

A Perceptually Based Embedded Subband Speech Coder

Benjamim Tang, *Member, IEEE*, Albert Shen, *Member, IEEE*,
Abeer Alwan, *Member, IEEE*, and Gregory Pottie, *Member, IEEE*

Abstract—A new scheme for robust, high-quality, embedded speech coding based on subband decomposition and perceptually optimized bit allocation and prioritization is presented. An infinite impulse response (IIR) quadrature mirror filterbank (QMF) performs subband decomposition. A perceptual model, computed using subband spectral analysis, optimizes the coder's perceptual quality. Dynamic bit allocation and prioritization is combined with embedded quantization resulting in little performance degradation relative to a nonembedded implementation. The coder output is scalable from high quality at higher bit rates to lower quality at lower bit rates, supporting a wide range of service and resource utilization. The lower bit-rate representation is obtained simply through truncation of the higher bit-rate representation. Since source-rate adaptation is performed through truncation of the encoded stream, interaction with the coder is not required, making the embedded coder ideally suited for rate-adaptive communication systems. Performance for both speech and music was verified through subjective listening tests.

Index Terms—Speech coding, subband coding, embedded coding, perceptual metrics

I. INTRODUCTION

VOICE transmission is currently the most widespread use of wireless communications. The shift from analog to digital transmission in today's cellular environment has driven efforts to lower speech coder bit rates in order to increase system capacity, resulting in sub-toll-quality performance and poor robustness to channel errors.

The addition of microcellular and mobile satellite access in the near future will significantly change the wireless communication environment. In particular, the microcellular environment will provide new capabilities and applications not available today. Among the many changes, there will be a need to support high-quality voice and audio for video conferencing, broadcasting, and multimedia applications under a variety of transmission conditions.

Future systems will either incorporate a set of different speech and audio coders optimized to work over a limited range of coding rate, delay, quality, and error rate requirements, or they will incorporate a modular coder that can be quickly configured under network control to provide good

performance over a range of conditions. The latter approach is clearly preferable, as long as the performance penalty compared to a set of individually optimized coders is not significant.

Current algorithms do not adequately address the issue of scalable performance [1]–[4]. Code excited linear prediction-based (CELP-based) coders and other low bit-rate techniques do not easily scale to produce high-quality speech, and have difficulty when the source does not closely match the all-pole speech production model. On the other hand, high bit-rate coders cannot be efficiently supported by a network without a significant waste of resources. Supporting both low and high bit rates through rate adaptation greatly improves the system quality, efficiency, flexibility, and robustness.

The benefits of rate adaptation have led to the development of several variable rate speech and audio coding schemes [5]–[17]. In particular, subband coding is proposed as a flexible scheme for robust speech coding [13]–[15], and high-quality audio compression [17]. A speech production model is not used, ensuring robustness to speech in the presence of background noise, and to nonspeech sources. High-quality compression can be achieved by incorporating masking properties of the human auditory system [1], [2] to spectrally shape the quantization noise.

In this paper, a new scheme for robust, high-quality, scalable, embedded speech coding is presented. A novel scheme for dynamic bit allocation and prioritization and embedded quantization optimizes the perceptual quality of the embedded bitstream, resulting in little performance degradation relative to a nonembedded implementation. The bit allocation is based on an interpolated noise mask threshold that results in improved performance relative to noise-to-mask ratio (NMR) optimization found in most perceptual coders. A subband spectral analysis technique was developed that substantially reduces the complexity of computing the perceptual model.

The encoded bitstream is embedded, allowing the coder output to be scalable from high quality at higher bit rates, to lower quality at lower rates, supporting a wide range of service and resource utilization. The lower bit-rate representation is obtained simply through truncation of the higher bit-rate representation. Since source-rate adaptation is performed through truncation of the encoded stream, interaction with the source coder is not required, making the coder ideally suited for rate-adaptive communication systems.

The source robustness, embedded bitstream, and rate adaptation provide improved performance and flexibility under various wireless communication system implementations and

Manuscript received August 2, 1995; revised June 24, 1996. This work was supported in part by the NSF under Grant IRI-9309418 and by ARPA/CSTO under Contract J-FBI-93-112. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. W. Bastiaan Kleijn.

B. Tang is with TRW Inc., Redondo Beach, CA 90278 USA.

A. Shen is with Intel Inc., Hillsboro, OR 97124 USA.

A. Alwan and G. Pottie are with the Electrical Engineering Department, University of California, Los Angeles CA 90095-1594 USA.

Publisher Item Identifier S 1063-6676(97)01897-X.

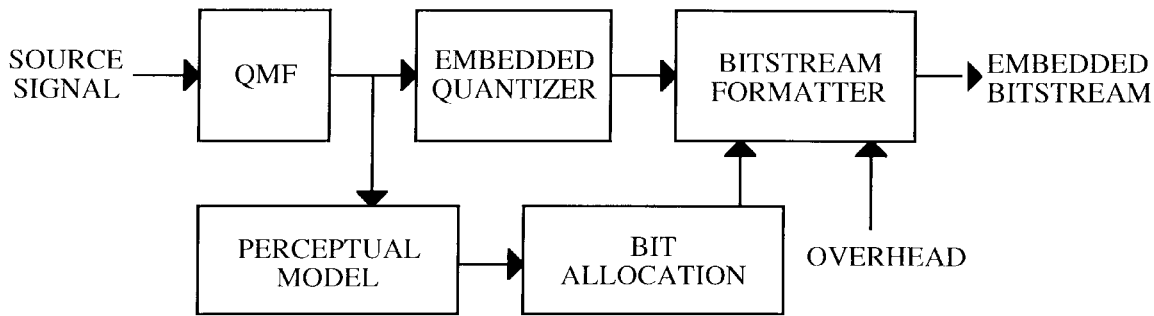


Fig. 1. Block diagram of the embedded subband speech encoder.

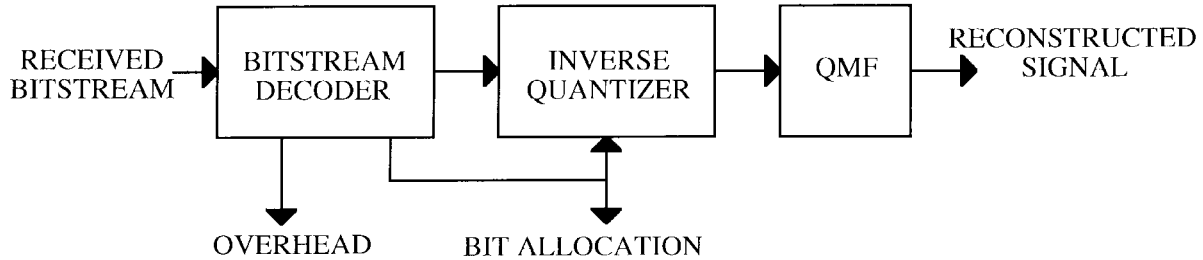


Fig. 2. Block diagram of the embedded subband speech decoder.

environments. The embedded bitstream improves robustness to transmission errors by allowing incorporation of unequal error protection or channel code rate adaptation. When combined with an embedded channel coding scheme, the embedded source code allows rate adaptation to occur without interaction with the source and channel coder, making it particularly attractive for evolving higher complexity wireless network topologies.

The benefits of embedded subband coding have been previously presented in [13] and [14] utilizing dynamic bit allocation and prioritization and embedded quantization in a scheme similar to the one described in this paper. Our approach, however, introduces several novel techniques, including

- 1) reduced complexity and delay through the use of an infinite impulse response (IIR) quadrature mirror filterbank (QMF);
- 2) improved perceptual bit allocation, and prioritization with the interpolated mask threshold cost measure;
- 3) reduced complexity of the perceptual model computation through the use of subband spectral analysis;
- 4) optimal embedded nonuniform scalar quantization; and
- 5) improved system robustness and flexibility using scalable, embedded source coding.

The speech coder algorithm is described in detail in Section II, highlighting the key blocks. The subjective listening test results are presented and discussed in Section III.

II. ALGORITHM DESCRIPTION

The coder performs subband decomposition of the input speech signal, spectrally shapes the quantization noise so the coding distortion is perceptually minimized for the human listener, and transmits the coded speech bitstream over an adaptive rate channel. The coder is scalable so that the bit rates

may be changed to optimize the quality of encoded speech for the given allocation of system resources and channel conditions.

The encoder, shown in Fig. 1, consists of five main components: analysis QMF, perceptual model, bit allocation, and prioritization, quantizer, and bitstream formatter. The decoder, shown in Fig. 2 consists of the bitstream decoder, inverse quantizer, and synthesis QMF. The source signal is 4 kHz filtered speech, or audio, sampled at 8 kHz. Nonoverlapping frames of 20 ms duration are used for processing and transmission.

The analysis/synthesis filterbank is an eight-channel, tree-structured IIR QMF. The filterbank is designed to provide good out-of-band rejection, alias cancellation, no amplitude distortion, minimal delay, and low phase distortion.

The perceptual model estimates the maximum noise level which still permits the noise to be masked by the signal. The coder takes advantage of the masking properties of the human auditory system to minimize the perceived degradation in quality introduced by the coding process. Since a sufficiently high bit rate may not be available to achieve an imperceptible noise level, a dynamic bit allocation and prioritization algorithm is used that attempts to maximize the perceptual quality for a given bit rate.

The subband samples are normalized and quantized using a nonuniform embedded scalar quantizer. The overall bit rate is made adaptive by specifying a maximum bit rate to be allocated or a target quality, and by allowing bit truncation of the quantization indices on a frame-by-frame basis to occur to reduce the bit rate. This allows various source-rate constraints to be met such as fixed bit-rate, variable bit-rate, or network-controlled source rate.

The adaptive nature of the source coding allows efficient utilization of system resources, providing consistent service under varying conditions. Various schemes for channel coding

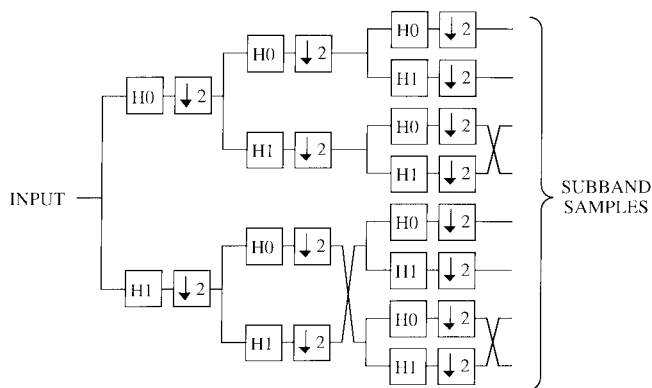


Fig. 3. Block diagram of the eight-band tree-structure QMF analysis filterbank.

and interleaving of the encoded bitstream can be easily implemented to take advantage of the scalability of the coder and provide the required robustness to transmission errors under various transmission and traffic conditions [6], [18], [19].

A. Analysis/Synthesis Filterbank

The QMF performs subband decomposition of the source signal. QMF's are a specific class of subband filters designed such that no aliasing is introduced even though the subband samples are critically decimated [20]. The synthesis filterbank is designed such that the aliasing introduced by the decimation process is canceled when the source signal is reconstructed.

FIR QMF's have traditionally been used in subband speech and audio coders. IIR QMF's have been avoided because of their nonlinear phase response. However, IIR filters allow low delay, sharp transition regions, and high stopband attenuation to be achieved with low complexity and low sensitivity to coefficient truncation [20].

Fortunately, the human auditory system is somewhat insensitive to phase distortion. Thus, it was reasonable to expect that an IIR QMF might be suitable for speech coding. A seventh-order elliptic IIR QMF halfband filter with 60 dB stopband attenuation was designed and implemented using the design methodology given in [20]. The eight-channel tree structure is shown in Fig. 3. Decimation of the high-frequency subband leads to spectral inversion, so the crossovers are necessary to maintain a low-frequency to high-frequency ordering of the subbands. The analysis filterbank response is shown in Fig. 4. The synthesis filterbank has a similar structure and response.

The imperceptibility of phase distortion in this filter design was verified through informal listening tests [21], [22]. Processing tones and sentences, both seventh- and ninth-order elliptic filters provided reconstructed signals free from audible distortion. The seventh-order design was selected based on its good performance and low complexity.

The complexity of the IIR QMF filtering operation is approximately 300 k operations/s and it introduces a delay of only 5 ms. In contrast, an eight-band 64-tap finite impulse response (FIR) QMF with a delay of 8 ms and complexity of 900 k operations/s provided only 40 dB of stopband attenuation [13]. Thus, the complexity of higher order FIR filters and the additional delay required limits the achievable

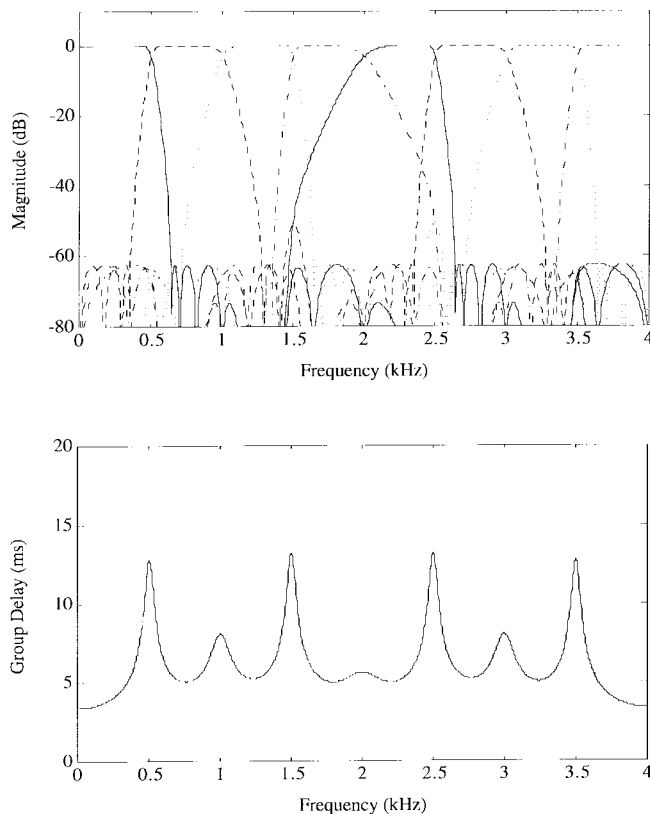


Fig. 4. Filterbank frequency response and group delay.

stopband attenuation. Poor stopband attenuation can become a significant source of perceptual distortion due to aliasing of quantization noise. Since the IIR QMF offered an improvement in stopband attenuation with no perceptible distortion and at lower complexity, it was the better choice to perform the subband decomposition.

B. Perceptual Model

A perceptual model is used to determine the optimal shaping of the quantization noise that minimizes the perceptual distortion of the reconstructed speech. The perceptual model takes into account the spectral masking properties of the human auditory system to compute the minimum level of an additive signal (the distortion) that is perceptible in the presence of masking from another signal. The minimum perceptible distortion level is defined as the just noticeable distortion (JND) [1]. The JND is usually referenced relative to the signal spectrum using the signal-to-mask ratio (SMR), rather than in absolute levels. Thus, if the distortion introduced by the coder is below the JND from the speech signal, the signal-to-noise ratio (SNR) exceeds the SMR, and the listener is not able to detect the distortion. Perceptual models are an integral part of several speech and audio coding schemes [16], [17].

An estimate of the JND is made by computing the signal spectrum in each frame using subband spectral analysis as shown in Fig. 5. Aliasing due to decimation of subband samples is eliminated by including the effects of the adjacent subbands in the analysis [23]. The frequency resolution and spectral leakage are nearly the same as if the transform had been performed on the input signal directly. When compared to

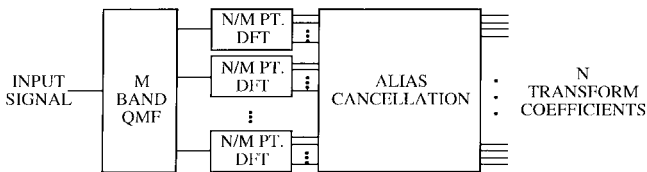


Fig. 5. Alias-canceled subband spectral analysis methodology.

performing the spectral analysis on the input signal directly, the analysis block length for subband spectral analysis with an eight-band filterbank is reduced by a factor of 8 and the complexity of a 256-point fast Fourier transform (FFT) computed every frame is reduced by 37.5%. This reduces the overall complexity of computing the perceptual model from 200 k to 125 k operations/s.

Once the source signal spectrum is obtained, both frequency and magnitude axes of the spectrum are transformed into dimensions that are more closely related to the characteristics of the human auditory system. The magnitudes of the spectrum are converted into sound pressure level (SPL) using calibration curves, and the frequencies are converted into critical bands, using the Bark scale [21]. The power of the spectral components over each of the 17 Bark bands in the 4 kHz signal bandwidth are then integrated, and masking curves are calculated from the signal energy [17]. The overall noise curve is obtained by power summing all the masking curves, giving the noise masking level in each of the bands. The Bark bands are then grouped according to the 500 Hz bandwidth of each subband. The JND is given by the minimum noise masking level, and the SMR by the ratio of the maximum signal level to minimum noise masking level for all the Bark bands in that subband.

Once the JND and SMR have been obtained, the perceptual quality metric usually optimized in perceptual coders is the NMR, defined as the ratio of distortion (noise) power to the JND masking threshold [17]. The assumption inherent in this metric is that a signal with lower NMR should be perceived as higher quality than one with higher NMR. Informal listening experiments, however, indicated that this was not necessarily true. Shaping noise based on the NMR often resulted in 0 bits being allocated to some subbands at low to medium bit rates, resulting in noticeable distortion. *Ad hoc* schemes that limited the maximum and minimum number of bits allocated to each subband resulted in improved perceptual quality at those rates [21].

An interpolated masking threshold (IMT) was devised as an alternative to NMR noise shaping. The IMT is a log-linear interpolation of the SMR, assumed to be the optimal spectrum for noise shaping for a given bit allocation. The IMT thus represents the achievable quantization noise level introduced by the coding process due to bit rate limitations. The IMT is computed by simply scaling the SMR such that $IMT = \alpha SMR$. If $\alpha = 1$, then the JND is met and the quantization noise is not perceptible. For $0 < \alpha < 1$, the quantization noise is perceptible, but the perceptual quality of the coded speech is maximized by achieving the highest possible α at a given bit rate.

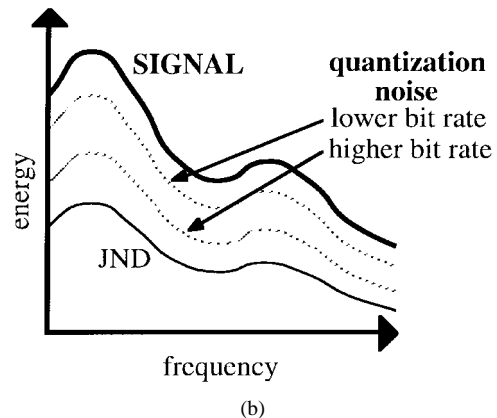
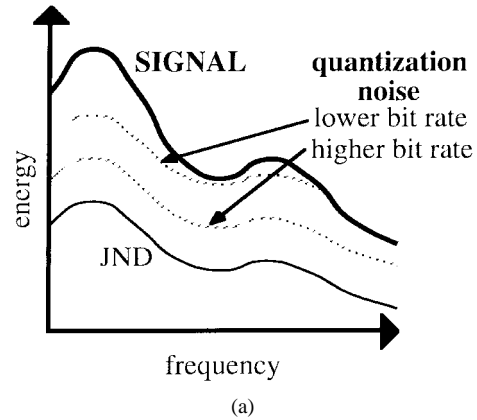


Fig. 6. Comparison of (a) NMR and (b) IMT shaping of quantization noise.

It should be noted that both the NMR and IMT are *ad hoc* schemes for assigning a cost measure to determine the optimum noise shaping when the JND cannot be achieved due to bit-rate limitations. Subjective listening tests, however, showed that IMT optimization resulted in higher perceptual quality than NMR optimization [21].

A comparison of NMR and IMT noise shaping is shown in Fig. 6. In both cases, if sufficient bits are available, the JND is the desired noise shape. If fewer bits are available, NMR attempts to distribute the additional quantization noise power evenly among the subbands. This leads to the classical reverse water filling solution for bit allocation. With IMT noise shaping, the desired noise shape is a log-linear interpolation between the source signal spectrum and the JND. Thus, the additional quantization noise power should be distributed in proportion to the required SNR in each band. This leads to the proportional bit allocation discussed in the next section.

C. Bit Allocation and Prioritization

The dynamic bit allocation and prioritization algorithm determines how many bits are allocated to quantize the samples in each subband for the given frame. The bit allocation determines the quantization noise power in each subband and, therefore, the overall quantization noise spectrum. The optimal bit allocation assigns sufficient bits in each subband such that the quantization noise spectrum is less than the JND. The bit prioritization algorithm determines how the bit allocation should be done if the full bit rate required is not available. The prioritization algorithm prioritizes each bit

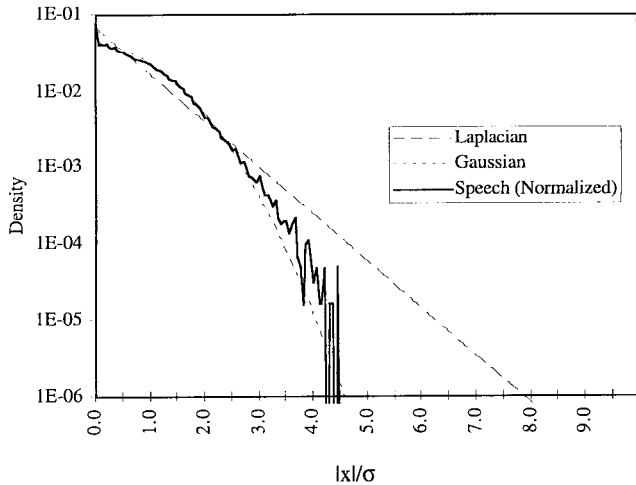


Fig. 7. Statistical distribution of normalized subband samples.

allocated (i.e., which subband should get the first bit allocated, which one should get the second, and so on), until all bits have been allocated. By prioritizing each allocated bit and using embedded quantization techniques, the encoded bitstream is assembled in a bit interleaved manner such that the bits are arranged in a perceptually prioritized manner. The bitstream then can simply be truncated in order to obtain a lower bit rate representation.

The IMT leads to a simple bit allocation and prioritization algorithm. The quantization noise is inversely related to the number of bits allocated to each subband. The signal-to-quantization noise ratio (SQNR) is approximately given by

$$\text{SQNR}_j = 6.02 \text{ dB} * n_j \quad (1)$$

where n_j is the number of bits per subband sample used to quantize subband j . This approximation holds fairly well, since the subband samples are scaled such that the energy in each frame is normalized to one. The statistical distribution of the normalized subband samples resembles a Gaussian distribution, as shown in Fig. 7, and the rate distortion limit for quantizing a memoryless Gaussian source is given by (1).

If sufficient bits are available, then each subband is allocated the number of bits required for the SQNR to exceed the SMR given by the JND. Thus, the bit-allocation algorithm to achieve perceptual transparency is

$$n_j > \frac{\text{SMR}}{6.02}. \quad (2)$$

If the number of bits available is not sufficient, then the bit allocation is based on the achievable IMT. Thus, the bit allocation is

$$n_j > \frac{\alpha \text{SMR}}{6.02} \quad (3)$$

where α is the IMT scale such that $\text{IMT} = \alpha \text{SMR}$. If $\alpha = 1$, then the JND is met and the quantization noise is not perceptible. For $0 < \alpha < 1$, the quantization noise is perceptible, but is *shaped* to optimize the IMT.

Since the number of bits allocated to each subband is proportional to the ideal bit allocation (the number of bits required to achieve the JND), we denote this algorithm the *proportional* bit allocation.

TABLE I

DYNAMIC BIT ALLOCATION AND PRIORITIZATION EXAMPLE. THE SMR DETERMINES THE NUMBER OF BITS, n_j , ALLOCATED TO QUANTIZE SAMPLES IN EACH SUBBAND. AT EACH STEP N , A BIT IS ALLOCATED, RESULTING IN THE BIT ALLOCATION AND PRIORITIZATION TO THE SUBBANDS IN THE FOLLOWING ORDER: 1 3 4 2 6 1 1 3 4 1 (10 B/SUBBAND SAMPLE TOTAL)

subband	1	2	3	4	5	6	7	8
SMR (dB)	22.0	4.0	12.0	8.0	0.0	4.0	0.0	0.0
n_j	4	1	2	2	0	1	0	0
N=1	1	0	0	0	0	0	0	0
N=2	1	0	1	0	0	0	0	0
N=3	1	0	1	1	0	0	0	0
N=4	1	1	1	1	0	0	0	0
N=5	1	1	1	1	0	1	0	0
N=6	2	1	1	1	0	1	0	0
N=7	3	1	1	1	0	1	0	0
N=8	3	1	2	1	0	1	0	0
N=9	3	1	2	2	0	1	0	0
N=10	4	1	2	2	0	1	0	0

The proportional bit allocation algorithm also implies a prioritization of each bit allocated. The bit prioritization allows the bitstream to be assembled in a prioritized manner, such that the reduction in bit rate can be achieved through truncation of the bitstream and the truncated bitstream still maximizes the IMT for the lower bit rate.

The implementation of bit prioritization can be simplified by using the ideal bit allocation directly instead of the IMT to compute the actual bit allocation. Thus, the IMT need not be calculated explicitly. Table I illustrates an example of the bit allocation and prioritization computation for a frame. First, the JND is computed by the perceptual model for the frame, giving the required SMR. The ideal bit allocation n_j is then determined from the SMR. The bits are prioritized one at a time based on maximizing the proportion of the assigned bits versus the ideal. If subbands have already had an equal proportion of bits assigned, priority is given to subbands with higher ideal bit allocations, then to lower frequency subbands. The bit allocation at each step N corresponds to the effective bit allocation if the bitstream is truncated at that point. The bit prioritization follows from the allocation at each step.

D. Quantization

In quantizing the subband samples, we attempt to minimize the quantization noise power for a given bit allocation. The quantization is done in 20 ms frames for each subband. The 20 subband samples in each frame are quantized in two stages, first by quantizing a gain or scale for the entire frame, then by quantizing each scaled subband sample. Nonuniform scalar quantizers are used for each stage. The subband frame energies are first computed for each of the eight subbands, and quantized with a logarithmic quantizer. The root mean square (RMS) value is then used to scale the subband samples. The statistical distribution of the scaled subband samples closely matches a normalized Gaussian distribution. The scaled subband samples are then quantized with an embedded nonuniform scalar quantizer matched to the Gaussian distribution.

The scale and bit allocation are transmitted as overhead, and the embedded quantization indices of the subband samples are

TABLE II
5-B GAUSSIAN EMBEDDED NONUNIFORM QUANTIZER
THRESHOLDS (T_k) AND RECONSTRUCTION VALUES (R_k)
AT 1-5 B/SAMPLE. ONLY POSITIVE VALUES ARE SHOWN

index	T_k	$R_k(5)$	$R_k(4)$	$R_k(3)$	$R_k(2)$	$R_k(1)$
00000	0.0000	0.0660	0.1318	0.2618	0.5082	0.7979
00001	0.1322	0.1983				
00010	0.2651	0.3318	0.3983			
00011	0.3995	0.4672				
00100	0.5364	0.6056	0.6745	0.8096		
00101	0.6768	0.7479				
00110	0.8217	0.8954	0.9687			
00111	0.9726	1.0497				
01000	1.1313	1.2128	1.2935	1.4467	1.6312	
01001	1.3001	1.3874				
01010	1.4824	1.5773	1.6704			
01011	1.6828	1.7883				
01100	1.9091	2.0298	2.1443	2.2929		
01101	2.1743	2.3188				
01110	2.5050	2.6912	2.8271			
01111	2.9764	3.2616				

transmitted according to the bit allocation and prioritization algorithm.

This scheme is based on the assumption that the subbands have sufficiently narrow bandwidth to decorrelate the subband samples. This assumption neglects the fact that speech and audio signals are nonstationary and that significant intraband correlation may exist. It is common, however, to assume a memoryless source model for subband coders [13].

Source statistics were obtained by analyzing male and female speech segments from the DARPA TIMIT Acoustic Phonetic Continuous Speech Corpus (NTIS PB91-505 065). The probability distribution for these source signals were more peaked (high probability density near zero and very slow tail decay) than Laplacian and Gaussian distributions. Even when the silence portions of speech are removed (by removing segments where the energy falls below a minimum threshold), the distribution of the samples is still highly peaked. The highly peaked distribution results in very poor quantization performance when a simple quantization algorithm such as uniform quantization is used. However, if the samples over a short frame are normalized (i.e., scaled by the energy), then the distribution of the normalized samples is not as peaked and more closely resembles a Gaussian distribution, as shown in Fig. 7. Since the normalized subband samples are nearly Gaussian, the SQNR closely follows the 6.02 dB/bit assumption in the bit allocation and prioritization scheme.

An embedded scalar quantization approach was chosen based on its low complexity in implementation. The principle behind embedded quantizers is that even a partial index should provide sufficient information to determine a suitable reconstruction value. This is a desirable property if the encoded pattern is not necessarily received in its entirety at the decoder. However, an embedded quantizer incurs a performance penalty relative to the optimal quantizer at any one rate. Some distortion tradeoff at the different effective bit rates must occur in the embedded quantizer design. The quantizer performance can be made optimal at one of the target bit rates, but the embeddability constraint implies that a suboptimal choice must be made at other rates. However, the performance penalty

can be minimized through proper design of the embedded quantizer, as will be shown.

The embedded scalar quantizer is fully described by the tree structure of the quantization indices and the quantization regions or thresholds at the highest bit rate. The indexing scheme must be tree structured such that the branch at each level is described by part of the index. A partial index allows decoding to a partial level of the tree. At the lowest level of the tree, the quantization regions are nonoverlapping and their union consists of the entire input range. The quantization region corresponding to each branch is the union of the quantization regions of its daughter branches. The optimal reconstruction value is the centroid of the quantization region.

Embedding a uniform quantizer incurs a very severe performance penalty. The reconstruction values are constrained to be the midpoint of the quantization region instead of the centroid. This is not a problem if the probability density is relatively uniform over the quantization region, but this is especially not true for the outlying regions, the tail of the distributions. This results in reconstruction values that are too far from the center of the distribution.

A second reason that embedded scalar quantizers are far from optimal is the difference in optimal scale required at each bit rate. Embedding the uniform scalar quantizer forces a single scale to be used for all bit rates. If the scale is too large, then the quantization regions are larger than ideal, and the SNR degrades because of loss of granularity. If the scale is too small, then the quantization regions are smaller than ideal, and the SNR degrades because of clipping or overloading. Since the effect of clipping is more drastic than loss of granularity, the scale should be optimized for the highest bit rate. However, using the optimal high rate scale results in poor SNR at the lower bit rates. The net result is that an embedded uniform scalar quantizer will have poor performance at low bit rates.

The optimal nonuniform embedded scalar quantizer does not suffer from the scaling problems at various bit rates. The optimal nonuniform scalar quantizer, the Lloyd-Max quantizer [24], [25], is approximated by an optimal compander, given by Bennett's companding function, followed by a uniform quantizer [26]. An embedded nonuniform quantizer is also represented by a companding function and an embedded uniform quantizer. The embedding constraint requires only that the same companding function be used at all bit rates and that the uniform quantizer have a single scale. Thus, the quantizer given by Bennett's companding function is also the optimal nonuniform embedded quantizer, and is a close approximation of the optimal nonembedded nonuniform scalar quantizer, the Lloyd-Max quantizer [27].

The relationship between Bennett's companding function and optimal embedded quantization is of significant importance. Embedded uniform scalar quantizers will always result in significant performance degradation for nonuniformly distributed sources. *Ad hoc* schemes for nonuniform embedded scalar quantizers, such as those proposed in [13], allow performance to be optimized at a single rate, but are suboptimal at other rates.

The design procedure for an optimal embedded nonuniform scalar quantizer is to use the Lloyd-Max quantizer at the

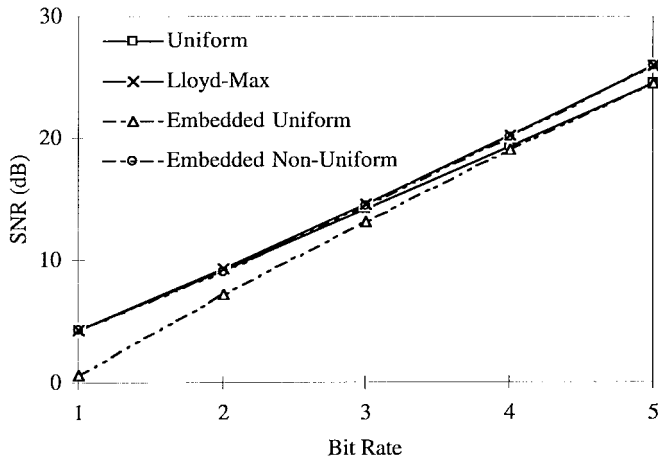


Fig. 8. Optimal embedded quantizer SNR versus uniform and Lloyd-Max quantizers for quantization of Gaussian sources.

TABLE III
OPTIMAL EMBEDDED QUANTIZER SNR VERSUS UNIFORM AND
LLOYD-MAX QUANTIZERS FOR QUANTIZATION OF GAUSSIAN SOURCES

Bit Rate	Quantizer SNR (dB)			
	Uniform	Lloyd-Max	Embedded Uniform	Embedded Non-Uniform
1	4.40	4.40	0.65	4.40
2	9.25	9.30	7.19	9.14
3	14.27	14.62	13.16	14.51
4	19.38	20.22	19.02	20.20
5	24.57	26.01	24.57	26.01
source	Max[24]	Jain[28]	Tang[27]	Tang[27]

full bit rate and to use the embeddability constraint to define the reconstruction levels at lower rates as the centroids of the quantization regions. The optimal 5-b embedded quantizer for normalized Gaussian distribution quantizer is defined by the quantization thresholds and reconstruction values as shown in Table II. Comparison to nonembedded and embedded uniform quantizers are summarized in Fig. 8 and Table III. As expected, the embedded quantizer performance approximately matches Lloyd-Max quantizers at all bit rates. The performance is substantially better than an embedded uniform quantizer, and the performance over the entire range of possible rates does not degrade significantly, allowing higher performance and more robust system design.

An additional issue considered in the quantizer design is robustness to source statistics. Robustness is the property of the quantizer being able to perform adequately if the statistics are different than those assumed. Most commonly, this is observed if the variance of the source does not match that assumed in the design of the quantizer. A robust quantizer implementation makes the error performance relatively insensitive to nonoptimal scaling. A nonuniform scalar quantizer offers significant improvement in robustness compared to uniform scalar quantizers for Gaussian sources [26], [27].

Embedded quantization with bit prioritization allows an embedded bitstream to be assembled as described in the following section. The bit prioritization ensures that the truncated bitstream's effective bit allocation is the same as would have been generated at the lower bit rate. The embedded quantizer

design ensures that the quantizer performance with a truncated bitstream is also approximately the same as if a nonembedded quantizer had been used at the same bit rate. Thus, the design methodology ensures that only a slight performance penalty has been paid in generating a scalable, embedded output.

E. Bitstream Formatting

Bitstream formatting arranges the coded bitstream for each 20-ms frame by perceptual significance. The overhead consisting of the subband bit allocation and scaling is the most important information in each frame. The bit allocation information consists of 24 b, three per subband. The scaling consists of 36 b, four for scaling the frame and four for scaling each of the eight subbands.

The subband samples are quantized with the embedded quantizer, as described in Section II-D. The bits in the indices for the quantized subband samples are interleaved according to the perceptual priority as determined by the bit allocation and prioritization algorithm. Thus if the bitstream is truncated, the transmitted indices constitute the most perceptually significant bits. The embedded quantizer then allows the subband samples to be reconstructed from the lower rate embedded index.

The subband sample quantization indices are grouped into 20 blocks (one per subband sample), and arranged in an interleaved manner based on the bit allocation and prioritization scheme. The variable-length bitstream frame structure is shown in Fig. 9. The first 60 bits correspond to the overhead bits, the next 20 bits are the most significant bits (MSB's) of the subband samples from the subband allocated the first bit, the next twenty bits are from the subband allocated the second bit, and so on, until all the bits are incorporated.

Each frame can be truncated as required to meet the bit-rate requirement. Since the bits have been prioritized, truncation effectively corresponds to the bit allocation of the lower bit rate. This allows the bit rate to be modified very easily by the system, since no additional interaction with the coder is required to specify the desired bit rate. The coder can thus be operated in several configurations as desired by the system, such as fixed bit rate, variable bit rate, or network controlled source-rate configurations.

In a fixed bit-rate configuration, the coder output is truncated to a particular rate in every frame, resulting in a fixed frame size. The bit allocation corresponds to a dynamic-frequency, fixed-time allocation minimizing the IMT on a frame-by-frame basis. This is the configuration in which most speech coders typically operate, minimizing a quality cost measure over each frame while constrained to a fixed use of system bandwidth.

In a variable bit-rate configuration, the coder operates in a perceptually transparent mode, producing varying frame sizes such that the number of bits is the minimum required to mask all the quantization noise in each frame. Alternatively, a fixed proportion of the bitstream is truncated, such that the coder operates in a variable-rate, fixed-frame quality mode. This is similar to variable-rate configurations that maintain a fixed quality while minimizing the average use of system bandwidth, but with the added capability of adjusting the quality as desired.

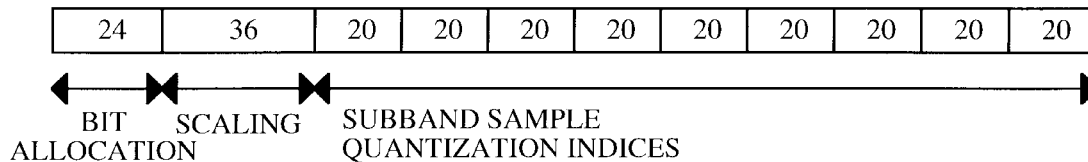


Fig. 9. Bitstream frame structure.

In a network-controlled source-rate configuration, the source rate is selected according to the status of the network. The source-rate adaptation provides the network with the capability to effectively deal with the transmission environment, which may be limited by interference, fading, congestion, or lack of resources. The coder provides an estimate of the signal quality, allowing the system to optimize the source and channel rate, transmission power, and bandwidth in a perceptually significant manner.

In all these configurations, no overhead is required to describe each frame, since the frame size does not need to be explicitly known. Since the bitstream is embedded, truncation of frames due to any rate-control algorithm affects only the frame size. The decoder treats each frame as a truncated frame and no information other than knowing the end of the received frame is required.

F. Algorithm Complexity and Delay

The algorithmic complexity is dominated by the QMF and perceptual model computations. The QMF analysis and synthesis filterbanks require 150 k operations/s each, and the perceptual model requires an additional 125 k operations/s. The total speech coder algorithm requires approximately 325 k operations/s for the encoder and 200 k operations for the decoder.

The algorithmic delay of 25 ms is due to buffering of one frame (20 ms) and the delay of the QMF (5 ms). This does not include the possible need to interleave frames to provide robustness to transmission, nor implementation delays.

III. PERFORMANCE EVALUATION

Listening tests using nine subjects were conducted. The signal sources consisted of speech, speech with additive background road noise, and music segments. Training sets were used to familiarize the subjects with the types and degrees of coding distortion in the listening tests. Signals were presented monaurally to simulate audition with a telephone handset or portable radio unit, using the Ariel ProPort model 656 16-b D/A and calibrated TDH-49 headphones in a double-walled sound chamber.

The speech sources consisted of four different phonetically balanced sentences, using three male and three female speakers, from the TIMIT database. The sentences were approximately 4 s in length, played three separate times in random sequence. The listening experiments were run with sentences coded without additive noise, and sentences coded with additive background road noise. The road noise, provided by Qualcomm Inc., was recorded in an automobile traveling at highway speeds with the windows rolled up. The sentences were presented at 85-dB SPL and the additive road noise measured at 79-dB SPL. We also conducted the listening

TABLE IV
LISTENING TEST RESULTS

Average MOS	Speech	Speech + Road Noise	Music
Subband (24 kbps)	3.7	3.6	3.5
Subband (16 kbps)	3.0	3.3	3.1
Subband (12 kbps)	2.5	3.0	2.5
G.728 LD-CELP (16 kbps)	3.7	3.2	3.0
GSM RPE-LTP (13 kbps)	3.6	2.9	2.3
G.726 ADPCM (16 kbps)	2.2	2.3	1.8

MOS Variance	Speech	Speech + Road Noise	Music
Subband (24 kbps)	0.6	0.4	0.4
Subband (16 kbps)	0.4	0.3	0.3
Subband (12 kbps)	0.2	0.3	0.3
G.728 LD-CELP (16 kbps)	0.5	0.4	0.4
GSM RPE-LTP (13 kbps)	0.6	0.5	0.4
G.726 ADPCM (16 kbps)	0.4	0.3	0.2

experiments with nonspeech sources, using two 4-s segments of baroque music from *Water Music* by Handel and *Autumn Concerto* by Vivaldi.

The embedded subband coder operating at fixed rates of 12, 16, and 24 kbps was compared to the G.728 LD-CELP (16 kbps), GSM RPE-LTP (13 kbps), and G.726 ADPCM (16 kbps) coders. The G.728 coder is the best performing of the comparison coders, and is generally considered to be “near-toll” quality. A five-point MOS scale was used, with the scale representing 1) bad, 2) poor, 3) fair, 4) good, and 5) excellent.

The listening test results are shown in Table IV. Without additive background noise, the subband coder at 24 kbps was comparable to the G.728 LD-CELP coder. At 16 kbps, the quality of the subband coder was inferior to both the G.728 and GSM coders. At 12 kbps, the quality was considerably poorer, but still superior to G.726 at 16 kbps.

However, the robustness of the subband coder is evident in that for additive background noise conditions and music sources, the performance relative to the other coders improved significantly. At the noise levels used in the experiment, the subband coder at 16 kbps was comparable in performance to G.728, and at 12 kbps it was as good as the GSM coder. For music, the coder at 16 kbps was comparable to G.728, and at 12 kbps it was comparable to the GSM coder.

The performance of the embedded subband coder is attractive in that, while it does not provide exceptionally good quality at low bit rates, it provides scalable quality at higher bit rates. In addition, the robustness it provides causes less severe degradation such that the effective quality is comparable to LD-CELP-type coders under adverse conditions. Finally, the advantage that a prioritized bitstream provides can be used by the system to effectively utilize a higher bit rate under the same channel throughput constraints, allowing higher quality transmission to be achieved.

IV. CONCLUSION AND FUTURE WORK

A highly flexible scheme for speech coding has been developed based on perceptually based subband coding and embedded bit prioritization and quantization. This scheme is well suited for high-quality speech and audio transmission over wireless communication channels, allowing the system to seamlessly adapt to changes in both the transmission environment and network congestion.

The embedded coder output allows the bit rate to be modified without interaction with the source coder. Systems can benefit from adapting the source rate, channel rate, transmission power, and transmission bandwidth to deal effectively with transmission degradations due to interference, fading, congestion, or lack of resources. Furthermore, the design of the bit prioritization and embedded quantization ensures a scalable design, with little performance penalty relative to a nonembedded design.

Although embedded subband coding and its system implementation benefits have been presented in the past, we introduce the following novel concepts:

- reduced complexity and delay through the use of an IIR QMF filterbank;
- improved perceptual bit allocation and prioritization with the interpolated mask threshold cost measure;
- reduced complexity of the perceptual model computation through the use of subband spectral analysis;
- optimal embedded nonuniform scalar quantization;
- improved system robustness and flexibility using scalable, embedded source coding.

The performance of the coder was verified through subjective listening tests. Results indicate that although the performance at bit rates below 16 kbps is inferior to G.728 LD-CELP and GSM RPE-LTP under nominal conditions, the performance is comparable under adverse conditions, such as additive background noise and nonspeech sources, and is scalable to higher quality at higher bit rates, thus providing superior quality for bit rates above 16 kbps.

Future work will include design of channel coding schemes to allow the benefits of embedded source coding to be incorporated into an adaptive communication system. In particular, the use of embedded channel coding along with embedded source coding will allow source and channel rate adaptation to be adaptively matched to the transmission channel, without requiring interaction with the source and channel coders.

ACKNOWLEDGMENT

The authors thank Dr. Kleijn and the reviewers for their useful comments and suggestions.

REFERENCES

- [1] N. Jayant, "Signal compression: Technology targets and research directions," *IEEE J. Select. Topics Commun.*, vol. 10, pp. 796–818, June 1992.
- [2] N. Jayant, J. Johnston, and R. Safranek, "Signal compression based on models of human perception," *Proc. IEEE*, vol. 81, pp. 1385–1421, Oct. 1993.
- [3] P. Noll, "Wideband speech and audio coding," *IEEE Commun.*, pp. 34–44, Nov. 1993.
- [4] A. Gersho, "Advances in speech and audio compression," *Proc. IEEE*, vol. 82, pp. 900–918, June 1994.

- [5] J. J. Dubnowski and R. Crochiere, "Variable rate coding of speech," *Bell Syst. Tech. J.*, pp. 577–600, Mar. 1979.
- [6] D. Goodman and C. Sundberg, "Combined source and channel coding for variable bit rate transmission," *Bell Syst. Tech. J.*, pp. 2017–2036, Sept. 1983.
- [7] A. Haoui and D. Messerschmitt, "Embedded coding of speech: A vector quantization approach," in *Proc. Int. Conf. Acoustics, Speech, and Signal Processing*, 1985, pp. 1703–1706.
- [8] I. Wassell, D. Goodman, and R. Steele, "Embedded delta modulation," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 36, pp. 1236–1243, Aug. 1988.
- [9] V. Karanam, K. Sriram, and D. Bowker, "Performance evaluation of variable bit rate voice in packet switched networks," *AT&T Tech. J.*, pp. 57–69, Oct. 1988.
- [10] M. Sherif, D. Bowker, G. Bertocci, B. Orford, and G. Mariano, "Overview of CCITT embedded ADPCM algorithms," in *Proc. Int. Conf. Acoustics, Speech, and Signal Processing*, 1990, pp. 1014–1018.
- [11] R. De Iacovo and D. Sereno, "Embedded CELP coding for variable rate between 6.4 and 9.6 Kbit/s," in *Proc. Int. Conf. Acoustics, Speech, and Signal Processing*, 1991, pp. 681–684.
- [12] S. Zhang, "An embedded scheme for regular pulse excited (RPE) linear predictive coding," in *Proc. Int. Conf. Acoustics, Speech, and Signal Processing*, 1995, pp. 37–40.
- [13] R. Cox *et al.*: "New directions in subband coding," *IEEE J. Select. Areas Commun.*, vol. 6, pp. 391–409, Feb. 1988.
- [14] R. Cox, J. Hagenauer, N. Seshadri, and C. Sundberg, "Subband speech coding and matched convolutional coding for mobile radio channels," *IEEE Trans. Signal Processing*, vol. 39, pp. 1717–1731, Aug. 1991.
- [15] K. Gould, R. Cox, N. Jayant, and M. Melchner, "Robust speech coding for the indoor wireless channel," *AT&T Tech. J.*, pp. 64–73, July 1993.
- [16] J. Johnston, "Transform coding of audio signals using perceptual noise criteria," *IEEE J. Select. Areas Commun.*, vol. 6, pp. 314–323, Feb. 1988.
- [17] K. Brandenburg and G. Stoll, "ISO-MPEG-1 audio: A generic standard for coding of high-quality digital audio," *J. Audio Eng. Soc.*, vol. 42, pp. 780–792, Oct. 1994.
- [18] J. Modestino and D. Daut, "Combined source-channel coding of images," *IEEE Trans. Commun.*, vol. COM-27, pp. 1644–1659, Nov. 1979.
- [19] J. Hagenauer, "Rate-compatible punctured convolutional codes (RCPC Codes) and their applications," *IEEE Trans. Commun.*, vol. 36, pp. 389–400, Apr. 1988.
- [20] Z. Jiang, A. Alwan, and A. N. Willson, Jr., "High-performance IIR QMF banks for speech subband coding," in *Proc. IEEE Int. Symp. Circuits and Systems*, June 1994, pp. 493–496.
- [21] A. Shen, "Perceptually-based subband coding of speech signals," Master's thesis, Dept. Elect. Eng., Univ. Calif., Los Angeles, CA, June 1994.
- [22] A. Shen, B. Tang, A. Alwan, and G. Pottie, "A robust variable rate speech coder," in *IEEE ICASSP-95*, Detroit, MI, pp. 249–252.
- [23] B. Tang, A. Shen, G. Pottie, and A. Alwan, "Spectral analysis of subband filtered signals," in *IEEE ICASSP-95*, Detroit, MI, pp. 1324–1327.
- [24] J. Max, "Quantizing for minimum distortion," *IRE Trans. Inform. Theory*, vol. IT-6, pp. 7–12, Mar. 1960.
- [25] S. Lloyd, "Least squares quantization in PCM," *IEEE Trans. Inform. Theory*, vol. IT-28, pp. 129–137, Mar. 1982.
- [26] A. Gersho, "Principles of quantization," *IEEE Trans. Circuits Syst.*, vol. CAS-25, pp. 427–436, July 1978.
- [27] B. Tang and G. Pottie, "Embedded nonuniform scalar quantizer for variable rate applications," in *ICSPAT-94*, Dallas, TX, pp. 456–460.
- [28] A. Jain, *Digital Image Processing*. Englewood Cliffs, NJ: Prentice-Hall, 1989.



Benjamin Tang (SM'94–M'96) was born in Sao Paulo, Brazil, in 1964. He received the B.S., M.S., and Ph.D. degrees in electrical engineering from the University of California, Los Angeles, in 1985, 1986, and 1995, respectively.

Since 1985, he has been with TRW Inc., Redondo Beach, CA, where he is currently a Department Staff with the Digital Products Center developing analog and digital electronics for satellite communication applications.

Dr. Tang is a member of Eta Kappa Nu and Tau Beta Pi.



Albert Shen (SM'92-M'94) received the B.S. degree in electrical engineering from the California Polytechnic University, Pomona, in 1992, and an M.S. degree in electrical engineering from the University of California, Los Angeles, in 1994.

He has worked as a Software Engineer in the Advanced Storage and Retrieval division of IBM, San Jose, CA, and as an evaluator of wireless speech compression algorithms at Qualcomm, Inc., in San Diego, CA. He is currently a DSP Software Developer with the Intel Architecture Laboratory,

Hillsboro, OR. He has published papers on robust compression of speech signals and subband spectral analysis. His current fields of interest include spatial audio and low bit-rate speech and audio coding.

Mr. Shen is a member of the Acoustical Society of America, Eta Kappa Nu, and Tau Beta Pi.



Gregory Pottie (S'85-M'88) was born in Wilmington, DE. He received the B.Sc. degree in engineering physics from Queen's University, Kingston, ON, Canada, in 1984, and the M.Eng. and Ph.D. degrees from McMaster University, Hamilton, ON, Canada, in 1985 and 1988, respectively.

From 1989 until 1991, he worked in the Transmission Research Department of Codex/Motorola, Mansfield, MA, with projects including high-speed digital subscriber lines and coding and equalization schemes for voice-band modems. He is presently an

Associate Professor in the Department of Electrical Engineering, University of California, Los Angeles. His research interests include channel coding, and systems design for personal communications systems and wireless distributed sensor networks.



Abeer Alwan (SM'89-M'92) received the Ph.D. in electrical engineering from the Massachusetts Institute of Technology, Cambridge, in 1992.

Since then, she has been an Assistant Professor in the Department of Electrical Engineering, University of California, Los Angeles, where she established the Speech Processing and Auditory Perception Laboratory. Her research interests include modeling speech production and perception mechanisms, and applying these models in speech coding, synthesis, and recognition systems.

Dr. Alwan is the recipient of the NSF Research Initiation Award (1993), the NSF Career Development Award (1995), the NIH FIRST Career Development Award (1994), and the UCLA-TRW Excellence in Teaching Award (1994). She is a member of Eta Kappa Nu, Sigma Xi, Tau Beta Pi, and the New York Academy of Sciences. She is a member of the Acoustical Society of America Technical Committee on Speech Communication, and the IEEE Signal Processing-Audio and Electroacoustics Technical Committee.