Scalar LPC Quantization Based on Formant JND's

ODED GHITZA AND JULIUS L. GOLDSTEIN, SENIOR MEMBER, IEEE

Abstract—**Efficient scalar quantization tables for LPC *k*-parameters were developed using a distortion measure based on just-noticeable-differences (JND's) in formant parameters of the speech spectrum envelope. Forty percent fewer bits were required than the 41/frame used in conventional approaches. An empirical technique was developed for relating perturbations in *k*-parameters and formant parameters. New estimates were obtained for the values of the formant JND's: they are about four times the steady-state values reported by Flanagan [6] and increase sharply above approximately 1.5 kHz.**

I. Introduction

SCALAR LPC vocoder technology recently has reached maturity after a decade of intensive development [24]. It is based upon an elegant discrete-signal formulation of the classical model for speech signals, comprising a glottallike excitation source driving a vocal tractlike filter [8], [1], [27]. To achieve the objective of data rate compression with minimal loss of speech quality, the parameters of the LPC model, estimated for each speech segment (10–30 ms), are quantized in some manner suited to perceptual tolerances. The methods extensively implemented seek to quantize the reflection coefficients of the autoregressive vocal-tract filter to obtain some measure of short-term spectral envelope distortion that is uniform throughout the coefficient range [25], [15], [19]. These techniques do not attempt to make full use of knowledge of speech sound perception. The present study was undertaken to examine how systematic application of psychoacoustical knowledge could improve scalar LPC quantization design.

Our approach builds upon Flanagan's [2]–[6] research to integrate speech perception knowledge into formant vocoder design. He combined Peterson and Barney's [20] vowel data on formant center frequencies with his measurements of just-noticeable-differences (JND's) of formant parameters for synthetic vowels. We represent Flanagan's conception of perceptually tolerable speech filter distortion in Fig. 1. Flanagan measured tolerances for the formant center frequency, bandwidth, and intensity, using sustained synthetic vowels (JNDF = 3–5 percent, JNDB = 20–40 percent, JNDI = 1–3 dB). The perceptibility of changes in intervalley intensity (JNDV = 10 dB) was far less than for formant intensity. Standard LPC quantization techniques make qualitative use of the greater perceptual sensitivity to spectral peaks than to valleys [17].

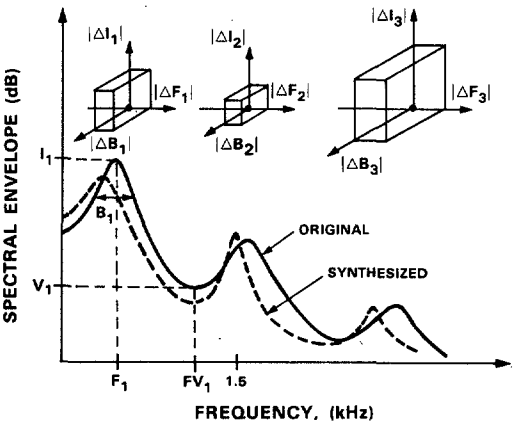


Fig. 1. Illustration of tolerable distortions of the speech spectral envelope as bounded by just-noticeable-differences (JND's) in formant center frequencies, bandwidths, and intensities. We assume the tolerances are most accurate around 1.5 kHz. Changes in interformant valley frequencies and intensities are perceptually far more tolerable than changes in the formant (spectral peak) parameters.

It is reasonable to adopt vowel tolerances for the design of quantization rules for all speech sounds, as the vowel tolerances are most stringent (e.g., Tremain [24]). However, knowledge of perception gained since Flanagan's work suggests significant modification of his JND tolerances. Klatt [16] showed that temporal dynamics of natural speech cause changes in the fundamental frequency to be less perceptible (JNDfo = 1.7 percent) than for unnatural stationary speech (JNDfo = 0.25 percent). Comparison to Flanagan and Saslow's [9] measurements of JNDfo (= 0.3 − 0.5 percent) for sustained speech sounds suggests that Klatt's finding may also apply to the formant JND's.

Indications that the formant JND's vary with frequency, with the third formant being least sensitive, are evident in vocoder development (e.g., Flanagan [7]). A more detailed guide to the unknown frequency dependence of formant JND's may be found in pitch theory studies. Goldstein [13] found from the psychophysics of fundamental pitch for complex-tone stimuli that precision in aural measurement of component stimulus frequencies depends upon frequency, as shown in Fig. 2. Experimen-

Manuscript received May 29, 1985; revised November 19, 1985. This paper is based on a Ph.D. dissertation submitted by O. Ghitza to the Tel-Aviv University, Tel-Aviv, Israel, June 1983, was presented in part at the 103rd Meeting of the Acoustical Society of America, Chicago, IL, 1982, and appeared in part in *Hearing—Psychophysical and Physiological Bases*, Klinke and Hartmann, Eds. (Berlin: Springer-Verlag, 1983).

O. Ghitza was with the Department of Electronic Communications, Control and Computer Systems, Tel-Aviv University, Ramat-Aviv, Tel-Aviv 69978, Israel. He is now with the Department of Acoustics Research, AT&T Bell Laboratories, Murray Hill, NJ 07974.

J. L. Goldstein is with the Department of Electronic Communications, Control and Computer Systems, Tel-Aviv University, Ramat-Aviv, Tel-Aviv 69978, Israel.

IEEE Log Number 8608139.

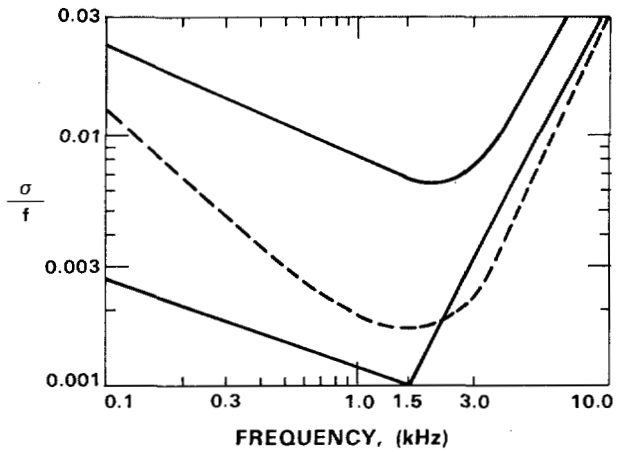


Fig. 2. Auditory precision in measuring component frequencies of complex tones: psychophysics of periodicity pitch (solid), computed from auditory-nerve spike interval statistics (dashed), linear approximation used in this study to describe the frequency dependence of the formant parameter JND's. The JND's at 1.5 kHz were taken as some multiple M of Flanagan's [6] JND's (JNDF = 3 percent, JNDB = 20 percent, JNDI = 1 dB, JNDV = 10 dB). (Two upper curves adopted from Goldstein [14].)

tal and theoretical studies of perceptual implications of auditory-nerve physiology [14], [28], [21], [23] suggest that a similar frequency dependence might be expected in auditory processing of speech signals.

In the study reported here, we examined the contribution of the suggested modifications of Flanagan's formant JND's to efficient quantization of LPC filter parameters. Our technique was to generate a family of quantization tables for the LPC reflection coefficients on the basis of various assumed values of the formant JND's (Section II), and then to choose the appropriate quantization by psychophysical experimentation (Section III). Following our suggested modifications of Flanagan's formant JND's, the JND's were assumed to be frequency dependent following the piecewise linear curve in Fig. 2. As a basis for comparison, quantization tables were also generated assuming frequency-independent JND's. In both cases, the actual values of the formant JND's were taken as multiples of the minimum values reported by Flanagan [6]. Examination in Sections IV–VI of the properties of the appropriate quantization table found in Section III supports our claim that more efficient scalar LPC quantization is obtained from systematic application of psychoacoustical knowledge.

II. Derivation of the Perceptually Based k-Parameter Quantization

In this section, we develop the computational procedures for relating JND-based spectral envelope tolerances to LPC parameter tolerances. We treat the allowable perturbation Δk_i of each LPC parameter as though it were a single-valued function of its original value k_i, and construct quantization tables for each presumed JND-based spectral envelope tolerance. These quantization tables are then used to generate synthetic speech in which all ten coefficients are simultaneously quantized. The choice of an appropriate quantization table is made in Section III on the basis of psychophysical evaluation of the quality of the quantized speech. Finally, in Section VI the appropriate formant JND's are ascertained by examining the actual formant perturbations produced by perceptually acceptable simultaneous quantization of the LPC parameters.

A. The JND Distortion Measure

Previous distortion measures suggested for speech spectral envelope encoding apply continuous cost functions on the spectral error (i.e., squared error). We used a *binary* cost function instead. Given the short-term spectral envelope of a speech frame calculated from the full precision LPC parameters, we define the class of spectra having *all* their formant parameters within the JND's of the original sound as being equivalent. Thus, inaccuracy in the LPC model parameters due to quantization is permitted until one of the deviated formant parameters crosses its JND boundary, each parameter being equally weighted. Fig. 1 illustrates the proposed measure. The JND's are different for each formant, and the decision is an OR decision: if any one of the three (for this particular example) difference vectors emerges from its box, the penalty is 1, otherwise it is 0.

A more general geometric formulation can be given.

Definition 1: Let NF be the number of formants in a given spectrum. Let $N = 4 * NF$. We define the *N-dimensional formant space* to be the space spanned by the $3 * NF$ formant parameters (its center frequency, bandwidth, and intensity) and the first NF interformant valleys.

Definition 2: Let NF and N be as in definition 1. Let JNDF_i, JNDB_i, and JNDI_i, be the suitable JND's of the ith formant. Let JNDV_i be the JND of the intensity of the ith valley. The *JND box* is the N-dimensional box, its size in the F_i dimension, $i = 1, \cdots, NF$, is $2 *$ JNDF_i; identical definitions hold for the JND box size in the B_i, I_i, and V_i dimensions.

Now let us look at the N-dimensional formant space, defined by a typical vowel spectrum. The full precision spectrum fixes a point S^o in this space, around which the appropriate JND box is constructed. In order to be equivalent, the point representing the quantized spectrum, S^q, should be inside the JND box. If this is the case, the penalty is 0; otherwise, it should be 1. The dimension of the formant space varies from one frame to another, according to the number of formants existing in the speech spectrum considered.

Our objective is to find the allowable perturbations of the LPC parameters that produce perceptually equivalent spectra. Clearly, we need to specify first the appropriate values of the formant parameter JND's. To every formant parameter a suitable JND curve is attached according to the piecewise linear curve in Fig. 2. Their behavior upon frequency is identical, the difference is in the value of the minimum point, JNDmin.

Definition 3: The JNDM set of curves is a set of four JND curves (JND*F*, JND*B*, JND*I*, JND*V*), their corresponding JNDmin points are the minimum values found by Flanagan (JND*F*min = 3 percent, JND*B*min = 20 percent, JND*I*min = 1 dB, JND*V*min = 10 dB), multiplied by *M*.

The value of *M* is to be found in Section III.

B. Single-Parameter Perturbation Analysis

For a given speech frame, there is a deterministic relation between the LPC parameter vector $\boldsymbol{k}$ and the spectral envelope which it represents. Recursive procedures for this relation are well known, and recently, a direct but complicated expression for the relation between the cepstral coefficients and $\boldsymbol{k}$ was proposed [22].

Because of the AR model structure, it is evident that the effect upon the spectral envelope caused by a perturbation of one of the components of $\boldsymbol{k}$ is dependent upon the values of the other components of $\boldsymbol{k}$. However, a multiparameter perturbation analysis would be unrealistic even in computer simulations because of the enormous number of computations it would require. Despite the interaction, which *a priori* is expected to be considerable, in the following we ignore it and calculate the statistics of the maximum perturbation of a single coefficient k_i for a presumed tolerable JND-based spectral envelope perturbation. In Section VI we examine the effect of the interparameter interactions due to simultaneous quantization of the coefficients.

The perturbation analysis procedure is described by the flowchart in Fig. 3. It was performed only on the voiced frames of the database. The signal analysis conditions are described at the end of this section. The procedure begins with the estimation of the reflection coefficients for the current frame. The resultant $\boldsymbol{k}$ vector, comprising P ordered reflection coefficients, is then processed through the two branches of the flowchart. In the left branch, the full precision LPC spectrum envelope is computed from $\boldsymbol{k}$ to be the basis for the extraction of the precise values of the formant parameter vectors $\boldsymbol{F}^o$, $\boldsymbol{B}^o$, $\boldsymbol{I}^o$, $\boldsymbol{FV}^o$, and $\boldsymbol{V}^o$ (formant frequencies, bandwidths, intensities, valley frequencies, and valley intensities, as in Fig. 1). Note that the superscripts have been added to the symbols of Fig. 1 to distinguish between the original and the perturbed values. The formant and valley frequencies $\boldsymbol{F}^o$ and $\boldsymbol{FV}^o$ are used to determine the appropriate ***JNDF***, ***JNDI***, ***JNDB***, and ***JNDV*** using the JND*M* set of curves, from which the JND box is constructed. The JND box is located in the formant space, around the full precision vector $\boldsymbol{S}^o$, defined by $\boldsymbol{F}^o$, $\boldsymbol{I}^o$, $\boldsymbol{B}^o$, $\boldsymbol{FV}^o$, and $\boldsymbol{V}^o$. Both $\boldsymbol{S}^o$ and the JND box are fixed in this position until the perturbation analysis of each of the reflection coefficients of the current frame is completed.

In the right branch, the perturbed spectrum envelope is created in the following way: being in the *i*th step of the procedure (i.e., examining the *i*th reflection coefficient), the perturbation coefficient is set to $k_i^* = k_i + \Delta$, while other coefficients remain at their original full precision values, $k_j^* = k_j$, $j \neq i$. The resulting perturbed vector $\boldsymbol{k}^*$ is passed through the same operators as in the full precision branch, to get $\boldsymbol{S}^*$, defined by $\boldsymbol{F}^*$, $\boldsymbol{I}^*$, $\boldsymbol{B}^*$, $\boldsymbol{FV}^*$, and $\boldsymbol{V}^*$.

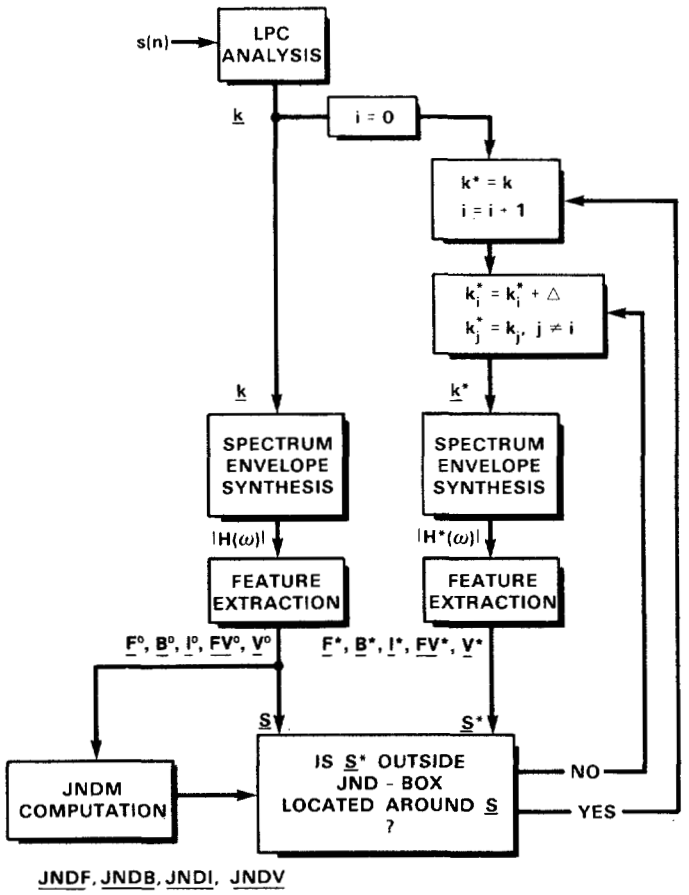


Fig. 3. Flowchart for the single-parameter perturbation analysis, to measure the maximum deviation in each LPC reflection coefficient that meets the assumed formant JND tolerances. A perturbation in each direction was separately examined.

Now the JND-distortion measure is applied in the formant space, checking whether the perturbed vector $\boldsymbol{S}^*$ is inside the JND box located around the full precision vector $\boldsymbol{S}^o$. If this is the case, k_i^* is further modified, while all the others remain constant (i.e., $k_i^* = k_i^* + \Delta$ and $k_j^* = k_k$, $j \neq i$), and the new $\boldsymbol{S}^*$ is again examined. The allowable deviation of k_i is defined as $dk_i = k_{io}^* - k_i$ where k_{io}^* is the value of k_i^*, which first causes $\boldsymbol{S}^*$ to exit the JND box.

After dk_i is found, the perturbed vector $\boldsymbol{k}^*$ is returned to its full precision value, $\boldsymbol{k}^* = \boldsymbol{k}$, and the next reflection coefficient k_{i+1} is examined, with the JND box and its location still unchanged. After analyzing the last component k_P, the next voiced frame is analyzed, yielding a new nonquantized $\boldsymbol{k}$ vector, on which the same analysis is performed.

Natural speech consisting of five speech segments, each 20 s long, spoken by three males and two females, provided the database. High-quality FM radio broadcasts of news programs in Hebrew were digitized at a 10 kHz sample rate, using a 12 bit A–D converter preceded by a fourth-order low-pass Butterworth filter set at 4 kHz.

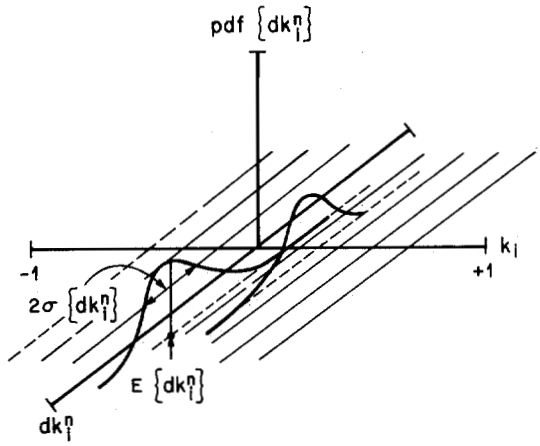


Fig. 4. Illustration of the statistics of the maximum allowable k-parameter deviations dk_i. The range of each parameter k_i is divided into N bins, and the measured allowable deviations in each bin are represented by their mean and standard deviation.

A tenth-order double-precision linear prediction with the autocorrelation method (e.g., [18]) was performed on a Hamming window weighted, 25 ms nonoverlapped, preemphasized frame. The preemphasis coefficient was set to be 0.9375. Since this study applied to voiced speech sounds, the analysis was carried out only on the voiced frames which were selected by a voiced–unvoiced decision made on energy observations only.

During the perturbation analysis, an adaptive step size searching algorithm was used, ending with a minimum step size of $\Delta = 0.005$. A perturbation in each direction was separately examined.

C. Perceptually Allowed k-Parameter Deviation Functions

In each cycle of the analysis, the allowable parameter perturbations dk_i, $i = 1, \cdots, P$, are determined for each voiced frame in the database. The measured perturbations as a function of the unperturbed value are organized in histograms, as illustrated in Fig. 4. The range of each parameter k_i is divided into N bins ($N = 40$), and the measured parameter perturbations are represented statistically by the mean and standard deviations for each bin, $E\{dk_i^n\}$ and $\sigma\{dk_i^n\}$ as given by

$$E\{dk_i^n\} = (1/L_n) \sum_{j=1}^{Ln} dk_{ij}^n \tag{1a}$$

$$\sigma^2\{dk_i^n\} = \text{var}\,\{dk_i^n\} = \left[(1/L_n) \sum_{j=1}^{Ln} (dk_{ij}^n)^2\right] - E^2\{dk_i^n\} \tag{1b}$$

where $n = 1, \cdots, N$, $i = 1, \cdots, P$, N is the number of bins in $(-1, +1)$, L_n is the number of occurrences of k_i in the nth bin, and dk_{ij}^n is the allowed deviation for the jth occurrence. It is assumed that the database is sufficiently large to cover most of the interrelations between the components of $\boldsymbol{k}$, so that $E\{dk_i^n\}$ and var $\{dk_i^n\}$ are accurate enough.

Next, for each coefficient k_i we smoothed the "measurements" $E\{dk_i^n\}$ by a polynomial function, using a *weighted* minimum-mean-square-error curve fitting. The polynomial order was selected to be Q, $Q = 1, 2, 3$, or 4, depending on the measurements. The weighting function was the pdf of the considered k_i, that is, $L_n/\Sigma_{n=1}^{N} L_n$ for the nth bin. A similar procedure was performed on the measured standard deviation $\sigma\{dk_i^n\}$, $n = 1, \cdots, N$.

Examples of the smoothed polynomial function $E\{dk_i\}$, together with its corresponding standard deviation zone (which is determined by the smoothed polynomial function var $\{dk_i\}$) are shown in Fig. 5(a)–(c), for the first three reflection coefficients, based upon our scaled formant perturbation tolerance JND4. Results for the remaining seven coefficients (with JND4) are given in Appendix A. Each plot of Fig. 5[(a)–(c)] contains:

1) the source distribution histogram pdfk_i;

2) the perceptually allowed "measured" average deviations $E\{dk_i^n\}$; and

3) the perceptually allowed smoothed deviation curve $E\{dk_i\}$ with its corresponding standard deviation zone.

The endpoints for each coefficient were determined by a range within which 95 percent of the area under the curve of pdfk_i is concentrated. It should be noted that each pdfk_i was measured on a large database containing five 1.25 min speech segments.

Several observations are noteworthy.

1) Each reflection coefficient has its unique perceptually allowed deviation curve. The shape as well as the amount of the deviation are different from one curve to another.

2) For a given k_i, a similar curve is obtained when the male speakers and the female speakers are evaluated separately.

3) The same curve is valid for a perturbation in both directions.

4) The standard deviation is relatively small, indicating that the allowable deviation is tightly related to its unperturbed value.

5) For different scaled values of the formant JND's, the curves for each k-parameter are nearly multiples of one another, increasing with the increasing JND scaling. Thus, the unique shape of each k-parameter is preserved over the range of the JND scalings.

D. Parameter Quantization Laws

The construction of the k-parameter quantization law from the perceptually allowed deviation function $E\{|dk_i|\}$ is as follows. Let us denote $E\{|dk_i|\} = g(k)$. At each point k_o, $g(k_o)$ is the amount of the perceptually allowed deviation in k around k_o, and $1/(g(k_o))$ gives the perceptually recommended relative density of the quantization levels for k_o. One may represent this nonuniform density function as a uniform density $1/(\nu(p))$ of a transformed parameter p. That is, it is desired to find a mapping ϕ: $k \to p$ such that the perceptually allowed deviation of p will be a constant, say, $\nu(p) = 1$. Let $p_i = \phi(k_i)$ and $pj = \phi(k_j)$. Then, for every k_i and k_j, the number of quantization lev-

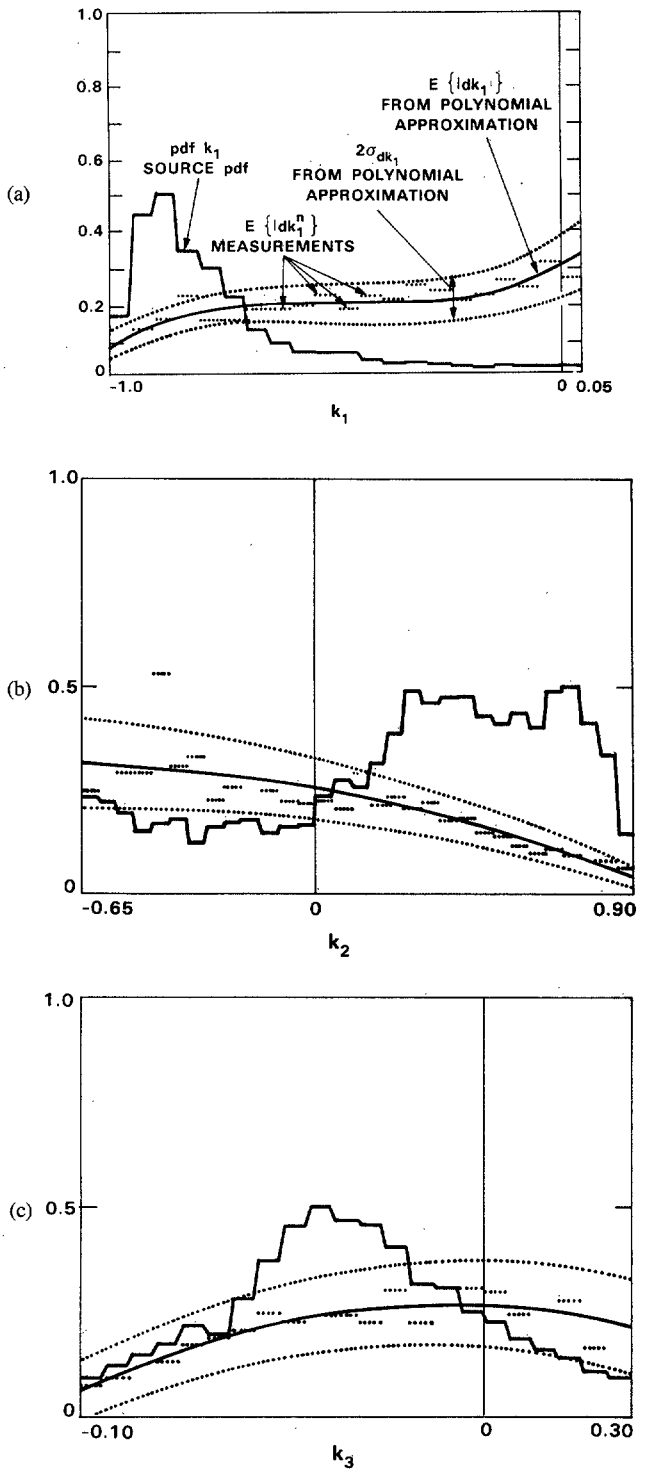


Fig. 5. Perceptually allowed k-parameter deviation functions, for k_1, k_2, and k_3, based on the scaled frequency-dependent formant tolerances JND4 (the minimum formant JND's at 1.5 kHz in Fig. 2 are four times Flanagan's values).

els between k_i and k_j is

$$\int_{k_i}^{k_j} \frac{dk}{g(k)} = \int_{p_i}^{p_j} \frac{dp}{\nu(p)} \tag{2}$$

if $\nu(p) = 1$, and if a domain-adjustment constraint is imposed on the mapping ϕ such that $k_{\min}$ is mapped onto $p = 0$, then from (2)

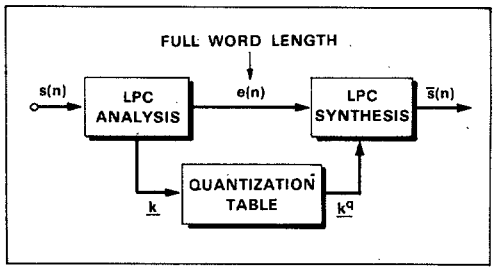


Fig. 6. Residual excited LPC-10 back-to-back coder: a special version for discrimination experiments.

$$\int_{k_{\min}}^{k} \frac{dx}{g(x)} = \int_{0}^{p} dp = p(k). \tag{3}$$

Equation (3) defines the desired mapping ϕ. Since the perceptually allowed deviation of p is 1, the perceptually recommended distance between every two successive quantization boundaries p_i, p_j is 1. That is, the quantization boundaries of p are $p^b = 0, 1, 2, \cdots, [p_{\max}]$ where $[p_{\max}]$ denotes the closest integer to $p_{\max}$, from above. In order to reduce the average quantization error, the corresponding quantization levels should be $p^l = 0.5, 1.5, 2.5, \cdots, [p_{\max}] - 0.5$.

In order to find the quantization law for k, the inverse mapping ϕ^{-1} (i.e., k as a function of p) has to be found. Recalling that $g(\cdot)$ is a polynomial function (see Section II-C), for polynomial order less than three an explicit solution for ϕ^{-1} exists; for higher orders, an iterative solution of (3) should be performed. Given the inverse mapping, the nonuniform quantization boundaries of k are the inverse values of the p^b's, while the quantization levels are the k's corresponding to the p^l's. The k-parameter quantization laws constructed in this manner are given in Appendix B for a formant tolerance scaling JND4.

III. Psychoacoustical Experiments I

A. Methods

As a first step in the psychoacoustical experiments, we create a library of speech segments containing an original speech segment followed by synthesized speech segments, each of which is a member of a different equivalent class of signals representing the original. Each of these classes is defined by the quantization laws for a different JND scaling of the formant perturbation tolerances. The technique we used to produce each synthesized segment was a special version of a back-to-back residual excited LPC-10 coder, shown in Fig. 6. In this version, the full precision (16 bit per sample) residual signal excites the LPC synthesis filter, defined by the appropriately quantized reflection coefficients.

The algorithm for the linear prediction yields both the current reflection coefficients vector $\boldsymbol{k}$ and its matched residual signal $e(n)$ (see Section II-B for further details on LPC procedures). Feeding the full precision filter (defined by the full precision vector $\boldsymbol{k}$), with the full precision $e(n)$, yields a synthesized output which is the original input frame, except for a negligible truncation error due to the

finite word length of the processor. The synthesized speech segments in our library were produced by feeding this full precision residual signal into filters with the same structure, but with the quantized reflection coefficient vector $\boldsymbol{k}^q$, according to the desired quantization tables. Two points are noteworthy.

1) Since $e(n)$ is the full precision inverse filter response to the original speech, it contains minimum information about the spectral envelope. This causes the synthesized speech spectral envelope to be affected mainly by the quantization error $\boldsymbol{k} - \boldsymbol{k}^q$.

2) Only the JND's for voiced frames were examined since the psychophysical data apply to such sounds. Larger JND's are expected for unvoiced sounds. Thus, no processing was performed on the unvoiced frames and each unvoiced frame was represented by its original waveform.

The database for the psychophysical experiments comprises natural speech consisting of four 6 s long speech segments spoken by two females and two males. Different speakers than those of the database in Section II-B were used. Special care was given to ensure that in each of the speech segments all the voiced phonemes would occur with equal probability. The analysis conditions in creating the database are as described at the end of Section II-B.

The speech material was presented to the subjects in two intervals, each of 6 s. The two intervals are the original segment, ORG, followed by one of the JND segments, which was randomly selected. The subject's task was to scale the quality of the second interval; the lowest mark was 1, the highest was 9. The ORG speech segment was also a candidate for the second interval, to provide a control on the reliability of the ratings given by each listener.

Each speech segment in the database was scaled by the subjects in a 140-trial experiment to provide data on the mean and its standard deviation. Four subjects participated in the experiments, two females and two males. No prelearning of the library segments was conducted.

B. Psychophysical Results I

The results of our basic psychophysical experiments are summarized in Fig. 7. Only the unshaded points (for ORG and formant tolerance scalings JNDM, $M = 3, 4, 5, 6$) are relevant to the task considered in this section, namely, the selection of the appropriate formant JND's for natural speech. We shall refer to the shaded points (for JND7flat) later, in Section V-A. As the subjects were not given prelearning sessions, the first 10 trials were not included in the average and standard deviation calculations.

A similar trend of the quality ratings is observed for all listeners and speakers. The average scores of ORG are slightly better than JND3 and JND4. However, there is a considerable overlapping of the distributions. An abrupt deterioration in perceived quality at JND5 points to the adequacy of JND4 scaling for preserving the original quality.

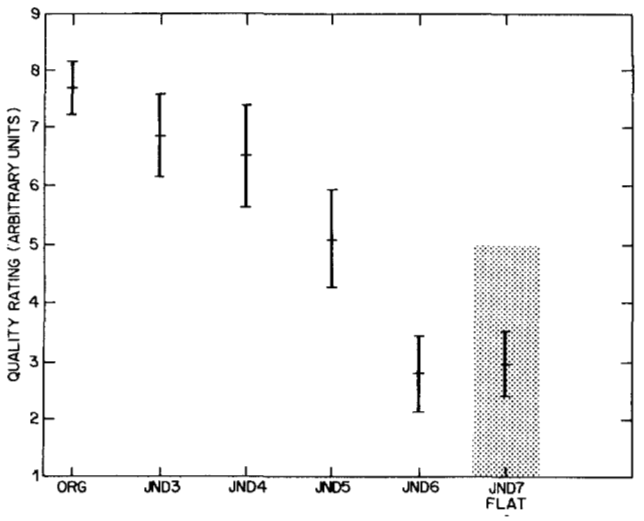


Fig. 7. Measurement of the appropriate quantization law for natural speech. Each entry is based on intersubject averages of quality ratings and standard deviations by four listeners to four 6 s speech segments.

It is important to emphasize that a stringent criterion was used by the subjects in their judgments. Since the subject was asked to rate the quality of a 6 s long speech segment, repetition of the examined segments in a 140-trial experiment focuses the subject's attention upon the finest details of the stimulus. Therefore, this experiment should be categorized as a *discrimination* experiment. However, this is not the situation in a natural speech conversation; indeed, all the listeners were unable to discriminate between ORG, JND3, JND4, and JND5 in the first 10–20 sessions. Thus, the scorings of Fig. 7 apply to a stringent psychophysical discrimination task, that is, the measurement of the JND's for natural speech sounds.

IV. Bit Allocation for JND4 Quantization

The significance of our psychophysical result (in Section III-B)—that good speech quality is preserved by JND4 k-parameter quantization—is that far fewer quantization levels are required than are used in conventional scalar quantization designs. Recall from Section II-D that the number of quantization levels for each k-parameter is given directly by the deviation function $p_i(k)$.

Table I shows the maximum values of $p(k)$, $p_{\max} = p(k_{\max})$, for each k_i. Rounding $p_{\max}$ upward to the nearest integer (shown in parentheses in Table I) gives the number of different levels that are distinguished for each k-parameter. The number of bits required to represent these levels is shown in the rightmost column of Table I. The total number of bits is 23.58. Thus, all 10 k-parameters could be specified with a binary word of 24 bits by grouping the coefficients as shown by the implementation in Table I. In contrast, 41 bits are used to encode the k-parameters of each frame in conventional LPC quantization [24].

Questions arise in judging whether the greater efficiency of the JND4 quantization implies that these quantization laws make better use of properties of perception than conventional approaches. It is possible that conven-

TABLE I
BIT ALLOCATION FOR JND4 QUANTIZATION

k	k_{min}	k_{max}	p_{max}	NO. OF BITS	IMPLEMENTATION
1	-0.995	-0.045	6.64 (7)	2.73 (2.82)	3
2	-0.645	+0.895	9.42 (10)	3.24 (3.32)	5
3	-0.795	+0.295	5.81 (6)	2.54 (2.58)	
4	-0.445	+0.595	4.26 (5)	2.09 (2.32)	5
5	-0.595	+0.395	4.68 (5)	2.23 (2.32)	7
6	-0.395	+0.495	3.66 (4)	1.87 (2.00)	2
7	-0.445	+0.445	4.24 (5)	2.08 (2.32)	
8	-0.295	+0.595	4.51 (5)	2.17 (2.32)	
9	-0.445	+0.395	3.70 (4)	1.89 (2.00)	2
10	-0.295	+0.395	2.97 (3)	1.57 (1.58)	
			TOTAL:	22.41 (23.58)	24

Complete quantization tables are in Appendix B.

tional quantization is conservative, having stopped far short of the minimum number of quantization levels that could be derived from conventional design logic. Therefore, we must compare a conventional quantizer, which has been well designed, using 24 bits with a JND4 quantizer. This is done in Section V.

A second question that arises concerns the role in the quantization laws played by the frequency dependence we imposed upon the format JND's. While it is reasonable that greater tolerances at higher formant frequencies could add to the efficiency of the quantization laws, what evidence have we that such greater tolerances exist? A complete treatment of this question is outside the scope of this report. However, it is demonstrated in the next section that the presumed greater tolerances at high formant frequencies contribute significantly to the JND4 quantization efficiency.

V. PSYCHOPHYSICAL EXPERIMENTS II

The objective of the experiments reported in this section is to evaluate the speech quality generated by back-to-back LPC synthesis in which the k-parameters for the resynthesized speech are quantized with a similar number of levels as in JND4 quantization, but having a different distribution of levels. The same methods are employed as described earlier in Section III-A.

A. LAR and JNDFlat Quantization

Conventional quantization will be represented by a log area ratio (LAR) quantizer. This method assumes that the form of the allowed k-parameter deviation function is [25]

$$g_i(k) = (1 - k^2)/C_i, \qquad -1 < k < 1. \tag{4}$$

Then the function that specifies the required number of quantization levels is

$$p_i(k) = \frac{C_i}{2}\left[\ln\frac{1+k}{1-k} - \ln\frac{1+k_{imin}}{1-k_{imin}}\right], \qquad k_{imin} < k < k_{imax}. \tag{5}$$

To construct a 24 bit quantization law, it is only necessary to find the value of the constants C_i that satisfies

$$\log_2 \prod_{i=1}^{10} p_i(k_{imax}) = 24. \tag{6}$$

Following Viswanathan *et al.* [26], we used the unequal step size scheme (C_i is a function of i) with the same range of k_i as in the JND quantization (Section II-D and Appendix B). We label this quantization law as LAR24.

The second quantization law we wish to evaluate is based upon the same perturbation analysis procedure described in Sections II-B and -C, with the single modification that the formant parameter tolerances are assumed to be insensitive to the formant frequency. Thus, Flanagan's formant JND's were scaled uniformly for all formants, and quantization tables were computed as before. Multiplication of Flanagan's formant JND's by 7 gave a quantization table with a similar number of levels as the JND4 quantization; specifically, 24 bits were required. We label this quantization law as JND7flat.

B. Psychophysical Results II

The quality judgments for the LAR24 quantization are summarized in Fig. 8. Recall (Section III-A) that the speech material was presented to the subjects in two intervals of 6 s each. The original speech segment was presented first, followed by a random selection of the LAR24 quantized version, the JND4 quantized version, or the unmodified original. The subjects' quality ratings on the second interval clearly indicate that LAR24 was inferior to JND4. Specifically, with reference to Fig. 7, it was judged as intermediate to the quality of JND5 and JND6.

The quality judgments for JND7Flat quantization are included in Fig. 7. These judgments were actually made during the same experimental sessions for which the JNDM data were collected; that is, during the second presentation interval of the basic experiment, JND7Flat was one of the possible speech signals. Of course, knowledge that 24 bits of JNDM quantization preserved good quality was obtained in preliminary experiments. Clearly, the

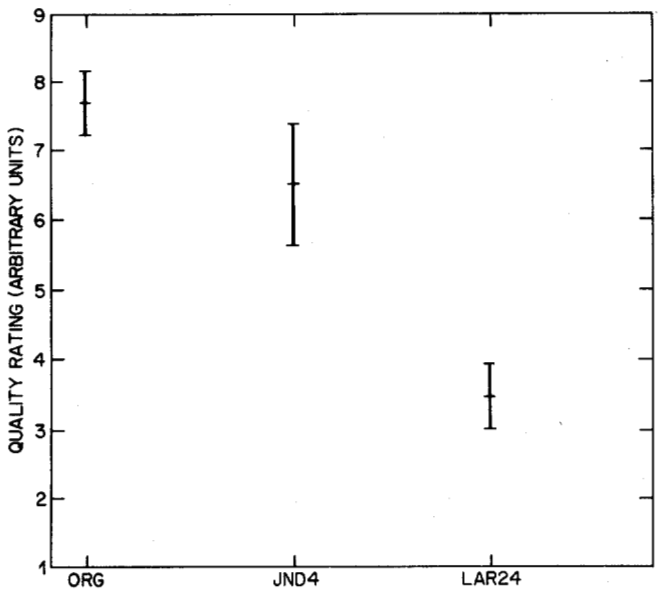


Fig. 8. Comparison of the JND4 and LAR24 quantized speech segments to the original. Each entry is based on intersubject averages of quality ratings and standard deviations by four listeners to four 6 s speech segments.

quality of JND7Flat was also judged inferior to that of JND4; specifically, it was similar in quality rating to JND6.

VI. Formant Parameter Perturbations Due to Simultaneous Quantization of LPC Parameters

The empirical computational and psychophysical procedures we used to determine that JND4 is an efficient quantization law were designed to answer two questions. 1) What are the formant JND's for dynamic speech? 2) What is the physical relation between k-parameter and formant perturbations? Yet, although we now have estimated perceptually acceptable tolerances for the k-parameters, we are not able to answer either of those questions. This is so because the perceptually allowed k-parameter deviation functions (Section II-C) relate formant parameter perturbations to individual k-parameter perturbations. The quantization tables were, of course, applied simultaneously to quantize all 10 k-parameters. Therefore, we cannot infer the formant JND's from the perceptual acceptability of JND4 quantization.

To estimate the formant JND's from our earlier results, the actual formant parameter perturbations produced by JND4 quantization were systematically measured. For this purpose, the procedure described in Fig. 9 was applied on a database consisting of four 1.25 min long speech segments spoken by the same speakers as in Section III-A. Each voiced frame in the database was LPC analyzed, yielding the reflection coefficient vector $\boldsymbol{k}$ to reproduce the simultaneously quantized $\boldsymbol{k}^q$. Formant parameters were then extracted from both spectrum envelopes evaluated from $\boldsymbol{k}$ and $\boldsymbol{k}^q$, providing the full precision and quantized formant parameters $\boldsymbol{F}$, $\boldsymbol{B}$, $\boldsymbol{I}$, $\boldsymbol{FV}$, and $\boldsymbol{V}$ and $\boldsymbol{F}^q$, $\boldsymbol{B}^q$, $\boldsymbol{I}^q$, $\boldsymbol{FV}^q$, and $\boldsymbol{V}^q$, respectively (see Fig. 11), from which the particular deviation in every dimension was

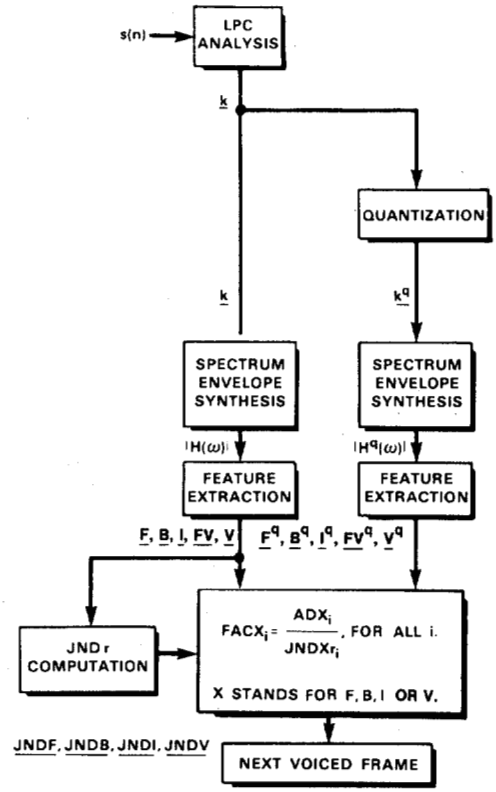


Fig. 9. Computation flowchart to measure the probability distributions of the actual formant parameter deviations for JND4 quantized speech normalized by the appropriate formant JNDr's (FACX_i, see text, Section VI).

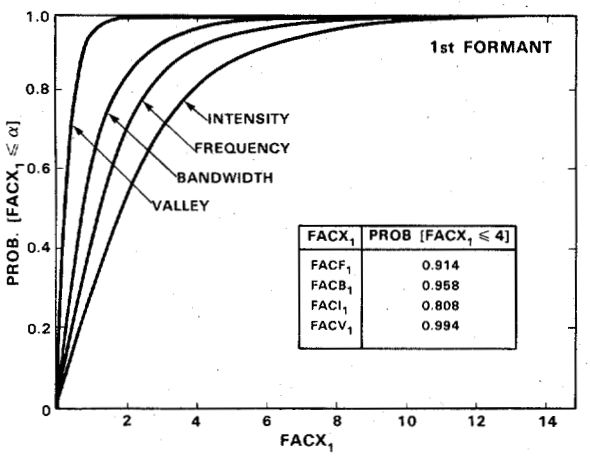
calculated. In parallel, from $\boldsymbol{F}$ and $\boldsymbol{FV}$, the *reference* JND's were computed using the frequency dependence shown in Fig. 2 with JNDFmin = 3 percent, JNDBmin = 20 percent, JNDImin = 1 dB, and JNDVmin = 5 dB. Note that, except for an inconsequential change in the last value, the minimum reference JND's are Flanagan's steady-state values. Finally, we calculate the absolute ratio factor (FACX_i in Fig. 9) between the actual deviation (ADX_i in Fig. 9) and the frequency-dependent reference JND's (JNDXr_i in Fig. 9) for every formant dimension ("X" in Fig. 9) and for each formant ("i" in Fig. 9).

Probability distribution functions were computed [Fig. 10(a)–(c)] for the first three formants. Only ratio factors less than 15 were included to eliminate possible influence of incorrect formant feature extraction (which occurred in less than 5 percent of the examined frames). It is clear that the actual deviation is only slightly above four times the reference JND's, especially in the formant intensity parameters.

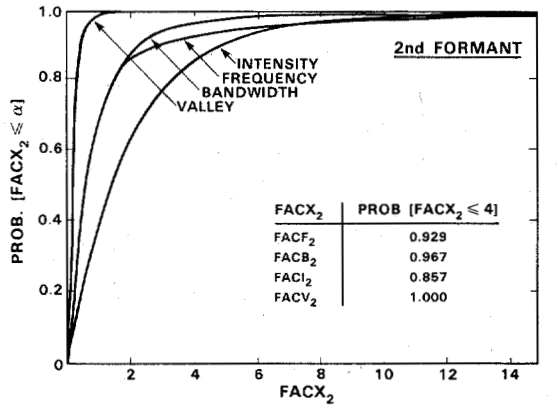
Thus, we now have approximate, although very useful answers, to the two questions posed earlier in this section: JND4 k-parameter quantization produces formant parameter perturbations of 4–5 times the reference JND's from which the quantization law was derived; and the formant JND's for dynamic speech are at least as large as 4 times the frequency-dependent reference JND's.

VII. Summary

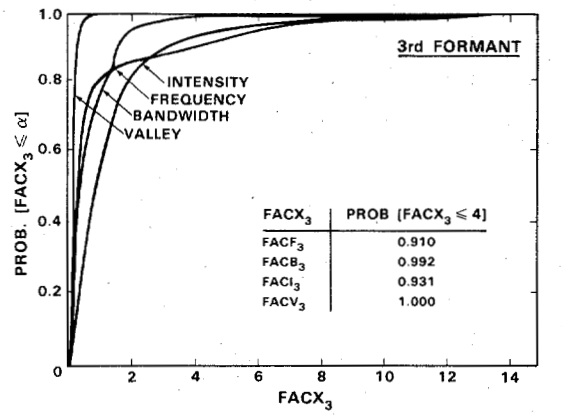
Systematic application of psychoacoustical knowledge yielded 40 percent more efficient scalar LPC quantization



(a)



(b)



(c)

Fig. 10. (a) Measured probability distributions of the normalized formant parameter deviations for the first formant of JND4 quantized speech. (b) As in (a) for the second formant. (c) As in (a) for the third formant.

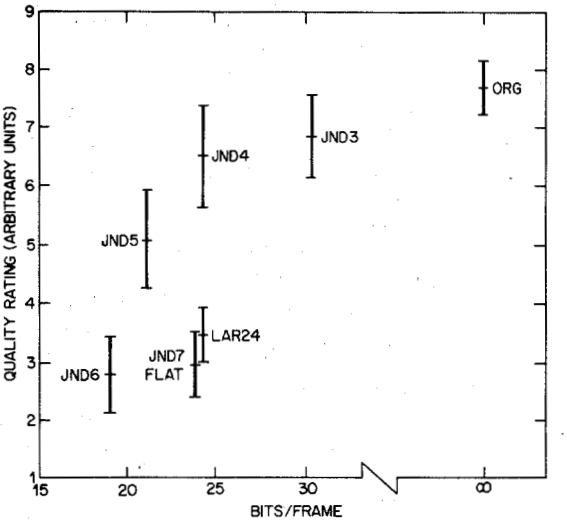


Fig. 11. Comparison of efficiencies of different scalar LPC quantization tables examined in this study. The entries are from Figs. 7 and 8.

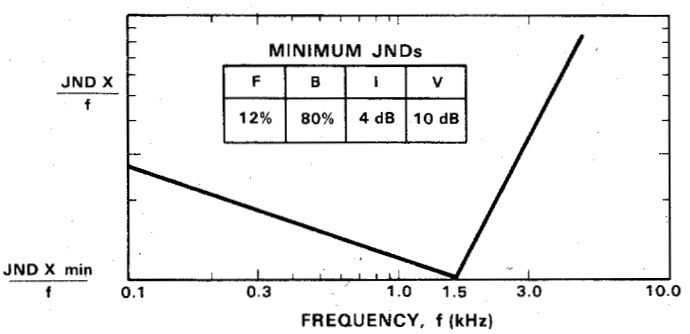


Fig. 12. Estimated values of the formant parameter JND's as a function of frequency. X represents the formant parameters: center frequency F, bandwidth B, peak power I, and valley power V.

than the 41 bits per frame of standard approaches. The psychoacoustical knowledge we applied comprised the perceptual tolerances to perturbations (i.e., just-noticeable-differences) in the formant parameters of the short-term spectral envelope of each speech frame, as first proposed by Flanagan for formant vocoder design. We measured new estimates for the values of the formant parameter JND's of about four times the steady-state measurements originally reported by Flanagan; and in addition, these values increase sharply above approximately 1.5 kHz. A new empirical technique, based on measured speech statistics, was developed for relating perturbations of the LPC k-parameters and formant parameters.

A graphical summary comparing the efficiencies of the different quantization laws is given in Fig. 11, while Fig. 12 summarizes our new measurements of the formant parameter JND's for the short-term spectral envelope of a speech frame. The latter data are more fundamental for speech system design, while the former data represent a particular speech system realization supporting the general thesis that speech system design benefits from systematic application of psychoacoustical knowledge.

Appendix A
k-Parameter Deviation Functions

The perceptually allowed k-parameter deviation functions, plotted in Figs. 5 and 13–19, are defined and discussed in Section II-C.

Appendix B
JND Quantization Tables

This Appendix contains the quantization tables obtained by applying the analytic procedure, discussed in

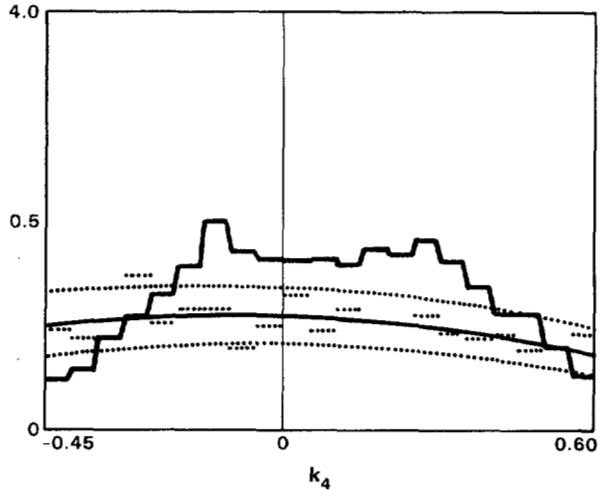


Fig. 13.

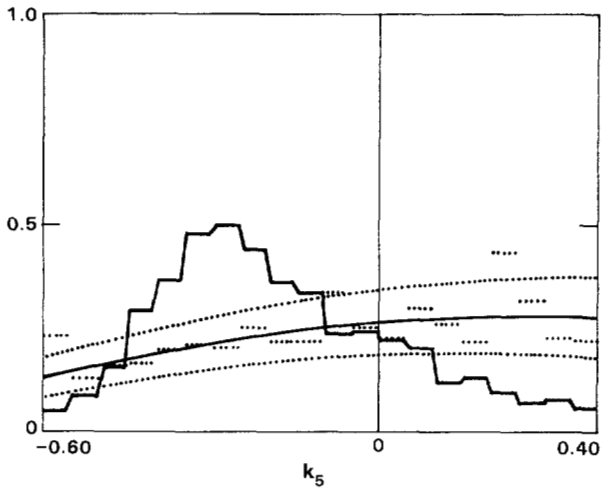


Fig. 14.

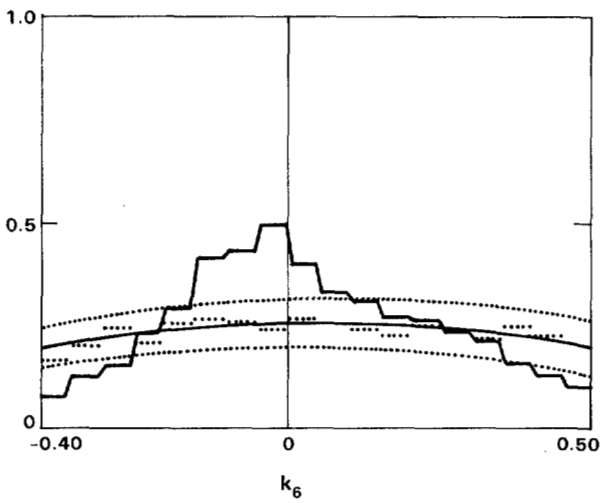


Fig. 15.

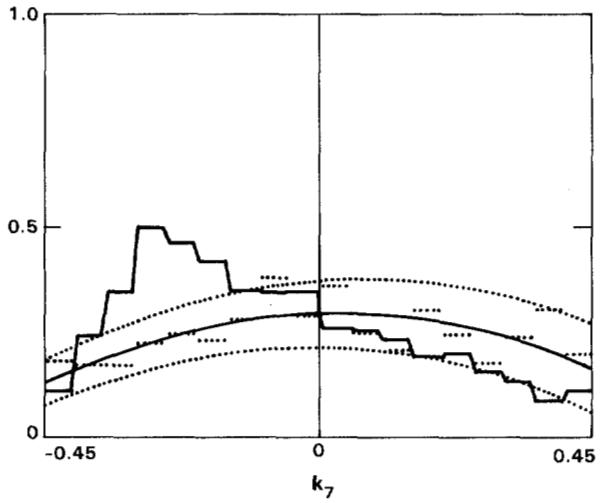


Fig. 16.

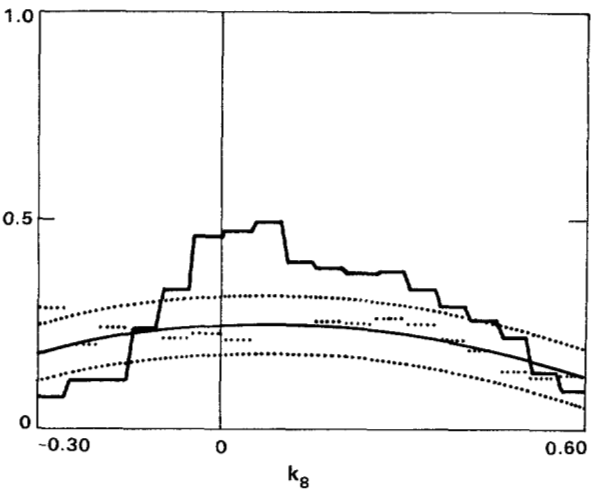


Fig. 17.

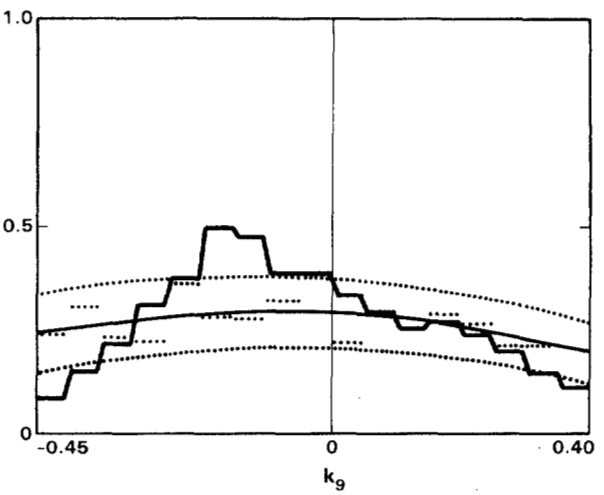


Fig. 18.

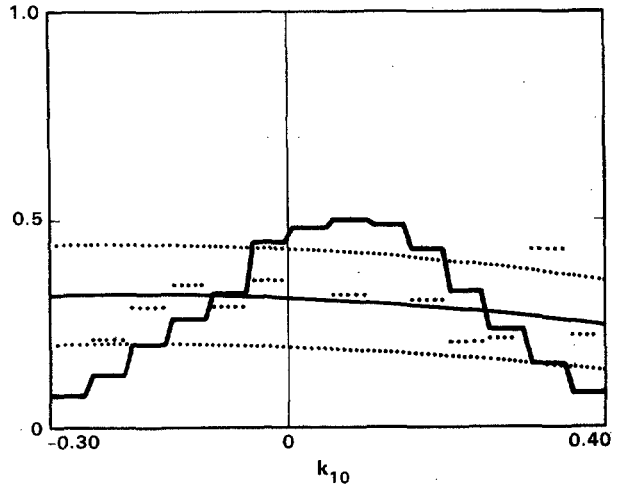


Fig. 19.

Section II-D, on the perceptually allowed k-parameter deviation functions of Appendix A and Fig. 5(a)–(c).

	From	*To*	*Is*
k 1	−1	−0.9469	−0.9741
	−0.9469	−0.8669	−0.9117
	−0.8669	−0.7425	−0.8109
	−0.7425	−0.5675	−0.6611
	−0.5675	−0.3534	−0.4637
	−0.3534	−0.1321	−0.2412
	−0.1321	+1	−0.0305
		* * *	
k 2	−1	−0.3525	−0.4979
	−0.3525	−0.0758	−0.2111
	−0.0758	+0.1696	+0.0515
	+0.1696	+0.3753	+0.2776
	+0.3753	+0.5396	+0.4625
	+0.5396	+0.6660	+0.6073
	+0.6660	+0.7604	+0.7168
	+0.7604	+0.8293	+0.7976
	+0.8293	+0.8778	+0.8561
	+0.8788	+1	+0.8979
		* * *	
k 3	−1	−0.7806	−0.8171
	−0.7806	−0.6741	−0.7337
	−0.6741	−0.5103	−0.6000
	−0.5103	−0.2854	−0.4049
	−0.2854	−0.0199	−0.1554
	−0.0199	+1	+0.1148
		* * *	
k 4	−1	−0.1346	−0.2675
	−0.1346	+0.1338	+0.0002
	+0.1338	+0.3842	+0.2627
	+0.3842	+0.5969	+0.4961
	+0.5969	+1	+0.6862
		* * *	
k 5	−1	−0.4433	−0.5260
	−0.4433	−0.2359	−0.3463
	−0.2359	+0.0152	−0.1143
	+0.0152	+0.2788	+0.1480
	+0.2788	+1	+0.4031
		* * *	
k 6	−1	−0.1035	−0.2282
	−0.1035	+0.1555	+0.0257
	+0.1555	+0.4014	+0.2819
	+0.4014	+1	+0.5112
		* * *	
k 7	−1	−0.3501	−0.4339
	−0.3501	−0.1125	−0.2420
	−0.1125	+0.1705	+0.0294
	+0.1705	+0.4031	+0.2978
	+0.4031	+1	+0.4840
		* * *	
k 8	−1	−0.0201	−0.1400
	−0.0201	+0.2342	+0.1077
	+0.2342	+0.4515	+0.3509
	+0.4515	+0.5968	+0.5333
	+0.5968	+1	+0.6443
		* * *	
k 9	−1	−0.1788	−0.3174
	−0.1788	+0.1060	−0.0353
	+0.1060	+0.3575	+0.2387
	+0.3575	+1	+0.4598
		* * *	
k 10	−1	+0.0204	−0.1379
	+0.0204	+0.3016	+0.1679
	+0.3016	+1	+0.4119

ACKNOWLEDGMENT

The authors wish to thank C. Searle, S. Seneff, and J. Tierney for reviewing the early stages of the manuscript and for their helpful comments.

REFERENCES

[1] B. S. Atal and S. L. Hanauer, "Speech analysis and synthesis by linear prediction of the speech wave," *J. Acoust. Soc. Amer.*, vol. 50, pp. 637–655, 1971.

[2] J. L. Flanagan, "A difference limen for vowel formant frequency," *J. Acoust. Soc. Amer.*, p. 613, May 1955.

[3] ——, "Difference limen for the intensity of a vowel sound," *J. Acoust. Soc. Amer.*, p. 1223, Nov. 1955.

[4] ——, "Bandwidth and channel capacity necessary to transmit the formant information of speech," *J. Acoust. Soc. Amer.*, p. 592, July 1956a.

[5] ——, "Difference limens for formant amplitude," Acoust. Lab., Mass. Inst. Technol., Cambridge, Quart. Rep., Sept. 1956b.

[6] ——, "Estimates of the maximum precision necessary in quantizing certain 'dimensions' of vowel sounds," *J. Acoust. Soc. Amer.*, p. 533, Apr. 1957.

[7] ——, "Focal points in speech communication research," *IEEE Trans. Commun.*, vol. COM-19, pp. 1006–1015, 1971.

[8] ——, *Speech Analysis, Synthesis and Perception*. New York: Springer-Verlag, 1972.

[9] J. L. Flanagan and M. G. Saslow, "Pitch discrimination for synthetic vowels," *J. Acoust. Soc. Amer.*, p. 435, May 1958.

[10] O. Ghitza, "Auditory-based criteria for parameter quantization of linear prediction coded speech," Ph.D. dissertation, Tel-Aviv Univ., Tel-Aviv, Israel, 1983.

[11] O. Ghitza and J. L. Goldstein, "Discrimination of formant frequency, intensity and bandwidth in natural speech," in *Proc. 103rd Meet. Acoust. Soc. Amer.*, Chicago, IL, 1982, p. 537.

[12] ——, "JNDs for the spectral envelope parameters in natural speech," in *Hearing—Psychophysical and Physiological Bases*, Klinke and Hartmann, Eds. Berlin: Springer-Verlag, 1983.

[13] J. L. Goldstein, "An optimum processor theory for the central formation of the pitch of complex tones," *J. Acoust. Soc. Amer.*, p. 1496, Dec. 1973.

[14] ——, "Mechanisms of signal analysis and pattern perception in periodicity pitch," *Audiol.*, vol. 17, pp. 421–445, 1978.

[15] A. H. Gray and J. D. Markel, "Quantization and bit allocation in speech processing," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-24, p. 459, Dec. 1976.

[16] D. H. Klatt, "Discrimination of fundamental frequency contours in synthetic speech: Implications for models of pitch perception," *J. Acoust. Soc. Amer.*, vol. 53, pp. 8–16, 1973.

[17] J. Makhoul, "Linear prediction: A tutorial review," *Proc. IEEE*, vol. 63, pp. 561–580, 1975.

[18] J. D. Markel and A. H. Gray, *Linear Prediction of Speech*. New York: Springer-Verlag, 1976.

[19] ——, "Implementation and comparison of two transformed reflection coefficients scalar quantization methods," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-28, p. 575, 1980.

[20] G. E. Peterson and H. L. Barney, "Control methods used in a study of vowels," *J. Acoust. Soc. Amer.*, p. 175, 1952.

[21] M. B. Sachs and E. D. Young, "Effects of nonlinearities on speech encoding in the auditory nerve," *J. Acoust. Soc. Amer.*, p. 858, Sept. 1980.

[22] M. R. Schroeder, "Direct (nonrecursive) relations between cepstrum and predictor coefficients," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-29, p. 297, Apr. 1981.

[23] P. Srulovicz and J. L. Goldstein, "A central spectrum model: A synthesis of auditory-nerve timing and place cues in monaural communication of frequency spectrum," *J. Acoust. Soc. Amer.*, p. 1266, Apr. 1983.

[24] T. E. Tremain, "The government standard linear predictive coding algorithm: LPC-10," *Speech Technol.*, p. 48, Apr. 1982.

[25] R. Viswanathan and J. Makhoul, "Quantization properties of transmission parameters in linear predictive systems," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-23, p. 309, June 1975.

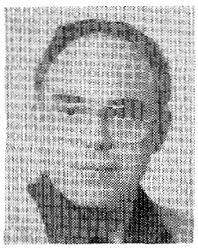
[26] R. Viswanathan, J. Makhoul, and A. W. F. Huggins, "Speech compression and evaluation," Bolt Beranek and Newman, Inc., Lexington, MA, Rep. 3794, Apr. 1978.

[27] H. Wakita, "Direct estimation of the vocal tract shape by inverse filtering of acoustic speech waveforms," *IEEE Trans. Audio Electroacoust.*, vol. AU-21, pp. 417–427, 1973.

[28] E. D. Young and M. B. Sachs, "Representation of steady-state vowels in the temporal aspects of the discharge patterns of populations of auditory-nerve fibers," *J. Acoust. Soc. Amer.*, p. 1381, Nov. 1979.

Oded Ghitza was born in Haifa, Israel, on September 24, 1948. He received the B.Sc., M.Sc., and Ph.D. degrees in electrical engineering from Tel-Aviv University, Tel-Aviv, Israel, in 1975, 1977, and 1983, respectively.

From 1980 to 1984 he worked at the Signal Corps Research Laboratory of the Israeli Defence Forces. During 1984–1985 he was a Bantrell postdoctoral Fellow at the Research Laboratory of Electronics, Massachusetts Institute of Technology, Cambridge, and a Consultant at the Lincoln Laboratory Speech Systems Technology Group, Lexington, MA. Currently he is with the Department of Acoustic Research, AT&T Bell Laboratories, Murray Hill, NJ, where he is studying auditory-based processing techniques for speech coding and speech recognition.



Julius L. Goldstein (S'56–M'62–SM'81) was born in Brooklyn, NY, on July 9, 1935. He received the B.E.E. degree from Cooper Union, Brooklyn, in 1957, the M.E.E. degree from the Polytechnic Institute of Brooklyn in 1960, and the Ph.D. degree from the University of Rochester, Rochester, NY, in 1965.

From 1957 to 1960 he was employed as an Electronic Circuit Design Engineer at Polytechnic R&D Co. in Brooklyn. Since his doctoral studies, his interests have turned to human sensory communication, and from 1965 to 1968 he held research appointments at the Institute of Perception Research, Eindhoven, The Netherlands, and at the Laboratory of Psychophysics, Harvard University, Cambridge, MA. He was a member of the EE Faculty at the Massachusetts Institute of Technology, Cambridge, from 1968 to 1973, where he conducted research on auditory psychophysics and physiology. Since 1973 he has been a member of the EE Faculty at the Tel-Aviv University, Tel-Aviv, Israel, where he was invited to develop and chair a new program in biomedical engineering. From 1976 to 1978 he served as Chairman of the Department of Electronics. His research interests include models of sensory communication, man–machine speech communication, and sensory aids.

Dr. Goldstein is a Fellow of the Acoustical Society of America and a member of the Collegium O.R.L.A.S. (Oto-Rhino-Laryngologicum Amicitias Sacrum).