined Masking Threshold

A combined masking threshold is computed by considering the effect of both temporal-masking and simultaneous-masking thresholds. There is a question as how to combine the masking due to these effects [23], [25]. We use a *power-law* rule as follows [23]:

$$\mathcal{M}_{\mathrm{net}} = \left(\mathcal{M}_1^p + \mathcal{M}_2^p\right)^{\left(\frac{1}{p}\right)} \tag{16}$$

where $\mathcal{M}_{\mathrm{net}}$ is the net masking threshold (in the linear domain) due to two masking thresholds $\mathcal{M}_1$ and $\mathcal{M}_2$. A value of 0.3 for parameter p is found to be a good match to experimental data [23].

D. Verification of the Masking Models

In order to verify the masking models, the masking thresholds for several audio signals were computed. After replacing the masked coefficients by zeros, there was no perceptual difference between the original and reconstructed signals. If we artificially increase the level of masking to have about 80% of the transform coefficients masked, the quality of the reconstructed signal is still good, i.e., with no annoying distortion. This experiment confirms that we should concentrate on reproducing the perceptually important transform coefficients.

V. Quantization of the Transform Coefficients

In the NPAC coder, we decompose subbands of transform coefficients into gains and shapes. Then a VQ scheme along with perception-based bit allocation is used to quantize the shape vectors. To quantize the gains (scale factors), a predictive/nonpredictive scheme is used.

A. Quantization of the Shape Vectors

One way to accomplish good quantization is to consider the characteristics of the hearing system such as masking phenomena and limited temporal and frequency resolution. Due to the limited number of bits available for coding the transform coefficients, vector quantization is used rather than scalar quantization. We quantize and transmit only unmasked transform coefficients. This approach would require additional bits to identify the masked/unmasked coefficients to reconstruct the audio signal at the receiver. Instead of doing so, we employ a split adaptive VQ scheme to quantize the transform coefficients. The bandwidth division is based on the critical bands. We incorporate the masking threshold while vector quantizing the transform coefficients without explicitly transmitting any information about the masking pattern.

We use a modified version of the LBG algorithm [26] with a distortion measure based on the audible noise energy to design the codebooks [10]. The same error criterion is used to select the best codewords.

For a normalized vector X_n and the jth codeword $\chi^{(j)}$, we define the following distortion measure:

$$\mathbf{d}(k) \triangleq \left|X_n(k) - \chi^{(j)}(k)\right|^2 - \mathcal{M}_n(k) \tag{17}$$

where $\mathcal{M}_n$ is the vector of normalized masking thresholds corresponding to X_n. The normalized energy of the audible noise is calculated from

$$D\left(X_n, \chi^{(j)}\right) = \sum_{k=1}^{K} \max\left(\mathbf{d}(k), 0\right) \tag{18}$$

where K is the dimension of X_n. The centroid of each Voronoi region is determined by minimizing the normalized energy of the audible noise as follows:

$$\chi_{\text{opt}}^{(j)} = \arg\min_{\chi^{(j)}} \sum_{i=1}^{I} D\left(X_n^{(i)}, \chi^{(j)}\right) \tag{19}$$

where I is the number of the vectors in region j.

At very low bit rates, it is not possible to have transparent coding. Since the quantization noise level often goes above the masking threshold, it is appropriate to shape the quantization noise inside each band too. Therefore, we may modify the error criterion as follows:

$$\mathbf{d}_w(k) = \frac{\left|X_n(k) - \chi^{(j)}(k)\right|^2 - \mathcal{M}_n(k)}{X_n^2(k) + \mathcal{M}_n(k)}$$
$$D_w\left(X_n, \chi^{(j)}\right) = \sum_{k=1}^{K} \max\left(\mathbf{d}_w(k), 0\right) \tag{20}$$

where D_w is the total weighted quantization noise above the normalized masking threshold. By making this modification, we allow the audible quantization noise to get shaped as a function of the distribution of energy inside a critical band. The reason for choosing $X_n^2(k)+\mathcal{M}_n(k)$ as the normalization factor is that we want to avoid giving a large weight to the masked coefficients which usually have a small magnitude. Note that the masking threshold is the same for all the coefficients in a critical band while the first term in the normalization factor, $X_n^2(k)$, is not the same. This normalization factor gives larger weights to the unmasked coefficients with small magnitudes.

1) Memory Reduction for Storage of the Codebooks: Vector quantization needs a lot of memory space to store the codebooks. Solutions to this memory problem have been proposed in the literature [27]. In the NPAC coder, bits are adaptively allocated to different critical bands. As such, codebooks with different length are needed to quantize shape vectors in each critical band. In this work, we have investigated the following methods to reduce the memory required to store the codebooks with little loss of quality. In all methods, codebooks with sizes equal to different powers of 2 are trained for each critical band. In one approach, we find the closest codewords in the largest codebook to the codewords of the second largest codebook (in the mean square sense). Then we reorder the codewords of the largest codebook to put the selected codewords on the top. We repeat the procedure for other codebooks to end up with embedded codebooks. With these embedded codebooks, the memory required is reduced by almost 50% for large codebooks, with very little loss of quality.

Another approach to creating an embedded codebook is to use the largest codebook to code a large set of training vectors. Then based on the frequency of selection of the codewords, we can reorder the codewords to have most-often selected codewords on the top of the codebook. The resulting codebook shows almost the same performance as when we use separate codebooks to quantize the subvectors. To further reduce memory, the bands with the same number of coefficients can share the same codebook with little loss of quality.

B. Predictive VQ of the Scale Factors

The transform coefficients in each critical band are normalized by the corresponding square-root-energy E_j which must be available to the receiver as side information. There exists a high level of correlation between the gain vectors. This similarity is partly due to the 50% overlap between successive frames. The interframe correlation can be efficiently exploited by applying a predictive scheme to quantize the scale factors.

Since processing several past frames to estimate the current vector makes the prediction scheme more vulnerable to channel errors, we use only one previous quantized spectrum. In the NPAC coder, a predictive/nonpredictive VQ scheme is used to quantize the scale factors in the log domain. We use the spectral distortion measure to choose the quantization scheme in the way that the predictive scheme is employed when the average spectral difference of the current and previous vectors is less than 6 dB. This strategy is compatible with the mechanism of the hearing system; in steady-state parts of the input signal such as voiced speech, we need finer quantization of both spectral shapes and gains, whereas for "unstructured" or noise-like parts more coarse quantization is adequate. This also can be justified by taking into account the masking property of the hearing system. Since the masking threshold in the case of tone-masking-noise is lower than that of noise-masking-noise, we need finer quantization for pseudo-periodic (tone-like) parts of the input signal. In the predictive scheme, we quantize the vectors containing the scale factors through the following steps (note that all these steps are performed in the log domain).

- Calculate the mean value of the scale factors

$$\mu_i = \frac{1}{17} \sum_{j=1}^{17} g_j^{(i)} \tag{21}$$

  where $g_j^{(i)}$ is the logarithm of E_j (square-root-energy) at time index i.
- Remove the mean value from the scale factors

$$\mathbf{g}_n^{(i)} = \mathbf{g}^{(i)} - \mu_i \tag{22}$$

  where $\mathbf{g}^{(i)}$ is the log-gain vector.
- Quantize μ_i using a differential quantizer.
- Predict the current mean-removed vector from the previous mean-removed vector using the best predictor matrix

$$\mathbf{P}_{\text{opt}} = \arg\min_{\mathbf{P}_k} \left\| \mathbf{g}_n^{(i)} - \mathbf{P}_k \hat{\mathbf{g}}_n^{(i-1)} \right\| \tag{23}$$
$$\tilde{\mathbf{g}}_n^{(i)} = \mathbf{P}_{\text{opt}} \hat{\mathbf{g}}_n^{(i-1)} \tag{24}$$

  where $\mathbf{P}_k$ is the predictor matrix, $\hat{\mathbf{g}}_n^{(i-1)}$ is the mean-removed quantized version of the previous vector and $\tilde{\mathbf{g}}_n^{(i)}$ is the prediction of the mean-removed current vector.
- Form the difference vector

$$\mathbf{d}_g = \mathbf{g}^{(i)} - \hat{\mu}_i - \tilde{\mathbf{g}}_n^{(i)} \tag{25}$$

  where $\hat{\mu}_i$ is the quantized mean value of the current vector.
- The difference vector will be quantized using a two-stage VQ.

This approach leads to fine quantization of the scale factors in steady-state parts of the input signal which is highly desired for high quality of the coded signal. A total of 37 bits is used to quantize the scale factors. For computational reasons, we limit codebooks size to 2048 codewords. For the predictive scheme, the coding is done as follows: 6 bits for the mean value, 9 bits for the predictor selection and 2×11 bits for the two-stage VQ of the difference vectors.

In the nonpredictive scheme, the vector of scale factors is mean-removed (in the log domain). The mean-removed vector is vector quantized using a codebook of 2048 codewords. In the next step the best estimator matrix is selected out of 64 matrices (requiring 6 bits) to estimate the current mean-removed vector based on the observation of the best codeword selected in the first step. Then the difference vector will be formed as described for the predictive scheme. Finally the difference vector will be quantized using a codebook of 2048 codewords. Note that 9 bits is spent to quantize the mean value.

For a set of 10 000 test vectors, the average spectral distortion for steady-state frames (using the predictive scheme) was less than 1.5 dB and for the rest of the frames (using the nonpredictive scheme) was 2.5 dB. The number of quantized vectors with spectral distortion above 4 dB was less than 0.5%.

1) Design of the Predictor Matrices: The predictor matrices are designed to minimize the average spectral distortion between the mean-removed gain vectors and the predicted vectors. We take a training set of 100 000 vectors and find the predictor matrices using a modified version of the Lloyd algorithm. First we design one predictor matrix for the whole training set. Then by splitting the first matrix and performing the Lloyd algorithm, we find new predictor matrices. This process continues until the desired number of predictor matrices are found. For each subset of the training set (corresponding to a predictor matrix), we find the optimal predictor matrix through the following optimization procedure:

$$\mathbf{P}_{\text{opt}}^{(j)} = \arg\min_{\mathbf{P}_k^{(j)}} \sum_{i \in R_j} \left\| \mathbf{g}_n^{(i)} - \mathbf{P}_k^{(j)} \hat{\mathbf{g}}_n^{(i-1)} \right\| \tag{26}$$

where R_j contains the time indexes of the vectors belonging to the jth region. Note that, in order to perform the optimization we need to have the quantized vectors. To overcome this problem, we use the quantized vector obtained through the nonpredictive method and then refine the predictor matrices by repeating the optimization procedure. In each iteration we use the finer quantized value for $\mathbf{g}_n^{(i-1)}$ obtained in the previous iteration.

By using the orthogonality theorem, we find the solution to the optimization problem as follows:

$$\mathbf{P}_{\text{opt}}^{(j)} = \mathbf{R}_{01}^{(j)} \left(\mathbf{R}_{11}^{(j)} \right)^{-1} \tag{27}$$

where $\mathbf{R}_{01}^{(j)}$ is the summation of the cross-correlation matrices of the current and quantized previous vectors in Voronoi region j. The matrix $\mathbf{R}_{11}^{(j)}$ is the summation of the autocorrelation matrices of the quantized previous vectors in the same region. We continue the iterations until the required number of predictor matrices is found and the change in the average spectral distortion becomes less than a threshold.

2) Simplification of the Predictor Matrices: The magnitude of the entries of the prediction matrices decrease as they are farther from the main diagonal. In effect, each component in the current vector will be predicted mainly by the corresponding and a few adjacent components of the previous quantized vector. We exploit this fact in order to set the far-off diagonal elements of the predictor matrices to zero. By doing so, we reduce the computation load and also the memory for the storage of the predictor matrices. In order to find the predictor matrices, we have to reformulate the optimization procedure. As an example, assume that the main diagonal and its adjacent diagonals are nonzero and the rest of matrix entries are set to zero. We can easily generalize the following formulation for any number of nonzero diagonals:

$$\tilde{\mathbf{g}}_n^{(i)} = \mathbf{P} \hat{\mathbf{g}}_n^{(i-1)} \tag{28}$$

where $\mathbf{P}$ is the predictor matrix defined as

$$\mathbf{P} = \begin{bmatrix} p_{1,1} & p_{1,2} & & & \\ p_{2,1} & p_{2,2} & \ddots & \mathbf{0} & \\ & \ddots & \ddots & \ddots & \\ & \mathbf{0} & \ddots & \ddots & p_{16,17} \\ & & & p_{17,16} & p_{17,17} \end{bmatrix}. \tag{29}$$

Rewrite (28)

$$\tilde{\mathbf{g}}_n^{(i)} = \mathbf{G}_i \mathbf{c} \tag{30}$$

where

$$\mathbf{G}_i^t = \begin{bmatrix} \hat{g}_{n1}^{(i-1)} & 0 & \cdots & 0 & 0 \\ \hat{g}_{n2}^{(i-1)} & 0 & & & \\ 0 & \hat{g}_{n1}^{(i-1)} & & & \\ & \hat{g}_{n2}^{(i-1)} & & \vdots & \\ & \hat{g}_{n3}^{(i-1)} & & & \vdots \\ & 0 & & & \\ \vdots & & & 0 & \\ & & & \hat{g}_{n15}^{(i-1)} & \\ & \vdots & & \hat{g}_{n16}^{(i-1)} & \\ & & & \hat{g}_{n17}^{(i-1)} & 0 \\ & & & 0 & \hat{g}_{n16}^{(i-1)} \\ 0 & 0 & \cdots & 0 & \hat{g}_{n17}^{(i-1)} \end{bmatrix} \tag{31}$$

and

$$\mathbf{c} = [p_{1,1}\ p_{1,2}\ p_{2,1}\ p_{2,2}\ p_{2,3} \cdots p_{16,15}\ p_{16,16}\ p_{16,17}\ p_{17,16}\ p_{17,17}]^t. \tag{32}$$

We have to find $\mathbf{c}$ to minimize the spectral distortion for each subset (Voronoi region) of the training set

$$\mathbf{c}_{\text{opt}}^{(j)} = \arg\min_{\mathbf{c}_k^{(j)}} \sum_{i \in R_j} \left\| \mathbf{g}_n^{(i)} - \mathbf{G}_i \mathbf{c}_k^{(j)} \right\|. \tag{33}$$

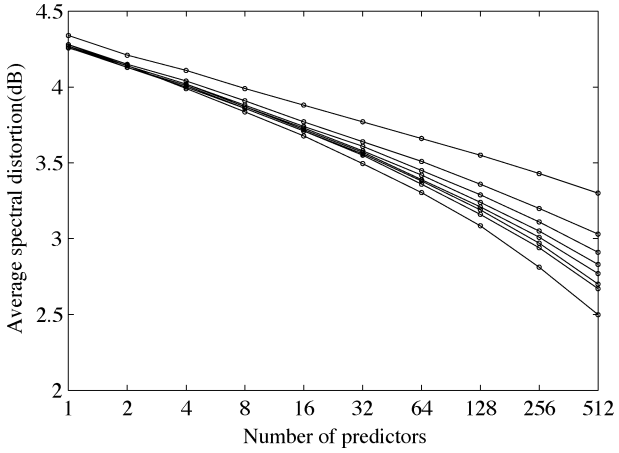


Fig. 3. Average spectral distortion versus the number of predictor matrices (from top to bottom) with 1, 3, 5, 7, 9, 11, 13, and 33 nonzero diagonals. The lowest curve corresponds to the predictor matrix with all-nonzero diagonals.

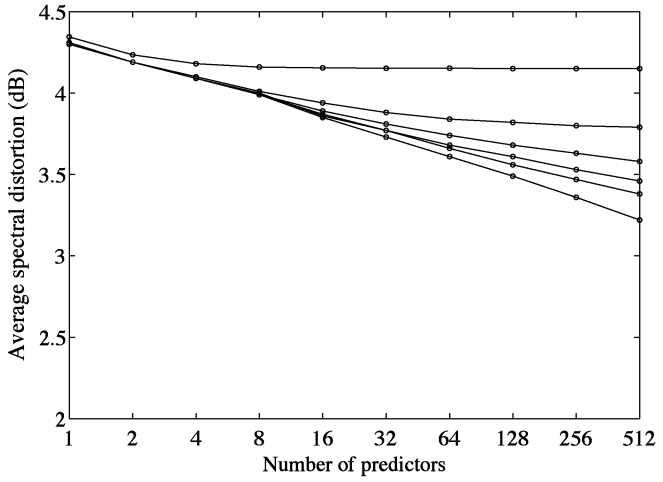


Fig. 4. Average spectral distortion versus the number of predictor filters with length (from top to bottom) 1, 3, 5, 7, 9, 17.

$\mathbf{c}_{\text{opt}}^{(j)}$ will be the solution to the following linear equations:

$$\mathbf{A}\mathbf{c}_{\text{opt}}^{(j)} = \mathbf{y} \tag{34}$$

where

$$\mathbf{A} = \sum_{i \in R_j} \mathbf{G}_i^t \mathbf{G}_i \quad \mathbf{y} = \sum_{i \in R_j} \mathbf{G}_i^t \mathbf{g}_n^{(i)}. \tag{35}$$

It is easy to show that $\mathbf{A}$ is a positive definite matrix and therefore we can use the Cholesky method to solve the linear equations. Fig. 3 shows the average spectral distortion for different predictor matrices as a function of the number of predictor matrices. There is a significant gap between the upper curve which corresponds to the diagonal predictor matrix and the other predictors. This is due to the fact that other predictor matrices exploit the lateral correlation among the components of the gain vector. At low rates, the performance of the predictors (except the single diagonal predictor) are almost the same, but as the number of bits increases, the performance of the predictor scheme can be enhanced at the cost of the higher computational load and larger memory storage for the predictors.

In order to lower the computational load and the required memory, we have investigated a special case of the above-mentioned procedure in which all the predictor matrix entries on the same diagonal are equal. Viewing this approach from a filtering perspective, we convolve the quantized previous gain vector with a noncausal FIR filter to estimate the current vector. The optimization procedure can be formulated in a fashion similar to the case above.

Fig. 4 shows the average spectral distortion for different predictor filters as a function of the number of predictors. Like the previous approach, there exists a gap between the upper curve which corresponds to the predictor filter with length 1 (single scalar predictor) and the other predictors. At low rates, the performance of the predictors (except the single diagonal predictor) are almost the same, but as the number of predictors increases, the spectral distortion saturates for short filters, but for long filters it decreases linearly with increasing the number of predictors. Also the difference between the prediction errors decreases as the filter length increases. That is expected as there is not a significant correlation between widely separated components of the gain vectors. Compared to the first approach, for the same number of bits, the prediction error is higher in the second approach.

In the NPAC coder, we have used the last approach with the predictor matrices having five nonzero diagonals.

C. Gain Adjustment

In low-rate coding, there are not enough bits to finely quantize the perceptually important coefficients. In this coder we propose the following procedure to reduce the quantization errors by adjusting the gain in each critical band:

$$\rho_{\text{opt}} = \arg\min_{\rho} \sum_{k=1}^{K} \max\left(\left(X(k) - \rho\hat{X}(k)\right)^2 - \mathcal{M}(k), 0\right) \tag{36}$$

where ρ is the gain adjustment factor, X and $\hat{X}$ are the original and quantized vectors of transform coefficients, $\mathcal{M}$ is the corresponding masking threshold and K is the dimension of the subvector. This optimization procedure gives the optimal ρ to minimize the audible difference between the input and the output vectors. Computation of the optimal adjustment factors requires an optimization over each critical band. We take a suboptimal approach to decrease the computation because in low-rate coding the quantization noise in most bands is above the masking threshold. First we ignore the masking threshold

$$\rho_{\text{opt}} \approx \arg\min_{\rho} \sum_{k=1}^{K} \left(X(k) - \rho\hat{X}(k)\right)^2. \tag{37}$$

The resulting ρ will be optimal in the squared-error sense but suboptimal in a perceptual one. Note that some critical bands are totally or partially masked and therefore there is no need to lower the quantization noise energy below the masking threshold. In those bands, the adjustment factors are sometimes found to be up to 1000. To overcome this problem and also limit the dynamic range of the adjustment factor, we confine ρ to a range of 0.5 to 2. With an overhead of less than 2 kb/s, the

adjustment factors (vectors of 17 components with the limited dynamic range) can be finely vector quantized.

The gain adjustment cannot be integrated into the gain-quantization block. The quantized gains are needed for the bit-allocation block whereas the gain-adjustment factors are found after performing the bit allocation and shape quantization. However, if the roughly quantized gains (output of the first stage gain quantization VQ) are used for the bit assignment, this block can be absorbed into the second stage VQ of the gain-quantization block. This way the rate can be reduced at the expense of the accuracy of the bit allocation.

D. Quantization of the Transform Coefficients in Short Frames

When a short window (80 points) is used, a number of changes occur. For a short window, 40 new input samples and 40 previous ones are used. The MDCT unit generates only 40 transform coefficients. Although the number of the critical bands remains constant at 17, the distribution of the MDCT coefficients within the bands changes. In this case, some low-frequency critical bands have only one coefficient. The 17 critical bands are combined into seven aggregated bands. This aggregation is performed so that the vector quantization in split VQ can always operate on code vectors of dimension greater than one. Changes in the quantization procedure are required to handle the aggregated bands. A single gain is calculated for each aggregated band. To quantize the spectral-shape vectors, the masking threshold is calculated as for the large frames and then used for the corresponding transform coefficients inside an aggregated band.

Since there is little or no similarity between the gain vectors of the consecutive short frames due to the transient behavior of the input signal, the gain vectors of dimension 7 are quantized in a nonpredictive manner.

VI. Adaptive Bit Allocation

In traditional low-rate adaptive transform speech coders, bit assignment is done based on the distribution of the signal energy in the frequency domain aiming at minimizing the total noise energy [28]. Since for most audio signals, energy is concentrated at low frequencies, few bits are assigned to high-frequency components. This leads to an output signal which suffers from lowpass effects. In addition to that flaw, the masking phenomena are not fully taken into account which often results in allocating bits to the transform coefficients which are masked.

The aforementioned argument underlines the importance of shaping the noise spectrum based on perceptual principles. In perceptual bit allocation which is used in state-of-the-art audio coders, the coding noise can be shaped to be less audible than a noise with the same energy without noise shaping. Note that noise shaping can provide high coding quality without requiring a high SNR.

In low-rate coding of audio signals, due to the scarcity of bits, unmasked quantization noise (audible noise) is often inevitable. The final goal in low-rate coding is to deliver acceptable quality with no annoying artifacts. Two different strategies can be considered to shape the audible noise spectrum [29]. In one approach, the quantization-noise spectrum is shaped to become parallel to the masking-threshold curve. An alternative approach is to generate a flat noise spectrum above the masking threshold. According to [29, pp. 427–428], these two approaches are different in terms of auditory object formation. In the first approach, the quantization noise is highly correlated with the input signal and the audible noise is equally audible in different frequency bands. Therefore, the input signal and the noise will be perceptually fused to form one auditory object. In the second approach, the noise is not correlated with the signal and is audible to various extents at different frequencies. This way, the noise remains perceptually distinct from the input signal. In this work, we have developed different bit allocation algorithms based on the above-mentioned approaches. The algorithms are presented and discussed in the following sections and the one which suits more to low-rate coding has been used in the NPAC coder.

In the NPAC coder, adaptive bit allocation is performed both at the transmitter and the receiver using the quantized scale factors. The masking thresholds are calculated from the quantized scale factors. For each band we need to specify the offset value which is subtracted from the excitation level (in the log domain) in order to obtain the simultaneous masking threshold. The procedure of adaptive bit allocation is discussed in the following sections.

A. Tone/Noise Discrimination

The masking offset depends on whether the spectrum in each band is tone-like or noise-like. At low bit-rates we cannot afford to code the offset value for each band. However we do distinguish between two cases. In one case the input block of data has a harmonic structure which implies that the spectrum is more tone-like. In the other case the input has a more noise-like spectrum.

In order to distinguish between the two cases, in our implementation we use the same flag which is used in gain quantization to select either the predictive or nonpredictive schemes. When the flag is on, we suppose that the input frame is tone-like. Since for many audio segments, the signal is more tone-like in the low-frequency bands than the high-frequency bands, we assume larger offset values (up to 18 dB) for the low-frequency bands. By doing so, we assign more bits to the low frequency bands to maintain the pitch structure of speech. In each band the distance between the energy and the masking threshold is upper bounded by the offset value (in dB). Hence, the maximum number of bits allocated to each band is determined by dividing the corresponding offset value (in dB) by the distortion reduction rate (see the following section). For those frames for which the flag is off, we suppose the input signal is mostly noise-like and set the masking threshold for all bands 8 dB below the excitation level. Fig. 5 shows the offset values and the maximum number of bits allocated to each transform coefficient in different frequency bands.

In the case of short frames, since the input signal contains a transient and therefore does not have any harmonic structure (purely noise-like), we set the masking threshold 6 dB below the spread Bark spectrum [1], [21].

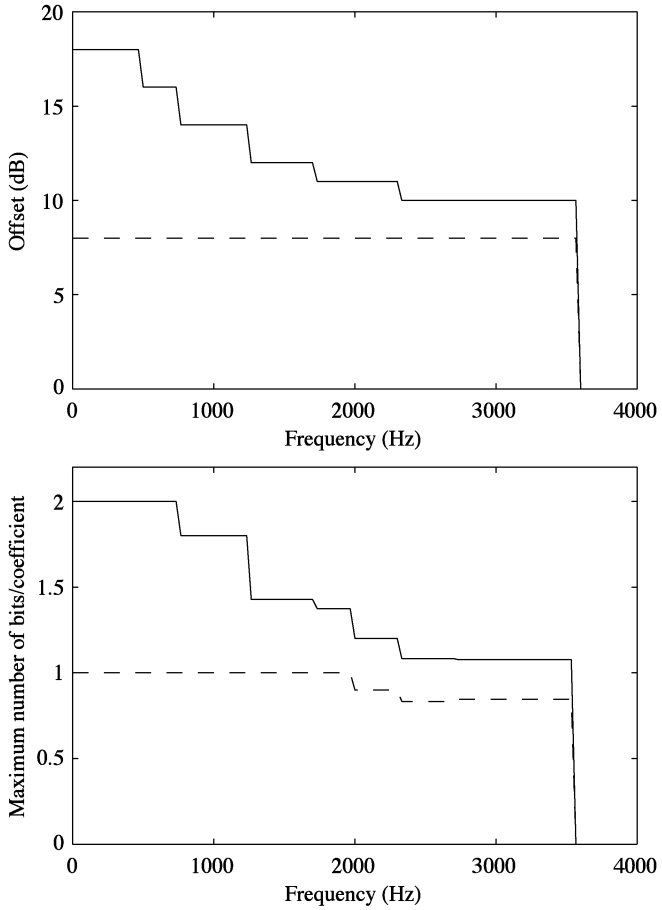


Fig. 5. Offset values for calculating the masking threshold (top) and corresponding maximum number of bits per coefficient (bottom) for tone-like frames (solid lines) and noise-like frames (dashed lines).

B. Critical Band Rate-Distortion Curve

In order to perform bit assignment we need the rate-distortion relationship for each codebook. A set of 100 000 normalized shape vectors was used to measure the average audible distortion (i.e., the average error above the masking threshold) for different numbers of bits. We note that the rate-distortion data can be well represented by a line fitted to the experimental data. As an example, Fig. 6 shows the rate-distortion data for the codebook corresponding to critical band 2 which contains three coefficients. The slope of the line which has been fitted to the data is -2.8 dB/bit.

Table I shows the slope of the lines fitted to the experimental data for the embedded codebook for each band. The high correlation between the experimental data and the fitted line verifies the accuracy of the linear approximation.

C. SMR-Based Bit Allocation

In this approach bit allocation is performed based on the signal-to-mask ratio (SMR). This way, the resulting noise spectrum will be parallel to the masking threshold curve. Each critical band is considered as a single entity with its corresponding SMR. The SMR is equal to the SNR when the quantization noise is at the threshold of audibility, i.e., when the noise level is at the masking threshold. The SMR for each band is calculated as

$$\mathrm{SMR}_j = \hat{g_j} - m_j \tag{38}$$

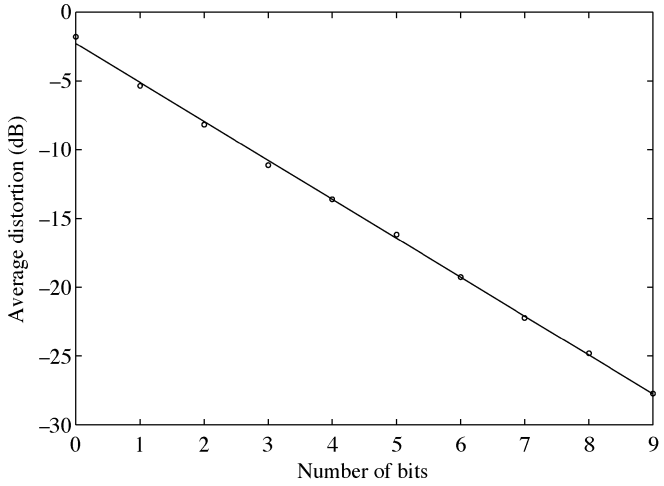


Fig. 6. Rate-distortion data for the embedded codebook corresponding to critical band 2 which contains three coefficients.

TABLE I
SLOPE OF THE RATE-DISTORTION LINE AND THE CORRELATION BETWEEN THE EXPERIMENTAL DATA AND THE LINEAR APPROXIMATION FOR DIFFERENT CRITICAL BANDS

Band	Number of coefficients	Slope (dB/bit)	Correlation Coefficient
1	2	-4.9	0.998
2	3	-2.8	0.999
3	3	-2.9	0.999
4	3	-2.9	0.999
5	3	-2.9	0.999
6	4	-2.1	0.999
7	4	-2.1	0.999
8	5	-1.6	0.998
9	5	-1.7	0.998
10	5	-1.7	0.999
11	7	-1.2	0.999
12	7	-1.2	0.998
13	8	-1.0	0.997
14	10	-0.9	0.998
15	12	-0.8	0.998
16	13	-0.7	0.999
17	13	-0.7	0.999

where $\hat{g_j}$ is the quantized log-energy in band j, and m_j is the logarithm of the masking threshold in that band. We assume that the initial distortion (in the log domain) for each band is equal to the corresponding SMR. A "greedy algorithm" using the rate-distortion data can be employed to assign one bit at a time to the band with the largest (updated) *noise-to-mask ratio (NMR)*. After assigning one bit to that band, its NMR on average decreases by the amount given by the corresponding rate-distortion data.

As a shortcut, a linear approximation of the rate-distortion data along with the values of SMRs can be used to allocate bits to each band according to the following formula:

$$b_j = \max\left(\frac{\mathrm{SMR}_j b_T}{\lambda_j \sum_{i\in\Omega}\left(\frac{\mathrm{SMR}_i}{\lambda_i}\right)}, 0\right) \tag{39}$$

where Ω contains the indexes of the bands with positive SMR and b_T is the total number of bits available to quantize the shape of the frequency spectrum within the critical bands. The slope of

the rate-distortion line, λ_j, indicates the approximate reduction in the NMR for one bit assigned to band j. Note that no bits are assigned to those bands whose SMR is negative. After the first round of bit allocation, the fractional parts of b_j's will be discarded to leave the integer parts. Therefore, the total number of bits allocated in the first step will be less than b_T. To allocate the remaining bits, the NMR is approximated for each band taking into account the bits already allocated in the first step

$$\text{NMR}_j = \hat{g_j} - m_j - \lambda_j b_j. \tag{40}$$

After calculating the value of NMR's, one bit at a time is allocated to the band with the largest value of the updated NMR. This process will continue until all the remaining bits are allocated.

D. Perceptual Energy-Based Bit Allocation

In the perceptual energy-based approach, bit assignment is based on the energy above the masking threshold. The distortion is considered as the audible part of the quantization noise, i.e., the noise above the masking threshold.

The level of audible noise will be relatively higher in the spectral valleys due to the fact that there is less energy above the masking threshold there than in regions corresponding to spectral peaks. In other words, for a given audible noise level across the spectrum, the local signal-to-noise ratio is lower in the spectral valleys. We consider two schemes to minimize the audible noise. In the first scheme the maximum of the distortion in the critical bands is minimized. In the second scheme the total audible noise is minimized.

Mini-Max Scheme: The mini-max bit assignment is done through the following optimization procedure:

$$\arg\min_{b_j} \left(\max\left(D_j(b_j)\right)\right) \quad \text{subject to} \quad \sum_{j=1}^{N_b} b_j = b_T \tag{41}$$

where N_b is the number of bands (i.e., 17 critical bands), b_T is the total number of bits available for each frame and D_j is the noise above the masking threshold.

We use a "greedy algorithm" to do the bit assignment. After each bit assigned, the distortion is updated. This way, one bit at a time is assigned to the band with the *largest updated distortion*.

Total Audible Distortion Minimization Scheme: This scheme minimizes the total audible distortion. Therefore, the optimization objective function changes to

$$\arg\min_{b_i} \sum_{i=1}^{N_b} D_i \quad \text{subject to} \quad \sum_{i=1}^{N_b} b_i = b_T. \tag{42}$$

One bit at a time is assigned to the band which results in the *largest reduction* in distortion. Alternately, an analytic approach may be employed. The energy above the masking threshold is related to the audible distortion through the following empirical formula:

$$D_i = c_i \mathcal{E}_i 2^{-\frac{b_i}{\beta_i}} \tag{43}$$

where D_i is the energy of the audible noise in band i, $\mathcal{E}_i$ is the total energy of the input vector above the masking threshold, c_i and β_i are constants found from the corresponding rate-distortion line for the codebook of critical band i. The linear approximation of the rate-distortion data in the log domain is re-expressed as an exponential relationship in the linear domain.

By using the mathematical expression for D_i in (43), the number of bits for each critical band is found to be

$$b_i = \max\left(\frac{\beta_i b_T}{\sum_{j=1}^{N_b}\beta_j} + \log_2\left(\frac{\mathcal{E}_i c_i}{\mathcal{E}_{gm}}\right), 0\right) \tag{44}$$

where

$$\mathcal{E}_{gm} = \left(\prod_{i=1}^{N_b}(c_i\mathcal{E}_i)^{\beta_i}\right)^{\left(\frac{1}{\sum_{i=1}^{N_b}\beta_i}\right)}. \tag{45}$$

The integer parts of the b_i's are kept and the remaining bits will be distributed one at a time to the band which reduces the *total distortion the most*.

E. Comparison and Subjective Evaluation of the Bit-Assignment Algorithms

Fig. 7(a) shows the power spectrum and the Bark power spectrum of a frame of voiced speech on the Bark scale. The Bark power spectrum is convolved with the spreading function to obtain the excitation pattern. The excitation and the masking curves are shown in Fig. 7(b). As it is seen in Fig. 7(b), the offset level, which is subtracted from the excitation pattern, is larger at the low-frequency critical bands. In bands 2, 4, 6, and 7, the energy falls below the masking threshold. The number of bits allocated to different critical bands using the two bit-assignment algorithms is shown in Fig. 7(c) and (d). Comparing the two bit allocation algorithms, we notice that the perceptual energy-based (i.e., mini-max) algorithm [Fig. 7(d)] allocates more bits to the bands with a large energy (for instance bands 1, 3, and 5). Both algorithms assign zero bits to the bands whose energy is below or almost below the masking threshold (bands 2, 4, 6, and 7). In this example, we have ignored temporal masking effects.

To evaluate the bit assignment algorithms, we conducted informal listening tests. Five listeners took part in the test and listened to two speech files (male and female) and two pieces of music (soprano and guitar) and the corresponding processed audio files. We used the perceptual bit-assignment schemes, i.e., perceptual energy-based approach (the mini-max scheme and the minimization of the total distortion scheme) and the SMR-based algorithm to compress the audio files. The test was run in an office environment and the subjects listened to the test materials over headphones.

For the tests, the NPAC coder was operating at 8 kb/s (120 bits per frame) and assigned 81 bits to 17 critical bands to quantize the spectral shapes. The unquantized adjusted gains were used to denormalize the quantized shape vectors.

As a preliminary test, we examined the impact of masking. In this test, we ignored any masking effect and performed the bit assignment based on the distribution of the signal energy. However, because the energy-based (ignoring masking) scheme allocates many bits to the low-frequency bands and relatively few bits to the high-frequency bands, the outputs suffered from incomplete coding of the high frequencies. This result verifies the importance of incorporating the masking effects into any bit-assignment algorithm.

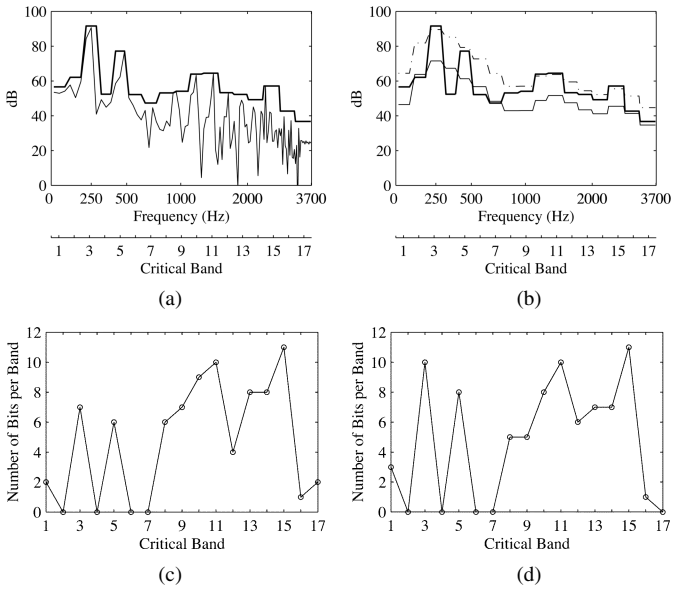


Fig. 7. Power spectrum, Bark power spectrum, excitation and masking curves for a frame of voiced speech. The lower plots show the bit allocation using the SMR-based and the Energy-based algorithms. (a) Power spectrum and Bark power spectrum (bold curve). (b) Bark power spectrum (bold curve), excitation curve (broken curve) and masking curve (thin curve). (c) SMR-based bit allocation. (d) Energy-based (mini-max) bit allocation.

The perceptual energy-based algorithm which minimizes the total audible distortion resulted in output quality similar to that for the energy-based (ignoring masking) bit allocation. Due to different dimensionality of different critical bands, the distortion reduction rate is higher for the narrower low-frequency bands. Moreover for many audio signals, the energy is concentrated in the low-frequency bands. Therefore, more bits compared to other perceptual schemes are assigned to the low-frequency bands. This results in finer quantization of low-frequency bands and coarser quantization of the high-frequency bands.

The other schemes (the SMR-based and the mini-max) delivered better quality with less high-frequency distortion. Both algorithms produced decoded signals which could be distinguished from the original narrow-band signals. The SMR-based algorithm gives less high-frequency distortion at the expense of a little degradation in the pitch structure which is perceived as an increased harshness. On the other hand, the decoded audio signals using the perceptual energy-based algorithm carried higher levels of high-frequency noise which sound like an echo along with the original signal. Although according to [29, pp. 427–428] the SMR-based algorithm should be less favored since the coding noise is fused with the signal, listeners showed a slight preference for this scheme over the mini-max scheme. Therefore, we use the SMR-based bit-allocation algorithm in the NPAC coder. However, for the future, we believe that the perceptually optimal bit-allocation algorithm for low-rate coding should be based on both the distribution of the audible noise and the SMR. This is a compromise between the schemes that might be better than either approach alone.

VII. Variable-Rate Coding of the Shape Vectors

Johnston [30] introduced *perceptual entropy* as the minimum bit rate for transmitting audio signals such that there is no perceivable difference between the original and coded signal. Based on the perceptual entropy criterion, it is possible to use a lossy compression scheme to code an audio signal without any perceivable distortion at a bit rate equal to its perceptual entropy.

TABLE II
Instantaneous Minimum, Average and Instantaneous Maximum Rates (kb/s) for Shape Quantization

File	Minimum	Average	Maximum
Female speech	0.0	7.2	11.5
Male speech	0.6	6.9	10.0
Piano	0.0	8.7	11.3
Orchestral	0.9	7.7	10.8

TABLE III
Bit Allocation to Code a Frame of Data

Data	Long Frame	Short Frame
Shape Quantization	81	25
Gain Quantization	37	14
Window Switching Flag	1	1
Gain Quantization Flag	1	0
Total	120	40

We conducted an experiment to calculate the number of bits required for each frame of data to achieve transparent quantization of the shape vectors. To estimate the number of bits needed to achieve transparent coding of the spectral shapes, we use the SMR in each critical band and determine the number of required bits by using the corresponding rate-distortion (approximated) line. Table II shows the instantaneous minimum, the average and the instantaneous maximum bit rates for the shape quantization of the transform coefficients for different audio signals. Note that some frames are temporally masked; therefore no bits are required to code the shapes.

McCourt in [31] reports that for a fixed-rate narrow-band coder, a minimum of 11 kb/s is required to perform transparent adaptive vector quantization of the shape vectors. Although the maximum rates shown in Table II are comparable to the minimum rate reported in [31], the average required rates are much lower than that rate. One conclusion from Table II is that the NPAC coder can provide high quality audio for narrow-band inputs in a source-controlled variable-rate scenario with a significant saving in average rate, but with a reasonable ceiling on the maximum number of bits.

VIII. Performance Evaluation

The NPAC coder has been designed to compress narrow-band audio signals. In the coder, different processing units have been designed to efficiently reduce the bit rate while maintaining acceptable audio quality free of annoying artifacts.

We have implemented the proposed coder in the C language. The source code was written for flexible experimentation and not optimized for execution speed. Nevertheless, the coder runs in real time on a computer using a 450 MHz Pentium processor.

The number of bits used to code each frame of the input data is 120 (for a long frame) and 40 (for a short frame), i.e., 1 bit per sample. Almost all of the bits were spent to code the normalized transform coefficients and the gains. Table III shows the bit

allocation for a long and a short frame of data when the NPAC coder operates at 8 kb/s (note that the coder operates without the gain adjustment module).

A. Subjective Evaluation

For many wide-band audio coders operating around 100 kb/s/channel, the compression process is transparent for most input materials. Crucial testing involves known difficult-to-code material. In low-rate coding of narrow-band audio signals, some distortion is inevitable. A wide range of material must be tested to ascertain that the distortion is not annoying. In our case, we chose a representative set of narrow-band audio files including various types of music, single instrumental music, single and multispeaker speech, and speech with background noise for testing the NPAC coder.

We have compared the quality of the coded signals using NPAC, the RealAudio[3] music coder operating at 8 kb/s, the RealAudio speech coder operating at 8.5 kb/s and the G.729 speech coder [32] operating at 8 kb/s. The quality of the coded signals were evaluated through informal listening tests. Eight test narrow-band audio signals (band-limited to 50–3600 Hz samples at 8000 Hz) including English female speech, English male speech, soprano, multispeaker, and various music types (piano, guitar, rock and orchestral music) were presented over headphones to five listeners. None of the test passages was used in training the quantizers of the NPAC coder.

In the listening test, the compressed signals were distinguishable from the original narrow-band signals. However, the purpose of the test was to ascertain whether the distortions in the output signals were annoying and to rank the outputs for the different coders for each audio passage. Due to narrow-band nature of the input, we expected the quality at the best of circumstances to be similar to that of AM broadcast radio.

At 8 kb/s, the listeners unanimously agreed that the NPAC coder delivered significantly better quality than the RealAudio music coder for most music passages and never performed worse than the RealAudio music coder. For all speech signals, the NPAC coder provided much better quality than the RealAudio music coder.

Compared to the G.729 coder and the 8.5 kb/s RealAudio speech coder, the listeners preferred the quality of almost all compressed signals using the NPAC coder. The exceptions were for the files containing a single speaker. Even for these cases, the quality was not far below that of the speech coders. Based on our experiments the NPAC coder works well as long as there is no strong harmonic structure due to voiced speech. In the case of the pseudo-periodicity in parts of the input signal, due to the sensitivity of the human ear to small variations of the harmonic structure, some distortion is perceived.

The NPAC met the expectations for the test passages, i.e., no annoying artifacts (e.g., block edge effects, annoying roughness, large Noise-to-Mask-Ratio in different critical bands). However, some enhancements should be made to the NPAC coder in order to achieve the same quality for single speaker passages as the quality delivered by speech-specific coders such as G.729.

[3]RealAudio is a trademark of RealNetworks, Inc.

IX. Conclusion

We have introduced the NPAC coder which is appropriate for a wide variety of narrow-band audio signals. In order to achieve high compression, this coder employs a variety of perception-based algorithms to account for the irrelevant parts of the input signal. Vector quantization is used to exploit intercoefficient dependencies in the scale factors as well as the normalized shape vectors. The new algorithms used in the coder include a perceptual error measure in training the codebooks and selecting the best codewords which takes into account the audible parts of the quantization noise, a perception-based bit-allocation algorithm and a predictive scheme to vector quantize the scale factors.

We have used the signal-to-mask ratio (SMR) measure and the empirically determined rate-distortion line to find the required bit rate for the transparent quantization of the spectral shapes in each critical band. The flexibility of the coder makes it possible to trade off quality versus rate for applications such as packet-based data networks. This coder can easily be modified to accommodate a wider range of input signals with different bandwidths and sampling rates.

Acknowledgment

The authors would like to thank the associate editor and the anonymous reviewers for their critical comments and valuable suggestions.

References

[1] E. Zwicker and T. Zwicker, "Audio engineering and psychoacoustics. Matching signals to the final receiver, the human auditory system," *J. Audio Eng. Soc.*, vol. 39, pp. 115–126, Mar. 1991.

[2] M. Sablatash and T. Cooklev, "Compression of high quality audio signals, including recent methods using wavelet packets," *Digital Signal Process.*, vol. 6, pp. 96–107, Apr. 1996.

[3] G. Davidson, L. Fielder, and M. Anil, "High quality audio transform coding at 128 kbps," in *Proc. IEEE Int. Conf. Acoustics, Speech, Signal Processing*, Albuquerque, NM, 1990, pp. 1117–1120.

[4] M. Bosi, K. Brandenburg, S. Quackenbush, L. Fielder, K. Akagiri, H. Fuchs, M. Dietz, J. Herre, G. Davidson, and Y. Oikawa, "ISO/IEC MPEG-2 advanced audio coding," in *Proc. 101st AES Conv.*, Munich, Germany, 1996.

[5] J. D. Johnston, D. Sinha, S. Dorward, and S. R. Quackenbush, "AT&T perceptual audio coding (PAC)," in *Collected Papers on Digital Audio Bit-Rate Reduction*, N. Gilchrist and C. Grewin, Eds. New York: Audio Eng. Soc., 1996, pp. 73–82.

[6] L. D. Fielder, M. Bosi, G. Davidson, M. Davis, C. Todd, and S. Vernon, "AC-2 and AC-3: low-complexity transform-based audio coding," in *Collected Papers on Digital Audio Bit-Rate Reduction*, N. Gilchrist and C. Grewin, Eds. New York: Audio Eng. Soc., 1996, pp. 54–72.

[7] M. Dietz, J. Herre, B. Teichmann, and K. Brandenburg, "Bridging the gap: Extending MPEG audio down to 8 kbit/s," in *Proc. 102nd AES Conv.*, Munich, Germany, 1997.

[8] M. Dietz, H. Popp, K. Brandenburg, and R. Friedrich, "Audio compression for network transmission," *J. Audio Eng. Soc.*, vol. 44, pp. 58–72, Jan. 1996.

[9] B. Feiten, R. Schwalbe, and F. Feige, "Dynamically scalable internet audio transmission," in *104th AES Conv.*, Amsterdam, The Netherlands, 1998.

[10] H. Najafzadeh-Azghandi and P. Kabal, "Perceptual coding of narrow-band audio signals at 8 kb/s," in *Proc. IEEE Workshop on Speech Coding*, Pocono Manor, PA, 1997, pp. 109–110.

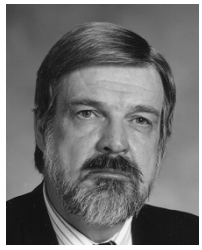
[11] ——, "Improving perceptual coding of narrow-band audio signals at low rates," in *Proc. IEEE Int. Conf. Acoustics, Speech, Signal Processing*, Phoenix, AZ, 1999, pp. 913–916.

[12] H. Najafzadeh-Azghandi, "Perceptual Coding of Narrow-band Audio Signals," Ph.D. thesis, McGill Univ., Montreal, QC, Canada, 2000.

[13] N. Iwakami and T. Moriya, "Transform Domain Weighted Interleave Vector Quantization (Twin-VQ)," in *Proc. 101st AES Conv.*, Los Angeles, CA, 1996.
[14] K. Brandenburg, G. Stoll, Y. Dehery, J. D. Johnston, L. V. Kerkhof, and E. F. Schroeder, "The ISO/MPEG audio codec: A generic standard for coding of high quality digital audio," *J. Audio Eng. Soc.*, vol. 42, pp. 780–791, Oct. 1994.
[15] J. P. Princen and A. Bradley, "Analysis/synthesis filter bank design based on time-domain aliasing cancellation," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. ASSP-34, pp. 1153–1161, Oct. 1986.
[16] N. S. Jayant and P. Noll, *Digital Coding of Waveforms*. Englewood Cliffs, NJ: Prentice-Hall, 1984.
[17] B. Edler, "Coding of audio signals with overlapping block transform and adaptive window functions," *Frequenz*, vol. 43, no. 4, pp. 252–256, 1989. In German.
[18] J. Herre and J. D. Johnston, "Enhancing the performance of perceptual audio coders by using temporal noise shaping TNS," in *Proc. 101st AES Conv.*, Los Angeles, CA, 1996.
[19] K. Pohlmann, *Principles of Digital Audio*. New York: McGraw-Hill, 1995.
[20] M. Schroeder, B. Atal, and J. Hall, "Optimizing speech coders by exploiting masking properties of the human ear," *J. Acoust. Soc. Amer.*, vol. 66, pp. 1647–1652, Jun. 1979.
[21] B. C. J. Moore, *An Introduction to the Psychology of Hearing*, 4th ed. New York: Academic, 1997.
[22] J. D. Johnston, "Transform coding of audio signals using the perceptual noise criteria," *IEEE J. Select. Areas Commun.*, vol. 6, no. 2, pp. 314–323, Feb. 1988.
[23] G. A. Soulodre, "Adaptive Methods for Removing Camera Noise From Film Soundtracks," Ph.D. dissertation, McGill Univ., Montreal, QC, Canada, 1998.
[24] B. Novorita, "Incorporation of temporal masking effects into bark spectrum distortion measure," in *Proc. IEEE Int. Conf. Acoustics, Speech, Signal Processing*, Phoenix, AZ, 1999, pp. 665–668.
[25] R. J. Beaton, J. G. Beerends, M. Keyhl, and W. Treurniet, "Objective perceptual measurement of audio quality," in *Collected Papers on Digital Audio Bit-Rate Reduction*, N. Gilchrist and C. Grewin, Eds: Audio Engineering Soc., 1996, pp. 126–152.
[26] Y. Linde, A. Buzo, and R. Gray, "An algorithm for vector quantizer design," *IEEE Trans. Commun.*, vol. COM-28, pp. 84–95, Jan. 1980.
[27] W. Y. Chan and A. Gersho, "Constrained-storage quantization of multiple vector sources by codebook sharing," *IEEE Trans. Commun.*, vol. 39, no. 1, pp. 11–13, Jan. 1991.
[28] R. Zelinski and P. Noll, "Adaptive transform coding of speech signals," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. ASSP-25, pp. 299–309, Aug. 1977.
[29] R. Veldhuis and A. Kohlrausch, "Waveform coding and auditory masking," in *Speech Coding and Synthesis*, W. B. Kleijn and K. K. Paliwal, Eds. New York: Elsevier, 1995, pp. 427–428.
[30] J. D. Johnston, "Estimation of perceptual entropy using noise masking criteria," in *Proc. IEEE Int. Conf. Acoustics, Speech, Signal Processing*, Apr. 1988, pp. 2524–2527.
[31] P. M. McCourt, "Critical band quantization analysis for masked distortion speech coding," in *Proc. IEEE DSP Workshop*, Loen, Norway, Sep. 1996, pp. 165–168.
[32] R. Salami, C. Laflamme, J.-P. Adoul, A. Kataoka, S. Hayashi, C. Lamblin, D. Massaloux, S. Proust, P. Kroon, and Y. Shoham, "Description of proposed ITU-T 8-kb/s speech coding standard," in *Proc. IEEE Workshop on Speech Coding for Telecommunications*, Annapolis, MD, 1995, pp. 3–4.

Hossein Najaf-Zadeh received the B.Sc. and M.Sc. degrees from Tehran University, Tehran, Iran, in 1986 and 1989, respectively, and the Ph.D. degree from McGill University, Montreal, QC, Canada, in 2000, all in electrical engineering.

In 2001, he joined the Advanced Audio Systems Group at the Communications Research Centre, Ottawa, ON, Canada, where he conducts research and development in audio source coding for multimedia applications.



Peter Kabal (S'70–M'75) received the Ph.D. degree in electrical engineering from the University of Toronto, Toronto, ON, Canada, in 1975.

He is a Professor of electrical and computer engineering at McGill University, Montreal, QC, Canada, and holds the NSERC/Nortel Industrial Research Chair. His current research interests focus on digital signal processing applied to speech and audio processing, adaptive filtering, and data transmission.