COMMITTEE T1

CONTRIBUTION

Document Number: T1Q1.5/91-134

****

STANDARDS PROJECT: Analog Interface Performance Specifications for Digital Video Teleconferencing/Video Telephony Service

****

TITLE: Technology-Independent, User-Oriented, Objective Classification of Voice Transmission Quality

****

ISSUE ADDRESSED: Objective Methods for Measuring Voice Quality of VTC/VT systems

****

SOURCE: NTIA/ITS - R. Kubichek

****

DATE: September 30, 1991

****

DISTRIBUTION TO: T1Q1.5

****

KEYWORDS: Voice Quality, Audio Performance Specifications, Objective Quality, Subjective Quality

****

DISCLAIMER:

****

Introduction

This contribution presents an overview of voice quality assessment methods being discussed and considered in T1Y1.2. The contribution is presented to the T1Q1.5 VTC/VT Sub-working Group to assist in the development of the audio performance specifications for the VTC/VT draft standard.

Committee T1Y1

T1Y1/91-083
T1Y1.2/91-015R2
T1Y1.2/90-061R1
July 31, 1991

Contribution to Working Group T1Y1.2

Project: T1Y1 20 / Objective Assessment of Voice Quality

Title: Technology-Independent, User-Oriented, Objective Classification of Voice Transmission Quality

Contact: R. Kubichek
ITS.N3
Institute for Telecommunication Sciences
National Telecommunications and Information Administration
U. S. Department of Commerce
Boulder, CO 80303
Phone (303) 497-3594
FAX (303) 497-5323

TABLE OF CONTENTS

ABSTRACT

This report examines techniques for automatic assessment of voice transmission quality. Such an "objective" measurement system could augment or replace the use of "subjective" listener panels in many situations. Potential benefits include improved procurement of voice systems and services, reduced equipment development costs, and wide accessibility of the system to users who otherwise might not afford extensive use of listener scoring.

Traditional objective systems have used analog-based parameters such as loss, noise, and talker echo to estimate transmission quality. Although effective for analog transmission systems, these models are not applicable to new digital voice technologies. Objective voice parameters such as cepstral distance, information index, and coherence function have been developed to better characterize the quality of digitally transmitted speech. This report summarizes the current state of objective voice quality assessment. A survey of objective parameters is presented, as well as a new pattern recognition-based method to estimate quality based on multiple objective parameters. The report includes a synopsis of recent test results comparing four objective assessment systems being considered by CCITT. The four methods are the Cepstral Distance (NTT, Japan), Information Index (France), Coherence Function (Bell Northern Research, Canada), and Pattern Recognition (NTIA, United States).

Each of these methods has produced excellent results for certain test cases. However, no method has demonstrated consistently good assessment accuracy across a wide variety of degraded speech, and algorithm performance appears to depend on the voice technology being used. Thus, as long as the scope of the test can be confined to speech degradations for which the objective measure is known to be effective, objective assessment techniques can be used reliably and accurately . A truly technology-independent objective measure remains a goal for future research.

1. Introduction

1.1 Motivation

1.1.1 Compelling Need for Objective Voice Quality Assessment

This Nation's once-unified telephone network is rapidly evolving into a collection of technologically-diverse, independently-operated, competing public and private networks. Effective interoperation of the evolving networks under a wide range of conditions is essential. It will require that network planners and operators have a common means of rapidly, accurately, and automatically assessing voice transmission quality as perceived by end users.

Effective procurement of equipment and services relies on using accepted performance measures to accurately compare competing products. Voice quality is an important measure of performance for voice equipment and services. Quality measures based on listener scores such as Mean Opinion Score (MOS) or Diagnostic Acceptability Measure (DAM) are well known. Unfortunately, the time and expense associated with designing and conducting these tests limit their wide application. As a result, procurement decisions may be based on criteria unrelated to end user satisfaction. The rapidly expanding market for voice transmission related services and equipment highlights an urgent need for a standard objective voice quality measure.

An important function of domestic and international standards organizations is to define standards for new and existing technologies in voice coding and transmission. Selecting a

single algorithm from among several candidates for standardization presents a very difficult task. Though listener scores are used successfully as a measure of algorithm performance, designing and implementing an acceptable test plan can strain the resources of member organizations. An accurate and automatic standard measure of voice quality could speed development time and reduce the costs of creating new standards. While current techniques may not be sufficiently accurate to replace all listener panel scoring, they can be extremely useful in the design and execution of subjective listener tests.

1.1.2 Benefits Expected From Such A Technology

Convenient access to a standard objective voice quality measure would furnish providers of voice equipment and services a tool to maximize their products' ability to compete in the marketplace. When listener scores are the sole means of gauging performance, the expense and time required for testing can impede product development. Relative low cost and fast turnaround of an automated measurement system could make voice quality assessment accessible to a wide range of users who currently cannot afford the costs associated with listener scoring.

Cost effective design and implementation of new voice technologies will benefit from efficient and accurate objective methods of assessing voice transmission quality. A voice quality measure can be used during the design phase of a product as an optimization parameter, i.e. the design can be iterated to achieve the best voice quality. Similarly, these system parameters can be adjusted during implementation and operation to maintain a consistent high level of voice quality performance.

Finally, as mentioned earlier, procurers of voice communication equipment and services would benefit from the availability of a *standard* objective performance measure. The standard would facilitate comparison of competing products in light of individual communication needs. Since it appears that objective measures currently lack the reliability of subjective test scores, such a standard would have to carefully define the scope of application to avoid misinterpretation of the objective scores.

1.2 Project Goals

1.2.1 Assess Current Objective Techniques

Objective parameters for voice quality estimation have been proposed and studied by numerous researchers. While some of the parameters were created specifically for voice quality assessment, many were originally developed in related fields such as speech recognition, speech coding, and speaker recognition. Excellent surveys of objective parameters are given by [1]-[5], [50] and are summarized in this report. In addition, four objective assessment methods being considered by CCITT SG XII for standardization are described in detail. These algorithms are compared using test results from a recent CCITT test which included a variety of impairment conditions and languages.

No widely accepted method exists for quantifying the performance of objective assessment techniques. Before performance goals can be set for objective techniques, the issue of performance measurement must be addressed. This report therefore proposes methods to measure objective assessment accuracy and suggests targets for acceptable performance.

The performance goals for objective assessment systems depend on their intended role. If the objective method is desired as a replacement for human listener panels, very stringent requirements must be set. On the other hand, alternate uses such as augmenting subjective test design may not demand such close precision. Several alternative uses for objective assessment are introduced.

Acceptable agreement between listener scores and objective quality scores must be demonstrated before the algorithm is useful. An objective of this project is to benchmark the performance of proposed algorithms and determine the limits of their applicability. Results of applying objective methods to three different speech databases are presented to determine how well current methods perform.

1.2.2 Advance the Technology Where Needed

Some proposed objective measures have shown good correlation with subjective scores for a variety of distortions but require further study to determine their effectiveness over a wide variety of conditions and impairments. To achieve sufficient robustness and accuracy, it may be necessary to use a multivariate approach integrating these and other parameters into a single framework. This would exploit the best features of each objective measure to obtain maximum assessment performance. Regression techniques are the best known approach to this problem, but have inherent disadvantages. These include the requirement for a-priori knowledge of the regression model, and inability to adjust automatically to new parameter sets. Section 3 describes a pattern recognition-based technique developed by NTIA that applies Bayes estimation techniques for improved multivariate assessments which avoids these problems.

1.2.3 Alternate Applications

The most obvious use for objective assessment is to replace subjective listener panels. Questions about reliability may limit this application. However, a number of alternate applications are suggested which may provide many more significant near-term benefits. In these applications, the objective measure would be used as a tool to aid in subjective test design and verification, as well as in the development of source speech material.

1.3 Synopsis of Results

This report discusses the motivation for objective voice quality assessment, and many of the proposed methods for accomplishing it. Many objective voice quality parameters have been proposed in the literature and are summarized in this report. The study by Barnwell et al [4]-[5] suggests, however, that most are not effective predictors of subjective quality when applied to broad classes of voice impairments. A few methods have shown promising results, however. One of the best simple objective measures is cepstral distance (CD). More complex algorithms use perceptual weighting functions in time or frequency to yield more accurate voice quality predictions. They include the information index, coherence function, and Bark spectral distortion measure.

Much of the focus of this report is on the four objective methods being considered by CCITT Study Group XII: cepstral distance, information index (II), coherence function (CF), and pattern recognition (PR). Each of these methods has produced excellent results for certain test cases. However, no method has demonstrated consistently good assessment accuracy across a wide variety of degraded speech. In other words, algorithm performance appears to depend on the voice technology being used. Thus, as long as the scope of the test is confined to speech degradations for which the objective measure is known to be effective, objective assessment techniques can be used reliably and accurately. A truly technology-independent objective measure remains a goal for future research.

Specific conclusions are summarized as follows:

- Methods to measure the performance of objective methods are needed to provide a meaningful basis for comparison. The squared correlation coefficient, ρ^2, has been used in the past as a measure of performance. To help clarify the meaning of ρ^2, a

measure of root mean-squared error (RMSE) is described in Section 4.2 which is directly comparable to the listener panel standard deviation values used to quantify subjective test precision. Under a few general assumptions, the correlation of objective assessments with subjective scores must have values of ρ^2=.95 or more to achieve accuracy comparable to a subjective test using 20 listeners.

- The cepstral distance, information index, coherence function, and pattern recognition methods have been tested on a variety of speech databases in an effort to measure objective assessment performance. Although somewhat inconclusive, the results can be summarized as follows: For the CCITT database containing North American English (NAE), Italian, and Japanese speech, the CD and II methods provided best results with typical ρ^2 values in the range of .85 to .95. For the Comsat and Bell Labs data (containing only English speech), the PR method performed best with ρ^2 of .8 to .9. When the PR method is trained using a small subset of the NAE data, correlations of .94 to .97 are achieved for 3 out of 5 languages. This assumes a parameter set consisting of the CD and II. Therefore, in tests where the impairments are related to the training data conditions, the PR approach may offer the accuracy needed to replace subjective testing.
- In addition to its potential role of replacing listener panels, three alternative applications for objective assessment methods are presented. These include assisting subjective test design, validating subjective test results, and guiding the design of source speech material. Significantly, these applications do not demand the high level of accuracy required for subjective testing, and represent immediate and important uses for currently existing objective measures.

It is the conclusion of this report that current objective assessment techniques may not possess the required precision or dependability to replace human listener panels in all cases. However, important alternative applications can immediately benefit from objective testing. It may be appropriate, therefore, to pursue a standard for objective voice quality assessment to address these alternative applications. Indeed, CCITT Study Group XII is likely to develop a recommendation consisting of some or all four of the proposed algorithms by the end of the Study Period, 1988-1992. Since the developers of all four assessment systems actively participate in CCITT SG XII, it is recommended that T1 closely monitor the ongoing work at CCITT in this area and participate in the development of international recommendations for objective voice quality.

2. Objective Voice Parameters

Objective parameters are sought that are useful for a broad range of voice applications. Ideally, the parameters should not make excessive computational demands, yet should correlate with human perception of quality for all distortions likely to be found in the communication system being tested. Different communication systems may call for different objective parameters to achieve best results. As mentioned earlier, a large number of parameters have been proposed and described in the literature. The parameters described below have been extracted from a variety of fields related to speech processing, voice compression, and speaker recognition. Surveys of these can be found in [1]-[5], [17], [24], and [50].

2.1 LPC Parameters

Linear Predictive Coding (LPC) coefficients are important in areas of spectral estimation and speech coding. By modeling the speech as an autoregressive process, speech frames may be represented using only LPC information. The basic model is given by:

$$y(t) = \sum_{i=1}^{m} a_i y(t - t_i) + e_i \qquad (1)$$

where y(t) is the voice signal at time t, and a_i represent LPC coefficients. Although speech can be encoded using only LPC coefficients, the quality of reproduced speech is not perfect. This indicates that some quality information is lost and that, by themselves, the LPC coefficients are not entirely adequate objective parameters. Furthermore, Pfeifer [8] showed that LPC coefficient distances yield poor results for speaker identification applications. Goodman et al [9] also reported marginal success using LPC distances for objective voice quality.

2.2 PARCOR Coefficients

Partial correlation coefficients, or PARCOR coefficients, can be derived directly from the LPC coefficients. They are equal to the negatives of reflection coefficients and have a physical interpretation as the parameters of an acoustical tube model of speech. The i-th PARCOR coefficient, k_i, also has meaning as the correlation between y(n) and y(n-i) with correlation of y(n-1) through y(n-i+1) removed.

The PARCOR coefficients are more popular for use in voice synthesis applications than LPC coefficients because of their improved numerical stability. There is ample evidence that these parameters contain important speaker dependent information not contained in fundamental frequency or gain related parameters [10].

PARCOR coefficients can be derived from LPC coefficients recursively as follows:

1) set $k_i = a_i$
2) $a_j = (a_j + a_i\, a_{i-j}) / (1 - k_i^2)$ for j = 1, 2, ..., i-1
3) Repeat steps 1 and 2 for i = 1, 2, ..., p, where p is the model order.

2.3 Cepstral Coefficients

The complex cepstrum is obtained by taking the inverse Fourier transform of the log magnitude of the Fourier transform of the speech frame. This process has the effect of separating glottal and vocal tract information from the excitation signal. In the cepstral domain, the first 3-4 ms corresponds to glottal and vocal tract effects, while the rest corresponds to pitch information. The cepstrum is used in pitch and formant estimation, and in elimination of fixed time echoes in speech signals.

The cepstral coefficients, h(n), can be derived directly from LPC coefficients [11]-[13]:

$$h(n) = a_n + \sum_{k=1}^{n-1} (k/n)\, h(k)\, a_{n-k} \quad \text{for } n>0 \qquad (2)$$

Cepstral distance measures can be computed as a difference between input and output speech cepstral coefficients. Cepstral distance equals the average distance between cepstrally smoothed log input and output spectra. The distance values are useful in speaker recognition problems and have been fruitful as voice quality parameters [2], [13]. The cepstral distance measure is further discussed in Section 2.10.1.

2.4 Log Area Ratio

Log area ratios correspond to the log of the area ratio of adjacent sections of a lossless tube model of the vocal tract [11, p440]. They display relatively flat spectral sensitivity to quantization error, that is, LPC spectrum distortion is uniform for all values of the log area

ratios [13, p131]. In contrast to reflection coefficients, which become unreliable as their magnitude approaches unity, the log area ratio is a nonlinear transformation expanding the region near k_i=1. This mapping reduces the likelihood that quantization error will move a pole onto or outside of the unit circle and cause instability. Because of this improved stability, log area ratios are sometimes preferred over other speech parameters, especially for small word size. They have not received much attention as voice quality parameters, though Barnwell and Voiers [4] stated that the best simple measure they found was a log area ratio. In addition, the log area ratio out-performed the best simple spectral distance measure in their tests.

Log area ratios are calculated from either LPC, a_i, or PARCOR, k_i, coefficients as follows [11]-[13]:

$$g_i = Log(a_{i+1}/a_i) = Log\left(\frac{1-k_i}{1+k_i}\right), \quad 1 \le i \le p \tag{3}$$

2.5 Inverse Sine

Inverse sine parameters are derived from the inverse sine of the reflection coefficients, k_i. They are similar to log area ratios in that they expand the region about k_i=1 [13, p131].

2.6 Autocorrelation Coefficients

Several parameters are obtained from the speech autocorrelation function. The log of the zero lag autocorrelation gives a measure of power in the frame. The ratio of the i-th lag to the zero lag autocorrelation measures the overall rate of decorrelation of speech samples. These time domain parameters are not generally used for voice applications, but have been used in other areas such as seismic pattern recognition [14] and myoelectric signals [51]-[52].

2.7 Spectral Measures

The human ear performs crude Fourier analysis of in-coming sound pressure waves [15]. Phase information is largely redundant and speech intelligence is carried predominately by the spectral envelope. This partially explains why frequency domain measures have better correspondence to subjective measures of quality than time domain measures. Further, among all frequency domain measures, spectral envelope parameters have been found to correspond best to perception of quality [15]. A number of spectral measures are described below.

2.7.1. Percentile Frequencies.

Percentile frequency measures describe the distribution of energy in the voice spectrum and are sensitive to the presence of vowels, consonants, and correlated noise. The parameters are derived as follows:

$$p_m = \sum_{i=1}^{m} S(f_i) \; / \; \sum_{i=1}^{N} S(f_i) \tag{4}$$

$$F25 = m \cdot \Delta f \quad such\ that\ p_m = .25$$
$$F50 = m \cdot \Delta f \quad such\ that\ p_m = .50$$
$$F75 = m \cdot \Delta f \quad such\ that\ p_m = .75$$

where N is the number of frequency values, S(f_i) is the power spectrum and F25, F50, and F75 are the resulting percentile frequency measures.

2.7.2. Average Power Weighted Frequency

A measure of the central mass of spectrum energy is given by:

$$AVPWF = \sum_i f_i \cdot P(f_i) \; / \; \sum_i P(f_i) \tag{5}$$

2.7.3. Spectral Flatness Measure

The degree of spectral flatness is an indicator of the maximum theoretical performance of redundancy removing coders [15]. The more peaky the spectrum is, the more predictable and redundant is the speech; flat spectra indicate little predictability or redundancy. Coders based on removing this redundancy will perform best for speech with non-flat spectra. A measure of spectral flatness is given by:

$$SFM = 1/N \cdot \sum_{k=1}^{N} S^2(f_k) \; / \left[\prod_{k=1}^{N} S(f_k) \right]^{1/N} \tag{6}$$

which is a ratio of the arithmetic mean to the geometric mean. Larger values of SFM correspond to vowels, while values near unity imply flat spectra such as occur with unvoiced fricatives.

2.8 Residual Measures

Examination of the LPC residual provides information about the adequacy of the LPC model. The residual, e(i), results from applying an inverse LPC filter to the data in each speech frame:

$$e(i) = \sum_k a_k \cdot y(i-k) \tag{7}$$

where y(i) is the speech signal. In some cases, the LPC model may not be adequate to fully represent the data y(i). This happens, for example, when model order is chosen too low and results in a correlated residual series. If the model is adequate to represent the data, the residual will in general be white Gaussian noise. Residuals with non-Gaussian statistics may indicate the type and degree of distortion. Potential voice quality parameters based on the residual signal include spectral flatness, average power weighted frequency, and percentile frequencies.

The kurtosis of the residual signal is based on the fourth central moment as follows:

$$kurtosis = E\left\{(e(i) - \overline{e(i)})^4\right\} / \left\{E\,(e(i) - \overline{e(i)})^2\right\}^2 \tag{8}$$

where R and R_m represent the residual and its mean, and E{·} is the expectation operation. Kurtosis is always 3.0 for Gaussian data and usually 4.0 or larger for voiced speech - this indicates a "peaked" distribution with long tails. For unvoiced speech, the residual's distribution is closer to Gaussian with kurtosis of 3.0 or less [16]. The presence of impulsive type noise causes kurtosis values to increase significantly beyond 3.0 due to outliers in the residuals. It is interesting that standard solutions of the LPC model assume Gaussian residuals. Hence, kurtosis is indicative of LPC coefficient accuracy. The change of kurtosis between the input and output signal provides a measure of distortion.

2.9 Distortion Measures

Distortion (or "distance") parameters directly measure the amount and type of distortion between input and output speech signals. They are calculated as a difference or ratio of

input and output speech parameters. Traditionally, these have been the focus of most research regarding objective parameters for speech quality.

Although distortion parameters have received the most attention, a case can be made for combining distortion parameters with parameters based purely on input or output speech. For example, a useful parameter set might combine an amplitude level measure (based on the input speech record) and a spectral distortion measure (derived from both the input and output speech records). This combination would allow the algorithm to give different weighting to spectral distortion of high amplitude frames, such as vowels, than to quieter frames such as unvoiced fricatives.

2.9.1 L_p Distances

Euclidean distances ("L_2 distances") have received much attention because of their convenient physical interpretation. A more general distance measure is the L_p norm distance. Barnwell and Voiers [4] studied a broad range of parameters using several values for p and determined that, for some conditions, p=8 provided the best assessment accuracy. Others [17] argue that, due to high correlation between Euclidean distance (L_2) and general L_p distances, use of p≠2 may not provide any significant advantage.

Parametric L_p Distances

Two types of parametric distances are described in [3]-[5], [17]. A "linear feedback" measure is given by:

$$d_1 = \left[1/m \cdot \sum_{i=1}^{m} |py(i) - px(i)|^p \right]^{1/p} \qquad (9)$$

where py(i) and px(i) represent output and input time domain parameters, and m is the order of the parameter model. Parameter sets used for this distance measure could include LPC, PARCOR, log area ratios, cepstrum, inverse sine, autocorrelation ratios, and others.

When cepstral coefficients are used in (9) with p=2, the distance measure has special importance. It can be shown [17] to be directly related to the log of the difference of the cepstrally smoothed spectra. A very efficient means of computing the log spectral distance without estimation of the spectrum results from using LPC derived cepstral coefficients. More importantly, studies show that this measure is an effective parameter for both speaker recognition and objective voice quality prediction [2],[17]. Cepstral distance is one of the objective parameters considered by CCITT and is discussed further in Section 2.10.1.

A "log feedback" measure is given by:

$$d_2 = \left[1/m \cdot \sum_{i=1}^{m} |20Log|py(i)/px(i)||^p \right]^{1/p} \qquad (10)$$

As with the linear feedback measure, LPC, PARCOR, log area ratios, cepstrum, inverse sine, and autocorrelation ratios can all be used in computing these distortion measures.

Spectral Distances

As already mentioned, most perceptual information is contained in the spectral envelope. Measures of spectral distance have been more effective in applications of speaker recognition and objective voice quality than time domain measures.

A "linear unweighted" L_p norm measure is given by:

$$D_1 = \left[1/L \cdot \sum_{i=0}^{L-1} [S_{yi} - S_{xi}]^p \right]^{1/p} \tag{11}$$

where S_{yi} and S_{xi} represent the output and input power spectra at the i-th frequency, L is the number of frequency points, and p is the distance norm. A "linear frequency weighted" measure gives more weight to distortion in the regions of greater spectral energy:

$$D_2 = \left\{ \frac{\sum_{i=0}^{L-1} S_{xi} \cdot |S_{yi} - S_{xi}|^p}{\sum_{i=0}^{L-1} S_{xi}} \right\}^{1/p} \tag{12}$$

"Log unweighted" and "frequency weighted log" measures are given in (13) and (14):

$$D_3 = \left\{ 1/L \cdot \sum_{i=0}^{L-1} |20 \cdot Log[S_{xi} / S_{yi}]|^p \right\}^{1/p} \tag{13}$$

$$D_4 = \left\{ \frac{\sum_{i=0}^{L-1} S_{xi} \cdot |20 Log[S_{xi} / S_{yi}]|^p}{\sum_{i=0}^{L-1} S_{xi}} \right\}^{1/p} \tag{14}$$

Banded spectral distance parameters can be developed using equations (11)-(14) with a preselected range of frequency indices. Such measures are sensitive to spectral envelope distortion in different bands and can be weighted according to the perceptual importance of individual bands. They can be used singly or in a multivariate scheme to estimate voice quality.

Bark Spectral Distortion

A related objective method has been developed by Wong et al at the University of California at Santa Barbara [55]. Prior to computing the spectral difference, the input and output spectra are perceptually weighted according to a model of the human auditory system. The method takes into account the human ear's nonlinear sensitivity to frequency and amplitude. The Euclidean distance of the transformed spectra provides an objective measure much more relevant to perceived quality than the untransformed spectrum. The analysis involves the following steps:

1. The speech is processed through a filter bank composed of 15 critical bands whose center frequencies and bandwidths were designed to mimic frequency resolution and masking characteristics of human hearing. Essentially, the filter bank produces a Bark transformation to create a frequency-warped 15 point spectrum.
2. Each element of the 15 point spectrum is weighted to adjust for frequency-dependent hearing sensitivity.

3. The resulting weighted spectrum is transformed from loudness level (in phons) to loudness (in sones). This additional nonlinear weighting function accounts for the fact that the amount of change in perceived loudness depends on the loudness level itself.

An objective voice quality parameter, called the Bark spectral distortion (BSD), is found by computing the Euclidian distance between the resulting 15 point vectors from the input and output speech. The BSD is then mapped to a MOS scale using quadratic regression. Excellent correlations of between .92 to .98 were achieved for speech coding rates of 2.4 to 64 kbit/s.

2.9.2 Signal to Noise Ratio

Signal-to-noise ratio (SNR) has long been used as an objective parameter for measuring voice quality. It can be computed as:

$$SNR = 10\ Log_{10} \frac{\sum_i x(i)^2}{\sum_i \left[y(i) - x(i)\right]^2} \tag{15}$$

Although SNR is a useful measure of degradation for many types of analog voice transmission, numerous studies have shown its performance is poor for modern digital communication techniques - especially low to medium rate voice codecs [9],[17]-[20]. A better parameter, commonly referred to as segmental signal-to-noise ratio (SNR_{seg}), is found as the average of SNR values computed over speech frames. SNR_{seg} provides better performance than SNR, but is not considered a reliable measure of voice quality for modern digital communications systems.

SNR is more difficult to use than one would expect. The SNR calculation requires accurate normalization of input and output speech levels using knowledge of the system gain. If the output signal is corrupted by added noise or non-linear distortions, accurate measurement of gain may not be possible.

The SNR measure has other serious problems, including sensitivity to delay estimation error and phase distortion. The delay between input and output must be measured and removed before extracting distortion parameters. If the delay error is on the order of one half of the dominant period, lower than expected SNR estimates could result even though the actual SNR is quite high. This is because output and input signals are nearly 180° out of phase (resulting in SNR as low as -6 dB, even with no distortion present). Additionally, SNR-based measures are ineffective for coding techniques which do not attempt to replicate the speech waveform, e.g., RELP or CELP coders.

2.9.3 Itakura Distance

Another important distance measure is the Itakura likelihood ratio [1],[4],[15],[22]-[24]. This measure is useful for speaker identification applications, and has been considered as an objective voice quality parameter. The likelihood ratio is defined as the ratio of two energy terms. The numerator is the residual energy of the output voice when processed by an inverse LPC filter determined from the input voice. The denominator is the residual of the output voice filtered by the inverse LPC filter determined from the output. The ratio can be written compactly as:

$$LR = \boldsymbol{a}_1^T \cdot R_2 \cdot \boldsymbol{a}_1 \,/\, \boldsymbol{a}_2^T \cdot R_2 \cdot \boldsymbol{a}_2 \tag{16}$$

where $\mathbf{a}_1$ and $\mathbf{a}_2$ are vectors of LPC coefficients measured from input and output speech frames, respectively. R_2 is the autocorrelation matrix measured from output speech frames.

When LR is unity, no distortion is present. A value of LR = 1.4 is a known threshold above which perceived changes between the input and output speech is significant.

2.9.4 Coherency Based Measures

A class of spectral distance measures can be developed from the squared coherency function. Bell Northern Research (BNR) has reported on the effectiveness of coherency measures for objective voice quality measurement [25], [32], [33]. Coherency is sensitive to nonlinearities and added noise in the channel, and is computed as:

$$\gamma^2(f) = \frac{|S_{xy}(f)|^2}{S_x(f)S_y(f)} \approx \frac{|\sum_n X_n^*(f)Y_n(f)|^2}{\sum_n |X_n(f)|^2 \sum_n |Y_n(f)|^2} \tag{17}$$

where $S_{xy}(f)$ is the cross-power density spectrum, $S_x(f)$ and $S_y(f)$ are the auto spectra of the input and output signals, and $X_n(f)$ and $Y_n(f)$ are FFTs of the n-th data frame of the input and output speech respectively. The function $\gamma^2(f)$ plays the role of a correlation coefficient defined at each frequency f_i. Objective parameters are obtained from coherency by using linear unweighted and linear frequency weighted distances as described in (13) and (14), with $\gamma^2(f)$ replacing the arithmetic difference. Banded coherency measures can also be obtained as described earlier.

A signal-to-distortion ratio measure (SDR) is computed as the ratio of coherent to non-coherent power in the output signal:

$$SDR(f)_{dB} = 10Log_{10}\frac{G_C(f)}{G_N(f)} \tag{18}$$

where coherent power, $G_C(f)$, and non-coherent power, $G_N(f)$ are defined as

$$\begin{aligned} G_C(f) &= \gamma^2(f)\,|S_y(f)|^2 \\ G_N(f) &= [1-\gamma^2(f)]\,|S_y(f)|^2 \end{aligned} \tag{19}$$

A number of voice quality parameters can be derived from SDR(f). For example, SDR_j, is the average signal-to-distortion ratio in the j-th band, b_j

$$SDR_j = \frac{1}{N_j}\sum_{f \in b_j} SDR(f) \tag{20}$$

2.9.5 Cross Residual Signal

The cross residual signal is created by filtering the *output* speech with an inverse LPC filter determined from the *input* speech. Assuming the LPC model adequately represents the input speech, the cross residual signal will be approximately white if the types of distortions present in the channel do not significantly change the voice statistics. Nonlinearities or correlated noise added to the signal should make the cross residual measurably non-flat. Potential voice parameters based on the cross residual include the spectral flatness measure, percentile frequencies, and average power weighted frequency.

2.10 Parameters Being Considered by CCITT

Three objective voice parameters are being considered by CCITT SG XII for possible standardization. They include the Cepstral Distance (NTT, Japan), the Information Index

(France), and the Coherence Function (BNR, Canada). A fourth technique being considered, NTIA's PR method, is not an objective parameter, but rather is a method for selecting and mapping objective parameters to predicted quality - it is discussed in Section 3.

2.10.1 Cepstral Distance (NTT)

The Cepstral Distance (CD) parameter proposed by NTT is a measure of overall difference between and output voice cepstra. Since a finite number of cepstral coefficients are used, CD is also the log difference of the cepstrally smoothed input and output spectra. It is defined as follows [1], [2], [26]-[29]:

$$CD = \frac{10}{Log_e 10}\sqrt{2\cdot\sum_{i=1}^{m}[C_x(i)-C_y(i)]^2} \qquad (21)$$

where $C_x(i)$ and $C_y(i)$ are the i-th cepstral coefficients of the input and output speech (derived from LPC coefficients as described in [11]), and m is the number of coefficients computed (currently 16). Before estimating C_x and C_y, the speech data is processed by a first order difference operation to emphasize high frequency information. This practice is sometimes used in LPC speech analysis to provide better representation of high frequency formants. In effect the differencing operation removes an assumed pole on the unit circle. Of course, not all speech frames possess these poles; however the effect of such frames on the cepstral distance has not been reported.

NTT has developed quadratic regression formulae to map CD values into predicted MOS, but they emphasize that these may not be valid for all types of data. Formulae for North American English, Italian, and Japanese are shown below [53].

$$\begin{aligned} MOS &= 3.70 - 0.05CD - 0.09CD^2 \quad (N.A.English), \\ MOS &= 3.05 - 0.19CD - 0.03CD^2 \quad (Italian), \\ MOS &= 3.26 - 0.39CD - 0.02CD^2 \quad (Japanese). \end{aligned} \qquad (22)$$

An alternative to using regression equations is to convert the CD values to an equivalent stationary noise level. This value is input to NTT's objective estimation model OPINE (Overall Performance Index model for Network Evaluation) to produce MOS estimates. When applied solely to a non-linear device and not a network, the former method is usually adopted.

2.10.2 Information Index (France)

The Information Index (II) was developed in France by J. Lalou, and accounts for multiplicative noise and types of distortion found in digital systems [30], [31]. The method may be used to compare different systems directly when such distortions are the main factors affecting transmission performance. The auditory system is modeled by dividing the spectrum into 16 critical bands, and applying empirical frequency weights and hearing thresholds for each bands. The basic form of the method is outlined here. See [31] for details about the theoretical development of the method and weighting functions used. See also [30] for additional details on implementation.

The signal-to-distortion ratio (SDR) in the i-th frequency band, denoted QS(i), is computed first:

$$QS(i) = 10\ Log_{10} \frac{\sum_{j \in b_i} |X(f_j)|^2}{\left| \sum_{j \in b_i} |X(f_j)|^2 - \sum_{j \in b_i} |Y(f_j)|^2 \right|}, \qquad (23)$$

where j ranges over all frequencies specified for the i-th band, b_i. Here, X(f) and Y(f) are Fourier transforms of a given input and output speech frame, and frequency bands b_i are tabulated. Treating the bands as separate, independent channels, the mutual information (i.e., the maximum channel capacity) of each band is computed, weighted, and summed to form an overall Information Index:

$$RII = \sum_{i=1}^{16} W_2(i) \cdot \frac{3}{0.1 + 10^{-[\overline{QS}(i) + W_1(i)]/10]}} \qquad (24)$$

where $\overline{QS}(i)$ is the average of QS(i) over all frames, and $W_1(i)$ and $W_2(i)$ are tabulated weighting functions accounting for critical bandwidth and perceptual importance of the i-th frequency band, respectively. Equation (24) is correct for MNRU noise, however, for typical digital systems an additional correction is used. The term $\overline{QS}(i) + W_1(i)$ in (24) is replaced by the following:

$$\begin{aligned}
&for\ V < -3.57: \\
&\quad V + d\tanh[0.07984(V) - 0.356325], \\
&for\ -3.57 < V < 0: \\
&\quad 4.3429\ln[exp\{0.23026(V+5.15)\} - 1] + d[0.276V + 0.3859], \\
&for\ V > 0: \\
&\quad 4.3429\ln[exp\{0.23026(V+5.15)\} - 1] + d\tanh[0.062715V + 0.3109255], \\
&where\ V = \overline{QS(i)} + W_1(i).
\end{aligned} \qquad (25)$$

Here, d is a correction factor, with d=0 for PCM and other digitized voice and d=-5.33 for natural voice. This correction is not an ad hoc correction from opinion tests, but results from communication theory and statistical properties of voice.

MOS is estimated from RII using the following mapping:

$$\begin{aligned}
RIT &= Log_e\left[\frac{RII}{27.6 - RII}\right] \\
YT &= 1.00356 \cdot RIT - 1.4027 \\
MOS &= \frac{3.4 \cdot e^{YT}}{1 + e^{YT}}
\end{aligned} \qquad (26)$$

The mapping functions in (26) have been developed based on a series of listener tests conducted in France. The regression equations can be adjusted for particular languages or applications to provide optimal assessment performance.

2.10.3 Coherence Function (BNR)[1]

BNR's Coherence Function (CF) provides an overall measure of signal-to-distortion between the input and output speech (see Section 2.9.5) [25], [32], [33], [57], [58]. The input and output signal are first processed to obtain the input and output spectra and the complex cross-spectrum. The power spectra and cross-spectrum are calculated using the Fast Fourier Transform algorithm operating on successive segments 256 samples long, with 50% overlap. Using a 125 μs sampling period (8000 samples/sec) each segment is 32 ms long. This is effectively further reduced by application of the Hamming window to about 22 ms. These segments are assigned to one of four quartiles of the segmental level distribution. The quartiles are produced by normalizing each input signal segment to the long term r.m.s. level and then dividing the resulting segmented level distribution into 4 parts, each containing a quarter of the segments. The average coherent (CP) and non-coherent power (NCP) for each quartile is computed as discussed in Section 2.9.5.

The next step is to normalize the coherent power to the "preferred" level (i.e., 82 dB rel. 20 μPa) Using power addition, the non-coherent power spectrum ("noise spectrum") is combined with the hearing threshold for continuous spectrum sounds to form a new masking noise spectrum (MNS). The hearing threshold is given in Table 1. From that, the sensation level Z is found by subtracting coherent power, CP, from masking power, MNS. this is then transformed, at every given frequency, to an additive index P(Z) using the modified growth functions:

$$Z<2.792\ dB: \quad P(Z) = 10^{(Z-6.646)/10}$$
$$Z\geq 2.792\ dB: \quad P(Z) = \left\{1-10^{(Z+0.5)/10}\right\}^{-0.7} \qquad (27)$$

TABLE 1
PARAMETERS OF THE LISTENING OPINION MODEL

Frequency f (Hz)	Hearing Threshold Bo-K (dB rel 20 μPa / Hz)	Frequency Weighting 10 Log B^a (dB)
100	+17.5	-35.8
200	+ 5.0	-34.2
300	0.0	-33.3
400	- 3.0	-32.9
500	- 5.0	-32.9
600	- 6.0	-33.0
800	- 8.0	-33.5
1000	- 9.0	-34.0
1250	- 8.5	-34.7
1600	- 8.0	-35.7
2000	- 9.0	-37.3

[1] Much of the text for this section has been taken directly from [57] and [58] with the kind permission of Bell Northern Research.

2500	-11.5	-39.4
3000	-14.0	-41.3
3500	-13.5	-42.9
4000	-13.0	-44.0
5000	-12.5	-45.5
6000	-11.5	-46.7
8000	- 9.0	-48.2

The product (or sum if expressed in decibels) of P(Z) and the frequency weighting factor B^a (where 10 $Log_{10}B^a$ is given in Table 1) is then integrated over the relevant frequency range to obtain the listening opinion index (LOI). In practice, assuming that the sensation level Z is approximately constant within suitably chosen narrow frequency bands, the integration is replaced by a summation of products: $B^a \bullet P(Z) \bullet \Delta f$.

Up to this point, all calculations were done separately for each individual quartile, yielding four values of listening opinion indices. Since these indices are assumed to be additive, they can be averaged using the following weighting factors:

0.19 - for the lowest quartile
0.21 - for the second quartile
0.53 - for the third quartile
0.07 - for the highest quartile

The final LOI may be transformed into the MOS using the following modified relationship:

$$MOS = \frac{1+5e^x}{1+e^x} \tag{28}$$

where

$$x = 1.145 Ln \frac{LOI}{0.885 - LOI} - 1.195 \tag{29}$$

2.11 Models for Voice Quality

Network-oriented models have been developed to assess quality assessment of voice transmissions over the switched public telephone network [34]-[45]. Examples include AT&T's loss, noise, and echo model (LNE), British Telecom's (BT) CATNAP model, and NTT's OPINE model. AT&T's LNE model and BT's CATNAP model are briefly described here to exemplify this approach.

The AT&T LNE model described by Cavanaugh, Hatch, and others (see [34] for example) examines circuit parameters such as loss, noise, echo path delay, and echo path loss. Experiments conducted by Bell Laboratories beginning in 1965 produced extensive listener satisfaction data for a broad range of network conditions. This information was normalized and combined using a single set of equations fitted to the data for each circuit parameter. The model predicts the Grade-of-Service as well as the percent of listeners rating the system as "Good" or "Excellent" and the percent rating the system as "Poor" or "Unacceptable".

British Telecom's CATNAP, and its earlier version CATPASS, models circuit loss, circuit noise, room noise, quantizing noise, attenuation-frequency distortion, and sidetone paths.

The CATNAP model requires overall sensitivity-frequency characteristic of each transmission path and of the sidetone path, the noise level spectrum at the listener's ear, the average speech spectrum, and the average threshold of hearing. With these values, CATNAP predicts loudness judgements, listening-effort scores, conversation-opinion scores, and vocal levels. In addition, the model predicts the Grade-of-Service and percent Good-or-Better and Poor-or-Worse.

Because these models (LNE, CATNAP, and OPINE), are strictly based on listener satisfaction data for conditions found in the public telephone network, they are not generally applicable to modern voice processing techniques such as low bit-rate coding. In addition, their use requires detailed knowledge of transmission channel parameters that may not be relevant to other applications such as digital mobile radio or voice codecs.

3. Pattern Recognition Based Assessment (NTIA)

Typically, regression curves are used to represent the relationship between the objective voice parameter and estimated quality. Estimates based on multiple voice parameters offer the potential of increased accuracy and robustness by taking advantage of the best features of each parameter. Parameters (x_i) can be combined using a multiple regression formula:

$$MOS = a_1 x_1 + a_2 x_2 + \ldots \quad (30)$$

Higher order terms can be added to account for nonlinear relationships. This approach has been studied by Barnwell et al [5] who found that significant improvement in voice quality estimation was possible in some cases.

Disadvantages of regression techniques include the necessity of knowing the form of the regression equation a-priori or determining it interactively by trial and error. In the event the data or the parameter set changes, a new regression equation must be found. An accurate regression equation becomes even more difficult to construct when multiple parameters are desired for increased accuracy and robustness.

The pattern recognition method (PR) is an alternative to regression, and uses Bayesian estimation to seek a nonlinear relationship between the parameters and objective quality. An advantage of this approach is that no explicit model is required a-priori as is the case for regression analysis. The non-linear model is determined automatically during training.

3.1 Requirements and Assumptions

3.1.1 Source Speech

Ideally, the source speech records will consist of male, female, and children (if appropriate) voices reciting preselected sentences. For best results, the same ensemble of source speech records should be used both during the training phase and for objective testing. This will allow development of an objective parameter set and training statistics that are tuned specifically for the test signal. In typical test scenarios, however, it is not often possible to use the same source speech in the evaluation as was used for training. Very good performance can still be achieved by the method in these cases.

3.1.2 Equipment Requirements

Any of the objective systems described in this report are simple enough to be implemented on a desktop computer. Since there is seldom a requirement for real time voice quality assessment, using several minutes of computer time to process each speech record is usually acceptable. The result is still inexpensive and fast relative to the alternative of subjective listener tests.

Collecting data for testing requires equipment to inject the source speech into the test channel and to record the processed output speech. Equipment for inputting the test speech into the test channel can be any of the following:

- A high fidelity tape recorder with source voice provided on audio tape.
- A Digital Audio Tape (DAT) drive with source voice provided on DAT tape.
- A Digital to Analog Converter (DAC) interfaced to a desk top computer, with source voice provided as a digital file on a floppy disk.

Recording the processed voice can be done using similar equipment, i.e. a high fidelity tape recorder, DAT drive, or Analog to Digital Converter (ADC) interfaced to a desktop computer or workstation. Since the computer must have access to the digitized speech, recordings made by audio tape or DAT must undergo analog to digital conversion by the computer-based ADC. (Software could conceivably be written to use the DAT digital information directly, and thereby remove this requirement.)

3.1.3 Training Speech Database

A representative database of source and processed speech is required to provide statistics about the speech degradations. Processed speech records are obtained by passing the source speech through systems relevant for a particular application. For example, a training database applicable to low bit rate codecs might include source voice processed through a variety of 2.4, 4.8, 8, and 16 kbit/s codecs, as well as multiple levels of quantizing error, tandeming distortion, environmental background noise, and bit errors.

Finally, the training database must include subjective test results for all source and processed records. Subjective information for each speech record should include:

- Number of listeners
- Fraction of listeners voting in each of 5 quality classes
- Mean opinion score

Reference information characterizing the listener panel may also be useful. This would consist of MOS as a function of Q (signal-to-quantization noise ratio) as determined from subjective tests of MNRU (Modulated Noise Reference Unit) distortion[2]. The importance of this information to objective assessment is still to be determined. Eventually, however, it may be useful in developing quality assessments that can be readily compared with scores from other listener panels.

3.1.4 Data Preconditioning

A number of additional processing procedures are required for effective objective assessment. These are important for conditioning both the digital speech data and the

---

[2] The MNRU, described in CCITT Recommendation P.70 (Red Book, Volume V), is a device to simulate the effects of quantizing distortion. It can be used as a transfer standard to allow comparisons between subjective tests (using different listener panels) of similar types of codec. A transfer curve is obtained by graphing MOS verses a range of Q levels (signal to quantization noise level - in dB) for a given listener panel. The quality of a circuit condition can now be expressed in terms of a Q-rating, which is the Q value corresponding to the circuit's MOS obtained from the transfer curve.

measured objective parameter information. The following procedures are applied to digitized speech prior to objective parameter measurement:

- Estimate relative time delay between input and output speech records. This is accomplished by finding the maximum magnitude of the crosscorrelation function. The delay is used to align the input and output speech records.
- Pauses are eliminated by rejecting frames with power level 40 dB or more below the peak RMS of the signal.
- Estimate the root mean-squared (RMS) amplitude of the input and output speech. This is used to normalize the output speech to the same average power level as the input. Power normalization reduces sensitivity of the objective algorithm to simple changes in amplitude level. For implementation efficiency, this step is done after pause elimination, even though RMS estimates are slightly affected by removal of the pauses.

The following procedures are applied to parameter measurements:

- The parameter values are scanned for outliers that occasionally result from ill conditioning of numerical algorithms. The presence of outliers is not detrimental to most stages of objective assessment, but can cause problems during training by skewing the training statistics. Therefore, outliers identified during training are removed from the parameter files.
- Parameter values are normalized by removing the parameter mean and dividing by the parameter standard deviation computed over the entire training database (or a carefully selected subset). This procedure ensures that all parameters are given equal a-priori weighting. For example, the cepstral distance measure may have magnitudes of between 0 and 7.0, while some spectral distortion parameters can range from 0 to 500. Without normalization, the spectral distortion measure will have more importance in a multivariate assessment than the cepstral distance parameter, even though it may be less important for estimating quality. With normalization, both parameters will tend to be zero mean with unit power, giving neither measurement more importance.

3.2 Selecting Effective Objective Parameters

Considering the large number of objective parameters being considered, it is important to identify a small subset that will provide the most accurate quality assessment. An exhaustive search of all possible parameter combinations would normally be required to identify the best parameter subset. Unfortunately, the number of possible combinations is enormous and an exhaustive search can be ruled out due to excessive computational requirements. An alternative to the exhaustive search is a "bottom-up" search which attempts to select a good parameter set but does not guarantee optimality [46].

The algorithm begins by examining the performance of each individual parameter, and selecting one providing lowest assessment error. In succeeding steps, new parameters are added to the best set if they cause a sufficient increase in performance. Parameters can also be removed if doing so does not significantly decrease performance.

3.2.1 Bottom-up Search Algorithm

1) Initialize the current best set to the empty set.
2) Evaluate all parameters individually, identify the best one.

3) Add the best parameter to the current best set if the new performance exceeds the old performance by threshold T_{add}.

4) Remove one parameter from the current set and evaluate the modified current set. Put this parameter back into the current set. Do this for all parameters in the current set (except for any parameter just added). Remove the parameter which causes the smallest reduction in performance, as long as performance is not less than threshold T_{sub}. This step tends to remove parameters whose usefulness has disappeared due to the addition of other parameters.

5) Repeat steps 3 and 4 until no parameters are added or removed.

3.2.2 Parameter Performance Metrics

The bottom-up algorithm attempts to identify parameters minimizing some error criterion. NTIA has designed several metrics specifically to measure voice quality prediction performance. In most of their testing, however, they have used the chi-squared metric as described below.

The chi-squared error between the predicted opinion score frequencies and the known listener panel opinion score frequencies gauges the overall accuracy of predicted opinion score frequencies. Assuming there are N listeners, the chi-squared error is

$$\chi^2 = N \cdot \sum_q \{P(\omega_q \mid \boldsymbol{x}) - H(\omega_q \mid \boldsymbol{X})\}^2 / H(\omega_q \mid \boldsymbol{X}). \quad (31)$$

In (31), $P(\omega_q \mid \mathbf{x})$ is the average of $P(\omega_q \mid \mathbf{x}_i)$ computed over all parameter vectors $\mathbf{x}_i$. The term $P(\omega_q \mid \mathbf{x}_i)$ is the estimated opinion score frequency for class ω_q conditioned on $\mathbf{x}_i$. Finally, the term $H(\omega_q \mid \mathbf{X})$ represents the actual listener panel opinion score frequency for class ω_q, given speech record $\mathbf{X}$[3].

3.3 Statistical Assessment Method

Bayes assessment of voice quality is accomplished by analyzing training data from representative types of distortions. Training data consists of parameter values measured from subjectively tested speech. Quality assessment is made by comparing parameters measured from the speech under test, i.e., of unknown quality, to the training data statistics. The following section describes these steps in detail.

3.3.1 System Training

The training database consists of processed speech records for N_d different types of impairments. Speech records are divided into 32 m.s. frames (256 samples taken at an 8 kHz sampling rate) and processed to create parameter measurements. The result is a set of parameter vectors (one vector per frame) for each distortion type.

The parameter conditional probability density function (cpdf) must be estimated for each distortion. The conditional probability density function of $\mathbf{x}_i$ for the m-th distortion can be estimated using a k-nearest neighbor method [48]:

$$p(\boldsymbol{x}_i \mid d_m) = (k-1) \,/\, (N \cdot v(\boldsymbol{x}_i)) \quad (32)$$

where i is the frame number, $\mathbf{x}_i$ is a vector of parameter measurements, N is the total number of frames per distortion, and $v(\mathbf{x}_i)$ is the volume of a hypersphere with radius equal

---

[3] To further clarify this notation, $\mathbf{x}_i$ is a parameter vector measured from the i-th frame of the speech record $\mathbf{X}$.

to the distance from $\mathbf{x}_i$ to the k-th nearest vector belonging to distortion d_m. The method makes heavy demands on system memory and is computationally intensive, but is completely automatic.

A quick multi-modal estimate of $p(\mathbf{x}_i \mid d_m)$ can be formed by modeling the density as a Gaussian mixture. K-means cluster analysis [49] is used to identify clustering in the parameter data for each distortion. Output of cluster analysis consists of mean vector $\mathbf{xm}_{mc}$ and covariance matrix $\mathbf{C}_{mc}$ for the m-th distortion and c-th cluster. The Gaussian mixture cpdf estimate is formed by fitting a Gaussian function to each cluster and forming a weighted sum of these functions:

$$p(x_i \mid d_m) = \sum_{c=1}^{N_m} \frac{N_{mc}}{N_m} \frac{1}{(2\pi)^{p/2} |C_{mc}|^{1/2}} exp\left\{ -\frac{1}{2}(x_i - xm_{mc})^T C_{mc}^{-1} (x_i - xm_{mc}) \right\} \quad (33)$$

where p is the number of parameters (dimensions) in vector $\mathbf{x}$, N_{mc} is the number of vectors assigned to the c-th cluster of the m-th distortion, N_m is the number of clusters in the m-th distortion, and N is the total number of training vectors. Typically, (32) is used during feature evaluation and selection, while the Gaussian mixture (33) is used to design the classifier.

The probability of distortion d_m is given by Bayes rule

$$P(d_m \mid x_i) = p(x_i \mid d_m) \cdot P(d_m) \; / \sum_{j=1}^{N_d} p(x_i \mid d_j) \cdot P(d_j) \quad (34)$$

where $P(d_m)$ is the a-priori probability of the m-th distortion and N_d is the number of distortions.

3.3.2 Opinion Score Probability

An estimate of the opinion score probability function $P(\omega_q \mid \mathbf{x}_i)$ can now be obtained. This is the probability of opinion score ω_q, where the classes (q) range from 1 (unacceptable) to 5 (excellent). This function can be interpreted as the predicted frequency of listener panel scores corresponding to test speech parameter vector, $\mathbf{x}_i$. The relationship is given by

$$P(\omega_q \mid x_i) = \sum_m P(d_m \mid x_i) \cdot P(\omega_q \mid x_i, d_m) \quad (35)$$

where $P(\omega_q \mid \mathbf{x}_i, d_m)$ is the probability that opinion score ω_q is chosen given distortion d_m and parameter vector $\mathbf{x}_i$. This function is approximated by the listener panel relative frequencies for distortion d_m, given by $H(\omega_q \mid \mathbf{X}_m, d_m)$. Here, $\mathbf{X}_m$ represents the standard source voice subjected to distortion d_m. Parameter vector $\mathbf{x}_i$ is measured from the i-th frame of $\mathbf{X}_m$.

3.3.3 Mean Opinion Score Prediction

Having estimated class probability (35), we can now obtain a classification of voice quality. This is given by ω_q which gives the largest value of $P(\omega_q \mid \mathbf{x}_i)$ and represents the quality level most likely to be selected by a listener panel member. In a similar fashion, (34) can be used to choose the most likely distortion.

The prediction of Mean Opinion Score for the i-th frame can also be found:

$$MOS_i = \sum_{q=1}^{5} q \cdot P(\omega_q \mid x_i) \quad (36)$$

$$MOS = E\{MOS_i \mid x_i\}.$$

Here, E{·} is the expectation operator and can be approximated by averaging over all frames of the speech record. As (35) shows, the estimate incorporates all of the listener panel information, not just the mean opinion value of the training data.

3.3.4 Discussion of the PR Method

A key advantage of this technique is its ability to use multiple parameters selected to maximize performance for a given application. After the voice parameter set is chosen, selected distortions representative of the application are used to train the algorithm. The ability of the method to "learn" based on training data allows it to handle very dissimilar types of distortion and adapt when new types of distortion are introduced.

As an example, Figure 1 shows objective MOS values computed using the cepstral distance method graphed against subjective scores. The distortions correspond to female speech from a database provided by Comsat containing speech processed through a variety of low to mid-rate codecs. As can be seen, the CD behaves differently for the MNRU-N condition ("M") and the 16 kbit/s condition ("+") than for the remaining conditions. Regression lines (dashed) for these three cases indicate that although the CD is clearly sensitive to distortion level, multiple mapping functions would be required to handle the diverse types of impairments.

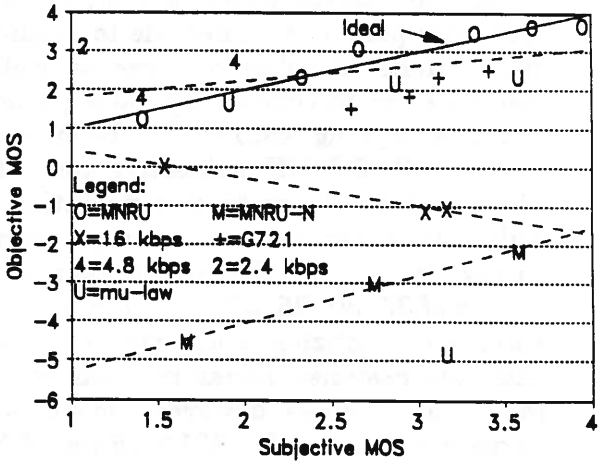

Figure 1 Cepstral distance MOS vs subjective MOS for Comsat dataset (female talker).

Figure 2 shows objective and subjective results for the PR method applied to the same female speech data using 3 objective voice parameters: banded spectral distortion as in (13), banded SDR as in (20), and LPC Euclidean distance as in (9). Significantly, parameter set selection and training were completed using *male* speech data from the Comsat database. In other words, training and testing was done using completely different speech records (i.e., male vs female voice), but using the same types of *impairments*.

Figures 1 and 2 show that the PR method provides much better correlation with subjective scores than the CD technique for this dataset. The reason is that the method uses estimates of the probability of distortion d_m, $P(d_m|\mathbf{x})$, to compute MOS. Since $P(d_m|\mathbf{x})$ is based on multiple parameters chosen specifically for this application, it can contain more information than simple distance measures. The mapping from

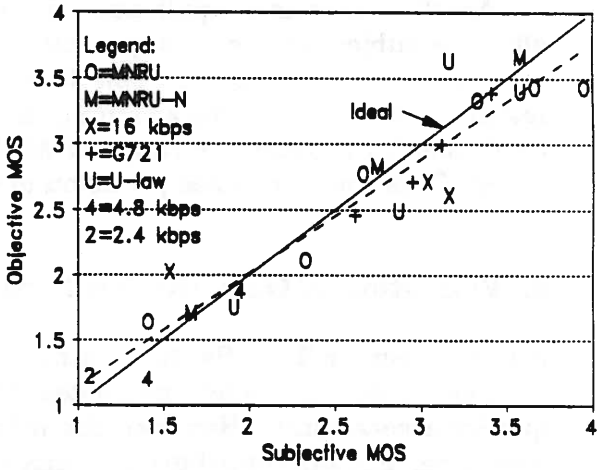

Figure 2 Pattern recognition MOS vs subjective MOS for Comsat dataset (female talker). Training and parameter selection completed using Comsat dataset, male talker.

objective parameter space to predicted MOS is thus automatically adjusted for distortion type, as shown in (35).

4. Alternative Applications

In addition to being a tool for assessing new algorithms and devices, an objective quality measure has other potential applications. For example, it could be used in the verification of subjective test results. Listener panel trials, which in some cases are quite complex, are vulnerable to statistical fluctuations of listener scores as well as human error in conducting the experiment and in analyzing test results. Results from the CCITT SG XII subjective test of 16 kbit/s codecs illustrate this potential use. Subjective mean opinion scores are graphed in Figure 1 for MNRU speech data with Q values of 35, 30, 25, 20, 15, 10, 5, and 0 dB signal-to-quantizing noise ratios. Rather than the assumed linear relation between MOS and Q values, however, the subjective curve is not linear for S/N ≥ 25 dB. Objective scores using the PR method applied to the same data are presented on the same graph, and show the same basic shape. This is indicative that the subjective scores are indeed reflective of degradations in the data, and are probably not due to problems in the subjective scoring procedure. In fact, it turns out that the curve reflects the presence of additional noise due to the A/D, and the tendency of listener scores to saturate at S/N values over 30 or 35 dB.

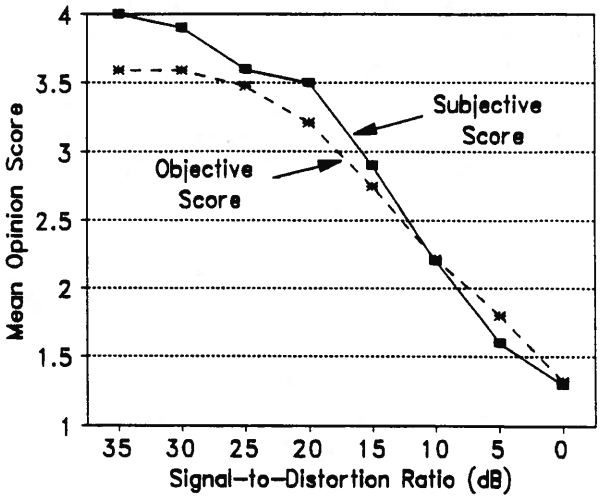

Figure 3 Averaged objective and subjective opinion scores for MNRU degraded speech data.

Another alternate application is to use objective MOS values to develop more cost effective subjective test designs [50]. For example, considerable effort is made to insert reference conditions into a subjective test spanning the entire range of possible MOS. In testing a new voice coding technique, it might be advantageous to concentrate reference conditions in a narrower range of MOS values bracketing the expected MOS of the new codec. This could increase precision of the test, reduce cost, or possibly both.

5. Validation of Objective Methods

5.1 Subjective Test Performance

Careful design of subjective listener tests provides reasonably accurate and reliable voice quality assessments. However, the inherent variability of listeners is a difficult problem to overcome. Factors contributing to this variability include differences in hearing ability, race, gender, geographical origin, emotional attitude during the test, and differing interpretations of the test criteria, such as "good", "fair", "poor", etc. Listener variability is measured by estimating the listener score variance for a given speech record:

$$\sigma_L^2 = \frac{1}{n-1}\sum_{q=1}^{5} n_q \cdot (q - MOS)^2, \qquad (37)$$

where n is the number of listeners and n_q is the number of listeners giving a score of q (on a 1 to 5 scale). Typical values for σ_L, the listener score standard deviation, range from .6 to .8.

Listener variability also affects the reliability of MOS estimates. In general, accuracy and repeatability are improved by using larger listener panels. The standard deviation of MOS estimates decreases as the number of listeners increases:

$$\sigma_{MOS} = \frac{\sigma_L}{\sqrt{n}}. \qquad (38)$$

This means that measured MOS values will fluctuate within a range of roughly $\pm\sigma_{MOS}$ if the test is repeated using a listener panel from the *same population* (i.e., from the same geographical location, same mix of male, female, race, backgrounds, etc.). For example, using the typical values σ_L=.7 and n=24 listeners, the range of error is approximately $\pm.7/\sqrt{24} \approx \pm.14$ opinion score points. In practice, however, the make up of listener panels is quite difficult to control, and σ_{MOS} values of .2 to .3 are more common. The value σ_{MOS} is a reasonable figure of merit for subjective listener scores.

5.2 Objective Test Performance

Measuring the accuracy of objective assessment techniques in a meaningful way is not straightforward. Accuracy may be sensitive to sentence material, talker voice, gender, and especially to distortion type. The performance of objective techniques must therefore be carefully defined in terms of the conditions and classes of distortion used in testing. For instance, an objective measure known to perform well for high bit-rate coders would not be applied to other types of voice communication such as single sideband radio without further validation tests for these new applications. As a very simple example, consider Figure 4, where the information index is used to predict MOS for speech corrupted with both band-limited Gaussian noise and speech-correlated noise. The objective MOS values are well correlated with subjective scores within each class of distortion, but the correlation is much worse when the distortions are considered together. The two regression lines in Figure 4 show how the mapping between subjective and objective scores can be adjusted as long as a priori knowledge exists about distortion types present in the system. (An exact correction to adjust for the type of additive noise is given in [31].)

A measure of performance can be computed using the objective and subjective scores measured from the same speech database. The squared correlation coefficient, ρ^2, has sometimes been used to measure accuracy, but it does not provide a meaningful basis for comparing objective

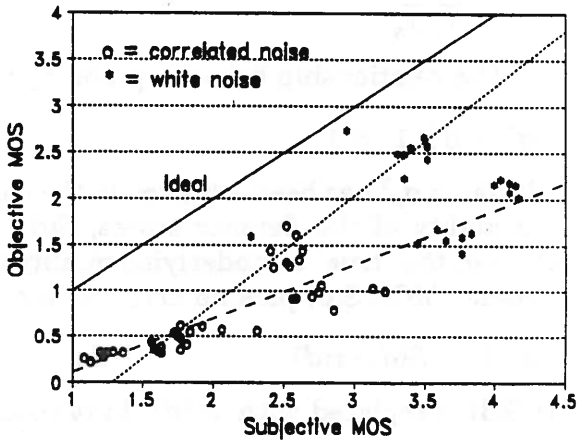

Figure 4 Information Index MOS predictions vs subjective values for added Gaussian noise and speech-correlated noise (Bell Labs database).

and subjective test accuracy. To explore this problem further, let the objective MOS for distortion d be O(d) and the subjective MOS be S(d). The error variance can now be written as

$$\sigma_e^2 = E\{(O(d)-S(d))^2\}. \tag{39}$$

where the expectation operator is taken over distortions and listeners. This value is not a fair figure of merit since the subjective scores being used as reference are characteristically different depending on listener race, geographic location, etc and are not absolutely accurate. A first order correction can be applied to the objective measure to adjust to the listener population (i.e., the set of listeners providing the S(d) scores):

$$O'(d) = a \cdot O(d) + b, \tag{40}$$

To demonstrate the need for a linear correction, suppose that MOS values from one listener panel were graphed against those from a different panel for the same speech data. It is quite unlikely that the values would lie on an ideal 45° line due to differences in the listeners. In no way does this imply that one of the listener panels is wrong. Likewise, MOS values from objective assessments should not be penalized for linear variation from the 45° line. The regression (40) simply adjusts the objective scores for listener panel differences to avoid this type of penalty.

A new error variance, $\sigma_{e'}^2$, can thus be written as:

$$\sigma_{e'}^2 = E\{(O'(d)-S(d))^2\}. \tag{41}$$

This error is minimized if a and b are chosen such that

$$a = \rho\frac{\sigma_S}{\sigma_O} \quad and \quad b = \overline{S} - a\overline{O}, \tag{42}$$

where σ_S and σ_O are standard deviations of subjective and objective MOS scores, respectively, $\overline{O}$ and $\overline{S}$ are their means, and ρ is the correlation coefficient defined as:

$$\rho = \frac{\sigma_{OS}^2}{\sigma_O \sigma_S}. \tag{43}$$

The relationship between ρ and $\sigma_{e'}$ is given by

$$\sigma_{e'}^2 = \sigma_S^2(1-\rho^2). \tag{44}$$

Although $\sigma_{e'}^2$ has been adjusted to the current listener population, it does not account for the variability of the listener scores, S(d), *within* this population. To help make this point clearer, the "true" or underlying quality of distortion d, S'(d) can be modeled as the measured listener MOS S(d) plus an error term ε(d):

$$S'(d) = S(d) + \varepsilon(d). \tag{45}$$

If S(d) is replaced with $S'(d)$ in (41), an expression of mean squared error is developed:

$$mse = \sigma_{e''}^2 = E\{(O'(d)-S'(d))^2\} \tag{46}$$

$$= E\{(O'(d)-S(d))^2\} - 2E\{\varepsilon(d)(O'(d)-S(d))\} + E\{\varepsilon^2(d)\} \tag{47}$$

$$\sigma_{e''}^2 = \sigma_S^2(1-\rho^2)+\sigma_L^2/n = \sigma_S^2(1-\rho^2)+\sigma_{MOS}^2, \quad (48)$$

where the following assumptions have been made:

- The errors ε(d) and $O'(d)-S(d)$ are uncorrelated and at least one of them is zero mean,
- $E\{\varepsilon^2(d)\} = E\{S(d)-S'(d))^2\} = E\{(S(d)-\overline{S(d)})^2\}$ = variance of listener scores, $\sigma_{MOS}^2 = \sigma_L^2/n$,
- The "true" quality of distortion d, S'(d), is equal to the global average of MOS values from all possible listener panels in the population: $S' = E\{S\} = \overline{S}$.

The value $\sigma_{e''}$ is a measure of root mean squared error, and can be used to assess objective performance. It is more meaningful than ρ^2 which does not directly provide information about expected error. Further, $\sigma_{e''}$ can be compared with σ_{MOS} which measures subjective listener score accuracy. However, since ρ^2 is easy to compute and often has been used to measure objective assessment performance in the past, it may actually be a more desirable metric. Equation (48) provides a means for determining what values of ρ^2 are needed to achieve a level of objective performance comparable to subjective testing methods.

Figure 5 shows the relationship between $\sigma_{e''}$, ρ^2, and n (the number of listeners) using typical values of σ_S^2=1.0 and σ_L=0.7. From this figure, an objective measure with ρ^2=.9 has a root mean squared error, $\sigma_{e''}$, in the range .3 to .35 when tested against a panel of 20 or more listeners. However, as discussed earlier, subjective error is typically less than this, often between .2 and .3. For the objective method to provide comparable accuracy requires a squared correlation coefficient of .95 or more. This is not an easy goal for objective techniques to meet, and may require that the domain of application be carefully defined for each objective parameter. In the case of the NTIA multivariate system, separate parameter sets and training databases may be needed in order to handle different applications.

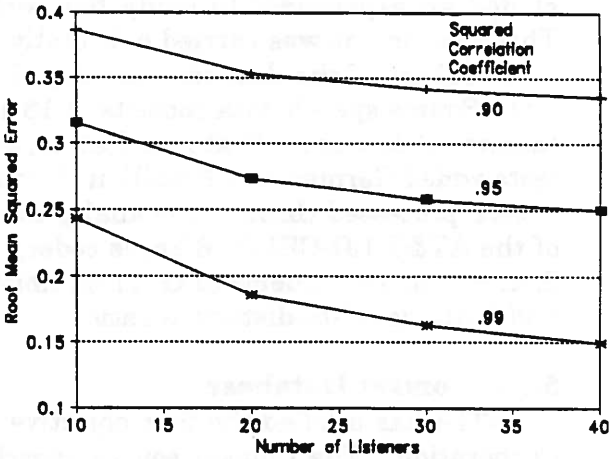

Figure 5 Relation of squared correlation coefficient, ρ^2, to RMS error,$\sigma_{e''}$, for different listener panel sizes. σ_S=1.0, σ_L=0.7.

Equation (48) says that *measured* objective assessment error is made up of two terms: one term accounts for difference between objective scores and the subjective reference, and a second term accounts for the fact that the subjective reference is inaccurate due to a finite number of listeners. In other words, even if an objective system seems to perform perfectly by providing MOS estimates identical to the listener panel scores, it will still have an expected level of error determined by the accuracy of the subjective test. This does not imply that objective performance is always poorer than subjective methods; rather it says that tests of objective methods should be performed using the most precise subjective data possible as reference. As Figure 5 shows, to achieve RMSE of .25, an objective method would have to demonstrate correlation of ρ^2=.95 if compared to results from a panel of 40 listeners, but would need ρ^2=.99 if only 10 listeners were used.

5.3 Test Results

Four objective assessment methods are being considered by CCITT for the Study Period ending in 1992. These are the cepstral distance (CD), information index (II), coherence function (CF), and pattern recognition method (PR) which are described in earlier sections. For these tests, the PR method was trained using the Comsat speech dataset (male talker only). The parameter set used in the PR method was selected to optimize assessment performance for Comsat male speech data and is described as follows:

1. Root mean-squared log spectral difference between input and output speech, as in (13).
2. Mean Euclidean distance between input and output LPC coefficients, as in (9).
3. Average signal-to-distortion ratio, as in (20).

Parameters 1 and 3 are computed in the frequency range 560 to 2820 Hz.

This section summarizes results of testing the performance of the four techniques on three different speech databases.

5.3.1 CCITT Database

The primary goal of the recent (1989-1991) CCITT 16 kbit/s codec evaluation was subjective testing of codecs for developing an international standard. The test plan also included an experiment to study the performance of four objective assessment techniques. This experiment was carried out jointly by NTIA and NTT.

A subset of the database developed for the codec evaluation was used in the objective test. Source speech data consists of 18 sentence pairs per talker using 2 male and 2 female talkers. This was replicated in Italian, Japanese, and North American English. Subsequent tests added German and Brazilian (Portuguese) speech data. Distortions consist of null (i.e., speech processed through the analog interfaces with no digital stages); 1, 2, and 4 tandems of the AT&T LD-CELP 16 kbit/s codec; 1, 2, and 4 tandems of G721 32 kbit/s ADPCM; 1, 2, 4, 8, and 16 tandems of G711 64 kbit/s PCM; and MNRU with 35, 30, 25, 20, 15, 10, 5, and 0 dB signal-to-distortion ratio.

5.3.2 Comsat Database

NTIA has applied the four objective methods to a speech database supplied by Comsat Corporation. The Comsat source speech consists of 4 sentences for 1 male and 1 female talker. Distortions include MNRU with 40, 35, 30, 25, 20, and 15 dB SDR; narrowband MNRU (noise lowpass filtered to approximately 28 kHz) with 35, 25, and 15 dB SDR; 1, 2, 3, and 4 tandems of G721 32 kbit/s ADPCM; 2.4 kbit/s LPC codec; 2 types of 4.8 kbit/s SELP codec; 3 types of 16 kbit/s codec; and 5, 6, 7, and 8 bit mu-law PCM.

5.3.3 Bell Labs Database

NTIA also applied the four objective methods to a speech database provided by AT&T Bell Labs. The Bell Labs source data consists of 3 sentences for 4 male and 4 female talkers. Distortions include added correlated noise at 5, 10, 15, 20, and 25 dBrnC; added band-limited Gaussian noise at 25, 35, and 45 dBrnC; and results of 2 field tests using the public switched telephone network in the San Francisco area.

5.3.4 Discussion of Test Results

Results of applying the four techniques to the three databases are summarized in Table 2, which shows the squared correlation coefficient between objective and subjective scores. In the CCITT column, a range of values is given corresponding to the results for 3 different languages. In cases where there were discrepancies between the NTT and NTIA results, the

higher correlation value is shown in the table. The Comsat and Bell Labs datasets were not processed by NTT, and the results shown are from NTIA tests.

TABLE 2
OBJECTIVE ASSESSMENT PERFORMANCE RESULTS

Objective Parameter	CCITT Dataset	Comsat Dataset	Bell Labs Dataset
Cepstral Distance	0.90-0.92	0.01	0.72
Coherence Fcn	0.86-0.92	0.67	0.75
Information Index	0.68-0.94	0.59	0.74
Pattern Recognition	0.84-0.89	0.89	0.80

The CD produced excellent correlation (.90 to .92) for the CCITT test in all 3 languages. Note that language specific regression functions are used to map CD to MOS, which improves the CD performance. The CD produced very poor results (ρ^2=.005) for the Comsat dataset, and only fair results (ρ^2=.72) for the Bell Labs dataset. The reason for the poor results on the Comsat data is not known, however, NTT feels it may be due to processing errors.

The coherence function performed well (ρ^2=.86 to .92) on the CCITT data, but only fair on the Comsat and Bell Labs datasets (ρ^2=.67 to .75). These datasets are both challenging for the objective algorithm. The Comsat dataset includes a wide variety of codec types, while the Bell Labs data uses a large number of talkers.

The information index produced fair to good correlation (ρ^2=.68 to .94) on the CCITT dataset, but poor to fair correlations for the Comsat (ρ^2=.59) and Bell Labs (ρ^2=.74). Results produced by NTT for the information index have been consistently better than those of NTIA suggesting possible implementation differences. Furthermore, the logistic function used to map information values to MOS is based on subjective tests made in France. This function could be tuned to achieve higher correlations by adjusting it for each language. Presumably, similar corrections could also be made for the coherence function.

The PR method performed well for all datasets (ρ^2=.80 to .89) for all datasets. It gave the best correlation for both Comsat (.89) and Bell Labs (.80) data. Subsequent tests of the PR method using a parameter set consisting of the CD and II, with training on North American female speech have been even more promising. Correlations of .94 to .98 were achieved for CCITT North American English, Italian, and Brazilian (Portuguese). Strangely, however, near-zero correlations were measured for two other datasets - Japanese and German. NTIA feels this problem will be corrected by a simple speech amplitude normalization procedure [59]. These results may be significant if further tests on other data show consistently high correlations.

Although each technique has produced good results on selected speech datasets (i.e., ρ^2>.9), none of the methods produces consistently good results across all tested data. Thus it appears that the ultimate goal of a technology independent assessment system has not yet been achieved. Current techniques are reliable and accurate only within these catagories of voice impairments for which the objective system has been well characterized. Applying the methods to speech impairments outside of these classes will yield assessments whose accuracy is unknown.

6. Standards Development

6.1 Standards Planning

Test results indicate that current objective assessment techniques may not have sufficient accuracy to replace subjective listener tests. However, there are several important applications of objective methods which require less precision. Developing a standard objective assessment technique will benefit users and developers of voice systems by providing a reference measurement for preliminary assessment of quality. In no way will the standard attempt to specify *levels* of voice performance, only methods of measuring performance.

The CCITT SG XII is considering three alternative approaches to developing a recommendation on objective voice quality. These are:

1. Choose the single best objective method.
2. Include all four of the methods being considered by CCITT.
3. Combine the best aspects of each into a single assessment algorithm.

Results of Section 5 make it clear that option 1 is not viable at this time since no one method stands out above the others for all types of distortions.

Option 2 would entail specifying each objective algorithm, along with the types of distortion and conditions for which it yields best results. This is an acceptable approach as it gives the user the option of selecting the method best suiting his needs.

Option 3 offers the potential advantage of improved assessment performance. Further, it removes from the user the burden of choosing which objective technique is best for his application.

Whether the methods can be successfully combined into a single assessment system is unclear. NTIA is currently studying this alternative by combining the CD, II, and CF parameters using the multivariate PR method. The PR method provides a framework for merging the three objective parameters into a single objective assessment and for training the system for optimal performance on specific voice applications. During training each parameter is automatically weighted based on its effectiveness in the current application.

Proceeding with option 3 will depend on how successfully the methods can be integrated. Assuming the results are acceptable, specification of the standard should be as straightforward as for the other two alternatives.

Specific elements of a proposed standard should include:

- Detailed specification of the algorithm.
- Scope of algorithm applicability. In other words, define the types of distortion for which the method is known to work and not work.
- Expected level of accuracy for each application and average performance in terms of squared correlation coefficient or RMSE for various distortions. Suggest recommended uses of the standard (e.g., augmenting subjective test design).

6.2 Development Effort

The expected level of effort for developing a standard depends on which alternative is chosen. Since the algorithms are fairly stable (i.e., have been defined and documented in the literature), option 1 or 2 would require minimal effort to specify the standard. The biggest workload will entail defining the scope of application for each objective method and compiling distortions and performance levels for which the method has been evaluated.

Work on combining the four algorithms as in option 3 is underway at NTIA. Modifications to some parameters may be necessary for them to be incorporated into the PR system.

If satisfactory results are achieved, NTIA will develop a specification for the resulting method.

6.3 Liaison

Close liaison should be maintained with CCITT Study Group XII, and in particular with the SG XII Speech Quality Experts Group (SQEG). Objective assessment methods are being considered for standardization within both Working Parties XII/1 and XII/3, with the SQEG being the primary focus of these efforts. Since the developers of all four objective algorithms actively participate in SG XII, this may be the most logical organization for developing objective voice quality standards. In this case, T1 Working Groups such as T1Y1, T1Q1, and T1M1 should monitor and contribute to the CCITT effort where appropriate.

As another option, it may be desirable to focus domestic standardization efforts on alternative applications of objective assessment techniques such as discussed in Section 4. In this case, liaison between T1Y1, T1Q1, and T1M1 would be necessary due to algorithm, performance, and measurement aspects of the project. Liaison with SG XII would also be important if this approach is adopted.

7.0 Future Work

The performance of several proposed objective methods is quite promising, and may justify their use in many applications where listener tests can be augmented or replaced completely. Standards defining these objective measures and their intended applications can be expected in the near future. However, a technique accurate and robust enough to replace human listeners in all situations does not currently appear feasible.

Several important areas of research related to objective voice quality assessment can be identified:

- It is doubtful that any single distortion measure, however complex, will be sufficient to accurately predict voice quality in many real world cases. For example, while spectral distortion parameters such as the coherence function or Bark spectral distortion appear to accurately model perceived quality for many impairments, they can not account for the effects of pure transmission delay. Delay has no effect on quality for one-way speech, but significantly impacts quality when two-way conversation is taking place. Echo and temporal warping (i.e., time varying delay) are other examples of impairments that are not accounted for by the types of objective parameters discussed in this report. This is not to say that delay or echo cannot be measured and used as objective parameters - they can. However, the assessment algorithm will need to incorporate information about all relevant degradations in the signal, including spectral distortion, delay, echo, and others. A much more involved psycho-acoustic model of hearing may be needed to address these impairments.
- Measures of voice quality are increasingly sought for vocoder, or synthesized, speech. Distortion measures such as discussed in this paper are not directly applicable, since generally no effort is made to reproduce the input voice signal. Only higher level information such as phonemes is transmitted.
- Application specific objective measures may be a better solution than attempting to identify a single technique for use in all situations. For example, voice transmitted via single sideband radio will be subject to significantly different types of degradation than found in telephone speech. Perceived quality is also highly dependent on application; an air traffic controller is more concerned with intelligibility than with tonal fidelity, while the opposite may be true for someone using the telephone to call home.

- More reliable and robust voice quality assessment may be possible using additional information contained in transmitted voice. For instance, amplitude and spectral properties are phoneme dependent and change markedly from frame to frame with the result that the type and degree of distortion imposed by various coding algorithms may be quite time dependent. Additionally, the perceptual importance of individual phonemes should be accounted for by weighting the distortion measure on a frame by frame basis. One approach is to segment the speech into phoneme related speech partitions prior to objective assessment. In this way, quality scores for each partition can be weighted according to the perceptual importance of phoneme related distortion.

These areas and others will see increasing interest as technical and economic pressure continues to build for dependable objective voice quality assessment techniques. Recommendations expected soon from CCITT will likely represent only a partial and temporary solution to this difficult problem, and continued research is crucial to develop more robust and reliable techniques.

REFERENCES

[1] K. Itoh, N. Kitawaki, and K. Kakehi, "Objective quality measures for speech waveform coding systems," Review Elec. Commun. Lab, vol. 32, no. 2, pp 220-228, Japan NTT, 1983.

[2] N. Kitawaki, K. Itoh, M. Honda, and K. Kakehi, "Comparison of objective speech quality measures for voiceband codecs," Proc. ICASSP, pp 1000-1003, Paris 1982.

[3] S. R. Quackenbush, T. P. Barnwell III, and M. A. Clements, Objective Measures of Speech Quality, Englewood Cliffs, N.J., Prentice Hall, 1988.

[4] T.P. Barnwell and W. D. Voiers, "An analysis of objective measures for user acceptance of voice communication systems," DCA Final Report DA100-78-C-0003, September 1979.

[5] T. P. Barnwell, III, "Objective fidelity measures for speech coding systems," J. Acoust. Soc. Am., vol. 65, no. 6, pp 1658-1662, Dec. 1979.

[6] R. F. Kubichek, E.A. Quincy, and K. L. Kiser, "Speech Quality Assessment using Expert Pattern Recognition Techniques," IEEE Pacific Rim Conference on Computers, Communication, and Signal Processing, June 1989.

[7] E.A. Quincy, "PROLOG-based expert pattern recognition system shell for technology independent, user-oriented, classification of voice transmission quality," Proceedings IEEE-ICC 87, Vol. 2, pp. 1164-1171, June 1987.

[8] L. L. Pfeifer, "Inverse filter for speaker identification," Speech Communications Res. Lab., Santa Barbara, CA, Final Rep. RADC-TR-74-214, 1974.

[9] D.J. Goodman, C. Scagliola, R.E. Crochiere, L.R. Rabiner, and J. Goodman, "Objective and subjective performance of tandem connections of waveform coders with an LPC vocoder," Bell Syst. Tech. J., vol. 58, no. 3, pp 601-629, March 1979.

[10] J.D. Markel, B.T. Oshika, and A. H. Gray, Jr, "Long-term feature averaging for speaker recognition," IEEE Trans. Acoust., Speech, and Signal Process., vol. ASSP-25, pp 330-337, Aug. 1987.

[11] Rabiner, L.R. and R.W. Schafer, Digital Processing of Speech Signals, Prentice-Hall, Englewood Cliffs, N.J., 1978.

[12] D. O'Shaughnessy, Speech Communication, Human and Machine, Addison-Wesley Publishing Co., 1987.

[13] P.E. Papamichalis, Practical Approaches to Speech Coding, Prentice-Hall, Englewood Cliffs, N.J. 1987.

[14] A. Sinvhal and K. Khattri, "Application of seismic reflection data to discriminate subsurface lithostratigraphy," Geophysics, vol. 48, no. 11, pp 1498-1513, 1983.

[15] J.L. Flanagan, M.R. Schroeder, B.S. Atal, R.E. Crochiere, N.S. Jayant, and J. M. Tribolet, "Speech Coding," IEEE Trans. on Communications, vol. COM-27, no. 4, pp 710-736, April 1979.

[16] J. L. Lansford and R. Yarlagadda, "Adaptive Lp approach to speech coding," Proceedings ICASSP 88, pp 335-338, 1988.

[17] A.H. Gray, and J. D. Markel, "Distance measures for speech processing," IEEE Trans. Acoustics, Speech and Signal Processing, vol. ASSP-24, no. 5, pp. 380-391, October 1976.

[18] P. Mermelstein, "Evaluation of a segmental SNR measure as an indicator of the quality of ADPCM coded speech," J. Acoust. Soc. Am., vol. 66, no. 6, pp 1664-1667, Dec. 1979.

[19] B.J. McDermott, C. Scagliola, and D. Goodman, "Perceptual and objective evaluation of speech processed by adaptive differential PCM," Bell Syst. Tech. J., vol. 57, no. 5, pp 1597-1618, May-June, 1978.

[20] M. Nakatsui and P. Mermelstein, "Subjective speech-to-noise ratio as a measure of speech quality for digital waveform coders," J. Acoust. Soc. Am., vol. 72, no. 4, pp 1136-1144, Oct. 1982.

[21] D.J. Goodman, B.J. McDermott, and L.H. Nakatani, "Subjective evaluation of PCM coded speech," Bell Syst. Tech. J., vol. 55, no. 8, pp 1087-1107, Nov. 1978.

[22] J.M. Tribolet, P. Noll, B. McDelmott, and R. E. Crochiere, "A study of complexity and quality of speech waveform coders," Proc. ICASSP, pp 586-590, April 1978.

[23] R.E. Crochiere, J.M. Tribolet, and L.R. Rabiner, "On the measurement of waveform coder distortion using the log likelihood ratio," Proc. ICASSP, pp 340-343, Denver, CO, April, 1980.

[24] N. Kitawaki, H. Nagabuchi, and K. Itoh, "Objective quality evaluation for low-bit-rate speech coding systems," IEEE Journal on Sel. Areas in Communications, vol. 6, no. 2, pp 242-248, Feb., 1988.

[25] Bell Northern Research, "Objective evaluation of non-linear distortion effects on voice transmission quality," Contribution to CCITT, Annex 3 to Question 13/XII, COM XII-1-E, pp 100-102, 1987.

[26] NTT, "Transmission performance objective evaluation model for fundamental factors," COM XII-174-E, Nov., 1983.

[27] NTT, "Proposal of objective quality measure for voiceband codecs," COM XII-8-E, April, 1985.

[28] NTT, "Suggestions on accelerating the study of an objective methodology for estimating the quality of digital transmission network," Annex B, (Reply to Question 18/XII), COM XII-R 17-E, May, 1987.

[29] NTT, "Comparison between four objective algorithms for non-linear distortion," COM XII-8-E, May, 1989.

[30] CCITT Blue Book, Volume V (Geneva 1989), Supplement No. 3.

[31] J. Lalou, "The information index: an objective measure of speech transmission performance," Ann. Telecommun., vol 45, no. 1-2, pp. 47-65, 1990.

[32] Bell Northern Research, "Evaluation of non-linear distortion via the coherence function," COM XII-No. 60-E, April, 1982.

[33] Bell Northern Research, Informal communication and Fortran code listing, Oct. 20, 1986.

[34] J.R. Cavanaugh, R.W. Hatch, and J.L. Sullivan, "Transmission Rating Model for use in Planning of Telephone Networks," Globecom '83, IEEE Global Telecommunications Conference Record, vol. 2, pp. 20.2.1-6, 1983.

[35] J.R. Cavanaugh, "A Model for the Subjective Effects of Quantizing Distortion in Digital Transmission of Speech," ICC '80 Conference Record, vol. 2, pp. 20.5.1-5, 1980.

[36] J.R. Cavanaugh, R.W. Hatch, and J.L. Neigh, "A Model for the Subjective Effect of Listener Echo on Telephone Connections," Bell System Technical Journal, Vol. 59, No. 6, pp. 1009-1060, 1980.

[37] J.R. Cavanaugh, R.W. Hatch, and J.L. Sullivan, "Models for the Subjective Effects of Loss, Noise, and Talker Echo on Telephone," Bell System Technical Journal, Vol. 55, No. 9, pp. 1319-1371, 1976.

[38] CCITT vol III.1 Recommendation G.114, Annex A, "Transmission delay and echo problems caused by multiple satellite hops," Geneva, 1985, pp. 132-134.

[39] CCITT, "Composite opinion model," Blue Book, Vol V, pp. 24-29.

[40] CCITT, "Transmission rating models," Blue Book, Vol V, pp. 248-329.

[41] CCITT vol V Recommendation P.11, Supplement 4, "Prediction of transmission qualities from objective measurements," Geneva, 1985, pp. 214-236.

[42] J. Gruber, "Estimating the Transmission Grade-of-Service of Telephone Connections with Multiple Echo Paths." Globecom '88, IEEE Global Telecommunications Conference Record, vol. 1, pp. 463-467, November 1988.

[43] J. Gruber, "Estimating the Transmission Grade-of-Service of Telephone Connections with Multiple Echo Paths." Submitted to IEEE Transactions on Communication.

[44] R.W. Hatch and J.L. Neigh, "Transmission Rating Model for use in Planning of Telephone Networks," National Telecommunication Conference Record, vol. II, pp. 23-2-5, November, 1976.

[45] R.D. Silverthorne, "IEEE Draft Standard: Methodology for Specifying and Evaluating Voiceband Channel Performance Criteria," Globecom '83, IEEE Global Telecommunications Conference Record, vol. 2, pp. 20.3.1-8, 1983.

[46] J. Kittler, "Feature selection and extraction," Handbook of Pattern Recognition and Image Processing, T. Y. Young and K. S. Fu, ed., Academic Press, Inc., Chapter 3, pp. 59-83, 1986.

[47] K. Fukunaga, "Statistical pattern classification," Handbook of Pattern Recognition and Image Processing, T. Y. Young and K. S. Fu, ed., Academic Press, Inc., Chapter 1, pp. 3-32, 1986.

[48] D.O. Loftsgaarden and C. P. Quesenberry, "A non-parametric estimate of a multivariate density function," Annals of Mathematical Statistics, 36:1049-1051, June 1965.

[49] J.T. Tou and R.C. Gonzalez, Pattern Recognition Principles, Reading, Mass.: Addison-Wesley Publishing Co., 1974.

[50] S. Dimolitsas, "Objective speech distortion measures and their relevance to speech quality assessments," IEE Proceedings, Vol. 136, No. 5, Oct. 1989.

[51] M. H. Sherif and R. J. Gregor, "Modelling myoelectric interference patterns during movement," Medical & Biological Engineering & Computing, 24 (1): 2-9 January 1986.

[52] M. H. Sherif, R. J. Gregor and, J. Lyman "Phasic relations in 90 abduction adduction of the arm: The ARIMA representation," J. Biomechanics, vol.17: 215-224, 1984.

[53] Nippon Telegraph and Telephone, "Provisional report on 16-kbit/s speech coder objective test," CCITT SG XII/1, Temporary Doc. No. 52-E, Geneva, Oct. 1990.

[54] P. Breitkopf and T. Barnwell III, "Segmental preclassification for improved objective speech quality measures," Proceedings ICASSP, pp 1101-1104, Mar. 1981.

[55] S. Wong, A. Sekey, and A. Gersho, "Auditory measure for speech coding," Proceedings ICASSP, May 1991.

[56] National Telecommunications and Information Administration, "Results of NTIA objective assessment tests," CCITT Speech Quality Expert's Group Meeting, Doc. SQ-56, Geneva, October, 1990.

[57] Bell Northern Research, Objective evaluation of nonlinear distortion effects on voice transmission quality," CCITT contribution COM XII-46-E, March 1986.

[58] Bell Northern Research, "Re-evaluation of the objective method for measurement of nonlinear distortion," CCITT contribution COM XII-175-E, June 1987.

[59] NTIA, "Reprocessing the CCITT 16 kbit/s data using objective voice quality assessment techniques," Contribution to CCITT SG XII Speech Quality Experts Group, doc. SQ 33.91, Florence, Italy, July 1991.