An Objective Measure for Predicting Subjective Quality of Speech Coders

Shihua Wang, *Member, IEEE,* Andrew Sekey, *Senior Member, IEEE,* and Allen Gersho, *Fellow, IEEE*
(Invited Paper)

Abstract—**A perceptually-motivated objective measure for evaluating speech quality is presented. The measure, computed from the original and coded versions of an utterance, exhibits statistically a monotonic relationship with the mean opinion score (MOS), a widely used criterion for speech coder assessment. For each 10 ms segment of an utterance, a weighted spectral vector is computed via 15 critical band filters for telephone bandwidth speech. The overall distortion, called Bark spectral distortion (BSD), is the average squared Euclidean distance between spectral vectors of the original and coded utterances. The BSD takes into account auditory frequency warping, critical-band integration, amplitude sensitivity variations with frequency, and subjective loudness. The effectiveness of the measure was validated by a regression analysis between the computed BSD values and actual MOS values obtained from a speech data set. In tests with speech distorted by a modulated noise reference unit (MNRU) and with speech coded at rates of 2.4–64 kb/s, a monotonic function of the BSD was found which predicted MOS ratings notably better than segmental SNR or cepstral distance. The standard error in estimating MOS scores with the new measure was 0.2–0.3, with the higher accuracy for low rate coders in the range of 2.4–8 kb/s. The measure offers a more consistent assessment of the effect of incremental changes in the parameter of a speech coder than is usually obtained by the designer who relies on his/her own informal listening. Preliminary results also indicate that the measure may be effective for the excitation search in analysis-by-synthesis coders.**

I. Introduction

THE evaluation of speech quality is of critical importance in the field of speech coding. Not only is it necessary to have consistent *subjective* tests for the comparative assessments of alternative coders, it is also essential to have an *objective* distortion measure which, during the development phase, can give the designer an immediate and reliable estimate of the anticipated perceptual quality of a particular coding algorithm. Moreover, such an objective measure should be coder-independent, so that it can compare the subjective qualities of various algorithms possibly entailing quite different types of distortion.

While the quality of classical waveform coding algorithms can be estimated by waveform matching via the signal-to-noise ratio (SNR) or the segmental SNR, these measures are of little relevance to the new generation of coding algorithms in the range of 2–8 kb/s where exact waveform preservation would be an unrealistic goal. Instead, such coders aim for the adequate rendering of only the perceptually significant aspects of the signal to preserve intelligibility and naturalness. Consequently, for these coders we must often depend on informal listening to make instant judgments of quality during the process of algorithm development. This is, unfortunately, a hazardous procedure since casual listening does not reliably reveal whether one version of an algorithm is better or worse than a slightly different alternative. Also, once a candidate algorithm has been developed there is no simple and reliable way to report on its overall quality short of the costly and time-consuming process of formal subjective testing.

The degradation introduced by different types of low-rate coders and by the effects of transmission errors on such coders are very diverse. It is indeed a challenging task for any one listener to predict how two coders with different kinds of distortion will rank in formal subjective tests. The reader might thus be skeptical about the prospects of finding a single objective measure to do this reliably. Yet we assert that this is indeed a feasible objective, for the following reason. The peripheral auditory system of humans, which is highly consistent from one person to another, preprocesses auditory information and sends highly compacted data to the higher-level brain functions. Subjective decisions are then based on these data. An adequate model of the auditory system should, therefore, be able to emulate this biological preprocessing and compare the reduced representations of original and coded speech.

Though we do not understand well the higher-level processing of the brain, we know that different groups of individuals judge the quality of degraded speech fairly consistently (see, for example, [1]). It is thus reasonable for an objective measure based on the compacted output of the auditory system to deliver ratings that are highly cor-

Manuscript received September 23, 1991; revised January 20, 1992. This work was supported in part by the National Science Foundation, the State of California MICRO program, Rockwell International Corporation, Bell-Northern Research, and Bell Communications Research, Inc. Preliminary versions of this paper were presented at the International Conference on Acoustics, Speech and Signal Processing, Toronto, Canada, 1991 and the IEEE Workshop on Speech Coding for Telecommunications, Whistler, Canada, 1991.

S. Wang is with Teknekron Communications Systems, Berkeley, CA 94704.

A. Sekey and A. Gersho are with the Department of Electrical and Computer Engineering, University of California, Santa Barbara, CA 93106.

IEEE Log Number 9107598.

related with subjective testing results. Of course, such a measure cannot eliminate the need for subjective testing but it can greatly aid the algorithm development process by allowing a preliminary evaluation of alternative candidate coders.

If indeed an objective measure can be found which accurately assesses the subjective quality of an entire utterance, it may also be able to measure the distortion of a short segment of speech. If so, it could serve as an alternative to the ubiquitous weighted MSE for excitation searching in analysis-by-synthesis coders.

Finding an objective measure for predicting the perceptual quality of coded speech is a pressing task in speech coding research. A few prior studies have attempted to find such a measure. An approach based on models of the human auditory system due to Schroeder, Atal, and Hall [2] was one of the main inspirations for this paper. Another perceptual distance measure was introduced by Paillard *et al.* [3] as a guide to wideband audio coding. Perhaps the first work to statistically correlate an objective model with a subjective rating is that of Kitawaki, Nagabuchi, and Itoh [4] who introduced the *cepstral distance* (CD) measure for assessing coders in the range of 16–32 kb/s. A comprehensive treatment of the objective distortion measures is presented in the book by Quackenbush, Barnwell, and Clements [5]. An interesting measure based on the distortion in the spectral peaks of the speech was proposed by Coetzee and Barnwell [6]. In another recent work, Kubichek *et al.* [7] estimate the MOS via a distortion measure computed from the output energies of a bank of filters operating on the error signal. Yet, none of the prior work has demonstrated a measure that is effective and reliable for evaluating low bit rate speech coders.

Any objective measure must ultimately be validated by comparisons with subjective assessments. We consider an objective measure "effective" if it can reliably predict the score of a generally accepted subjective quality rating scheme. Several such measures are in use, such as the mean opinion score (MOS) and the diagnostic acceptability measure (DAM). In this paper, we use MOS as the basis for validating our objective measure.

Our main purpose in developing a new, perceptually-oriented objective measure is to be able to reliably and rapidly assess the performance of a coder and to compare the overall performance of different coders. A secondary purpose is to improve the error criterion by which excitation codebooks are searched in analysis-by-synthesis coders, i.e., vector excitation coding (VXC) or code-excited linear prediction (CELP). The first application requires a *global* distortion measure, which provides an overall evaluation: by aggregating segmental distortions, it rates the reconstruction quality of entire utterances. The second is a *local* measure, which must be rapidly computable for segments of a few milliseconds duration. The success of a measure for global evaluation of coder performance should also raise hope for its use for the second proposed application. This paper focuses primarily on the first application but we briefly report some results on the second problem.

In the rest of this paper, we shall discuss subjective measures as well as traditional objective ones, describe and justify our proposed measure, and critically compare its performance in predicting MOS scores of a speech database to that of other objective measures.

II. Subjective Distortion Measures

In order to develop an objective measure that correlates well with subjective quality assessments, we need to be cognizant of subjective measures which are the ultimate arbiters of speech quality. Two commonly used measures are the mean opinion score (MOS) [8] and the diagnostic acceptability measure (DAM) [9].

MOS scores require lengthy subjective testing [4], [1], but are widely accepted as a norm for comparative rating of different systems. The automatic prediction of MOS scores directly from the speech signals and without human subjects could, therefore, be of great practical value.

The rating scale employed in MOS testing is illustrated in Table I along with a general description of the levels of distortion typically associated with each numerical score. (These descriptions are not given to the subjects who perform the ratings). An MOS score is a mapping of perceived levels of distortion into either the descriptive terms "excellent, good, fair, poor, unsatisfactory," or into equivalent numerical ratings in the range 5–1. The numerical mapping is clearly a mixed blessing. On the one hand, it permits the ranking of coders and direct comparisons with objective measures. On the other hand, it lumps different kinds of distortion together, which gives little insight into the causes of distortion.

The diagnostic acceptability measure (DAM) is almost the opposite: it is highly descriptive but needs to be reduced to a single parameter for comparative ratings. DAM descriptors are highly suggestive of the *kind* of distortion observed, but they are both numerous and nonnumerical. DAM tests are even more laborious to perform, partly because listeners must be trained to conform to recognized meanings of the descriptors. The automatic prediction of individual DAM descriptor ratings is entirely out of reach with our present knowledge. However, we conjecture that our BSD measure may well be able to predict a single-parameter reduction of the DAM score, but this needs further investigation.

III. Traditional Objective Measures

The most common objective measure is the mean squared error (MSE) between original and coded speech waveforms. This leads to signal-to-noise ratio (SNR), a term that is inaccurate because the "noise" does not exist as a physical entity; rather, it is manifested perceptually as the discrepancy between what the listener hears and what she expects to hear. The SNR treats the entire speech sample as a single vector, as if the listener made a single comparison after storing the entire utterance, clearly an

TABLE I
DESCRIPTIONS IN THE MEAN OPINION SCORE (MOS)

Rating	Speech Quality	Level of Distortion
5	Excellent	Imperceptible
4	Good	Just perceptible but not annoying
3	Fair	Perceptible and slightly annoying
2	Poor	Annoying but not objectionable
1	Unsatisfactory	Very annoying and objectionable

TABLE II
COMPARATIVE PERFORMANCE OF OBJECTIVE MEASURES

Objective Speech Quality Measure	$\lvert\rho\rvert$
Time domain measure	
SNR	0.24
Segmental SNR	0.77
LPC-based measures	
Log area ratio	0.62
Log likelihood ratio	0.50
Cepstral distance	0.63
Frequency variant log spectral distance	
LPC-based	0.68
Filter bank	0.72
Weighted-slope spectral distance	0.74

unreasonable proposition. In a more realistic variant called *segmental* (SNRSEG), SNR power ratios over short segments are determined and their *geometric* mean is computed. This seems to correspond much better to the auditory experience.

In an improved measure, a perceptually weighted mean squared error (PWMSE) was studied [2] where the weighting depends on the time-varying spectral envelope of the original speech. This measure is widely used for the analysis-by-synthesis in VXC/CELP coders. Recently, it has also been proposed as the basis for a modified SNR measure [10]. However, even with perceptual weighting such a distortion measure is still fundamentally waveform-based and focuses on approximating the sample-to-sample variations of the signal, rather than on the perceptual closeness of the reconstructed speech to the original.

Other objective measures are derived from LPC coefficients or other time- or frequency-domain parameters [4], [5], [1], but none of these have been totally satisfactory. In particular, the cepstral distance (CD) measures the disparity between the original and coded spectral envelopes, and is determined by generating cepstral coefficients from the LPC parameters. In [4], a high correlation coefficient of 0.93 was claimed between CD and MOS for speech coded in the range of 16–32 kb/s. However, for a broader range of bit rates, Quackenbush *et al.* [5] reported a much lower correlation of 0.63.

The limited ability of some of these measures to predict MOS ratings reported in [5] is shown in Table II, where ρ is the correlation coefficient between MOS ratings and corresponding objective measures. The value $\rho = 1$ would indicate that the measure predicts MOS perfectly, while $\rho = 0$ could be obtained even by randomly guessing the MOS. As is seen, conventional SNR does not perform much better than this, but others do. For example, segmental SNR reaches 0.77, with LPC-based measures trailing closely behind it.

Objective measures based on auditory models [11], [12] are also used in the field of speech recognition. Yet their goal there is different: to measure the phonetic distance between received and stored utterances, usually without regard to the speaker's identity. In contrast, high-quality speech coding demands the minimization of psychoacoustic distance, of which speaker identity is an integral part.

IV. A PSYCHOACOUSTICALLY MOTIVATED OBJECTIVE MEASURE

Our starting point in seeking a new measure was the observation that the PWMSE or the traditional segmental SNR (SNRSEG) distance measures apply a sufficient, but not necessary, condition for good quality by demanding that original and reconstructed speech *waveforms* be identical. Yet counterexamples abound: Hilbert-transformed speech and phase-equalized speech [13] sound almost indistinguishable from the original, even though the waveforms differ widely. Still, we know we cannot totally ignore phase effects since, for example, speech reconstructed with random phase definitely sounds distorted. Fortunately, in practical systems such as low-rate coders, drastic phase changes are typically accompanied by corresponding variations in the magnitude spectrum, which do get sensed by the BSD measure.

Our aim for the objective distortion measure was to emulate several known features of perceptual processing of speech sounds by the human ear, specifically:

- frequency scale warping, as modeled by the Bark transformation, and critical band integration in the cochlea;
- changing sensitivity of the ear as the frequency varies;
- difference between the loudness level and the subjective loudness scale.

These effects and the transformations incorporating them will be discussed in turn; they will lead us to a meaningful new distance measure in the perceptual space.

The manner in which the measure is determined is shown in Fig. 1. Both the original speech $x(n)$ and its coded version $y(n)$ are separately processed by identical operations, leading to what we shall refer to as *Bark spectra* $L_x(i)$ and $L_y(i)$, respectively. The subjective quality measure is then an appropriately defined distance between these spectra.

Unlike many objective measures, we do *not* process the error signal (between original and coded speech) to measure distortion. The substantial nonlinearity of the auditory model implies that, for a meaningful auditory measure, a separate processing of each speech signal that the ear would actually hear is necessary.

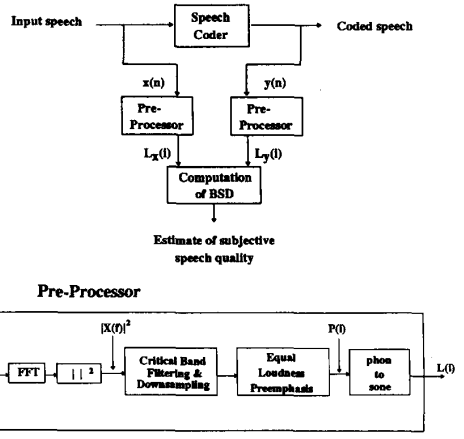


Fig. 1. Computing the objective measure.

The processing operation itself begins with a computation of the magnitude squared FFT spectrum to generate the power spectrum $|X(f)|^2$, followed by several stages which will be explained in turn.

A. Critical-Band Filtering

The human auditory system is known to have poorer discrimination at high frequencies than at low frequencies. This, together with observations on masking of tones by noise, led to modeling the peripheral auditory analysis by *critical-band filters*. The model postulates that sounds are preprocessed by a band of such filters, with center frequency spacings and bandwidths increasing with frequency, as shown in Fig. 2. These filters may be viewed as the tuning curves of auditory neurons; their spacing corresponds to 1.5 mm steps along the basilar membrane [14].

Observe that the frequency scale in Fig. 2 is logarithmic, and that at higher frequencies the critical-band filter shapes are identical (on a linear scale, they would appear as ever widening). On a true Bark scale, the spacing between the filter maxima is a constant 1 Bark, and their points of intersection are 3 dB down from maximum.

It is conceptually convenient to think of critical-band filtering as a two-stage process.

Stage 1: Hertz-to-Bark transformation via the relation [15]:

$$f = Y(b) = 600 \sinh (b/6) \qquad (1)$$

where f is frequency in Hz and b is its equivalent on the Bark-scale.[1] The mapping $Y(b)$ has been called the "critical-band density" [2].

Stage 2: "smearing" the density function by the prototype critical-band filter. This is a straightforward convolution, since on the Bark scale the shapes of the critical-band filters $F(b)$ are identical and are given by [16]:

$$10 \log_{10} F(b) = 7 - 7.5(b - 0.215) - 17.5 [0.196 + (b - 0.215)^2]^{1/2}. \qquad (2)$$

The continuous spectrum resulting from this operation is, thus, $D(b) = F(b) * Y(b)$ and has been called the "excitation pattern" [17], since it corresponds to the distribution of stimulation in the auditory nerve.

[1]Note that other approximations, e.g., [2], use slightly different constants in a similar formula.

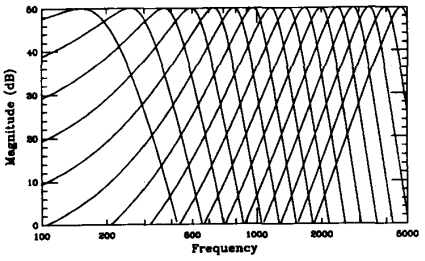


Fig. 2. Bank of critical-band (CB) filters (after [16]).

B. Perceptual Weighting of Spectral Energy

The spectral modifications described so far account for the frequency resolution of the ear (Hertz-to-Bark transformation) and the nonlinear smoothing effect of critical-band cochlear filtering. We next incorporate the fact that the ear is not equally sensitive to stimulations at different frequencies. For example, a 100 Hz tone may need to be up to 35 dB more intense than a 1000 Hz tone, for the two to sound equally loud. The well-known equal loudness level curves [18] shown in Fig. 3 express this graphically. Each curve shows how the intensity level of a tone must be varied with frequency in order to maintain a constant level of *perceived* loudness by human listeners. The acoustic intensity level in these curves is measured at the listener's ear in dB units, relative to a standardized reference level set to the threshold of hearing at 1 KHz. This intensity scale is referred to as sound pressure level (SPL). For a tone of a particular frequency and intensity, its loudness level in *phones* is defined as the intensity in dB of a 1 KHz tone which sounds equally loud.

In order to account for the auditory processing represented by the equal loudness curves, we must convert intensity levels in dB to loudness levels in *phons*. In the region of interest in telephone speech (300–3 400 Hz, 40–80 dB intensity level), highlighted by the box in Fig. 3, the response can be equalized by the Bark-domain equivalent of a bilinear preemphasis filter in the form $H(z) = (2.6 + z^{-1})/(1.6 + z^{-1})$. This moderately boosts frequencies above 1800 Hz, which is also justified by a noted increase in sensitivity to distortions in this region.

For convenience, the computation is performed in the linear (rather than Bark) frequency domain, where we denote the excitation pattern by $D'(z)$. The weighted spectrum then becomes $P'(z) = H(z)D'(z)$. The Bark-domain equivalent of this, $P(b)$, is then sampled at 1 Bark intervals to yield the discrete function $P(i)$.

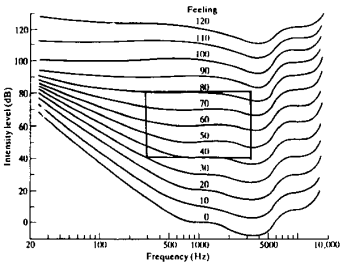


Fig. 3. Contours of equal loudness level. The numbers on the curves indicate the loudness in phons (after [18]).

C. Subjective Loudness

The spectrum available at this point is loudness equalized so that the relative intensities at different frequencies correspond to relative loudness in *phones* rather than relative acoustical levels. As a last step, we must include yet another perceptual nonlinearity: the increase in phons needed to double the subjective loudness is not a constant, but varies with the loudness level [19]. For example, at an average level (40 phons) an extra 10 phons will double the loudness, while near threshold the same 10 phons will increase the loudness tenfold. The phon scale may thus be converted to a truly perceptual scale of *sones*. A sone is, by definition, the increase in power which doubles the subjective loudness. Fig. 4 shows the relation between subjective loudness in sones and loudness level in phons. We hypothesize that in comparing different signals, listeners respond to differences in their levels expressed in sones. The phon (P)-to-sone (L) conversion is modeled in [17] by:

$$L = \begin{cases} 2^{(P-40)/10} & \text{if } P \geq 40 \\ (P/40)^{2.642} & \text{if } P < 40. \end{cases} \tag{3}$$

Observe that, to apply this equation, one must be able to determine what intensity level corresponds to the threshold of 40 phons. Since the level in phons refers to the loudness level at the subject's ear, the proper procedure requires calibration via an artificial ear. However, this is not always practical since one must frequently work with speech files obtained from others, for which such calibration was not performed.

An alternative approach is as follows. Noting that the average speech threshold (the lowest level at which speech is barely audible and intelligible) is about 20 dB SPL [20] and that typical comfortable listening levels are around 55–60 dB above the speech threshold, we can assume that most MOS files would have been presented to subjects at around 75–80 dB SPL. For example, the AT&T speech files in our database were presented at an average level of 78 dB [21].

However, given that speech seldom falls more than 35 dB below its average value, and when it does, distortions are no longer significant, we may safely assume that the part of the curve lying below 40 phons is of no importance, and we use the upper part of (3) throughout. We

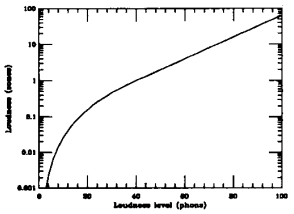


Fig. 4. Conversion from phons to sones (after [19]).

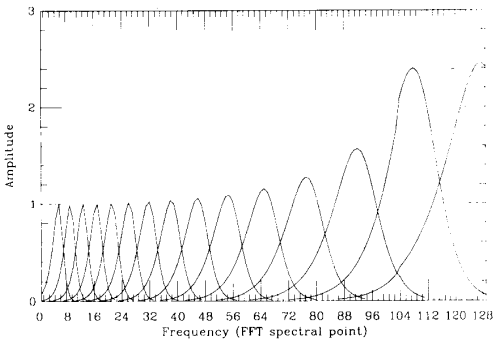


Fig. 5. The fifteen weighting functions used for computing fifteen samples of a Bark spectrum from a power spectrum.

have also experimented with a modified distortion measure where this transformation was omitted but we found it generally inferior.

D. Computational Considerations

While conceptually simple, nonlinear frequency warping is not computationally convenient since it would be necessary to resample and interpolate the DFT-based spectrum to generate a Bark-domain spectrum. Instead, the operations described above were mapped back into the linear frequency domain following Hermansky's method [22]. Bark-scale convolution with critical-band filter shape was then substituted by weighted averaging in the linear frequency domain with a precomputed weighting function shown in Fig. 5. Since "smearing" the spectrum with the critical-band filter reduces its true dimensionality to the point where it could be sampled at the critical-band rate, it is actually sufficient to perform the averaging only at points spaced 1 Bark apart. All information is then included in a low-dimensionality vector which we call the *Bark spectrum*, and which represents the energies collected by the critical-band filters. For the 3.4 KHz bandwidth speech which is of interest to us, we set the vector dimension at $N = 15$. However, our method could be readily extended to wideband speech or audio by increasing the dimensionality of the vector to represent the needed frequency range.

The above operations are carried out frame by frame. With the frame length set to 80 samples, the best was found in the range 30–400 samples. Each frame is weighted by a Hamming window, and consecutive frames

overlap by 50%. However, on the assumption that very low distortion values may correspond to silent intervals or are otherwise perceptually insignificant, if in a given frame the signal was found to fall below a preset power threshold level, the contribution of that frame to the average distortion was set to zero.

E. Distortion Computation

The Bark spectrum $L(i)$ reflects the ear's nonlinear transformations of frequency and amplitude, together with important aspects of its frequency analysis and spectral integration properties in response to complex sounds. We may use the squared Euclidean distance between two Bark spectral vectors, called the *Bark spectral distortion* or BSD, as an objective measure of the distortion between the speech segments from which they were derived. For the kth segment, this is given by:

$$\mathrm{BSD}^{(k)} = \sum_{i=1}^{N} [L_x^{(k)}(i) - L_y^{(k)}(i)]^2 \tag{4}$$

where

N–number of critical bands

$L_x^{(k)}(i)$–Bark spectrum of kth segment of original speech

$L_y^{(k)}(i)$–Bark spectrum of kth segment of coded speech.

The overall unnormalized distortion is then the average BSD, or $\mathrm{BSD}_u = \mathrm{Avg}\ [\mathrm{BSD}^{(k)}]$, i.e., the mean Euclidean distance of the spectral vectors in sones, taken over successive frames in an utterance.

However, this must still be normalized since all values computed so far depend on the arbitrary numerical representation of speech samples in the computer. For example, suppose the A/D converter sensitivity is increased by 20 dB so that all sample amplitudes are multiplied by 10. Then, from the upper formula in (3), the value of BSD will increase by $2^{20/10} = 4$. Yet the distortion should not depend on such arbitrary scalings. Consequently, we divide the unnormalized BSD_u by a number representing the average Bark energy of the original signal:

$$E_{\mathrm{Bark}} = \underset{k}{\mathrm{Ave}} \sum_{i=1}^{N} [L_x^{(k)}(i)]^2. \tag{5}$$

The final value, $\mathrm{BSD} = \mathrm{BSD}_u / E_{\mathrm{Bark}}$ is invariant under scaling of the input signal.

Fig. 6 shows examples of spectra as a function of frequency at different stages of the computation illustrated in Fig. 1. Observe that the conversion from loudness level to subjective loudness magnifies differences at spectral peaks, while it hardly affects spectral valleys. This agrees with the well-known phenomenon in speech perception that peaks are more important than valleys.

V. Testing the New Objective Measure

As a preliminary evaluation of our measure, we first performed some simple tests by applying the measure to four types of distorted speech with predictable results: i) Hilbert-transformed speech; ii) FIR and IIR all-pass filtered speech; iii) DPCM speech with linear or nonlinear prediction, with five-bit quantization and 40 kb/s data rate [23]; and iv) speech subjected to precedence (Haas) effect with over 1 ms delay. (The Haas effect refers to the inability of the ear to recognize a slightly delayed and attenuated version of a signal as a separate entity, when presented together with the original.) The effects of i)–iii) are waveform distortions of the type which are known to cause no significant deterioration in quality, as also confirmed by informal listening. For these test samples the SNRSEG values decreased, which could falsely be interpreted as an indication of audible distortions, yet the BSD remained essentially unchanged. In contrast, in precedence effect experiments, for delays increasing over the range $\tau = 1.25$–32 ms and with the delayed component attenuated by 0.5, audible distortions increased. This was well reflected by a steady increase of the BSD from 7.7 to 14.2, while SNRSEG showed no consistent trend, merely fluctuating between 5.8–6.8.

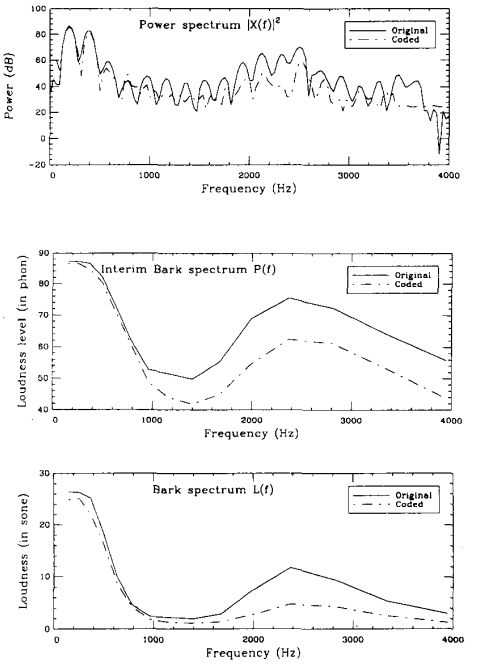


Fig. 6. Spectra at different stages of the preprocessor.

Encouraged by these observations, we applied the measure to a database containing two types of distorted speech for which MOS ratings were available. One class consisted of speech utterances corrupted by correlated noise as generated by a modulated noise reference unit (MNRU) circuit, while the other contained speech coded by seven different coders. These included a representative selection ranging from 2.4 to 64 kb/s. Test sentences were spoken by four male and four female speakers, and the panel of 42 subjects listened to the samples with handsets.

In order to assess the ability of our measure and that of others to predict MOS ratings, we fitted second-order polynomial predictors to the various scatterplots by least

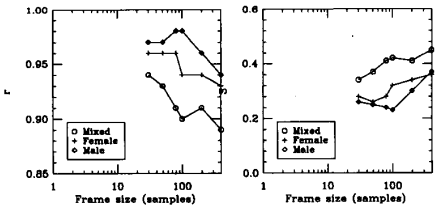


Fig. 7. Bark spectral analysis window size versus prediction performance.

squares linear regression:

$$EMOS = a + by + cy^2 \tag{6}$$

Here y is the objective measure being studied, and *EMOS* is the corresponding estimated MOS value as "predicted" by the fitted function.

We computed the BSD measure frame by frame, with the frame length set to 80 samples. This figure was selected after examining the effect of frame length on the distortion measure, as shown in Fig. 7. With window lengths ranging from 30 to 400 samples, the highest correlation for mixed (male and female) speakers is attained when the frame length is 80 samples. Each frame was weighted by a Hamming window, and consecutive frames were overlapped by 50%.

We conjectured that spurious distortion values may arise at low power levels, which might contaminate our results. As a precaution against this, we eliminated very small distortion values which may correspond to silent intervals or are otherwise perceptually insignificant. Thus, if in a given frame the signal was found to fall below a preset power threshold level, the contribution of that frame to the average distortion was set to zero (see also Section VI).

We compared the performance of our measure to that of two popular distortion measures: segmental SNR (SNRSEG) and cepstral distance (CD) [4]. SNRSEG is widely used for assessing distortion and recently a perceptually weighted version has been proposed [10]. Cepstral distance is the average distance between cepstrally smoothed input and output spectra. It is one of three objective voice parameters, well-correlated with human perception of quality, currently being considered by the CCITT for assessing digital voice telecommunications quality.

The results are summarized in Figs. 8–11. They show four objective measures as follows.

SNRSEG–segmental SNR

PW–SNRSEG—perceptually weighted SNRSEG

CD–cepstral distance

BSD–Bark spectral distortion

Also shown are second-order polynomial predictors fitted by least squares linear regression to the scatterplots.

The numbers shown in the figures have the following meanings. r is the correlation coefficient between the actual and predicted MOS values. Perfect prediction would yield the value $r = 1$. The variable s is the standard de-

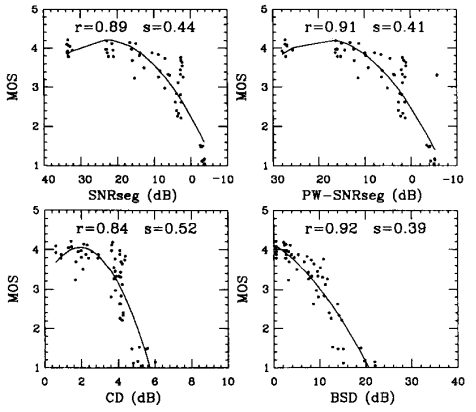


Fig. 8. Relationship between measured MOS values and values estimated via the distortion measures as shown. Mixed speakers, coding rates 2.4–64 kb/s.

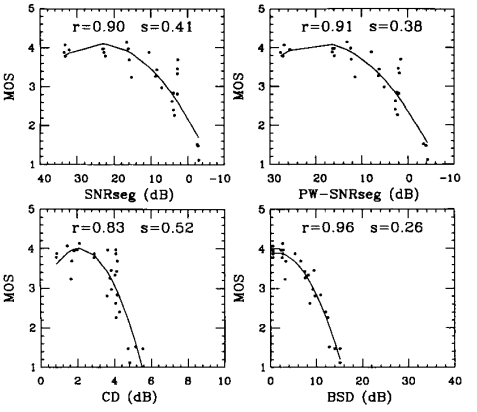


Fig. 9. Relationship between measured MOS values and values estimated via the distortion measures as shown. Female speakers, coding rates 2.4–64 kb/s.

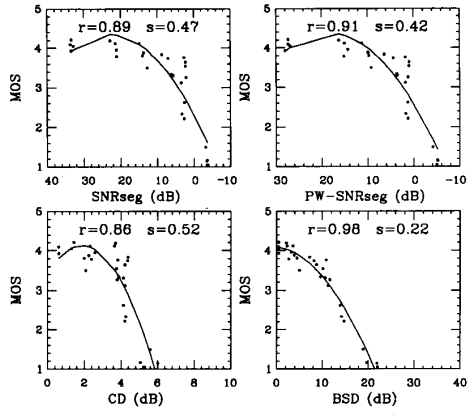


Fig. 10. Relationship between measured MOS values and values estimated via the distortion measures as shown. Male speakers, coding rates 2.4–64 kb/s.

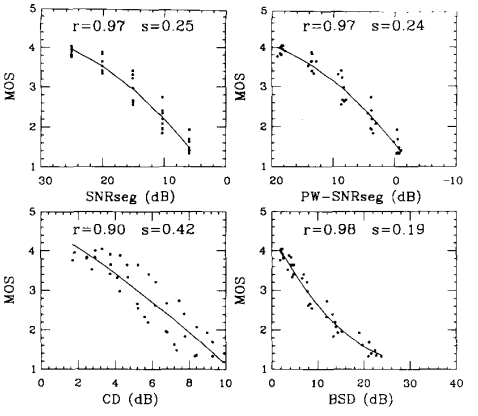


Fig. 11. Relationship between measured MOS values and values estimated via the distortion measures as shown. Mixed speakers, MNRU 5–25 dB.

viation of the prediction error, and in ideal conditions it would be zero.

A comparison of the figures leads to the following tentative conclusions:

i) While SNRSEG and CD are largely insensitive to whether the speaker is male or female, BSD performs best for male speakers followed by female, and much less well when the two sets of data are mixed. This suggests that BSD responds differently to some generic male versus female acoustic speech feature than do MOS ratings. Steps towards reducing this discrepancy are described below.

ii) SNRSEG does moderately well for high MOS values, corresponding to high bit rate (waveform) coders, but it can err by a whole unit of MOS or more at low bit rates. In contrast, BSD improves towards low MOS values, as seen in Table III where we compare correlation coefficients and standard deviations of predictor curves fitted by linear regression for different measures and with bit rates between 2.4–8 kb/s.

iii) We have checked a few "outliers" from the BSD regression curve and noticed that those with an MOS rating lower than the predicted value tended to be samples with speakers with raspy or breathy voices. We speculate that the effects of the coder and the speaker voice quality are *confounded*, in a statistical sense, in determining the MOS.

iv) The least-squares regression fit to CD data yields a "hook" shape for low CD values, unreasonably linking a further reduction in cepstral distance with a *lowering* of the MOS. This, together with the fact that CD is much lengthier to compute than SNRSEG yet performs no better, raises doubts about its usefulness for predicting MOS.

v) Contrary to expectation, perceptually weighted SNRSEG is only marginally superior to its unweighted version.

vi) The above results generally hold also when the source of the distortion is speech-correlated noise (MNRU).

TABLE III
PERFORMANCE MEASURES FOR LOW-RATE CODERS

	Mixed		Female		Male	
Measure	r	s	r	s	r	s
SNRseg	0.82	0.55	0.80	0.54	0.85	0.57
PW-SNRseg	0.84	0.52	0.83	0.51	0.87	0.53
CD	0.84	0.52	0.79	0.55	0.88	0.51
BSD	0.85	0.51	0.93	0.33	0.98	0.22

VI. IMPROVING THE BSD MEASURE

Experience with the BSD measure led us to two modifications, both resulting in small improvements. First, comparisons of plots of BSD with corresponding original and coded speech waveforms revealed that frequently the BSD attains large values in regions of low-energy unvoiced fricatives. This is not unexpected, since in such regions a CELP-type coder, in attempting to match quasi-random waveforms, will frequently commit gross errors in the frequency domain, which will be detected by the BSD. An example of this is illustrated in Fig. 12, showing the waveform and BSD trace for the sentence "*The shaky barn fell with a loud crash.*" Observe that the most prominent BSD values occur for the two /sh/ phonemes. Yet intuition as well as experience suggests that errors in such low-energy regions have little effect on speech quality. The BSD, thus, leads one to an unduly pessimistic estimate of the MOS.

To circumvent this problem, we eliminate unvoiced portions from the BSD calculation. This is accomplished by a voiced/unvoiced detector followed by an on/off switch in the path of the BSD integrator.

Second, we speculated on the causes of the male/female discrepancy and asked ourselves why, when faced with a male and a female file with the same MOS, would the BSD be larger for the male than for the female. Put differently, given the same spectral distortion as measured by BSD, why are listeners more severe in their judgement against female speakers?

A partial explanation may be tendered by noting the sensitivity of CELP-type coders to overall speaker volume. Male speakers, with their larger speech apparatus, generally tend to be louder, and their closely spaced pitch harmonics are more assured to excite formants close to their peaks. As a result, a rapid increase in average power, such as after the release of a stop consonant, will generally confront a coder with a larger energy step for a male speaker than a female. The coder's response will usually then be less prompt for the male than for the female.

An illustration of the behavior of the BSD during onsets is shown in Fig. 13, for the word *fast*. Note how the BSD peaks at the onset, where the coded version clearly fails to track the original. Yet, when listening to this word, the distortion in the /*f*/ is hardly audible, while the vowel /*ae*/ is definitely inferior in the coded version.

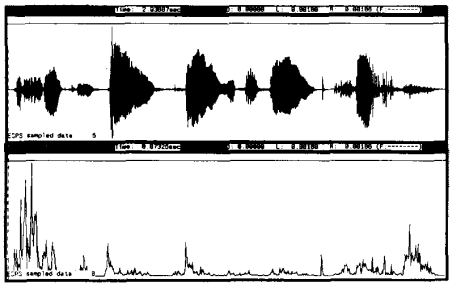


Fig. 12. Waveform and BSD track of the sentence "*The shaky barn fell with a loud crash.*"

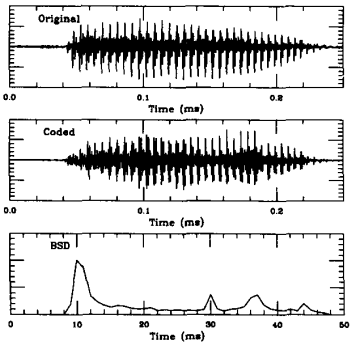


Fig. 13. Waveforms and BSD trace of the word *fast*.

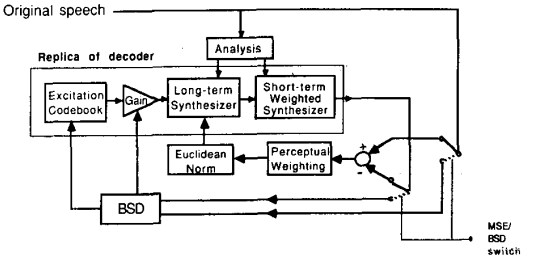


Fig. 14. Hybrid MSE/BSD excitation selection in CELP/VXC.

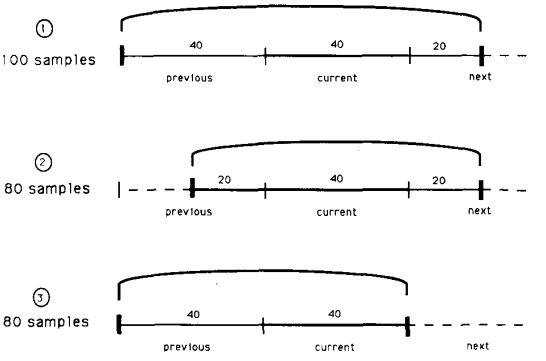


Fig. 15. BSD analysis frame positioning.

However, it is well-known that during rapid transitions the acuity of auditory spectral analysis is subordinated to temporal resolution, i.e., the ear changes its behavior from acting like a bank of narrowband filters to that of wideband filters [24]. Hence, whatever the spectral distortions in such regions, they may be relegated to secondary importance when compared to distortions in the steady state. We thus applied yet another filter to the BSD: whenever the difference between adjacent Bark spectral vectors exceeded an empirically set threshold, we ruled that frame to be a "fast transient" and omitted its BSD value from subsequent calculations.

VII. Applying the BSD Measure to Analysis-by-Synthesis Coding

The success of meeting the first objective encouraged us to pursue our second objective, namely the excitation codebook search in analysis-by-synthesis coding. The coder we tested was the VXC/CELP type, and the experimental arrangement is shown in Fig. 14. As is seen, the excitation selection is performed in a hybrid fashion. First, the long-term synthesizer parameters are determined with a MSE criterion. Second, the gain for each excitation vector in the codebook is computed, also with the MSE criterion. However, in the third step the actual excitation vector and its gain are selected on the basis of minimizing the BSD.

Some care must be taken with the size and positioning of the analysis frame in BSD computations. Three possibilities were explored, as shown in Fig. 15. In each case, only the ringing part from the current subframe is considered in order to include part of the next subframe, as if the next excitation vector would be identically zero. All windows extend over subframe boundaries, assuring that potential waveform discontinuities at boundaries are prevented, since the measure detects the energy splatter that would be caused by such discontinuities. Also, with at least 80–100 samples for the frame, at least one pitch cycle is included for male speakers and several for females.

An illustration of this hybrid selection method is given in Fig. 16, corresponding to the word *and*. Clearly, the waveform coded with MSE tracks the original better than the one with BSD. In contrast, when the short-term spectra for the same segments are compared (Fig. 17), the BSD shows a distinct advantage in preserving the magnitudes of the first few harmonics. Moreover, when listening to the entire sentences, we noted a slight superiority of the BSD-coded version.

In another result (not shown), for sustained vowels in female speech, the hybrid BSD/MSE method provided a better looking spectral match than when MSE was used for both searches, especially near the onset of the vowel. These preliminary results are encouraging and appear to indicate that the BSD measure selects perceptually more satisfying random code vectors than does the weighted MSE waveform matching measure.

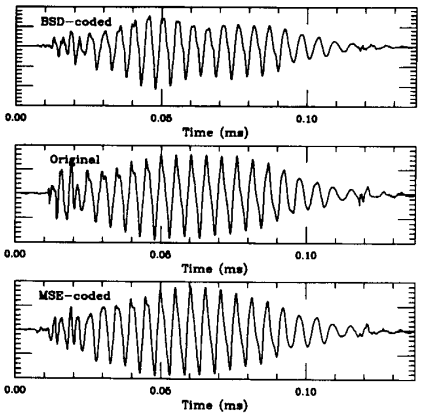


Fig. 16. Excitation selection by MSE versus BSD—waveform comparison.

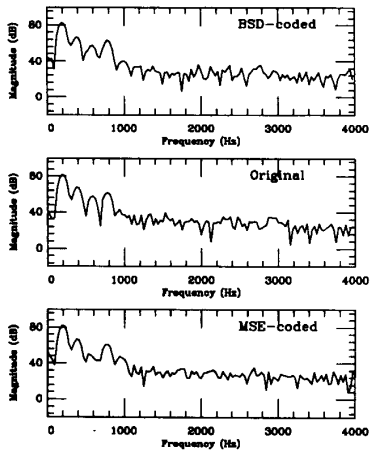


Fig. 17. Excitation selection by MSE versus BSD—spectral comparison.

VIII. Concluding Remarks

The first objective of our proposed distortion measure appears to have been met to a fair extent, though there is still room for improvement. First, the remaining difference between male and female speakers in predicting MOS score needs to be further studied. Second, the resolution of Bark spectra, held at 1 Bark in all our experiments, would be improved by spacing the Bark filters closer. The possible effect of this on the ability of the BSD measure to predict MOS scores should be investigated. Third, the BSD measure does not, at present, take into account temporal masking effects yet we know that high energy speech portions perceptually dominate low energy portions. This suggests that time-dependent weighting factors should be incorporated into the measure to account for *forward/backward masking*. Finally, we believe that with a much larger database of MOS-rated speech we could test the proposed objective measure more extensively and obtain statistically more reliable regression curves.

The pilot experiment of basing the selection of VXC random excitation codevectors on BSD, reported in the previous section, indicates that BSD may have much potential to function in this capacity. Specifically, the fact that BSD is sensitive to fine structure changes suggests that it may even be suitable, with further refinements, for the adaptive codebook method (closed loop pitch searching) in analysis-by-synthesis coders. If so, the evolution of BSD as an alternative to MSE may open up a new direction in low-rate coding by freeing designers from the waveform matching requirement.

In summary, we find that the BSD measure as currently computed appears to be remarkably effective in predicting MOS scores for low bit rate coders. While more work is needed, this study suggests that we are approaching the stage when a standardized objective measure may become available, which would be a valuable aid to algorithm developers and would provide simple and rapid means for assessing, comparing, and specifying speech coder performance.

References

[1] D. Goodman and R. Nash, "Subjective quality of the same speech transmission conditions in seven different countries," *IEEE Trans. Commun.*, vol. COM-30, pp. 642–654, Apr. 1982.

[2] M. R. Schroeder, B. S. Atal, and J. L. Hall, "Objective measure of certain speech signal degradations based on masking properties of human auditory perception," in *Frontiers of Speech Communication*. New York: Academic, 1979.

[3] B. Paillard, J. Soumagne, P. Mabilleau, and S. Morissette, "A perceptual distance criterion for the coding of speech or music signals," in *Signal Processing IV: Theories and Applications*, J. L. Lacoume *et al.*, Ed. New York: Elsevier Science, 1988, pp. 1093–1096.

[4] N. Kitawaki, H. Nagabuchi, and K. Itoh, "Objective quality evaluation for low-bit-rate speech coding systems," *IEEE J. Select. Areas Commun.*, vol. SAC-6, pp. 242–248, Feb. 1988.

[5] S. Quackenbush, T. Barnwell, and M. Clements, *Objective Measures of Speech Quality*. Englewood Cliffs, NJ: Prentice Hall, 1988.

[6] H. J. Coetzee and T. P. Barnwell III, "An LSP based speech quality measure," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, Glasgow, Scotland May 1989, pp. 596–599.

[7] R. F. Kubichek and D. J. Atkinson, "NCS voice quality project final report," NTIA, U.S. Dept. of Commerce, Inst. Telecommun. Sci., Boulder, CO, Sept. 1990.

[8] "IEEE recommended practice for speech quality measurements," *IEEE Trans. Aud. Electroacoust.*, pp. 227–246, Sept. 1969.

[9] W. Voiers, "Diagnostic acceptability measure for speech communication systems," in *Proc. IEEE Int. Conf. Acoust., Speech, and Signal Process.*, pp. 204–207, May 1977.

[10] Y. Be'ery, Z. Shpiro, T. Simchony, L. Shatz, and J. Piasetzky, "An efficient variable-bit-rate low-delay CELP," in *Advances in Speech Coding*, B. S. Atal *et al.*, Eds. New York: Kluwer, 1990.

[11] D. Klatt, "Prediction of perceived phonetic distance from critical-band spectra: a first step," in *Proc. IEEE Int. Conf. Acoust., Speech, and Signal Process.*, 1982, pp. 1278–1281.

[12] H. Wakita, "Linear prediction voice synthesizers: Line spectrum pairs (LSP) is the newest of several techniques," *Speech Technol.*, pp. 17–22, Fall 1981.

[13] T. Moriya and M. Honda, "Speech coder using phase equalization and vector quantization," in *Proc. IEEE Int. Conf. Acoust., Speech, and Signal Process.*, pp. 1701–1704, 1986.

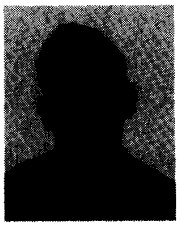
[14] D. O'Shaughnessy, *Speech Communication*. New York: Addison-Wesley, 1987.

[15] A. Fourcin, "Speech processing by man and machine—Group report," in *Recognition of Complex Acoustic Signals*, T. Bullock, Ed. Life Sciences Res. Rep. 5 of the Dahlem Workshops, Berlin, Germany, 1977.

[16] A. Sekey and B. Hanson, "Improved one-Bark bandwidth auditory filter," *J. Acoust. Soc. Am.*, vol. 75, pp. 1902–1904, June 1984.
[17] R. Bladon, "Modeling the judgment of vowel quality differences," *J. Acoust. Soc. Am.*, vol. 69, pp. 1414–1422, May 1981.
[18] D. Robinson and R. Dadson, "A redetermination of the equal-loudness relations for pure tones," *Brit. J. Appl. Phys.*, pp. 166–181, 1956.
[19] H. Fletcher and W. Munson, "Relation between loudness and masking," *J. Acoust. Soc. Am.*, vol. 9, pp. 1–10, 1937.
[20] D. Dirks, R. Stream, and R. Wilson, "Speech audiometry: Earphone and sound field," *J. Speech and Hearing Disord.*, vol. 37, Feb. 1972.
[21] P. Kroon, personal commun.
[22] H. Hermansky, "Perceptual linear predictive (PLP) analysis of speech," *J. Acoust. Soc. Am.*, vol. 87, pp. 1738–1752, Apr. 1990.
[23] S. Wang, E. Paksoy, and A. Gersho, "Performance of nonlinear prediction of speech," in *Proc. Int. Conf. Spoken Lang. Process.*, Kobe, Japan, Nov. 1990.
[24] A. Sekey, "Short-term auditory frequency discrimination," *J. Acoust. Soc. Am.*, vol. 35, pp. 682–690, May 1963.

Shihua Wang (S'89–M'92) was born in Shanghai, China. She received the B.S. and M.S. degrees in electrical engineering from the Northwest Telecommunication Engineering Institute, Xi'an, China (now renamed Xidian University) in 1982 and 1984, respectively, and the Ph.D. degree in electrical engineering from the University of California, Santa Barbara, in 1991.

From 1984 to 1985, she was a Teaching Assistant in the Department of Information Engineering at the Northwest Telecommunication Engineering Institute, China. From 1986 to 1991, she performed research in high-quality speech coding at low bit rates, perceptual effects in speech coding, and subjective/objective speech quality assessment at the Center for Information Processing Research, Department of Electrical and Computer Engineering, University of California, Santa Barbara. Since September 1991, she has been with Teknekron Communications Systems, Inc., Berkeley, CA, working on digital cellular communications. Her current research interest is in wireless communication systems, with emphasis on speech coding.



Andrew Sekey (M'63–SM'83) studied at the Technical University of Budapest, and obtained the Ph.D. degree in electrical engineering (psychoacoustics) from the Imperial College, University of London, in 1962.

He worked on signal transmission problems at the Post Office Research Station, London, and then at Bell Laboratories in New Jersey. Between 1968 and 1974, he taught at City College, New York, NY, and at Tel-Aviv University, Israel. From 1974 to 1976, he developed an introductory self-study course in technology at Everyman's University, Tel-Aviv, Israel. Since 1976, he has been with the University of California, Santa Barbara. He also worked part-time at the Speech Technology Laboratory, Santa Barbara, from 1982 to 1983. In 1985, he founded SignalCraft, a consulting company producing multimedia instructional packages and engineering short courses. In addition to publishing over 30 papers, he was Editor of *Electroacoustic Analysis and Enhancement of Alaryngeal Speech* (Thomas, 1982), and Guest Editor of the December 1983 special issue on speech communications of the *IEEE Communications Magazine*, of which he is also Technical Editor. He authored the IEEE Individual Learning Packages *Digital Signal Processing* (now in its second edition) and *Introduction to Digital Speech Processing*, for which he was awarded the 1989 Meritorious Achievement Award in Continuing Education Activities by the IEEE Educational Activities Board.

Allen Gersho (S'58–M'64–SM'78–F'82) received the B.S. degree from the Massachusetts Institute of Technology in 1960 and the Ph.D. degree from Cornell University in 1963.

He is Professor of Electrical and Computer Engineering at the University of California, Santa Barbara (UCSB), and Director of the Center for Information Processing Research at UCSB. He was at Bell Laboratories from 1963 to 1980. His current research activities are in the compression of speech, audio, images, and video signals. He holds patents on speech coding, quantization, adaptive equalization, digital filtering, and modulation and coding for voiceband data modems.

Dr. Gersho served as a member of the Board of Governors of the IEEE Communications Society from 1982 to 1985 and is a member of the Communication Theory Technical Committee and the Signal Processing and Communications Electronics Technical Committee of the IEEE Communications Society. He served as Editor of the *IEEE Communications Magazine* and Associate Editor of the IEEE TRANSACTIONS ON COMMUNICATIONS. In 1980, he was awarded the Guillemin–Cauer Prize Paper Award from the Circuits and Systems Society. In 1983, he received the Donald McClennan Meritorious Service Award from the IEEE Communications Society, and in 1984 he was awarded an IEEE Centennial Medal. In 1987 and 1988, he received NASA Tech Brief Awards for technical innovation.