AD-A078 208 ROME AIR DEVELOPMENT CENTER GRIFFISS AFB NY F/G 17/2
INTELLIGIBILITY THRESHOLD LEVEL (ITL) RATINGS OF SOME CURRENT D--ETC(U)
JUL 79 C P SMITH
UNCLASSIFIED RADC-TR-79-215 NL

1 OF 1
AD A078208

END
DATE FILMED
1-80
DDC

LEVEL II

RADC-TR-79-215
In-House Report
July 1979

INTELLIGIBILITY THRESHOLD LEVEL (ITL) RATINGS OF SOME CURRENT DIGITAL VOICE COMMUNICATIONS PROCESSORS

Caldwell P. Smith

ROME AIR DEVELOPMENT CENTER
Air Force Systems Command
Griffiss Air Force Base, New York 13441

This report has been reviewed by the RADC Information Office (OI) and is releasable to the National Technical Information Service (NTIS). At NTIS it will be releasable to the general public including foreign nations.

RADC-TR-79-215 has been reviewed and is approved for publication.

APPROVED:

J. P. VETRANO, Chief
COMSEC Engineering Office
Deputy for Electronic Technology

APPROVED:

PATRICK J. O'TOOLE, Colonel, USAF
Deputy for Electronic Technology

FOR THE COMMANDER:

JOHN P. HUSS
Acting Chief, Plans Office

If your address has changed or if you wish to be removed from the RADC mailing list, or if the addressee is no longer employed by your organization, please notify RADC (ETC), Hanscom AFB MA 01731. This will assist us in maintaining a current mailing list.

Do not return this copy. Retain or destroy.

Unclassified

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

REPORT DOCUMENTATION PAGE		READ INSTRUCTIONS BEFORE COMPLETING FORM
1. REPORT NUMBER RADC-TR-79-215	2. GOVT ACCESSION NO.	3. RECIPIENT'S CATALOG NUMBER
4. TITLE (and Subtitle) INTELLIGIBILITY THRESHOLD LEVEL (ITL) RATINGS OF SOME CURRENT DIGITAL VOICE COMMUNICATIONS PROCESSORS		5. TYPE OF REPORT & PERIOD COVERED In-House Report 6. PERFORMING ORG. REPORT NUMBER
7. AUTHOR(s) Caldwell P. Smith		8. CONTRACT OR GRANT NUMBER(s)
9. PERFORMING ORGANIZATION NAME AND ADDRESS Deputy for Electronic Technology (ETC) Hanscom AFB Massachusetts 01731		10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS 33401F 78200202
11. CONTROLLING OFFICE NAME AND ADDRESS Deputy for Electronic Technology (ETC) Hanscom AFB Massachusetts 01731		12. REPORT DATE July 1979 13. NUMBER OF PAGES 58
14. MONITORING AGENCY NAME & ADDRESS(if different from Controlling Office) N/A		15. SECURITY CLASS. (of this report) Unclassified 15a. DECLASSIFICATION/DOWNGRADING SCHEDULE

16. DISTRIBUTION STATEMENT (of this Report)

Approved for public release; distribution unlimited.

17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report)

Same

18. SUPPLEMENTARY NOTES

N/A

19. KEY WORDS (Continue on reverse side if necessary and identify by block number)

Voice communication
Intelligibility
Speech processing
Diagnostic rhyme test
Test and evaluation

20. ABSTRACT (Continue on reverse side if necessary and identify by block number)

Intelligibility scores for digital voice communications processors have been found to be typically characterized by highly significant differences among speakers, as well as highly significant differences among scores for the various phonetic features.

Distributions of intelligibility scores for speech processors are not normally distributed, but highly skewed. → next page

DD FORM 1 JAN 73 1473 EDITION OF 1 NOV 65 IS OBSOLETE

Unclassified

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

309 050

A negative binomial probability distribution was found to give good agreement with empirical data distributions of speech intelligibility scores.

A new performance rating for voice communications devices, termed an intelligibility threshold level (ITL), was conceived as a means of taking these findings into consideration in establishing a measure of intelligibility performance that is an estimate of an intelligibility value that the majority (rather than the simple average) of intelligibility scores for a voice processor will equal or exceed, at a specified confidence level established in relation to the sample size used in obtaining the rating.

It is proposed that an ITL rating is a more meaningful assessment of the degree of risk involved in misunderstanding voice messages or causing time to be lost in requiring messages to be repeated.

It was shown that ITL's can be determined by two alternative methods: by rank-ordering the intelligibility scores for a processor and constructing the cumulative distribution of data and its confidence band, or by using a negative binomial probability model for the data distribution.

Chi-squared tests indicated that in most cases the negative binomial probability model gave a reasonable approximation to the data distribution.

Intelligibility threshold levels (ITL's) estimated with the negative binomial probability model differed by at most one quantum value (3.125) from the ITL values determined from the empirical distributions.

It is recommended that future speech intelligibility tests and evaluations of digital voice communications processors and systems include an analysis of the data to determine the 80 percent ITL's at 95 percent probability, that is, determine the intelligibility level for which there is a 95 percent probability that 80 percent of the population of intelligibility scores (for individual speakers and phonetic features) will equal or exceed.

Contents

Contents

Illustrations

Illustrations

Tables

Intelligibility Threshold Level (ITL) Ratings of Some Current Digital Voice Communications Processors

1. INTRODUCTION

Over the past several years there has been an opportunity to study speech intelligibility data obtained from tests and evaluations of a wide variety of processors for digital speech communications applications. These studies have led to a conclusion that average intelligibility scores currently used to specify intelligibility performance fall short of providing adequate ratings of intelligibility, and that a need exists for an alternative rating for speech intelligibility that takes into consideration dispersion and highly skewed distributions of scores that characterize data obtained with multiple speakers and diagnostic intelligibility testing. This report presents some of those findings and presents a new concept for a speech intelligibility rating to supplement or replace average scores that rate the intelligibility performance of speech communications systems and devices.

1.1 Diagnostic Rhyme Test (DRT)

The intelligibility performance of speech communications processors is for the most part evaluated with the Diagnostic Rhyme Test or DRT. This is a test that grew out of research on very-low-data-rate digital speech communications by the method of voice pattern-matching. A speech intelligibility test method was needed to provide diagnostic intelligibility data, that is to assess separately the

(Received for publication 17 August 1979)

intelligibility obtained for each of the categories of phonetic events, as well as assessing overall or "total" intelligibility. This need was critical for evaluating performance and guiding the research in connection with the voice pattern-coding technique, as it was considered essential that the library of spectral patterns of speech should fully accommodate the range of allophones of conversational speech. A diagnostic intelligibility test method was required in order to assess whether this objective had been met and to identify any deficiencies in analysis and synthesis of the various speech sounds. The Diagnostic Rhyme Test or DRT was developed to fulfill this need. It quickly became apparent that the DRT combined unique properties of resolving power and sensitivity for assessing speech intelligibility performance, as well as being extremely efficient and economical for obtaining detailed intelligibility analyses with minimum investments in processing time and listener crew time. An initial single-speaker version of the DRT which was used in a first survey of the intelligibility performance of vocoder technology in 1967 was subsequently expanded to a multiple-speaker version that was used in a second survey of intelligibility performance of vocoder technology in 1972. Since that time additional multiple-speaker versions of the DRT including men and women speakers have been recorded with close-talking dynamic, carbon, and pressure-gradient (noise-cancelling) microphones; recording conditions have included quiet environments and talkers in various simulations of acoustic noise environments of interest in the Department of Defense (DOD). These intelligibility test recordings have been widely used in the DOD for evaluating intelligibility of a variety of digital and analog speech processing techniques and hardware. Intelligibility data obtained in these tests has provided the opportunity to closely examine such questions as intelligibility differences found among different speakers and variations in intelligibility of the various phonetic features.

1.2 Speaker Variability

A general finding has been that intelligibility scores are typically characterized by significant differences among speakers. Examples of this effect are presented in Figures 1 and 2 which present rankings of nine speakers (six male and three female) obtained from tests of three categories of voice processors, consisting of approximately a dozen each of narrowband, mediumband, and wideband devices. The speaker scores presented are the averages for all processors of each group. The Newman-Keuls test of significant differences among means was utilized in determining significant differences among the mean scores, leading to the rankings shown by the brackets. Data from the tests with the high-quality dynamic microphone indicated that the speakers fell into four categories which overlapped except in the case of the narrowband processors. Data obtained with

Nine - Speaker Diagnostic Rhyme Test

2400 BPS Systems		3600 & 4800 BPS Systems		Wideband Systems	
87.7	CH	88.0	CH	89.2	RH
85.2	RH	86.1	RH	88.8	CH
83.5	LL	86.0	LL	88.5	LL
81.0	PK	83.4	PK	86.3	PK
79.7	MP(Fem.)	82.1	BV	85.2	LS(Fem.)
79.7	BV	80.5	LS(Fem.)	84.0	MP(Fem.)
79.5	JE	79.8	MP(Fem.)	83.5	JE
78.6	LS(Fem.)	79.4	JE	83.3	JS(Fem.)
75.7	JS(Fem.)	78.4	JS(Fem.)	82.3	BV

The Newman - Keuls test of differences between means indicated that Scores within brackets did not differ significantly.

Figure 1. Mean Speaker Scores Obtained with a Dynamic Microphone

Nine - Speaker Diagnostic Rhyme Test

2400 BPS Systems		3600 & 4800 BPS Systems		Wideband Systems	
83.1	CH	85.0	CH	91.4	CH
82.5	RH	83.8	RH	89.9	LL
79.8	LL	81.4	LL	89.0	RH
77.9	PK	80.2	PK	87.4	PK
76.1	MP(Fem.)	78.6	MP(Fem.)	86.9	LS(Fem.)
75.0	JE	78.5	LS(Fem.)	86.0	JS(Fem.)
74.7	LS(Fem.)	77.8	JE	85.9	BV
72.6	BV	75.7	BV	85.8	MP(Fem.)
70.8	JS(Fem.)	72.1	JS(Fem.)	83.9	JE

The Newman - Keuls test of differences between means indicated that Scores within brackets did not differ significantly.

Figure 2. Mean Speaker Scores Obtained with a Carbon Microphone

the carbon microphone had the affect of accentuating the significant differences among the speakers as shown in Figure 2.

The data serve to highlight the necessity for standardizing the speakers used in intelligibility testing. They leave unanswered the question of the degree to which differences in intelligibility performance might occur in a large and diverse group of speakers that might be involved in using a military voice communications system. These tests represented a very limited sample of speakers in terms of factors such as age, regional dialects, fundamental pitch, and other factors that can be involved in speech quality. Considerable research remains to be done before it can be determined whether tests with a small group of speakers such as these provide intelligibility performance data that would be typical of a much larger population of speakers.

Variations in intelligibility performance of different speakers are presented in other contexts as shown in Figures 3 and 4, which were obtained from tests of a linear-predictive vocoder algorithm operating at 2400 bits per second (BPS). Figure 3 summarizes the effects of random bit errors superimposed on the data stream; linear regression equations were determined for the relation between overall intelligibility and the bit error rate for six male speakers. The intelligibility performance obtained for the different speakers tended to diverge as the bit error rate increased; analysis of variance confirmed that differences between slopes of the regression lines for the different speakers were highly significant ($a = 0.001$). The results emphasize the hazards of using the trend in average intelligibility scores (across all speakers) for predicting intelligibility expected for any one speaker.

In Figure 4 the degradation of intelligibility caused by the speakers being located in a noise environment is summarized. In this experiment the effects of the ambient noise in the cabin of a jet aircraft were simulated by electrically mixing a recording of noise measured in the aircraft, with the recorded speech signal, prior to processing the speech with the vocoder algorithm. Intelligibility tests were performed at various signal-to-noise ratios and the resulting intelligibility data were analyzed to determine 2d order regression models relating the overall intelligibility and the signal-to-noise ratio. These data need to be interpreted with caution, since the procedure of electrically mixing the noise signal does not accurately encompass two significant effects that would be present in a "real world" situation: a speaker would alter his performance in order to try to compensate for the effect of the noise, and the frequency spectrum of the interfering noise would undergo changes due to the response of the pressure gradient microphone to a far-field source of sound. While these effects would probably alter the values shown for the regression coefficients, the results show significant differences in intelligibility for the six speakers, differences that are obscured in a single, overall intelligibility score.

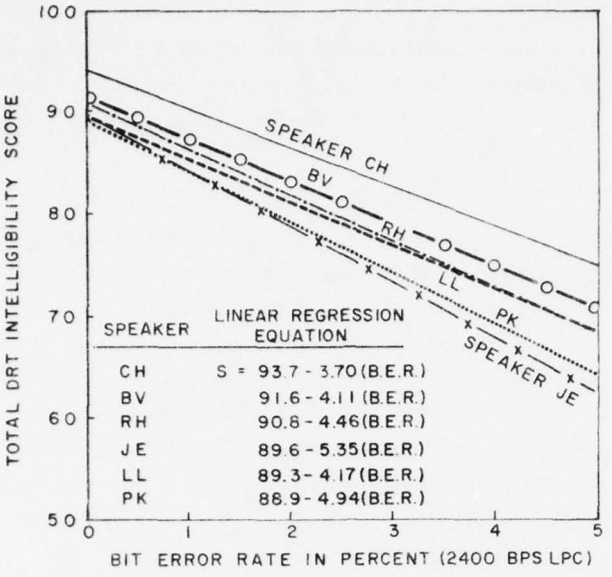

Figure 3. Total Intelligibility as a Function of Bit Error Rate for an LPC Vocoder at 2400 BPS (Six Male Speakers)

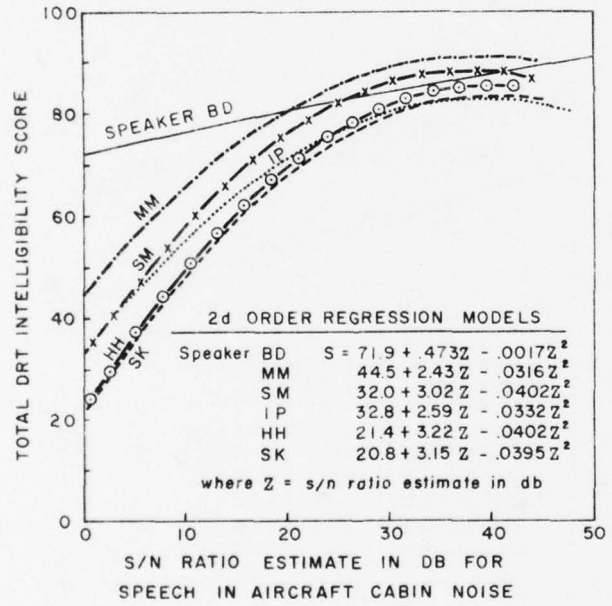

Figure 4. Total Intelligibility as a Function of Speech-to-Noise Ratio in Jet Aircraft Cabin Noise for an LPC Vocoder at 2400 BPS (Six Male Speakers)

Significant differences in intelligibility attained by different speakers, and significant speaker/processor interactions have been found to be the rule rather than the exception in speech intelligibility performance.

1.3 Phonetic Feature Variability

It has been shown by examples how large differences are found among speaker intelligibility scores. Even larger variations have been found to commonly occur among scores for the various phonetic features. Examples are presented in Figures 5 and 6 showing phonetic feature scores obtained in intelligibility performance of a linear-predictive vocoder (LPC) operating at 2400 bits per second. For these examples, separate trends are shown for the six primary phonetic features tested with the Diagnostic Rhyme Test: Voicing, Nasality, Sustention, Sibilation, Graveness, and Compactness. (It will be shown later in this report that each of these features is further subdivided by the test into four feature states, among which even larger variations occur.)

In Figure 5 linear regression lines are shown that represent the variation in intelligibility score for each of these features, in relation to the bit error rate. The spread of the regression lines serves to indicate how much deviation is

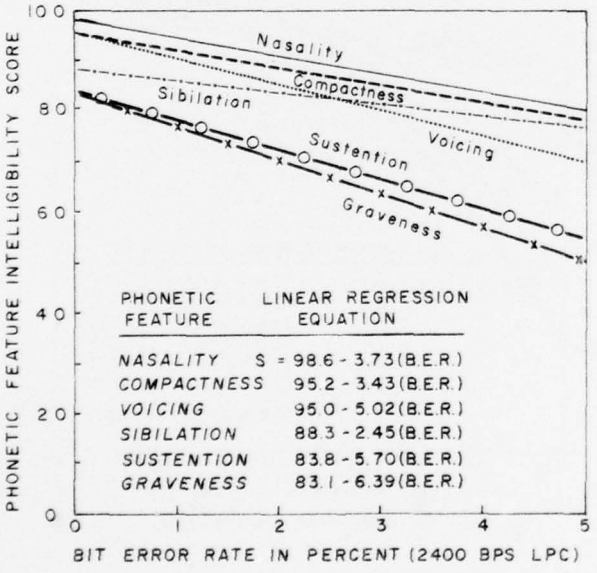

Figure 5. Phonetic Feature Intelligibility as a Function of Bit Error Rate for an LPC Vocoder at 2400 BPS

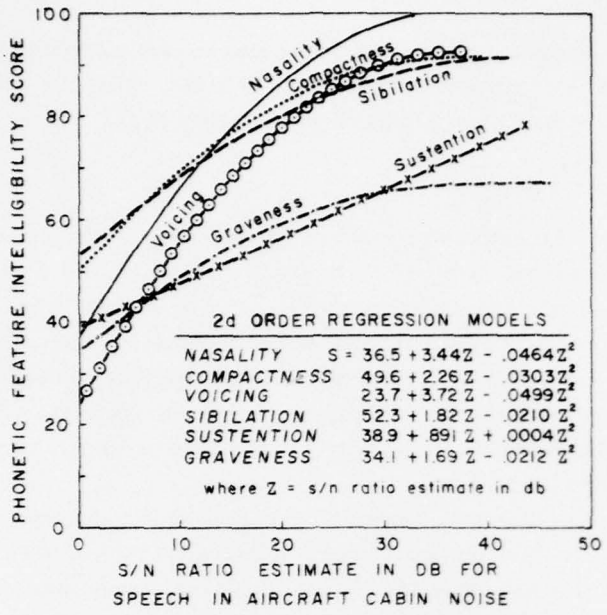

Figure 6. Phonetic Feature Intelligibility as a Function of Speech-to-Noise Ratio in Jet Aircraft Cabin Noise for an LPC Vocoder at 2400 BPS

possible with regard to the average or "total" intelligibility variation for the ensemble, as is the case in Figure 6 showing 2d order regression curves for phonetic feature intelligibility in relation to the speech-to-noise ratio estimate for speech in simulated jet aircraft cabin noise.

2. INTELLIGIBILITY THRESHOLD LEVEL (ITL) RATING

It has been shown that typically the population of intelligibility scores, the aggregate of scores representing the intelligibility of individual speaker/phonetic feature combinations, involves considerable dispersion. Perhaps because there is so much detailed information from a single multi-speaker intelligibility test, it has been the practice to condense the results to the average score and its standard error.

The proposed ITL rating involves an interpretation based on the distribution of scores obtained in a test. From analysis of the distribution, the ITL rating estimates the intelligibility value that will be equaled or exceeded by a specified percentage of the population of scores, at a specified confidence level. For example, the distribution of scores might lead to an ITL estimate that there is

95 percent probability that 80 percent of the scores (for individual speakers and phonetic features) will equal or exceed 70. (This ITL value has been observed in connection with average intelligibility scores around 90.)

2.1 Rationale for the ITL Rating

The proposed ITL rating derives from inspection of many distributions of intelligibility scores and observations of the large differences in speaker and phonetic feature scores that have been cited. It also derives from a premise that in critical military voice communications (and perhaps in other critical speech communications such as air traffic control) an intelligibility assessment is required that is made in terms of the level of performance to be expected for the majority rather than for the average: the majority of the speech events, and the majority of the speakers and listeners.

The average intelligibility scores currently used for specifying performance could be expected to state a value equaled or exceeded by half the underlying population of scores (for speakers and phonetic features) assuming that the scores were normally distributed.

However, it has been found that the populations of scores from diagnostic intelligibility tests typically are not normally distributed, but highly skewed; the higher the average score, the greater the extent to which the distribution of scores tends to be skewed.

The Lilliefors test for conformity with a normal distribution has been applied to many distributions of intelligibility data. Where the sample population is made up of the phonetic feature scores, whether of one or several speakers, the null hypothesis (for conformity with the normal distribution) has invariably been rejected.

2.2 Examples of Distributions of Speech Intelligibility Data

An example of a DRT data summary is shown in Figure 7 consisting of phonetic feature scores for a single speaker (CH) obtained with a test of continuous-variable-slope delta modulation (CVSD) at 16 kilobits per second. The numbers of listener errors (based on evaluation with a crew of eight listeners) that provided the basis for each score are shown in parentheses after each score. (A multi-speaker intelligibility test results in a set of scores such as this from each speaker of the test, plus an overall summary listing.)

The listing of Figure 7 indicates how each of the six features is tested in terms of four feature states, that is the feature Voicing involves separate, independent assessments of Voicing Present (Frictional), Voicing Present (Non-Frictional), Voicing Absent (Frictional), and Voicing Absent (Non-Frictional), these being

Diagnostic Rhyme Test Summary: CVSD 16 kbps Speaker CH

	Feature Present		Feature Absent		Feature Total	
Voicing	98.45	(1)	98.45	(1)	98.45	(2)
Frictional	100	(0)	96.9	(1)	98.45	(1)
Non-Frictional	96.9	(1)	100	(0)	98.45	(1)
Nasality	100	(0)	100	(0)	100	(0)
Grave	100	(0)	100	(0)	100	(0)
Acute	100	(0)	100	(0)	100	(0)
Sustention	90.65	(6)	95.35	(3)	93.0	(9)
Voiced	100	(0)	96.9	(1)	98.45	(1)
Unvoiced	81.3	(6)	93.8	(2)	87.55	(8)
Sibilation	87.5	(8)	95.3	(3)	91.4	(11)
Voiced	78.1	(7)	90.6	(3)	84.35	(10)
Unvoiced	96.9	(1)	100	(0)	98.45	(1)
Graveness	71.9	(18)	78.15	(14)	75.0	(32)
Voiced	84.4	(5)	100	(0)	92.2	(5)
Unvoiced	59.4	(13)	56.3	(14)	57.85	(27)
Compactness	100	(0)	92.2	(5)	96.1	(5)
Voiced	100	(0)	90.6	(3)	95.3	(3)
Unvoiced	100	(0)	93.8	(2)	96.9	(2)

Total DRT Intelligibility Score: 92.325 (59)
(Nr of Listeners: 8)

Figure 7. Summary of Diagnostic Rhyme Test (DRT) Intelligibility Scores Obtained for Speaker CH with CVSD at 16 kbps

averaged four ways to obtain scores for the effects of voicing being present and absent, and for frictional and non-frictional effects of voicing. An overall average estimates the total intelligibility for voicing. These details help diagnose specific deficiencies and help in identifying possible causes and remedies.

The 24 scores for the four states of each of the six features, shown within dotted boxes in Figure 7, portray the fine details of intelligibility performance of a processor. It is among these details that the differences in voice processor performance are usually identified, and deficiencies are highlighted.

When the 24 feature-state intelligibility scores are rank-ordered and plotted in the form of a cumulative distribution, a plot of the type shown in Figure 8 results. Here each of the 24 scores is represented by a vertical line segment representing 1/24th of the data population and the ends of adjacent segments have been connected to form a cumulative distribution starting with the lowest score which in this example was for Graveness Absent (Unvoiced), the next lowest score which was for the feature state Graveness Present (Unvoiced), etc. Across the top of the figure is shown a scale corresponding to the total number of listener errors associated with each score. A normal ogive is also shown corresponding to the values of the mean (92.3) and the standard deviation (12.4) associated with

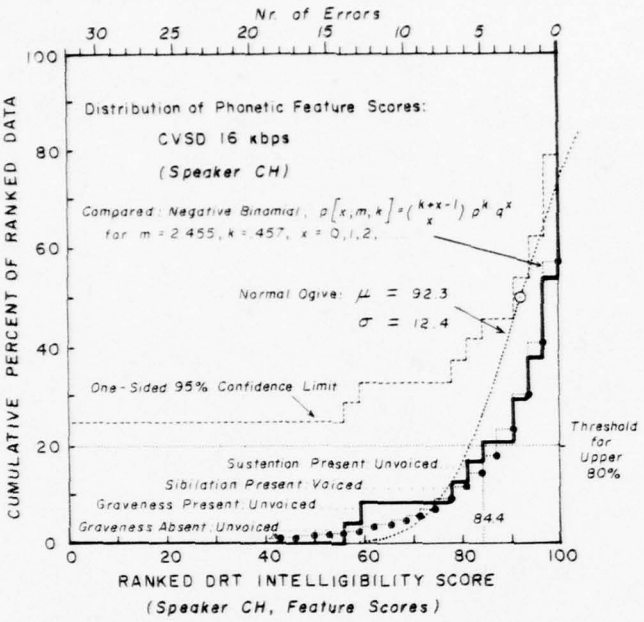

Figure 8. Distribution of Phonetic Feature Scores Obtained for Speaker CH with CVSD at 16 kbps

this group of data. The plot indicates that 80 percent of the actual data equaled or exceeded a score of 84.4. A one-sided 95 percent confidence band for the data distribution was constructed by the method of Kolmogorov, shown as a dashed profile. The confidence limit associated with the cumulative data distribution does not intersect the dotted horizontal line defining 80 percent of the data sample; thus we are unable to make a statement about an estimated 80 percent ITL at p = 0.95.

This example illustrates an important feature of the ITL rating: it is assessed with respect to a confidence limit, taking into consideration the size of the sample used in arriving at the estimate. Data from a single speaker did not provide an adequate sample size for estimating the value for an 80 percent ITL at p = 0.95.

Also shown in Figure 8 are points representing values that result from a negative binomial probability distribution model for the distribution of listener responses in terms of listener errors, rather than conventional DRT scores. In these studies of intelligibility data distributions, comparisons were made with the normal, binomial, Poisson, and negative binomial forms. Of these, only the negative binomial probability distribution was found to provide a reasonable approximation to the speech intelligibility data distributions.

2.3 Negative Binomial Probability Distribution Model

The negative binomial probability model is summarized in Figure 9. The distribution is defined by two parameters, consisting of the mean value m (the mean number of listener errors that established feature-state scores) and a parameter k for which an estimate can be made based on the mean and variance of the data population (in units of "nrs. of errors"). These values are obtained by a simple transformation on the mean and variance values obtained from conventional scores. No adequate tables of values of the negative binomial probability distribution have been found in the literature; however, values can be readily calculated with a programmable calculator. A program for calculating negative binomial values with a Wang Model 720 Calculator is presented in Appendix B.

Figure 10 presents the distribution of scores for the feature states obtained from testing CVSD at 16 kbps, in which the data for Speaker CH presented in Figures 7 and 8 has been combined with data for eight additional speakers. Shown for comparison is a normal ogive based on the mean score (90.3) and the standard deviation (14.9) for this sample population. These values equate to m = 3.116 representing the average number of listener errors per feature state, and to a value of k = 0.496 representing the parameters for a negative binomial probability distribution model for which points are shown plotted in comparison with the data distribution.

It can be observed from the figure that 80 percent of the actual data population equaled or exceeded a value of 81.3. An ITL is shown in relation to the confidence band for the data distribution: it discloses an estimate that at p = 0.95, 80 percent of the data population will equal or exceed a score of 71.9. The difference between the ITL value and the actual 80 percent data value takes into consideration the sample size that was the basis for the determination.

$$\Pr[X = x] = \binom{k+x-1}{x} p^k q^x \qquad x = 0, 1, 2, \ldots\ldots$$

$$0 < p < 1$$

$$\text{Parameters} \begin{cases} E(x) = m \\ k \end{cases} \qquad k \text{ pos.}$$

$$k^* = \left| \frac{m^2}{s^2 - m} \right| \qquad (m = \text{Mean nr. of Errors})$$

$$p = \left(\frac{k}{m+k} \right) \qquad q = \left(\frac{m}{m+k} \right)$$

Figure 9. Negative Binomial Probability Distribution Parameters

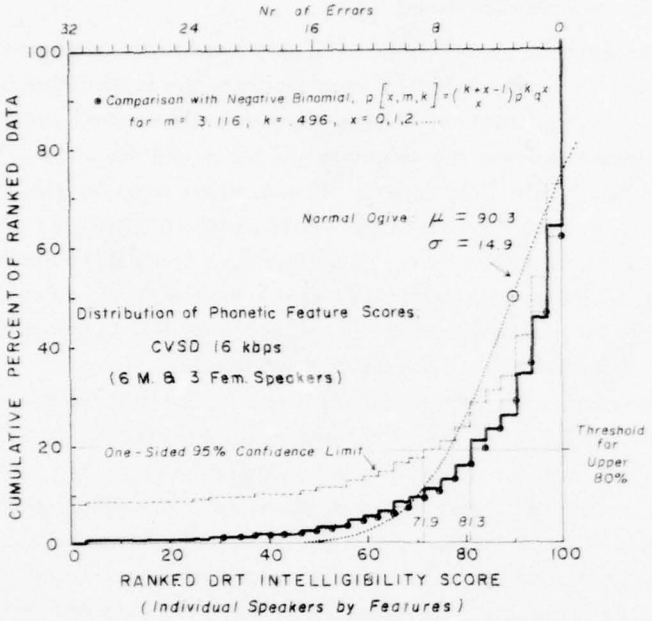

Figure 10. Distribution of Phonetic Feature Intelligibility Scores Obtained for Nine Speakers, with CVSD Processing at 16 kbps

This determination of the ITL and the resulting value obviously do not take into account the identity of the "bad" intelligibility scores that made up the bottom 20 percent of the distribution. On the other hand, in critical voice communications, poor intelligibility performance for any of the phonetic features in combination with any of the speakers presents a risk in terms of possible consequences because of messages being misunderstood, or time lost while messages are repeated. An ITL rating provides a more meaningful assessment of the degree of that risk than a conventional score specifying average intelligibility performance. The distributions of scores reveal that voice systems can have relatively high average scores and still have a significant proportion of low scores; the ITL focuses on the critical low 20 percent (in these examples) and provides information about the lower tail of the distribution of scores.

If the identity of "bad" scores is required for the purpose of guidance for research in improving intelligibility performance, or for comparing different processors, that information is readily available from conventional listings of intelligibility data.

It can be observed in Figure 10 that the negative binomial model gave good agreement with the actual data distribution. The data frequencies are compared

with theoretical frequencies based on the negative binomial model in Table 1. A chi-squared test comparing the data and the model resulted in a value of p = 0.395; the null hypothesis (for conformity with a negative binomial model) was not rejected for this data obtained from testing CVSD at 16 kbps.

In Figure 11 the distribution of scores obtained from testing CVSD at 32 kbps is presented together with the 95 percent upper confidence limit for the empirical data distribution. The data indicate that there is 0.95 probability that 80 percent of the population of intelligibility scores will equal or exceed 87.5. (In the actual data, the threshold was 93.8.) The negative binomial probability model based on the parameters observed in connection with this data led to an identical ITL estimate: 87.5. The data frequencies are compared with the theoretical negative binomial model in Table 2; a chi-squared test comparing the data with the model resulted in a value of p = 0.44.

A third example of the distribution of intelligibility scores, from a test with nine speakers, is presented in Figure 12, representing a summary of a test of CVSD operating at 9.6 kbps. In this case the ITL was considerably lower than the two previous examples: it is estimated that at a confidence level of 0.95, 80 percent of the population of scores will equal or exceed 53.1. When the average intelligibility scores are compared for 16 kbps and 9.6 kbps CVSD, the difference was 10.2 points (90.3 vs 80.1). However, the difference in ITL values was 18.8 points (71.9 vs 53.1).

The data frequencies are compared with the theoretical negative binomial model based on the parameters derived from the data for 9.6 kbps CVSD in Table 3. Again in this case, a value of p was obtained indicating that the negative binomial probability distribution gave a reasonable approximation to the data distribution. As with the previous examples, the negative binomial model led to an identical ITL value (53.1) as obtained with the empirical data distribution.

Intelligibility data frequencies obtained from testing an LPC-10 vocoder algorithm (LPC-23*) operating at 2400 BPS, with six male speakers and two independent presentations to the listener crew are summarized in the distribution of scores presented in Figure 13. Also shown are the normal and negative binomial forms based on the parameters for this distribution and the upper one-sided 95 percent confidence belt for the empirical data distribution. From the confidence limit, an ITL estimate for this processor was obtained: there is a 95 percent probability that 80 percent of the population of intelligibility scores will equal or exceed 75.0.

Table 4 compares the data frequencies and the theoretical negative binomial frequencies based on the parameters calculated for this set of data. The value of chi-squared and corresponding value of p = 0.098 indicate not to reject the null hypothesis for conformity of the data distribution with the negative binomial form.

Table 1. Comparison of Intelligibility Data Frequencies and Negative Binomial Probability Distribution (m = 3.116, k = 0.4962, N = 216)

DRT Score	Nr of Listener Errors	Theoretical Probability p(x;m,k)	Theoretical Frequency pN	Cumulative Probability Σ p	Cumulative Frequency Σ pN	Data Frequency f_s	$\frac{(f_s - pN)^2}{pN}$	
100.00	0	.373	80.619	.373	80.619	76	.265	
96.87	1	.160	34.513	.533	115.133	41	1.219	
93.75	2	.103	22.276	.636	137.409	25	.333	
90.62	3	.074	15.992	.710	153.401	18	.252	
87.50	4	.056	12.059	.766	165.461	6	3.044	
84.37	5	.043	9.356	.809	174.818	5	2.028	
81.25	6	.034	7.394	.844	182.213	11	1.759	
78.12	7	.027	5.921	.871	188.134	6		
75.00	8	.022	4.786	.893	192.921	4	.0467	
71.87	9	.018	3.898	.911	196.820	4		
68.75	10	.015	3.194	.926	200.014	2	.168	
65.62	11	.012	2.629	.938	202.644	4		
62.50	12	.0101	2.173	.948	204.818	2	.299	
59.37	13	.0083	1.802	.957	206.621	2		
56.25	14	.0069	1.499	.964	208.120	3		
53.12	15	.0058	1.250	.969	209.371		.029	
50.00	16	.0048	1.044	.974	210.415	1		
46.87	17	.0040	.874	.978	211.290	2		Chi-Squared= 9.474
43.75	18	.0034	.733	.982	212.024			with 9 d.f.
40.62	19	.0029	.616	.984	212.640			p=.395
37.50	20	.0024	.518	.987	213.158	1		
34.37	21	.0020	.436	.989	213.594			
31.25	22	.00170	.367	.991	213.962			
28.12	23	.00143	.310	.992	214.273	1		
25.00	24	.00121	.262	.993	214.535	1		
21.87	25	.00102	.221	.994	214.756			
18.75	26	.00086	.187	.995	214.944		.0308	
15.62	27	.00073	.158	.996	215.103			
12.50	28	.00062	.134	.9965	215.237			
9.37	29	.00052	.114	.9970	215.351			
6.25	30	.00044	.096	.9974	215.448			
3.12	31	.00038	.082	.9978	215.530	1		
0	32	.00032	.069	.9982	215.600			

CVSD at 16 kbps. Mean DRT Score:90.26 S.D.:14.902 N=216 (6M & 3Fem.Speakers)

Table 2. Comparison of Intelligibility Data Frequencies and Negative Binomial Probability Distribution (m = 1.598, k = 0.3962, N = 216)

DRT Score	Nr of Listener Errors	Theoretical Probability p(x;m,k)	Theoretical Frequency pN	Cumulative Probability Σ p	Cumulative Frequency Σ pN	Data Frequency fs	$\frac{(f_s - pN)^2}{pN}$
100.00	0	.527	113.855	.527	113.855	109	.207
96.87	1	.167	36.147	.694	150.002	44	1.706
93.75	2	.094	20.221	.788	170.224	22	.157
90.62	3	.060	12.943	.848	183.168	10	.669
87.50	4	.041	8.806	.889	191.974	8	
84.37	5	.029	6.205	.917	198.179	4	.604
81.25	6	.021	4.472	.938	202.652	3	
78.12	7	.0152	3.274	.953	205.927	4	.718
75.00	8	.0112	2.426	.965	208.353	3	
71.87	9	.0084	1.813	.973	210.167	3	.732
68.75	10	.00632	1.365	.979	211.533	2	
65.62	11	.00478	1.034	.984	212.567	1	
62.50	12	.00364	.787	.988	213.354		
59.37	13	.00278	.601	.991	213.956		
56.25	14	.00213	.461	.993	214.417	1	
53.12	15	.00164	.354	.994	214.772		
50.00	16	.00126	.273	.996	215.046	1	.005
46.87	17	.00097	.211	.997	215.257		
43.75	18	.00075	.163	.997	215.421	1	

Chi-squared = 4.798
with 5 d.f.
p=.441

CVSD at 32 kbps. Mean DRT Score: 95.01 S.D.: 8.865 N=216(6M & 3 Fem.Speakers)

Table 3. Comparison of Intelligibility Data Frequencies and Negative Binomial Probability Distribution ($m = 6.366$, $k = 0.8676$, $N = 216$)

DRT Score	Nr of Listener Errors	Theoretical Probability $p(x;m,k)$	Theoretical Frequency pN	Cumulative Probability Σp	Cumulative Frequency ΣpN	Data Frequency f_s	$\frac{(f_s - pN)^2}{pN}$
100.00	0	.159	34.305	.159	34.305	48	5.467
96.87	1	.121	26.193	.280	60.498	29	.301
93.75	2	.100	21.525	.380	82.024	17	.951
90.62	3	.084	18.107	.464	100.132	9	4.580
87.50	4	.071	15.408	.535	115.540	10	1.898
84.37	5	.061	13.201	.596	128.742	15	.245
81.25	6	.0526	11.361	.649	140.104	11	.011
78.12	7	.0454	9.809	.694	149.913	7	.804
75.00	8	.0393	8.490	.733	158.404	7	
71.87	9	.0341	7.362	.767	165.766	4	1.485
68.75	10	.0296	6.393	.797	172.159	8	.404
65.62	11	.0257	5.558	.823	177.718	6	.035
62.50	12	.0224	4.838	.845	182.556	7	.966
59.37	13	.0195	4.214	.865	186.771	3	
56.25	14	.0170	3.673	.882	190.445	4	.100
53.12	15	.01483	3.204	.897	193.649	6	
50.00	16	.01294	2.797	.909	196.446	2	.666
46.87	17	.01130	2.442	.921	198.889	2	
43.75	18	.00987	2.133	.931	201.022	3	.039
40.62	19	.00863	1.864	.939	202.887	3	
37.50	20	.00754	1.630	.947	204.517	2	.649
34.37	21	.00659	1.425	.953	205.943	3	
31.25	22	.00577	1.247	.959	207.190	1	Chi-squared = 19.707 with 15d.f.
28.12	23	.00505	1.091	.964	208.281	1	
25.00	24	.00442	.954	.969	209.236	2	1.105 p=.183
21.87	25	.00387	.835	.973	210.072		
18.75	26	.00338	.731	.976	210.804	4	
15.62	27	.00296	.641	.979	211.445		
12.50	28	.00259	.561	.982	212.006		
9.37	29	.00227	.491	.984	212.498		
3.12	31	.00174	.377	.988	213.307	1	.001
(-28.1)						1	

CVSD at 9.6 kbps. Mean DRT Score: 80.11 S.D.: 22.77 N=216 (6M & 3Fem.Speakers)

Table 4. Comparison of Intelligibility Data Frequencies and Negative Binomial Probability Distribution (m = 3.501, k = 0.7274, N = 288)

DRT Score	Nr of Listener Errors	Theoretical Probability p(x;m,k)	Theoretical Frequency pN	Cumulative Probability ΣP	Cumulative Frequency ΣpN	Data Frequency f_s	$\frac{(f_s - pN)^2}{pN}$
100.00	0	.278	80.054	.278	80.054	89	1.000
96.87	1	.167	48.213	.445	128.267	48	.001
93.75	2	.120	34.477	.565	162.745	27	1.622
90.62	3	.090	25.952	.655	188.697	28	.162
87.50	4	.070	20.023	.725	208.720	15	1.260
84.37	5	.054	15.674	.779	224.395	8	3.757
81.25	6	.043	12.388	.822	236.783	17	1.717
78.12	7	.034	9.857	.856	246.641	11	.133
75.00	8	.027	7.883	.884	254.525	10	.569
71.87	9	.022	6.329	.906	260.855	5	
68.75	10	.018	5.097	.923	265.953	3	1.027
65.62	11	.014	4.116	.938	270.069	8	3.665
62.50	12	.012	3.330	.949	273.400	5	
59.37	13	.009	2.699	.959	276.100	4	1.464
56.25	14	.008	2.191	.966	278.292	2	
53.12	15	.006	1.781	.972	280.073		
50.00	16	.0050	1.450	.978	281.524	1	
46.87	17	.0041	1.181	.982	282.705		.870
43.75	18	.0033	.963	.985	283.669	2	
40.62	19	.0027	.786	.988	284.455	2	Chi-Squared = 17.350 with 11 d.f. p = .098
37.50	20	.0022	.642	.990	285.097		
34.37	21	.00182	.524	.992	285.622	1	
31.25	22	.00148	.429	.993	286.051		
28.12	23	.00121	.351	.994	286.402		
25.00	24	.00099	.287	.995	286.689	1	.103
21.87	25	.00081	.235	.996	286.924	1	

LPC-10 Vocoder Algorithm (LPC-23*) at 2400 BPS. Mean DRT Score: 89.06
Standard Deviation: 14.10 N = 288 (Six Male Speakers, two presentations)

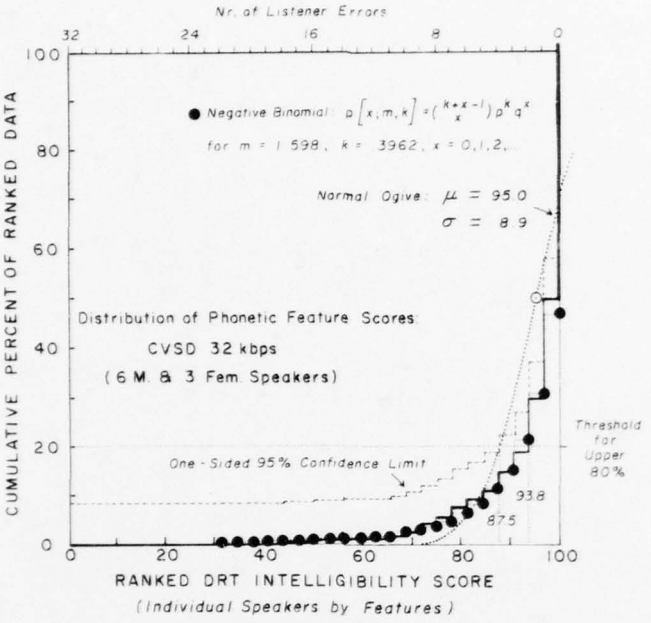

Figure 11. Distribution of Phonetic Feature Intelligibility Scores Obtained for Nine Speakers with CVSD Processing at 32 kbps

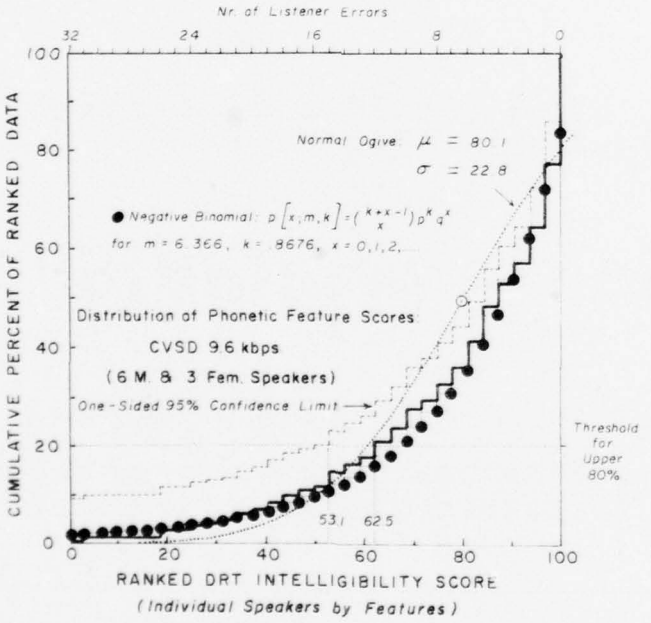

Figure 12. Distribution of Phonetic Feature Intelligibility Scores Obtained for Nine Speakers with CVSD Processing at 9.6 kbps

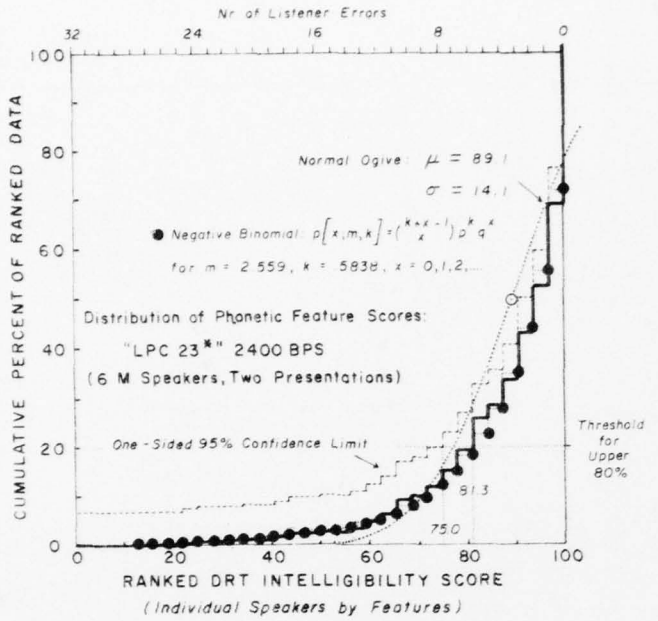

Figure 13. Distribution of Phonetic Feature Intelligibility Scores Obtained for Six Speakers (Two Presentations) with an LPC Vocoder

The intelligibility threshold level (ITL) estimate calculated with the theoretical cumulative probability associated with the negative binomial model resulted in an identical value (75.0) as that obtained from the empirical data distribution.

An example based on intelligibility data obtained with another type of voice processor, a sixteen-channel vocoder operating at 2400 BPS tested with six male and three female speakers, is presented in Figure 14. Comparing the data distribution for the LPC-10 vocoder algorithm (Figure 13) illustrates again that differences in ITL ratings tend to be greater than those observed for mean intelligibility scores. The channel vocoder, with a mean score of 83.0, was approximately six points lower than the mean intelligibility score obtained for the LPC-10 vocoder algorithm. However, the difference in ITL ratings for the two processors was more than 15 points (75.0 vs. 59.4).

The values associated with the data and the negative binomial model are listed in Table 5, together with the results of a chi-squared test comparing the empirical data frequencies and the negative binomial model. In this case, the value of p was 0.025, and the null hypothesis was rejected. Although the data represented a poor fit to the negative binomial form, the ITL calculated from the negative binomial model differed by only one quantum level in the data (ITL of 62.5 from the model vs. 59.4 from the actual data distribution).

Table 5. Comparison of Intelligibility Data Frequencies and Negative Binomial Probability Distribution (m = 5.455, k = 0.9910, N = 216.

DRT Score	Nr of Listener Errors	Theoretical Probability p(x;m,k)	Theoretical Frequency pN	Cumulative Probability P	Cumulative Frequency pN	Data Frequency f_s	$\frac{(f_s - pN)^2}{pN}$
100.00	0	.156	33.771	.156	33.771	49	6.868
96.87	1	.131	28.321	.287	62.093	24	.659
93.75	2	.110	23.859	.398	85.952	14	4.074
90.62	3	.093	20.130	.491	106.083	17	.487
87.50	4	.079	16.997	.570	123.081	13	.940
84.37	5	.066	14.358	.636	137.440	15	.029
81.25	6	.0562	12.133	.692	149.573	9	.809
78.12	7	.0475	10.254	.740	159.827	15	2.197
75.00	8	.0401	8.668	.780	168.496	16	.205
71.87	9	.0339	7.328	.814	175.824	13	.390
68.75	10	.0287	6.196	.843	182.021	3 }	
65.62	11	.0243	5.239	.867	187.260	3 }	2.583
62.50	12	.0205	4.430	.887	191.691	5	.073
59.37	13	.0173	3.747	.905	195.438	6 }	
56.25	14	.0147	3.168	.919	198.607	1 }	.001
53.12	15	.0124	2.680	.932	201.287	3 }	
50.00	16	.0105	2.266	.942	203.554	2 }	.001
46.87	17	.00887	1.917	.951	205.472	3)	
43.75	18	.00750	1.621	.959	207.093	1)	
40.62	19	.00635	1.371	.965	208.465	1)	
37.50	20	.00537	1.160	.970	209.626	1)	.157
34.37	21	.00454	.981	.975	210.607	)	
31.25	22	.00384	.830	.979	211.438	5)	Chi-Squared= 24.733 with 13 d.f. p = .025*
28.12	23	.00325	.702	.982	212.140	1)	
25.00	24	.00275	.594	.985	212.735	)	
21.87	25	.00232	.502	.987	213.237	)	
18.75	26	.00196	.425	.989	213.663	)	
15.62	27	.00166	.359	.991	214.022	1)	
12.50	28	.00140	.304	.992	214.327	)	1.260
9.37	29	.00119	.257	.993	214.584	)	
6.25	30	.00100	.217	.994	214.802	)	
3.12	31	.00085	.184	.995	214.986	)	
0	32	.00072	.155	.996	215.142	)	
-3.12	33	.00061	.131	.997	215.274	1)	

Sixteen-Channel Vocoder at 2400 BPS. Mean DRT Score: 82.95 S.D.: 18.62
N = 216 (Six Male & three Female Speakers)

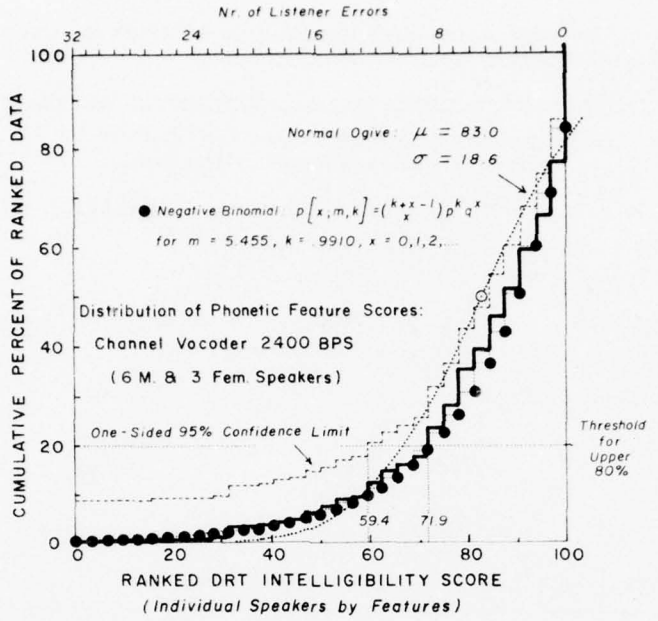

Figure 14. Distribution of Phonetic Feature Intelligibility Scores Obtained for Nine Speakers with a Sixteen-Channel Vocoder at 2400 BPS

2.4 Results of Chi-Squared Tests for Conformity of Intelligibility Data Distributions with Negative Binomial Probability Model

To further test conformity of intelligibility data distributions with a negative binomial probability model, a series of chi-squared tests were performed comparing empirical data distributions with the negative binomial probability model on 81 sets of speech intelligibility data on hand from prior tests and evaluations of various speech processor algorithms and hardware. These tests included data from the combinations of speakers and processors summarized in Table 6, ranging from data for a single speaker (both male and female speakers) to as many as twelve male speakers, the processors including LPC and channel vocoders and CVSD at three data rates. Intelligibility data were analyzed separately for voiced and unvoiced feature data as well as total data summaries. The results of this exploratory study are summarized in Figure 15 which presents the distribution of the values of probability based on the values of chi-squared in conjunction with the degrees of freedom. Almost one-fourth of the 81 cases resulted in values of p less than 0.05, for which the null hypothesis would be rejected. Thus the agreement with the negative binomial probability model was far from perfect. Many of

Table 6. Summary of Intelligibility Data Sets Tested for Conformity with the Negative Binominal Probability Model

DRT Intelligibility Data Tested for Conformity with the Negative Binomial Distribution	
Talkers:	Single Speaker (M;Fem) Three-Speaker (3 M; 3 Fem) Six M. Speakers (two versions) Nine Speakers (6 M, 3 Fem) Twelve M. Speakers
Processing:	LPC Vocoders; Channel Vocoders; APC; CVSD (3 rates); Hybrid Vocoders.
NR of Listeners:	Eight
Data Sets:	Feature scores (by individual speakers) -Voiced features -Unvoiced features -All features Total DRT scores (by listeners/speakers)
Total Data Groups Used for Chi-Squared Tests: 81	

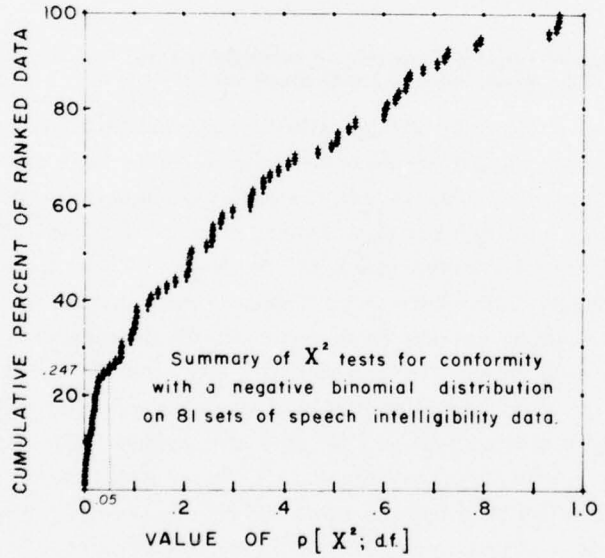

Figure 15. Summary of Chi-Squared Tests for Conformity of Speech Intelligibility Data Distributions with a Negative Binomial Probability Model

Table 7. Deviations in Conformity with the Negative Binomial Probability Model Found with Replication of Listening Tests

COMPARISON OF PROBABILITIES ESTIMATING THE CONFORMITY WITH A NEGATIVE BINOMIAL PROBABILITY DISTRIBUTION, FOR DIFFERENT DATES OF PRESENTATION TO LISTENERS (WITH ASSOCIATED PARAMETERS).

Processor	Test Nr.	Date of Presentation	Mean DRT Score	DRT Standard Deviation	Negative Binomial Model m	k	X^2	d.f.	$p(X^2, d.f.)$
LPC	#2050	26 Apr	90.58	13.31	3.015	.6014	9.603	7	.212
		27 May	91.25	11.97	2.799	.6603	17.742	7	.013*
LPC at	#2051	29 Apr	85.22	16.62	4.730	.9498	13.849	9	.128
1% BER		27 May	86.74	14.95	4.244	.9668	19.535	9	.021*
PLPC	#2043	8 April	92.16	12.39	2.508	.4760	6.790	7	.451
		20 May	92.62	11.269	2.362	.5241	9.549	7	.216
PLPC at	#2047	22 April	87.65	13.661	3.952	1.031	7.139	7	.415
1% BER		29 April	88.37	14.191	3.723	.8201	17.523	9	.041*
LPC at	#2052	6 May	76.65	20.62	7.473	1.549	17.657	13	.171
3% BER		13 May	77.52	19.47	7.195	1.637	16.928	13	.203
LPC at	#2053	6 May	68.79	23.06	9.987	2.244	11.637	12	.475
5% BER		13 May	68.31	24.38	10.140	2.028	20.686	16	.191
PLPC at	#2048	22 April	79.79	19.26	6.466	1.3272	11.100	12	.520
3% BER		20 May	81.36	19.15	5.966	1.1267	5.298	11	.916
PLPC at	#2049	22 April	70.46	22.04	9.452	2.2175	16.467	13	.225
5% BER		20 May	71.89	22.09	8.994	1.9743	25.365	14	.031*
LPC with	#2069	21 July	88.56	14.61	3.660	.7361	7.547	8	.479
"new"		4 Aug	89.67	13.56	3.306	.7042	12.276	9	.198
speakers		1 Sept	89.71	14.11	3.292	.6345	13.318	7	.065

the data sets represented composite results from two separate presentations to the listener crews. It was subsequently found that a curious trend was present in the data sets in which the composite data (for two or more presentations to listeners) showed poorer agreement with the model; data from the first presentation to listeners agreed with the negative binomial model in all of these cases. Further details of this anomaly are discussed in the following section.

A comparison of ITL values obtained from the cumulative data distributions, and ITL's obtained with the use of the negative binomial probability model, showed good agreement, even for data sets that deviated significantly from the model.

2.5 Phenomenon Observed in Connection with Replication of Listening Tests

In the speech intelligibility data analyzed for conformity with the negative binomial probability model many of the data sets represented composite results of presentation of the DRT recordings to the listener crew on two separate occasions a week or more apart; one set of data involved three presentations at intervals of a month. The initial chi-squared tests reported in the previous section were performed on composite data combined over all presentations of a particular DRT recording. Subsequently in the studies, separate assessments were made for the data resulting from each separate presentation to listeners. A pattern was found in the results: agreement of the intelligibility data with the negative binomial form was almost always higher for the first presentation to listeners, than for subsequent presentations. These findings are presented in Table 7.

The listener tests were performed "blind," that is, the listening crew had no knowledge as to the identity of any particular speech processor or the processing conditions. Any given test was always interspersed with other tests in a random manner. The listeners had much prior experience with these scramblings of the DRT word lists, and while there was a slight tendency for scores to increase at the second presentation, in most instances the change was not statistically significant.

Because of these considerations it is difficult to conceive any explanation as to why the intelligibility data distributions tended to show increasing deviation from the negative binomial probability form with the second and subsequent presentations.

3. DETERMINATION OF INTELLIGIBILITY THRESHOLD LEVELS (ITL's)

It was established that the distributions of diagnostic intelligibility scores were sufficiently close to the negative binomial probability model that two alternative procedures for determining ITL's were available.

3.1 Determination of ITL's from Cumulative Data Distributions

In this procedure the sample population of diagnostic intelligibility scores is rank-ordered, and the ranked data are utilized in constructing a cumulative table of scores in relation to cumulative percentiles of the data population. In these studies a program was written for the Wang Model 720 Calculator to perform the ranking. Data values were entered into an auxiliary memory via the calculator keyboard, and a program subsequently rearranged the table in rank-order. Each datum was paired with a code number to maintain identification of the phonetic feature and speaker associated with each value.

A confidence band for the distribution was established by the method of Kolmogorov, in which a one-sided confidence band is defined by a simple offset of population percentiles by an appropriate quantile that is a function of the value of p and the sample size. Kolmogorov's method assumes a random sample X1, X2, ..., Xn of size n associated with some unknown distribution function F(x). For the confidence coefficient to be exact, the method requires that the samples are from a continuous distribution; however, if the random variables are discrete (as in this case) the confidence band is conservative, that is, the "true" but unknown confidence coefficient is greater than the stated one.

Kolmogorov's method is most readily used by constructing a tabular listing or a graphical representation of the empirical distribution function. In a graphical representation, as in Figures 8 and 9 and the examples in Appendix A, each datum can be considered as a vertical segment that establishes 100/n percent of the total population; end points of adjacent segments are joined to form the distribution function S(x) which is terminated at zero and 100 percent of the sample. A tabular listing as in Table 4 can be utilized for listing cumulative percentiles representing the end points or boundaries of each segment (datum).

A confidence band with confidence coefficient 1-a is created with the use of the 1-a quantile from a table of the Kolmogorov test statistic (Appendix D). In determining ITL's, the upper one-sided confidence band is of interest, which is formed by vertical displacement of the empirical distribution function graph by the value of Q_{1-a} from a table of the Kolmogorov test statistic. Thus the confidence boundary is an exact replica of the empirical distribution, offset by an appropriate amount, and terminated at the values Q_{1-a} and 100 percent illustrated in the examples of Figures 8 and 9. Alternatively, the value of Q_{1-a} is added (or subtracted for a lower one-sided limit) from the cumulative percentiles of the tabular listing. If the confidence limit is denoted by U(x),

$$U(x) = S(x) + Q_{1-a}$$

forming the boundary of a one-sided 1-a confidence band which completely contains the "true" F(x).

Values for ITL's for various percentages of the data population are determined in relation to the upper one-sided confidence limit, either graphically as illustrated in Figure 9, or with more accuracy from a tabular listing as in the example of Table 4.

Diagnostic intelligibility data populations presented here involved 24 intelligibility scores (for the phonetic feature states) from each of the speakers in an intelligibility test. With two and more speakers, the number of samples exceeded 40, and the approximation for calculating the one-sided quantile of the Kolmogorov test statistic at p = 0.95 was used:

For n > 40,

$$Q_{0.95} = 1.22/\sqrt{n} \ .$$

3.2 Determination of ITL's from a Negative Binomial Probability Model

After a mean DRT score and its variance have been determined for a set of intelligibility data, conversions to parameters m and k that characterize a negative binomial probability model are as follows:

Mean nr. of listener errors = $m = (100 - \bar{D})/3.125$

where $\bar{D}$ = mean DRT score

Listener error variance = $s_e^2 = S_D^2/(3.125)^2$

where S_D^2 = DRT score variance

$$\text{Estimate of } k = k^* = \left| \frac{m^2}{s_e^2 - m} \right|$$

Here is an example of the derivations for the data of Figure 9 and Table 1, from intelligibility data from testing CVSD at 16 kbps with nine speakers.

Mean DRT score = $\bar{D} = 90.26$ DRT score variance = $S_D^2 = 222.06$ $n = 216$

Consequently

$$m = \frac{100 - 90.26}{3.125} = 3.12 \text{ (mean listener errors per score)}$$

$$s_e^2 = \frac{222.06}{(3.125)^2} = 22.74 \text{ (listener error variance)}$$

leading to the estimate for k:

$$k^* = \frac{(3.12)^2}{22.74 - 3.12} = 0.496$$

Using these values in calculating probabilities based on a negative binomial distribution:

$$p = \frac{k}{m + k} = 0.137 \quad p^k = 0.373 = \text{probability of no errors}$$

$$q = \frac{m}{m + k} = 0.863$$

$$p[x;k, m] = \binom{x + k - 1}{x} p^k q^x \quad x = \text{nr of errors} = 0, 1, 2, \ldots$$

$$= \binom{x - 0.504}{x} (0.3731)(0.863)^x$$

resulting in the probability values and distribution function presented in Table 1.

A confidence band for the negative binomial distribution function can be formed with the same Kolmogorov test statistic and procedure used with the empirical data distributions; the Kolmogorov test statistic is valid without regard to the form of the distribution. Thus a confidence band for the negative binomial distribution function is formed by a displacement of the percentiles associated with the distribution function by an appropriate value of Q_{1-a} for the Kolmogorov test statistic.

A number of intelligibility data distributions are shown in Appendix A, together with 95 percent one-sided confidence bands and ITL values.

4. INTELLIGIBILITY THRESHOLD LEVEL (ITL) RATINGS FOR SOME VOICE PROCESSORS

Conventional intelligibility scores are compared with ITL values in Tables 8, 9, and 10, and illustrate that differences in intelligibility scores usually become magnified in differences in ITL's for the same intelligibility data. The data also illustrate that typically there are large differences between the intelligibility scores

Table 8. Intelligibility Threshold Level (ITL) Ratings for LPC Vocoder Algorithms Operating with Random Bit Errors

Effects of Random Bit Errors on Serial 2400 BPS Linear Predictive Vocoders

based on six male speakers and two presentations to eight listeners

Test Configuration	Mean Score	80% Intelligibility Threshold Level (p = 0.95)*		
		Voiced	Unvoiced	Total
LPC Vocoder				
Zero BER	90.9	81.3	68.8	75.0
1%	86.0	68.8	53.1	65.6
3%	77.1	53.1	40.6	53.1
5%	68.6	40.6	31.3	40.6
Piecewise-LPC Vocoder				
Zero BER	92.4	78.1	71.9	78.1
1%	88.0	71.9	65.6	71.9
3%	80.6	62.5	43.8	62.5
5%	71.2	50.0	31.3	43.8

*There is 95% Confidence that 80% of the specified population of diagnostic intelligibility scores (feature scores, by Speakers) will equal or exceed the stated value.

Table 9. Comparisons of Conventional Intelligibility Scores and ITL's from Diagnostic Rhyme Test Scores, 6 Male and 3 Female Speakers and 8 Listeners

Processor	Mean Score	80% Intelligibility Threshold Level (95% Confidence)*		
		Voiced	Unvoiced	Total
CVSD 9.6 kbps	80.1	62.5	25.0	53.1
CVSD 16 kbps	90.3	81.3	56.3	71.9
CVSD 32 kbps	95.0	90.6	71.9	87.5
Ch. Vocoder 2400	83.0	62.5	37.5	59.4

*There is 95% confidence that 80% of the population of diagnostic intelligibility scores (feature scores, by speakers) will equal or exceed the stated value.

Table 10. Comparisons of ITL's Derived from Empirical Data Distributions, and Obtained with the Negative Binomial Probability Model

Comparison of 80% ITL's (p = 0.95)
Values from data distribution, vs. Negative Binomial model
(Data from 6 M. & 3 Fem. Speakers, 8 Listeners)

	Voiced Features		Unvoiced		Total	
Processor	Data	(Model)	Data	(Model)	Data	(Model)
CVSD 9.6 kbps	62.5	(65.6)	25.0	(28.1)	53.1	(53.1)
CVSD 16 kbps	81.3	(81.3)	56.3	(53.1)	71.9	(75.0)
CVSD 32 kbps	90.6	(90.6)	71.9	(71.9)	87.5	(87.5)
Ch. Voc. 2400	62.5	(65.6)	37.5	(40.6)	59.4	(62.5)
Ch. Voc. 2400	65.6	(65.6)	56.3	(56.3)	65.6	(65.6)
APC-4 6400	68.8	(68.8)	40.6	(40.6)	59.4	(62.5)

for the voiced and for the unvoiced speech sounds, in comparison with the total ensemble, and highlight the fact that the greatest potential payoff in improving speech intelligibility will come through improving the fidelity of the unvoiced speech events.

In Table 8, two linear-predictive (LPC) vocoder algorithms are compared in terms of their intelligibility performance in the presence of random bit errors as could be caused by interference or low-grade transmission channels, when no measures are provided for error protection.

Table 9 compares conventional intelligibility scores and ITL's for continuous variable-slope delta modulation (CVSD) at three data rates, and a conventional channel vocoder.

Table 10 compares the ITL's obtained from the empirical data distribution with the values obtained with the use of the negative binomial probability model. The greatest discrepancy in ITL values was a one-quantum change in the value (3.125 points); over half of these eighteen comparisons gave perfect agreement in ITL values assessed by the two method.

5. CONCLUSIONS AND RECOMMENDATIONS

Intelligibility scores for voice processors have been found to be typically characterized by highly significant differences among speakers, as well as highly significant differences among scores for the various phonetic features.

Distributions of intelligibility scores are not normally distributed, but highly skewed.

A negative binomial probability distribution was found to give good agreement with empirical intelligibility data distributions.

A new performance rating for voice communications devices, termed an Intelligibility Threshold Level (ITL), was conceived as a means of taking these findings into consideration in establishing a measure of performance that is an estimate of an intelligibility value that the majority (rather than the simple average) of intelligibility scores for a voice processor will equal or exceed, at a specified confidence level established in relation to the sample size used in obtaining the rating.

It is proposed that an ITL rating is a more meaningful assessment of the degree of risk involved in misunderstanding voice messages, or causing time to be lost in requiring messages to be repeated.

It was shown that ITL's can be determined by two alternative methods: by rank-ordering the intelligibility scores for a voice processor and constructing the cumulative distribution of data and its confidence band, or by using a negative binomial probability model for the distribution.

Chi-squared tests indicated that in most cases the negative binomial probability model gave a reasonable approximation to the data distribution.

Intelligibility Threshold Levels (ITL's) estimated with the negative binomial model differed by at most one quantum value (3.125) from ITL's determined from the empirical distributions.

It is recommended that future speech intelligibility tests and evaluations of digital voice communications processors and systems include a determination of the 80 percent ITL's at 0.95 probability, that is determine the intelligibility level for which there is a 95 percent probability that 80 percent of the population of intelligibility scores (for individual speakers and phonetic features) will equal or exceed.

Bibliography

1. Atal, B.S. and Hanauer, S.L. (1971) Speech analysis and synthesis by linear prediction of the speech wave, Jour. Acous. Soc. Am. 50(No. 2):(Part 2).
2. Bishop, Y.M.M., Fienberg, S.E., and Holland, P.W. (1975) Discrete Multivariate Analysis: Theory and Practice, MIT Press, Cambridge, Mass.
3. Conover, W.J. (1971) Practical Nonparametric Statistics, J. Wiley & Sons, NY.
4. Feller, W. (1957) An Introduction to Probability Theory and Its Applications, J. Wiley & Sons, NY.
5. Lilliefors, H.W. (1967) On the Kolmogorov-Smirnov test for normality with mean and variance unknown, Jour. Amer. Stat. Assoc. 62:399-402.
6. Roberts, J.E. and Wiggins, R.H. (1976) Piecewise Linear Predictive Coding (PLPC), Conf. Record, ICASSP, IEEE Cat. No. 76CH1067-8 ASSP, 470-473.
7. Smith, C.P. (1969) Perception of Vocoder speech processed by pattern-matching, Jour. Acous. Soc. Am.
8. Smith, C.P. (1977) Intelligibility Performance of Narrowband Linear Predictive Vocoders in the Presence of Bit Errors, ESD-TR-77-328, AF Electronic Systems Division (AFSC), Hanscom AFB, Mass.
9. Smith, C.P. (1979) Talker Variance and Phonetic Feature Variance in Diagnostic Intelligibility Scores for Digital Voice Communications Processors, Conf. Record, ICASSP, IEEE 79CH1379-7 ASSP, 456-459.
10. Voiers, W.D. et al (1973) Research on Diagnostic Evaluation of Speech Intelligibility, AFCRL = 72-694, AF Electronic Systems Division (AFSC), Hanscom AFB, Mass.
11. Voiers, W.D. (1977) Diagnostic Evaluation of Speech Intelligibility, in Speech Intelligibility and Speaker Recognition, M. Hawley, Ed., Dowden Hutchinson and Ross, Stroudsburg, Penn.

12. Voiers, W.D. and Smith, C.P. (1972) Diagnostic Evaluation of Intelligibility in Present-Day Digital Vocoders, Conf. Record, 1972 Conf. on Speech Comm. and Processing, AFCRL-72-0120, 170-174.

13. Winer, B.J. (1971) Statistical Principles in Experimental Design, McGraw-Hill Book Co., NY.

Appendix A

Voice Processor Intelligibility Data Distributions and
Intelligibility Threshold Levels (ITL's)

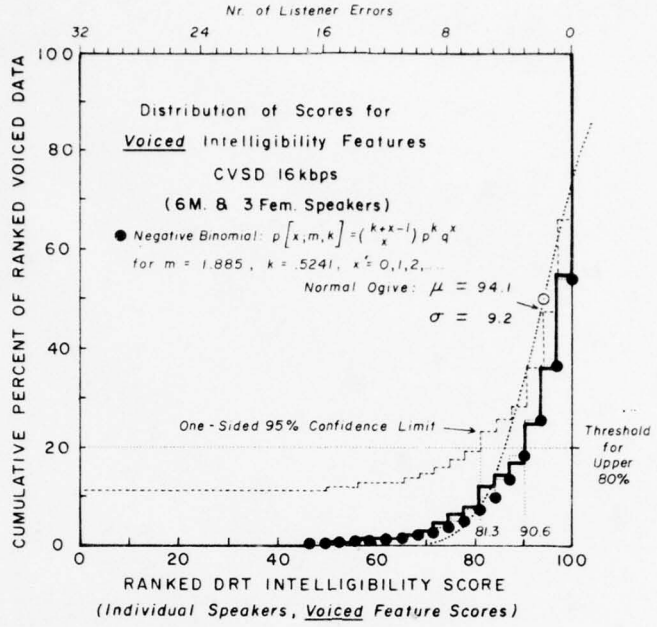

Figure A1.a. CVSD at 16 kbps: Voiced Intelligibility Features

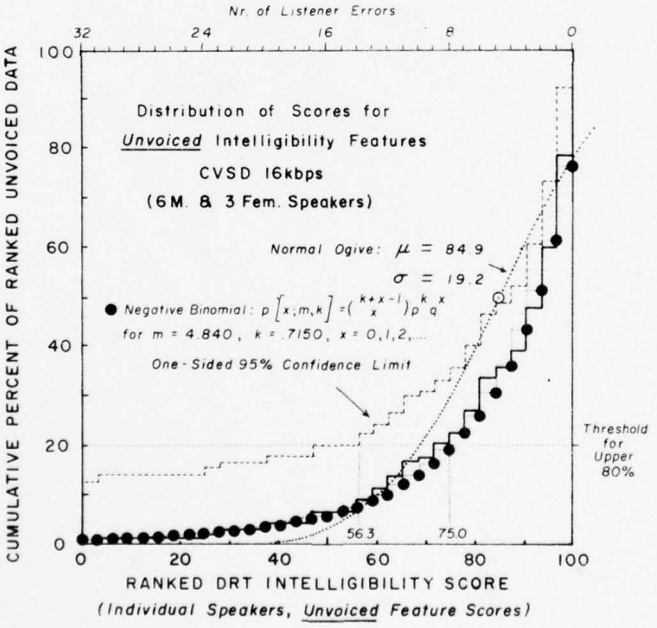

Figure A1.b. CVSD at 16 kbps: Unvoiced Intelligibility Features

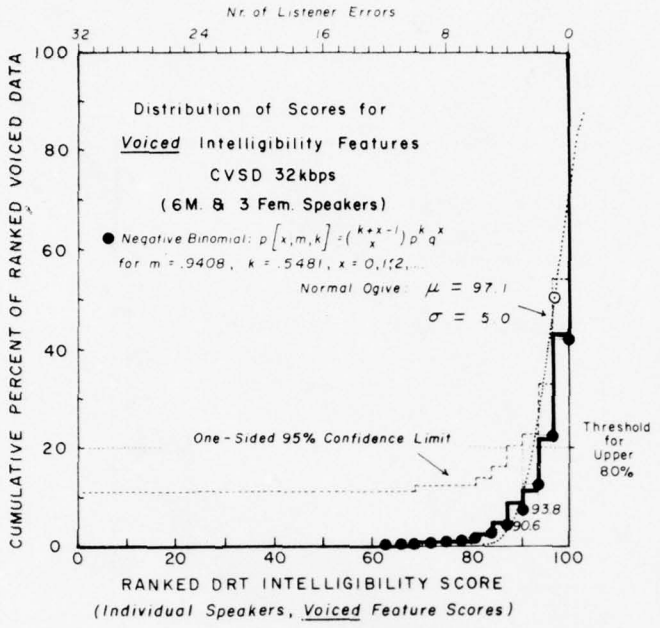

Figure A2.a. CVSD at 32 kbps: Voiced Intelligibility Features

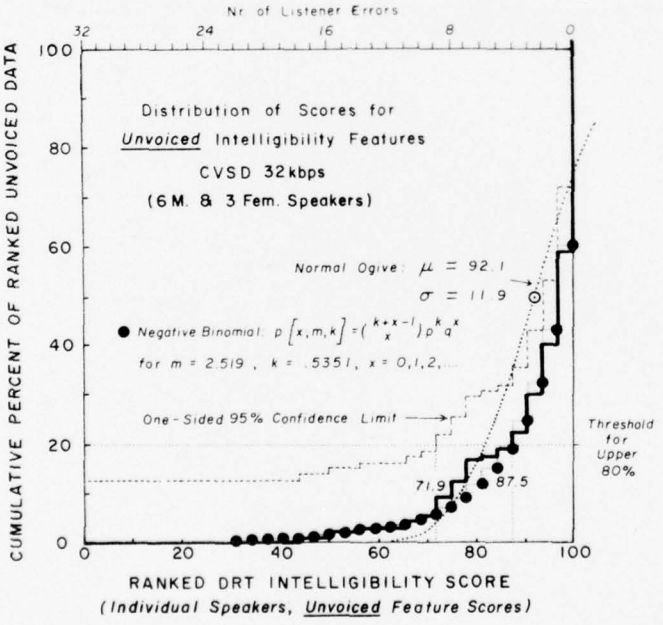

Figure A2.b. CVSD at 32 kbps: Unvoiced Intelligibility Features

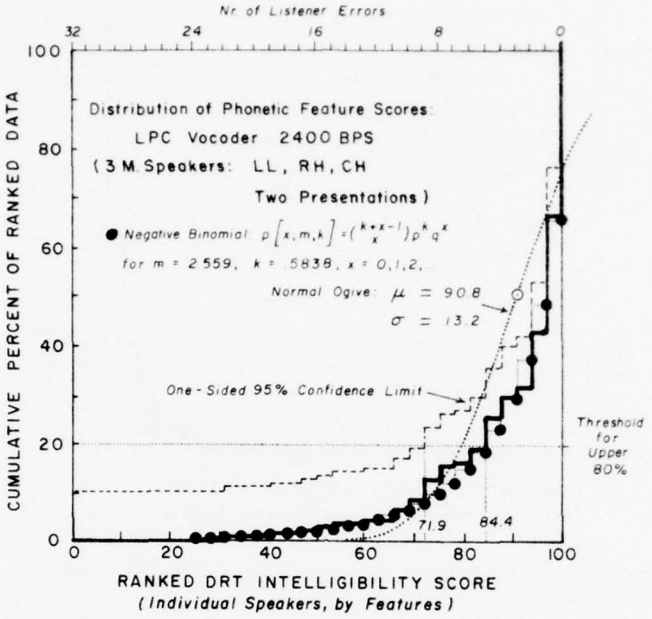

Figure A3. LPC Vocoder at 2400 BPS: Three Male Speakers (LL, RH, CH)

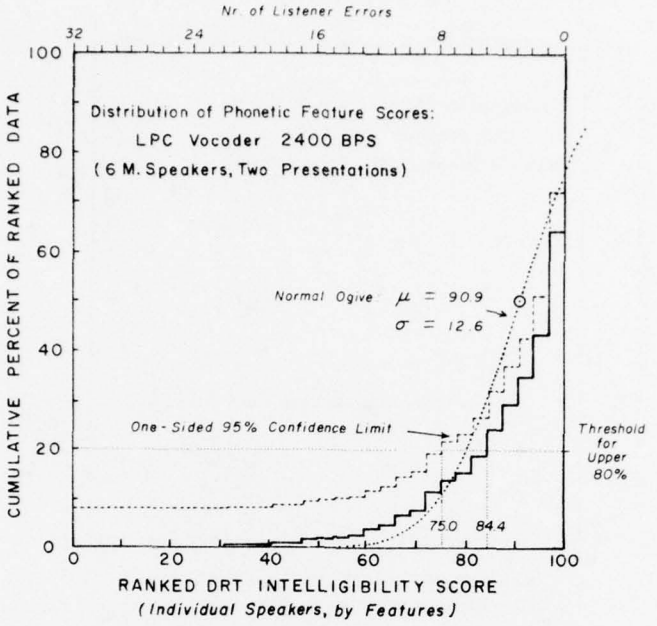

Figure A4.a. LPC Vocoder at 2400 BPS: Six Male Speakers

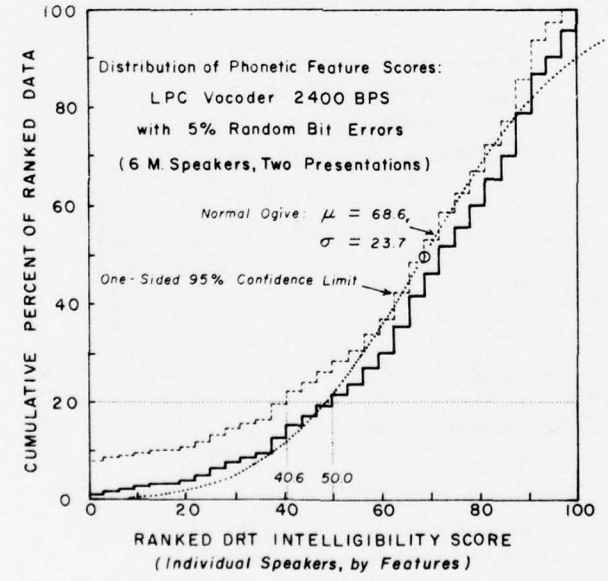

Figure A4.b. LPC Vocoder at 2400 BPS with 5 Percent Random Bit Errors (Six Male Speakers)

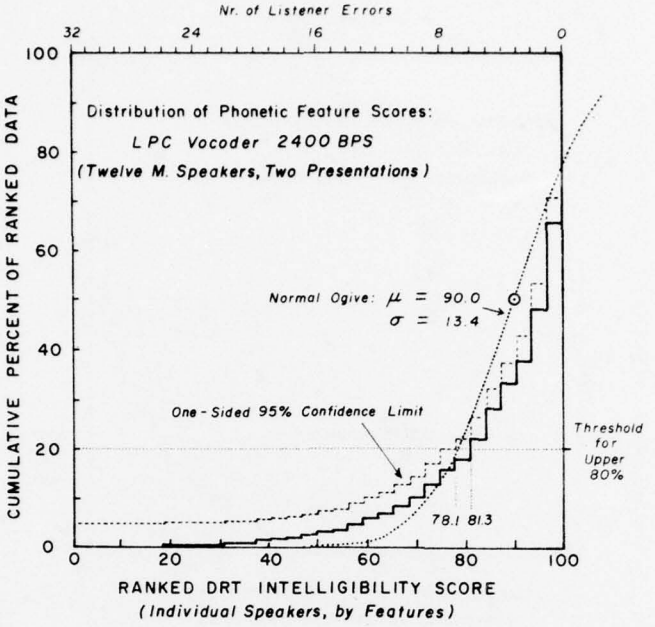

Figure A5.a. LPC Vocoder at 2400 BPS: Twelve Male Speakers

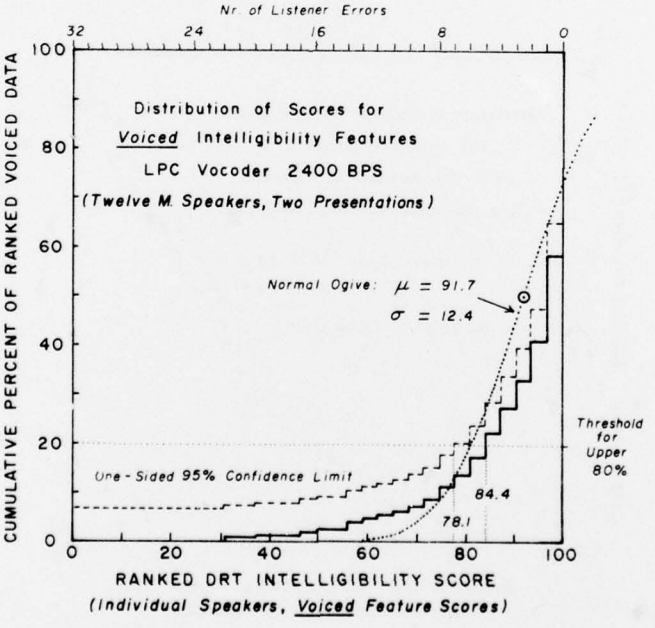

Figure A5.b. LPC Vocoder at 2400 BPS: Voiced Intelligibility Features (Twelve Male Speakers)

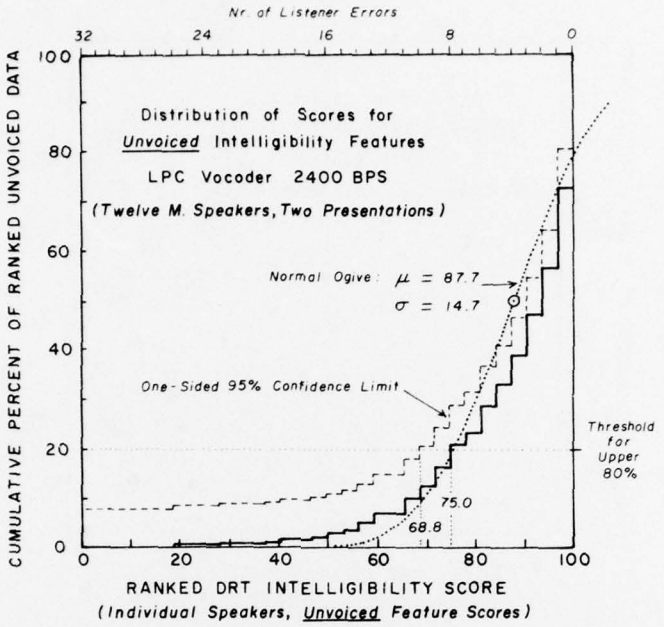

Figure A5.c. LPC Vocoder at 2400 BPS: Unvoiced Intelligibility Features (Twelve Male Speakers)

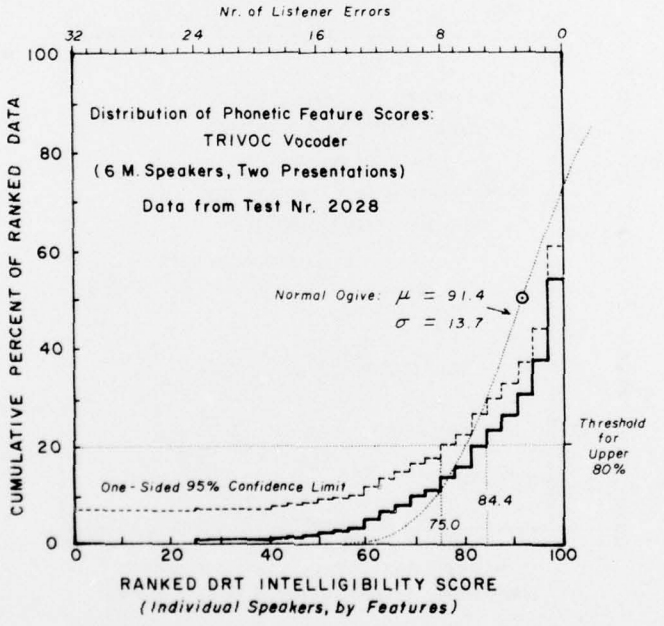

Figure A6.a. TRIVOC Vocoder at 2400 BPS: Six Male Speakers

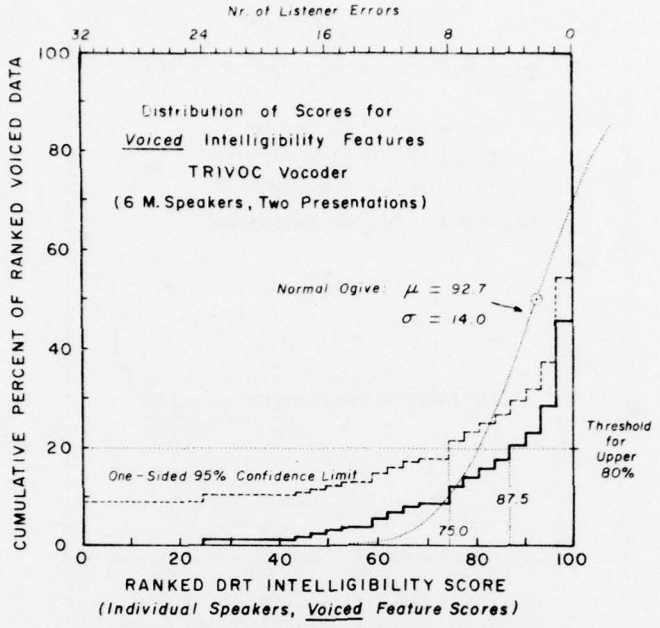

Figure A6.b. TRIVOC Vocoder at 2400 BPS: Voiced Intelligibility Features, (Six Male Speakers)

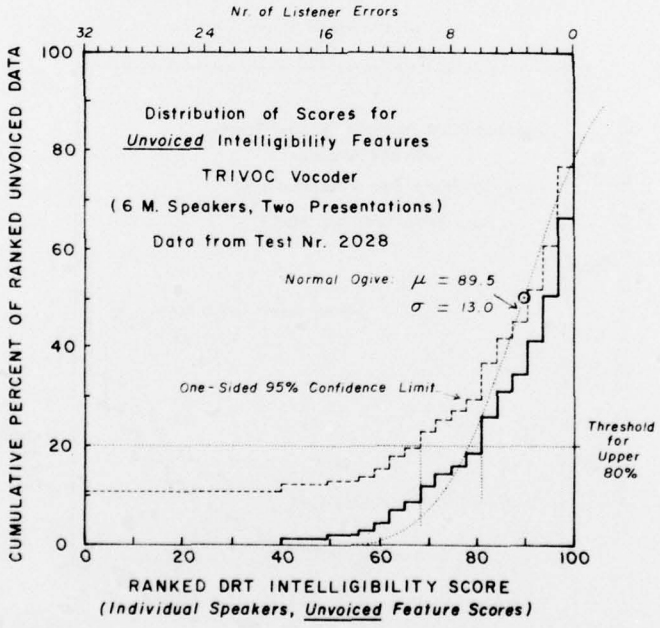

Figure A6.c. TRIVOC Vocoder at 2400 BPS: Unvoiced Intelligibility Features, (Six Male Speakers)

Appendix B

Wang Model 720 Calculator Program for Negative Binomial Probability Distribution

This program sequence is designed to operate in conjunction with some utility subroutines for storing and retrieving the contents of the calculator x&y registers, and for calculating and executing plots in conjunction with the Wang Model 702 Printer/Plotter.

To initialize, values of m, s, n, and the constant for converting between listener errors and DRT score, are entered from the keyboard as follows:

m into register 006,
s (standard deviation in listener errors) into reg. 044
n into register 002
c (for converting error count to DRT score) into reg 001

Load a zero into reg 032 to obtain a listing.
Load a 2 into reg 032 to obtain a plot.

Position the Model 702 Plotter Printer at the top of the page for a listing; at the origin for a plot.

Execute "Search 1515".

Wang Calculator Program - continued.

```
0408   MARK    Start Neg.
1515   1515    Binomial Routine
0700   0
0404   store dir
0008   008
0405   recall dir
0404   044     s
0713   square
0604   ↑
0405   recall dir
0006   006     m
0601   -
0713   square
0606   exch. x&y
0603   divide
0605   ↓
0607   |x|
0604   ↑
0414   store Y
0001   001     k*
0405   recall dir
0006   006     m
0606   exch. x&y
0600   +
0606   exch. x&y
0603   divide
0605   ↓
0611   Ln x
0404   store dir
0003   003     Ln (p)
0701   1
0606   exch. x&y
0601   -
0605   ↓
0611   Ln x
0404   store dir
0004   004     Ln (q)
0408   MARK
1412   1412
0405   recall dir
0009   009     Xi
0404   store dir
0012   012
0415   recall y
0003   003         ln (p)
0405   recall dir
0001   001         k*
0602   multiply
0414   store y
0011   011
0405   recall dir
0012   012         Xi
0412   Write A     Skip if
0711   Ch Sign     Xi = 0
0407   Search
1411   1411
0415   Recall Y
0004   004         Ln (q)
0602   multiply
0605   ↓
0400   + direct
0011   011
0408   MARK
1415   1415
0415   recall y
0001   001         k*
0405   recall dir
0012   012         Xi
0600   +
0701   1
0601   -
0605   ↓
0611   Ln x
0400   + direct
0011   011
0405   recall dir
0012   012         Xi
0611   Ln x
0401   - direct
0011   011
0701   1
0401   - direct
0012   012       decr. Xi
0415   recall y
0012   012         Xi
0701   1
```

Wang Calculator Program - continued.

```
0508    Skip if y<x
0407    Search
1415    1415
0405    rec direct
0011    011
0614    e^x
0400    + direct
0008    008
0415    rec Y
0008    008
0405    rec direct
0009    009
0004    SR(store x&y) 0004
0415    rec Y
0302    032
0702    2
0508    Skip if Y<X
0100    Execute SR0100(calc.plot)
0114    Execute SR0114(execute plot)
0415    rec Y
0302    032
0702    2
0508    Skip if Y<X
0407    Search   (Jump if flag < 2)
1514    1514
0015    Execute SR0015 (CR/LF)
0005    Execute SR0005(recall x&y)
0604    ↑
0405    rec direct
0000    000 (score incr.)
0602    multiply
0701    1
0700    0
0700    0
0606    exch x&y
0601    subtract
0605    ↓
0411    Write
0302    (3.2) (Equivalent DRT Score)
0005    Execute SR0005(recall x&y)
0411    Write
0200    (2.0) (Nr of listener errors)
0405    rec direct
0011    011
0614    e^x
0411    Write
0105    (1.5)  p(m,x,k)
0604    ↑
0405    rec direct
0002    002     N
0602    multiply
0606    exch x&y
0411    Write
0403    (4.3) Neg.Binomial Frequency
0405    rec direct
0008    008
0411    Write
0105    (1.5) Cumulative p
0602    multiply
0605    ↓
0411    Write
0403    (4.3) Cumulative Frequency
0408    Mark
1514    1514
0701    1
0400    + direct
0009    009     incr. x
0415    rec Y
0203    023    Xmax
0600    +
0600    +
0600    +
0600    +  (Xmax + 4)
0405    rec direct
0009    009      x_i
0508    Skip if Y<X
0407    Search  (loop for
1412    1412     next x_i)
0515    STOP
```

Appendix C

The Intelligibility Threshold Level (ITL): A new Approach for Evaluating Performance of Digital Speech Communications Processors

Caldwell P. Smith

Reprinted from ASA*50 Speech Communication Preprint Experiment,
Acoustical Society of America, June 1979

Intelligibility performance of voice processors has been typically specified in average intelligibility scores: values presumably equaled or exceeded by half the underlying populations of scores, assuming a normal distribution. However, extensive multi-speaker testing of a wide variety of processors over the past decade has shown conclusively that (1) populations of scores for processors typically deviate significantly from normal; (2) mean scores of individual talkers in multi-speaker tests typically show highly significant differences ($a = .001$), and (3) significant differences among the mean scores for phonetic features are also typical. To further an implied objective of intelligibility testing: estimation of a level equaled or exceeded by the majority of scores, for example, 80% of a population of scores for talkers and phonetic features, a new approach has been established for evaluation, in which intelligibility data is analyzed to estimate intelligibility threshold levels (ITL's): levels equaled or exceeded by specified fractions of populations of scores, at a specified confidence level. The method, based on non-parametric statistics of the Kolmogorov-Smirnov type, involves rank-ordering a population of scores and constructing a cumulative distribution and its confidence band, from which ITL's can be readily assessed. Intelligibility data for various processors has typically shown larger differences in ITL's than occur with mean intelligibility scores.

INTRODUCTION.

Diagnostic intelligibility testing (Voiers,1973;1977) has been applied extensively to test and evaluation of a variety of speech communications processors and systems over the past decade; numerous test results have been published in the literature (for example, Voiers and Smith, 1972; Smith,1977; 1979). Testing has usually served multiple objectives of providing a basis for comparing different speech processor algorithms or hardware, and guiding research to "fine tune" algorithms to obtain superior intelligibility or correct deficiencies, but also for estimating intelligibility predicted for the processor when used in a "real world" environment for support of voice communications for some population of talkers and listeners, presumably not too different from those used in conducting the tests. The average intelligibility scores customarily cited for voice systems carry the implication of representing values that would be equaled or exceeded by 50% of an underlying aggregate of scores for individual talkers, listeners, and phonetic features, a population sampled in the process of intelligibility testing, presumably normally distributed and representative of a "real world" communications environment.

RATIONALE FOR THE INTELLIGIBILITY THRESHOLD LEVEL (ITL) PERFORMANCE RATING

Even if populations of intelligibility scores obtained in multi-speaker tests were normally distributed - and there is considerable evidence that this is customarily not the case (Smith, 1979) - it would seem appropriate to reassess the practice of specifying performance in terms of average intelligibility scores. Instead, intelligibility levels estimating the values attained or exceeded by a majority in the populations of scores, for example 80% of the scores,at a specified confidence level, would seem to be more meaningful and relevant performance ratings,especially so for assessing voice systems required to support critical communications involving brief, terse messages and where misunderstandings, or time lost in requiring messages to be repeated, could result in severe cost penalties. From this perspective, a performance rating in the form of an intelligibility threshold level (ITL) is proposed: an experimentally determined intelligibility level (based on testing with appropriate talkers, listeners, and test items, such as current multi-speaker versions of the Diagnostic Rhyme Test) specifying a forecast of the intelligibility level that will be equaled or exceeded by a specified percentage of a population of intelligibility scores, at a stated confidence level.

EXAMPLES OF INTELLIGIBILITY THRESHOLD LEVELS (ITL's)

Some examples of ITL's determined from distributions of diagnostic intelligibility scores of some well-known digital speech processors follow. Each was determined by utilizing the 24 basic diagnostic scores, i.e. the four scores ascertained for each of the six primary consonantal features tested with the Diagnostic Rhyme Test, forming a composite data table composed of feature scores of all speakers in the test. The composite table was then rank-ordered and used to construct a cumulative distribution. Except for the fact that each datum (score) represented an average of the responses of eight listeners, there was no averaging: for example, separate scores were included for voicing present (frictional), voicing present (non-frictional), voicing absent (frictional) and voicing absent (non-frictional) for each speaker, etc. The details of phonetic feature scores that typically reveal idiosyncracies and deficiencies peculiar to a voice processor algorithm are retained in the cumulative plots of scores. While the distributions do not reveal specific identities of features and speakers involved in particular "bad" scores, that information is readily available from the conventional listing of diagnostic scores when needed for the purpose of diagnosing specific deficiencies.

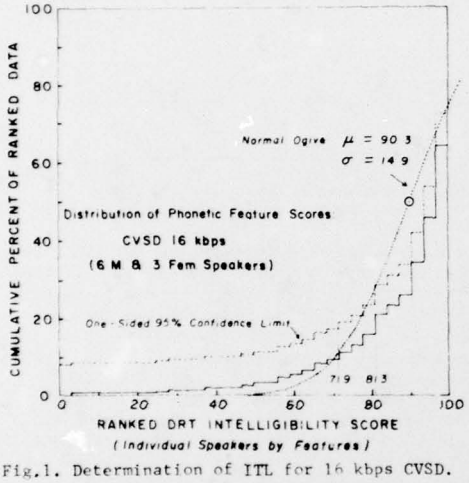

Fig.1. Determination of ITL for 16 kbps CVSD.

The cumulative plots summarize in a compact and efficient way all of the information contained in the detailed scores. They can facilitate formal or informal assessment of the proportion of "bad" scores, something not easily determined by inspecting the listing of scores. Confidence bands for the distributions were constructed by the method of Kolmogorov (Conover, 1971), these providing the basis for ascertaining ITL's. Cumulative distributions also provide a basis for other non-parametric tests for comparisons with other standard or experimental discrete or continuous forms. One such test is the Lilliefors test for conformity with a normal distribution (Lilliefors, 1967). Application of this test to distributions of diagnostic intelligibility scores, including single and multi-speaker data for a variety of processors, has in all cases indicated that the null hypothesis (for conformity with a normal distribution) should be rejected. Distributions of total intelligibility scores (individual listeners) have been found approximately normal for a single speaker, but deviating significantly with multiple-speaker data.

Fig. 1 presents the distribution of scores from evaluation of continuous variable-slope delta modulation (CVSD) operating at 16 kilobits per second. A one-sided 95% confidence band, and the normal ogive for the mean and variance found in this data set from tests with nine talkers (6 male and 3 female) are shown. The ITL estimate from this data suggests that, at the 95% confidence level, 80% of the scores in populations for which the data is a random sample will equal or exceed 71.9. The 80% value in the actual data distribution was 81.3.

ALTERNATIVE DATA BASES FOR ITL's

Assuming that multi-speaker diagnostic intelligibility data has been established by formal intelligibility testing, several alternative methods can be considered for selecting the data base used in determining ITL's, differing in the basis of selecting an aggregate of scores for the assessment. One procedure has already been described above. Some other possibilities are (2) separating the population of scores previously described into sub-groups consisting of voiced and unvoiced scores respectively, and determining ITL's separately for the two categories; (3) averaging across phonetic feature scores obtained with each listener, thus creating a population of total intelligibility scores, one for each speaker/listener combination, and (4) utilizing the separate scores obtained for every speaker/feature/listener combination.

Application of method (2) is illustrated in Fig. 2. The diagnostic scores for 16 kbps CVSD of Fig. 1 were separated into groups representing voiced and unvoiced features; separate rankings and ITL's were established for the two groups. The data represent a trend found in scores for voice processors, in which values of ITL's for voiced features are higher, and unvoiced features lower, than the ITL values for the total population of scores. Further comparisons are shown in

Table I, which presents ITL ratings from total aggregates of scores (as described in the first method) and separate ITL's for the data separated into populations of voiced and unvoiced feature scores. The comparisons suggest that the greatest potential for improving intelligibility will lie in improving the modeling of unvoiced speech events.

Method (3), using distributions of total intelligibility scores (for individual listeners and speakers) was found to result in distributions approximately normal with a single speaker, but deviating significantly with multiple speakers (on a basis of Lilliefors test results). However, in all cases these distributions were much less skewed than distributions of diagnostic scores for phonetic features. Method (3) reveals variations in listener performance not revealed by the first method described; however, listener variations have been found to be much smaller than variations due to phonetic features or due to speakers. A major limitation of method (3) is the failure to reveal significant deficiencies among the feature scores. For this reason, evaluation by this grouping of data is considered to have limited value.

Method (4) would offer composite information about talkers, phonetic features, and listeners. However, it poses a major shortcoming: the 192-word Diagnostic Rhyme Test includes only eight tokens for each of the 24 feature states. Consequently method (4) would result in distributions with extremely gross quantization of the scale (nine possible values for the scores) resulting in inadequate resolution.

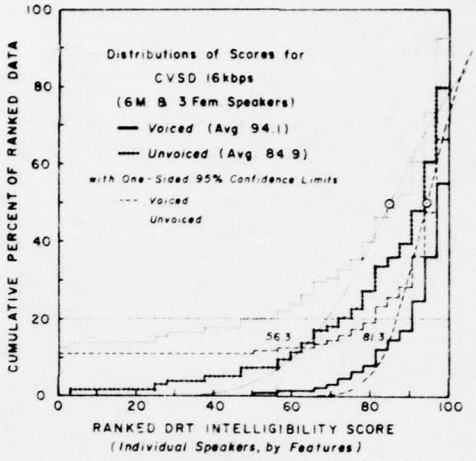

Fig. 2. Separate distributions and 80% ITL's for voiced and unvoiced scores from the data of Fig. 1. The data suggest that improvements in intelligibility could best be found through improved modeling of unvoiced consonant events.

Steps in determining ITL's (using the first method described) are as follows: 1. Rank-order the population of intelligibility scores comprised of the 24 feature-state (or "sub-feature") scores of each talker, combined into an aggregate data population. 2. Using the ranked table, construct the cumulative distribution of scores. For this purpose, each datum is interpreted as a segment of (1/n X 100)%. 3. Construct a confidence band, using an appropriate quantile from a table of values of the Kolmogorov test statistic (Conover,1971) or the value $Q = 1.22/\sqrt{n}$ (for a one-sided band at $p = .95$, and $n > 40$). (Note that with 24 feature scores for each of six speakers, Q = .0983, or 9.83%). In reference to the figures, the value of Q designates an amount of vertical displacement of the data distribution required to define the specified confidence band. 4. To obtain desired ITL's, read from the confidence band profile the corresponding intelligibility level. This can be done graphically, or for greater accuracy, from the listing of ranked scores and associated cumulative percentages (as done with the examples of ITL's presented here).

PROPOSED ADVANTAGES OF THE INTELLIGIBILITY THRESHOLD LEVEL (ITL) RATING.

It is proposed that ITL ratings have several advantages over average intelligibility scores for rating intelligibility performance of voice communications processors: 1. An ITL rating assesses intelligibility performance attained for the majority of scores rather than the average score. The rating provides information as to whether a voice processor caused a significant proportion of "bad" scores for any combinations of speaker and phonetic features. 2. A confidence level is defined for the ITL rating. 3. The ITL rating, based on non-parametric properties of the distribution of scores, is valid without regard to the form of the distribution of scores, and is not affected by departure from normality. 4. The rating is inherently compensated for the number of speakers and/or the number of replications of a test. 5. ITL ratings for various processors have been found to show greater differences among processors than revealed by average scores. 6. The ITL does not require any new testing method; given detailed data, it is applicable for evaluating intelligibility data not only for the Diagnostic Rhyme Test or DRT, but also for the Modified Rhyme Test or MRT (House et al, 1965) or the Consonant Recognition Test or CRT (Preusse, 1959). However, details of diagnostic scores by speakers and features are known only for the Diagnostic Rhyme Test.

TABLE I COMPARISONS OF INTELLIGIBILITY THRESHOLD LEVELS (ITL's)

from Diagnostic Rhyme Test Scores, 6 Male & 3 Female Speakers and 8 Listeners

PROCESSOR	Mean Score	80% Intelligibility Threshold Level (95% Confidence)*		
		Voiced	Unvoiced	Total
CVSD 9.6kbps	80.1	62.5	25.0	53.1
CVSD 16 kbps	90.3	81.3	56.3	71.9
CVSD 32 kbps	95.0	90.6	71.9	87.5
Ch Vocoder 2400	83.0	62.5	37.5	59.4

*There is 95% confidence that 80% of the population of diagnostic intelligibility scores (feature scores, by Speakers) will equal or exceed the stated value.

It is hypothesized that ITL ratings may be more closely correlated with user acceptance of digital speech processors than are average intelligibility scores; however, no data has been available to permit a test of this hypothesis.

CONCLUSIONS.

The relatively new techniques of multi-speaker diagnostic intelligibility testing have tended to overwhelm the evaluator with the mass of information contained in typical intelligibility test results. Perhaps because of the difficulties of evaluating and interpreting fine details of performance, there has been a tendency to reduce results to single numbers: the average intelligibility scores, values that provide no information as to variations among individual speakers and among scores for phonetic features, even though the salient differences in various speech processor algorithms are usually revealed in these details. Preparation of cumulative distributions of scores can provide a means for summarizing entire populations of scores in meaningful but compact form that contains the significant variations and highlights deficiencies. Preparation of confidence bands for distributions can permit forecasts of intelligibility threshold levels (ITL's) for selected proportions of scores at specified confidence levels. Studies of ITL ratings of intelligibility performance obtained with various speech processors operating under a variety of conditions should provide guidance for determining minimum ITL standards that would be appropriate performance criteria for various applications in communication.

REFERENCES

Conover, W.J.(1971), Practical Non-Parametric Statistics, Wiley & Sons, New York.

House, A.S.,Williams,C.E.,Hecker,H.L.,and Kryter,K.D.(1965), "Articulation Testing Methods: consonantal differentiation with a closed response set," J.Acous.Soc.Am. 37, 158-166.

Lilliefors,H.W. (1967), "On the Kolmogorov-Smirnov test for normality with mean and variance unknown," J.Amer.Stat.Assoc. 62, 399-402.

Preusse, J.W. (1959), "The Consonant Recognition Test," US Army Electronics Command, ECOM-3205, Ft. Monmouth, New Jersey.

Smith, C.P. (1977), "Intelligibility Performance of Narrowband Linear Predictive Vocoders in the Presence of Bit Errors," ESD-TR-77-328, AF Electronic Systems Division, Hanscom AFB, Mass.

Smith, C.P. (1979), "Talker Variance and Phonetic Feature Variance in Diagnostic Intelligibility Scores for Digital Voice Communications Processors," Conf.Record, 1979 IEEE Intl.Conf. on Acoustics,Speech & Signal Processing, IEEE 79CH1379-7 ASSP, 456-459.

Voiers, W.D. et al, (1973), "Research on Diagnostic Evaluation of Speech Intelligibility," AFCRL-72-694, AF Electronic Systems Division, Hanscom AFB, Mass.

Voiers, W.D. (1977), "Diagnostic Evaluation of Speech Intelligibility," in Speech Intelligibility and Speaker Recognition, M.Hawley,Ed., Dowden Hutchinson & Ross, Stroudsburg, PA.

Voiers, W.D. and Smith,C.P. (1972), "Diagnostic Evaluation of Intelligibility in Present-Day Digital Vocoders," Conf.Record, 1972 Conf.on Speech Comm.& Proc.,AFCRL-72-0120, 170-174.

Appendix D

Table of the Kolomogorov Test Statistic

Table D1. Quantiles of the Kolmogorov Test Statistics

One-Sided Test	p=.90	.95	.975	.99	.995		p=.90	.95	.975	.99	.995
n=1	.900	.950	.975	.990	.995	n=21	.226	.259	.287	.321	.344
2	.684	.776	.842	.900	.929	22	.221	.253	.281	.314	.337
3	.565	.636	.708	.785	.829	23	.216	.247	.275	.307	.330
4	.493	.565	.624	.689	.734	24	.212	.242	.269	.301	.323
5	.447	.509	.563	.627	.669	25	.208	.238	.264	.295	.317
6	.410	.468	.519	.577	.617	26	.204	.233	.259	.290	.311
7	.381	.436	.483	.538	.576	27	.200	.229	.254	.284	.305
8	.358	.410	.454	.507	.542	28	.197	.225	.250	.279	.300
9	.339	.387	.430	.480	.513	29	.193	.221	.246	.275	.295
10	.323	.369	.409	.457	.489	30	.190	.218	.242	.270	.290
11	.308	.352	.391	.437	.468	31	.187	.214	.238	.266	.285
12	.296	.338	.375	.419	.449	32	.184	.211	.234	.262	.281
13	.285	.325	.361	.404	.432	33	.182	.208	.231	.258	.277
14	.275	.314	.349	.390	.418	34	.179	.205	.227	.254	.273
15	.266	.304	.338	.377	.404	35	.177	.202	.224	.251	.269
16	.258	.295	.327	.366	.392	36	.174	.199	.221	.247	.265
17	.250	.286	.318	.355	.381	37	.172	.196	.218	.244	.262
18	.244	.279	.309	.346	.371	38	.170	.194	.215	.241	.258
19	.237	.271	.301	.337	.361	39	.168	.191	.213	.238	.255
20	.232	.265	.294	.329	.352	40	.165	.189	.210	.235	.252
			Approximation for n>40:				$\frac{1.07}{\sqrt{n}}$	$\frac{1.22}{\sqrt{n}}$	$\frac{1.36}{\sqrt{n}}$	$\frac{1.52}{\sqrt{n}}$	$\frac{1.63}{\sqrt{n}}$

MISSION
of
Rome Air Development Center

RADC plans and executes research, development, test and selected acquisition programs in support of Command, Control Communications and Intelligence (C^3I) activities. Technical and engineering support within areas of technical competence is provided to ESD Program Offices (POs) and other ESD elements. The principal technical mission areas are communications, electromagnetic guidance and control, surveillance of ground and aerospace objects, intelligence data collection and handling, information system technology, ionospheric propagation, solid state sciences, microwave physics and electronic reliability, maintainability and compatibility.

Printed by
United States Air Force
Hanscom AFB, Mass. 01731