IEEE Recommended Practice for Speech Quality Measurements

Acknowledgment

The *IEEE Recommended Practice for Speech Quality Measurement* is the product of roughly six years of study, discussion, writing, and rewriting by a diligent team of scientists representing a broad range of disciplines and research institutions. We will attempt to recognize here all those who participated, ranging from the dogged individuals who were with the committee for its entire life to those who participated in at least one meeting.

The charter for the Methods of Subjective Measurement Subcommittee of the Audio and Electroacoustics Group Standards Committee (then of the IRE) was drawn in January of 1963, and later that year the committee was formed. The committee was initially chaired by Knut S. Nordby of IBM, and consisted of Anderson P. Evans, Jr., CBS; F. T. Andrews, Jr., Bell Telephone Laboratories (BTL); W. D. Chapman, IBM; and Newman Guttman, BTL; with E. H. Rothauser, IBM (then of the Vienna Science Group) as Consultant. In early 1964, the focus on speech quality measurement had been established, and the membership evolved to include Herman R. Silbiger and John L. Sullivan of BTL, and David Speaker (then of Haskins Laboratories). By late 1964, Michael H. L. Hecker (then of Bolt, Beranek and Newman) had joined the committee as the leading exponent of the technique which later evolved into the Relative Preference Method. At the same time ,Victor E. McGee (then of Dartmouth College) joined us and led the cause for the Facet Analysis Method, In 1965, Martin Weinstock of USAECOM, Fort Monmouth, joined us and made many valuable contributions through to the final drafts of the document. Other contributors who met with the committee at least one time during its lifetime were B. B. Bauer, CBS Laboratories; Katherine S. Harris, Haskins Laboratories; W. D. Voiers (then of Sperry-Rand Research Center); Joseph Tierney, Lincoln Laboratories; Franz Lenk (then of the IBM Vienna Science Group); Gunther Urbanek, Technical University of Vienna; Sanford E. Gerber, Hughes Aircraft Company; and Douglas E. Phyfe, USAECOM, Fort Monmouth. Ernst Rothauser took over the chairmanship in 1966.

After general completion of the document, Herman R. Silbiger assumed editorial responsibility, assisted by W. D. Chapman and E. H. Rothauser. The final touches to the document were made by these individuals along with the able contributions of Messrs. Guttman and Hecker. We offer our grateful thanks to all of the above individuals for their contributions of time and talent, and to William W. Lang and Reg A. Kaenel of the G-AE Standards Committee for their patience and guidance.

The Editor

Standards Publication No. 297, available from IEEE.

FOREWORD

The increasing variety of speech transmission systems has created a new problem for communication engineers. This problem consists of the subjective evaluation of the speech quality produced by these systems. For investigators interested only in the utilization of some procedure applicable to this problem, it was difficult to make a choice among the wide range of documented approaches.

The IEEE Audio and Electroacoustics Group therefore formed a Subcommittee on Subjective Measurements, which has been charged with writing a Recommended Practice for speech quality measurement. The intent of the Subcommittee became to write a document that would facilitate the choice of method by reducing the variety of procedures to several which had been successful and looked promising for the future, and which could handle a variety of signals. By describing these methods in some detail, it would be possible for different laboratories to obtain comparable results and thus aid in the exchange of information.

Because of the relatively early stage of the development of speech quality evaluation, the present document is intended to be a Recommended Practice, but not a long term standard, which would require speech quality measurement only according to the proposed methods.

It is hoped that the document will generate additional investigation into problems of this type of measurement, and allow comparison of test data obtained in different laboratories. Thus, a body of experimental data may be generated that will be useful to future investigations as well as the improvement of the present practices. The Subcommittee expects that feedback from users of the practices will allow the revision and updating of the practices within a relatively short time period.

It should be pointed out that the present practices evaluate only speech signals as they are generated by one-way speech communication systems. Systems, especially two-way systems, in which interactive communication takes place, cannot validly be evaluated by these methods. One of the directions, therefore, in which these practices might be updated in the future is by adding material on the evaluation of communications systems.

It is the hope of the Subcommittee that the present document will be helpful in some way to people confronted with the task of making speech quality measurements. In order to guide its future work, the Subcommittee would appreciate that any technical comments about the usefulness of the documents, or any other comments or information about speech quality measurement, be sent to its Chairman, c/o the IEEE Audio and Electroacoustics Group, IEEE Headquarters.

ACKNOWLEDGMENT

The Institute wishes to acknowledge its indebtedness to those who have so freely given of their time and knowledge, and have conducted experimental work on which many of the IEEE publications are based.

This publication was prepared by the Subjective Measurements Subcommittee of the Standards Committee of the IEEE Audio and Electroacoustics Group. The membership of the Subcommittee was:

E. H. Rothauser, *Chairman*
W. D. Chapman, *Vice-Chairman*

N. Guttman
M. H. L. Hecker
K. S. Nordby
H. R. Silbiger
G. E. Urbanek
M. Weinstock

Consultants

V. E. McGee, U. P. Pachl, and W. D. Voiers

IEEE Recommended Practice for

SPEECH QUALITY MEASUREMENTS

1. INTRODUCTION

1.1 Scope. Speech quality is based on the subjective appraisal of speech. Speech may be appraised on the basis of preference, loudness, intelligibility, and recognizability of properties of the original speaker's voice, of the transmission channel, and/or in the case of artificial speech, of the speech synthesizer.

The IEEE Subcommittee on Subjective Measurements, charged with writing an engineering practice for the measurement of speech quality, has concluded that a single method should not now be recommended. The present Recommended Practice is concerned only with preference measurements for which three methods will be tentatively outlined. These are the Isopreference Method, the Relative Preference Method, and the Category-Judgment Method.

1.2 Review of Past Work

1.2.1 Surveys of work on measurement of speech quality have been given by Munson and Karlin (1962) and by Hecker and Guttman (1966).* Munson and Karlin divided methods into "indirect comparisons" (circuits assessed singly) and "direct comparisons" (circuits assessed in pairs). Examples of the former are category tests and intelligibility tests. Direct comparison usually implies a pair-comparison. Hecker and Guttman took cognizance of newer methods and purposes, and distinguished two broad categories, analytic and utilitarian. The analytic methods aim at discovering the psychological attributes of the speech signal. Selected degradations of the speech signal are among the means of investigation, and correlations of physical parameters of degradations with attributes are of great interest. Choices of psychometric methods are arbitrary. Typically, analytic investigations have distinguished more than one significant psychological dimension. The utilitarian methods are distinguished by prior assumption of psychological dimensions, by ease of administration and evaluation, and by effective reduction of measures to a unidimensional scale.

1.2.2 Appendix B contains an annotated bibliography of some of the literature. No attempt has been made to be complete. Items are classified as analytic or utilitarian.

1.3 Background of Recommendations

1.3.1 It has been clear to the Subcommittee since the beginnings of its deliberations that no generally applicable method of preference measurement has been developed. In fact, the existence of the Subcommittee has tended to stimulate research. The shortcomings of the various existing methods are not easy to document, and for the sake of simplicity and brevity, the Subcommittee will not elaborate at length on why it rejected some approaches. Furthermore, it cannot prove that details in the recommended practices are optimum. In short, the present report represents the collective best judgment of the Subcommittee, where "best judgment" consists of "common sense" and "educated guesses."

1.3.2 The utilitarian methods appear to be best suited for engineering practice. Three particular approaches seem to be the most promising at this time: a) the Isopreference Method; b) the Relative Preference Method; and c) the Category-Judgment Method.

1.3.2.1 The Isopreference Method—was originated by Munson and Karlin (1962). When fully developed in practice for any system of interest, the Isopreference Method would include the construction of a set of isopreference contours drawn on a plane having the coordinates of speech level and noise level. These contours are given a metric in terms of Transmission Preference Level and Transmission Preference Unit scales that are anchored to degradations produced by noise added to live speech. In order to calibrate the responses to a test system, it is necessary to experimentally compare the test system to a known one. The Isopreference Method was advanced by Munson and Karlin as an exploratory effort. So far, satisfactory replication and validation of the full method has not been made available. There appear to be a number of doubtful features. a) Attempts to replicate transitivity have not been entirely successful. b) It is not clear why speech signals (for example, vocoder speech) should be studied on a speech level/noise level plane. c) It is not clear why a wide range of loudness must be examined.

A simplification of the Munson–Karlin method (Rothauser) is given in Section 3. In the simplified method, signal-to-noise ratio is the measure of speech quality. This is nearly equivalent to the Munson–Karlin Transmission Preference Level. At present, attempts are being made to derive a difference-limen scale such as the Transmission Preference Unit scale.

1.3.2.2 The Relative Preference Method—outlined in Section 4 appears to be simpler, but more validation is required on this method. The degree of degradation applied to the several reference conditions is of critical importance. The method is outlined in Hecker and Williams (1966) and is abstracted in Appendix A.

1.3.2.3 The Category-Judgment Method—has listeners compare the speech test signal to some subjective standard of speech quality. It is undoubtedly the most common one currently in use as it is described in the CCITT Redbook (1962). The serious difficulty with the method is that it may be influenced by context effects, i.e., the judgments are easily biased by the set of signals judged [Torgerson

* References may be found in Appendix B.

(1958) pp. 78–82]. The method is discussed by Richards and Swaffield (1959) with specific reference to two-way telephone communications. It has also been advocated for the evaluation of television transmission [Prosser, Alnatt, and Lewis (1964)]. By this method quality will eventually be characterized by an adjective such as "good" or "bad."

1.4 Basic Considerations in Preparing the Practices

1.4.1 It is important to organize methods in such a way that the effects of quality attributes other than preference are neutralized. Therefore, word intelligibility should be high, about 95 percent. For loudness requirements, see paragraph 3.2. For talker requirements, see paragraph 3.4.

1.4.2 The results should be quantifiable on a unidimensional scale.

1.4.3 The methods must produce highly reproducible results on retest and at different laboratories.

1.4.4 The testing procedures should allow reasonably rapid measurements. Similarly, analysis of raw data should be rapid and simple enough not to require an electronic computer.

1.4.5 Reference signals in the comparison methods should be easily realizable and reproducible.

1.4.6 The methods should be adaptable to standardized speech recordings.

1.4.7 The methods should allow principally the evaluation of one-way communication systems. (Two-way communication systems introduce complexities that may require evaluation by behavioral indices.)

1.4.8 The ultimate goal of preference measurement is the evaluation of specific speech transmission systems, speech synthesizers, etc. The specifications of such systems will have a strong influence on the selection of the most appropriate test method and on all details of the test specifications. In Appendix A, the format for such pertinent system specifications can be found.

Using one of these methods, it is possible to produce an estimate of the quality of a speech signal; but only under special conditions, here unspecified, can one deduce therefrom the characteristics of the system under consideration.

2. DEFINITIONS

2.1 Speech Signals. Utterances in their acoustical form or electrical equivalent.

2.2 Speech Quality. A characteristic of a speech signal that can be described in terms of subjective and objective parameters. For the purposes of this Practice, speech quality is evaluated only in terms of the subjective parameter of preference.

2.3 Preference. The proportion of a listening group expressed in percent that prefers the speech test signal to the speech reference signal as a source of information.

2.4 Isopreference. Two speech signals are isopreferent when the votes averaged over all listeners show an equal preference for the speech test and speech reference signals.

2.5 Preference Level. The signal-to-noise ratio (S/N) of the speech reference signal when it is isopreferent to the speech test signal.

2.6 *A*-Weighted Sound Level. A weighted sound pressure level obtained by the use of a metering characteristic and the weighting A specified in USAS S1.4-1961 (General Purpose Sound Level Meters).

2.7 Speech Level. For the purpose of this practice the speech level shall be defined and measured subjectively by comparison of the speech signal with a signal obtained by passing pink noise through a filter with A-weighting characteristics that has been judged to be equal to it in loudness. The value of the speech level is defined to be the A-weighted sound pressure level of this noise [dB(A)]

2.8 Pink Noise. A random noise whose spectrum level has a negative slope of 10 decibels per decade.

2.9 Noise Level. The A-weighted sound level of the noise.

2.10 Signal-to-Noise Ratio (S/N). In decibels of a speech signal, the difference between its speech level and the noise level.

2.11 High-Fidelity Signal. A signal transmitted over a system comprised of a microphone, amplifier, and loudspeaker or earphones. A tape recorder may be part of the system. All components should be of the best quality the state of the art permits.

2.12 Speech Reference Signal. Used as a standard of reference for the purpose of preference testing, a speech signal which is artificially degraded in a measurable and reproducible way.

2.13 Speech Test Signal. A speech signal whose speech. quality is to be evaluated.

2.14 Listening Group. A group of persons assembled for the purpose of speech quality testing. Number, selection, characteristics, and training of the listeners depend upon the purpose of the test.

2.15 Trained Listening Group. Six to ten listeners who understand thoroughly the purpose of the test and respond properly throughout the test. All persons of the group shall meet the requirements on auditory acuity as described by USAS S3.2-1960 (Monosyllabic Word Intelligibility). The training of the listeners will depend on the special type of tests to be conducted.

3. ISOPREFERENCE METHOD

3.1 Synopsis. The speech test signal is compared in a forced-choice test procedure directly with a speech reference signal that is subjected to variable degrees of degradation. Thus the isopreference level of the speech test signal is defined as the signal-to-noise ratio of the speech reference signal at which preference votes of a listener group are equally divided. The test and reference signals are then "isopreferent." The preference of the test signal is described by the signal-to-noise ratio in decibels of the degraded reference signal.

3.2 Test Procedure. A testing session may consist of a number of runs on different speech test signals. Each speech test signal should be evaluated during a separate run. During each run, the listeners are presented with repeated signal pairs in the order A B A B. The listeners are asked to indicate the signal in each repeated pair that they would prefer as a source of information. Figure 1 shows the time pattern and the variation of the signal-to-noise ratio for a segment of a typical run. The listeners hear the samples of A and B alternately for 5 seconds each. The interval between adjacent signal samples is 0.5 second. To allow time for a listener to make his response, an interval of at least 3 seconds is provided after each repeated signal pair (A B A B). Figure 1 also illustrates the variation of the signal-to-noise ratio of the speech reference signal from one repeated signal pair to the next. For each repeated signal pair, the signal-to-noise ratio for signal is constant. The signal-to-noise ratio for successive signal pairs is varied (at random) within a restricted range defined below. In evaluating a speech test signal, a preliminary test run is performed to establish the approximate value of the isopreference level. In this preliminary run, the incremental steps of the signal-to-noise ratio of signal A may be large, e.g., 6 decibels. The preliminary run is also performed to determine the lower and upper limits on the signal-to-noise ratio for signal A at which all listeners prefer either signal A or signal B. In successive runs, the incremental steps of the signal-to-noise ratio of signal A should be made small enough, e.g., 1 decibel, to cause inconsistent decisions by some listeners in the vicinity of their isopreference level.

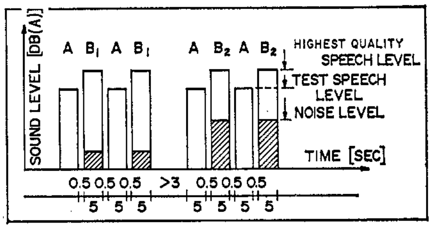


Figure 1. Example of isopreference test pair sequence.

3.3 Loudness of Speech Signals. For the specific purpose of preference testing, all speech signals should be presented to the listeners at a speech level that is the average of all individually preferred levels. A paired-comparison procedure is proposed to determine this average speech level. (Allowing the listeners to adjust their own levels is not a reliable method.) This procedure simply consists in the alternate presentation of the speech signal to be tested with different speech levels. Two samples of the same speech signal A and B, which have a constant level difference, e.g., 5 decibels, are presented in repeated pairs A B A B to the listeners who are requested to choose the preferred sample within the signal pair. In consecutive presentations, sample A is varied in a random order over a suitable loudness range with incremental steps of 1 decibel. The presentation of A B A B and B A B A is also randomized so that there is an interval within the entire loudness range in which all signals are presented twice. The individually preferred speech level is defined as the mean value of all those levels L_i of speech samples, that are preferred by the respective listener for both presentations, i.e., for comparisons with $L_i + 5$ decibels and $L_i - 5$ decibels. The preferred speech level for the listening groups is found by averaging the individually preferred speech levels of all listeners.

3.4 Speech Material. The type of speech material for test and reference signals should be the same. Two types of material are suitable, narrative and short homogeneously structured sentences. Extracts from novels or newcasts are recommended as narrative material. As an alternative to this practice, lists of sentence material may be used such as shown in Appendix C. In any case, the language and the semantic content should be intelligible to the listeners. Whenever possible none of the speech material should be repeated during a given test; furthermore, each test should contain different material.

3.5 Talkers. To check on possible talker–system interactions, a variety of talkers should be used in each test.

3.6 Listeners

3.6.1 Selection. The selection of listeners for the purpose of speech quality testing must be guided by the purposes and applications of the test, i.e., the listening group should have some specifiable relation to the population that will use the system.

Some speech communication systems, such as radio and telephone, may require testing with large untrained listening groups. In such cases, the listeners may well include persons inexperienced in subjective measurements or unfamiliar with special speech signals and acoustic devices, or possessing some degree of hearing loss. In cases where it is necessary to make speech quality evaluations of speech processing and transmission systems used only by a very limited and well-experienced group of individuals, such as special space communication systems, measurements may be based on the decisions of a small, selected group of listeners familiar with the system.

3.6.2 Training. The amount and type of training of the listening group should correspond to that of the presumed users of the system. Special training of the listening group is mandatory only if the system specifications require trained personnel. In the training procedure, each member of the group is presented with a representative sample of words or messages which he will hear during the test. The kind of response which is expected of him is described. Training continues until the testing procedure has become routine to the listening group. In intelligibility testing, the criterion of this condition is the lack of improvement with continued training. Unfortunately, this is not true for preference testing. Listeners in preference testing are affected far more seriously by learning, by adaptation effects, and by previous uncontrolled experiences. Thus it may happen that when preference tests are repeated on different days, some of the listeners will have changed their minds. In view of these uncertainties, it is recommended that the method of training be documented.

3.6.3 Screening. For the purpose of preference testing, the ideal listener should be motivated, and he should understand the purpose of the test, the test question, and the test procedure. Furthermore, he should be capable of discriminating speech quality and of expressing his preference in a consistent way.

In isopreference tests, those listeners whose preference votes are random in a range as great as 6 decibels around their apparent individual isopreference level should be excluded.

3.6.4 Estimates of Minimum Group Size. The listening group should be representative of the relevant characteristics of the user population, such as sex, age, technical, or socioeconomic background. A group of untrained listeners probably should consist of no less than 50 listeners. When the user population is limited, a trained listening group may be rather small, consisting of 6–10 listeners.

3.7 Reference Signals. The present state of knowledge does not support the selection of any specific reference signal. Two reference signals that are fundamentally different with regard to generation and perceptual effect are proposed. However, the user of this Practice is free to choose one of them.

The multiplicative noise signal has been found especially useful when the signal to be evaluated has been digitally processed. A reference signal $r(t)$ is generated by adding a certain amount k of a degradation signal $d(t)$ to the ideal speech signal $s(t)$.

$$r(t) = s(t) + k \cdot d(t).$$

Given this principle, the choice of a reference signal reduces to the choice of a proper degradation signal. The two degradation signals which have been used are the following.

(a) *Additive Random Noise*: The respective speech reference signal will be referred to as "additive reference." The noise signal $n_0(t)$ should be pink noise weighted by the A curve described in USA Standard S1.4-1961 (A-Weighted Pink Noise)

$$d(t) = n_0(t).$$

(b) *Multiplicative Rar.dom Noise*: The respective reference signal will be referred to as "multiplicative reference." The degradation signal is then constructed by multiplying the ideal speech signal with A-weighted pink noise. This can be done by means of an electronic analog multiplier or by means of an electronic switch which inverts the speech signal at a fast rate.

$$d'(t) = s(t) \cdot n_0'(t).$$

Figure 2 illustrates the comparison between the two reference signals.

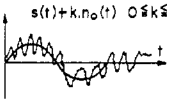
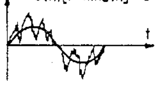
ADDITIVE REFERENCE	MULTIPLICATIVE REFERENCE
$s(t)+k.n_o(t) \quad 0 \leq k \leq 1$ 	$s(t).[1+k.n_o'(t)] \quad 0 \leq k \leq 1$ 
$s(t) + k.n_o(t)$	$s(t)+k.s(t).n_o'(t)$
$S/N = 20 \log \frac{S}{k.N_o}$	$S/N = 20 \log \frac{S}{k.S\,N_o'}$
S/N = 20 log 1/k -LEVEL OF NOISE SIGNAL $n_o(t)$ +LEVEL OF SPEECH SIGNAL $s(t)$	S/N = 20 log 1/k - LEVEL OF NOISE SIGNAL $n_o'(t)$

Figure 2. Isopreference test reference signals.

3.8 Analysis of Results. The average isopreference level for a speech test signal is found using a graphical method. For each of the signal-to-noise ratios of the reference signal, the percentage of listeners who prefer the test signal is plotted on cumulative normal-probability paper. Since the plotted preferences should approximate a normal-probability distribution, it should be possible to fit a straight line to the points. Large deviations from a normal distribution between 10-percent and 90-percent preference indicate that insufficient test data have been obtained. For a given number of listeners, the amount of data may be increased by test repetitions on different days. The isopreference level is the interpolated signal-to-noise ratio corresponding to 50-percent preference.

4. RELATIVE PREFERENCE METHOD

4.1 Synopsis. In this method, the quality of the test signal is measured on a quality continuum defined by several reference signals that represent different types of speech distortion. The selected reference signals are placed on an arbitrary rating scale by considering how often each reference signal is preferred to the other reference signals. The test signal is located on the same rating scale by considering how often it is preferred to any reference signal.

4.2 Test Procedure. A number of preference tests are constructed in the paired-comparison format, each test featuring speech material recorded by a different speaker. Each test contains all possible forward and reverse order comparisons between the test signal and the reference signals, and also all possible comparisons among the reference signals. When n reference systems are employed the test therefore contains a random sequence of $2n + n(n - 1)$ items or a multiple thereof. Typically, five reference systems are used. Listeners are provided with 30-item answer sheets and informed that they will hear two speech samples for each test item. They are instructed to express their preference for the first or second sample.

For each listener, two preference matrices are prepared from the results of each test. These matrices correspond to the forward and reverse orders of all possible system comparisons. The degree to which the listener's responses are consistent may be judged by correlational techniques and used to eliminate listeners whose responses produce highly inconsistent data.

The reference signals are then plotted on a rating scale by considering only those preference judgments obtained for the 20 comparisons among the reference systems (i.e., responses involving the test system are here ignored). Since each reference signal is presented exactly eight times to the listeners, the number of times that a given system was preferred, when divided by the number of listeners, can be expressed as a number between 0.0 and 8.0. The numbers thus calculated for the reference signals are multiplied by 1.25 and the products plotted. Ideally, the reference signal having the lowest preference is never preferred to any other reference signal, and the reference signal having the highest preference is always preferred to the other reference signals so that these two signals are located on the rating scale at 0.0 and 10.0, respectively. The remaining three reference signals are ideally located at 2.5, 5.0, and 7.5.

The test signal is plotted on the rating scale by considering only those preference judgments obtained for the ten comparisons involving the test signal. Since the test signal is presented exactly ten times to the listeners, the number of times that it was preferred, when divided by the number of listeners, can be expressed as a number between 0 and 10. The number thus calculated X for the test signal is not plotted directly but entered into the formula

$$Y = 1.25(X - 1)$$

to determine the position Y of the test signal on the rating scale. The formula provides a numerical coincidence on the rating scale for the case where the test signal and a reference signal are judged to be of equal preference.

4.3 Loudness of Speech Signals. The comments made under this heading for the Isopreference Method also apply to this method.

4.4 Speech Material. The comments made under this heading for the Isopreference Method also apply to this method.

4.5 Talkers. The comments made under this heading for the Isopreference Method also apply to this method.

4.6 Listeners. The comments made under this heading for the Isopreference Method also apply to this method.

4.7 Reference Signals. It is proposed to employ signals that cover the whole range of quality from "highest quality" to "insufficient." The two primary requirements to be considered in selecting the reference signals are that when the signals are compared to each other a) they should sound appreciably different and b) they should be consistently preferred in a certain order. The first requirement can usually be satisfied by generating reference signals through devices that produce fundamentally different types of speech distortion. To determine the extent to which the second requirement is satisfied, it is advisable to construct preliminary 20-item preference tests containing all possible comparisons among five candidate signals. The frequency distribution of the preference judgments obtained for each candidate signal may be used to check the adequacy of a particular set of signals.

Other requirements are c) that the devices used to generate the reference signals can be described accurately and without ambiguity and d) that such devices can be easily assembled or constructed. If the last requirement cannot be met, it should be at least possible to make the chosen reference signals available to other investigators, who may wish to duplicate certain measurements.

For particular applications of the relative preference method, it may be desirable to impose additional requirements on the reference signals. For example, the test system may be a device that can only be compared meaningfully with similar devices. In such a case, an effort should be made to restrict the choice of reference signals accordingly, and still meet the four requirements cited above. In other applications, it may only be necessary to measure speech quality within a limited range. A more closely spaced set of reference signals will reduce the margin by which listeners prefer one system to another and hence will provide a more accurate measurement of the quality of the test system, as long as criterion b) is sufficiently met.

4.8 Analysis of Results. For each preference test conducted, means and standard deviations for the positions of the test signal and the reference signals on the rating scale are to be presented. If earlier data on the positions of the reference systems are not available, the frequency distributions of the preference judgments among the reference signals shall also be presented.

5. CATEGORY-JUDGMENT METHOD

5.1 Synopsis. The quality of the speech test signal is described in terms of a number of categories which have an intuitive meaning to the listeners.

5.2 Test Procedure. A number of simple categories are used by listeners to describe their impressions of the quality of the speech test signal. Examples of those are Unsatisfactory, Poor, Fair, Good, Excellent.

The experiment consists of two phases, familiarization and evaluation.

5.2.1 Familiarization. In this phase, the categories to be used in rating are given to the listener group. In order to establish a point of reference for listeners' responses, it is strongly recommended that one or two speech reference signals be presented. This is commonly termed "anchoring." The experimenter defines the quality category of these signals to the listeners. For example, a recording of a telephone signal is played and assigned a rating of Good. Another possibility is to play two speech reference signals defined by the experimenter as representative of the extreme categories.

It may furthermore be desirable to present all or a representative sample of the speech test signals. The purpose is to orient the listener to the range of qualities he will encounter in the evaluation phase.

5.2.2 Evaluation Phase. The speech test signals are presented in random order. There should be an adequate exposure time to each signal and adequate time for the listeners to respond to the signals. Responses consist of marking the category that corresponds to the impression of quality.

If there are many speech test signals to be evaluated, it is advisable to intersperse identified speech reference signals to refresh the standard of reference of the listeners.

5.3 Loudness of Speech Signals. The comments made under this heading for the Isopreference Method also apply to this method.

5.4 Speech Material. The comments made under this heading for the Isopreference Method also apply to this method.

5.5 Talkers. The comments made under this heading for the Isopreference Method also apply to this method.

5.6 Listeners. The comments made under this heading for the Isopreference Method also apply to this method.

5.7 Reference Signals. Since this method traditionally relies on previous experience of the listeners with all kinds of speech signals, speech reference signals should here, whenever possible, reinforce such prior listening experience.

The reference signals should be selected on the basis of the context in which the listeners are to make their judgments. All reference signals used should be described along with the test results.

5.8 Analysis of Results. The evaluation of the test signal may be accomplished by means of Mean Opinion Scores or by Cumulative Preferences.

5.8.1 Mean Opinion Scores. Each category used in the test is given a number, e.g.,

Unsatisfactory	= 1.
Poor	= 2.
Fair	= 3.
Good	= 4.
Excellent	= 5.

The number of listeners choosing each category is multiplied by the number assigned to that category. These totals are then added, and divided by the total number of listeners. The number then obtained is the mean opinion score (MOS) for that condition. The systems under test can then be listed in order of preference on the basis of their MOS. Although this is an apparently simple way of arriving at a single preference score for a test signal, the numbers are basically only labels for the categories and may not be equidistant along a preference scale. Therefore, the value of an arithmetic operation such as averaging is doubtful.

5.8.2 Cumulative Preferences. This method is especially useful when the systems can be ordered along one or more parameters, such as "added noise," "clipping," etc. The listeners' responses for each value of the parameter under test are tabulated as a function of the percentage votes for each category. A plot is drawn with percentage votes on the ordinate and test condition on the abscissa. Normal cumulative preference curves are then fitted to the equal category points (see Figure 3). From this graph the value of the condition under test for which the desired criterion is reached (for example, 95 percent judged good or better) can then be read off. Additional information on the relation of preferences for the test signals to a known reference signal may be obtained by adding one or more reference signals in the array of signals under investigation. The cumulative preferences for the test signals may then be interpreted in the light of the cumulative preferences for the reference signals.

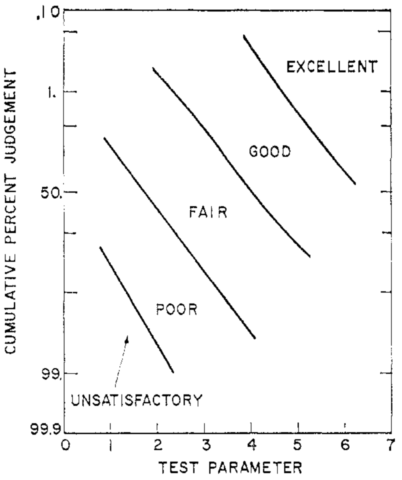


Figure 3. Example of cumulative preference plot on normal probability paper.

APPENDIX A

Documentation Requirements

For the choice of meaningful test conditions a description has to be given of the signal and system properties, and of the users of the system. The system objectives, specifications or other descriptions shall include the following.

(a) Typical applications of the system, e.g., telephone circuits, general purpose dictation, quotation service.

(b) Applications which are considered to be particularly severe from a speech quality standpoint. The intelligibility necessary under such circumstances shall be specified.

(c) Typical messages which will be handled.

(d) Complementary equipment from other sources, e.g., telephone equipment, including handsets.
(e) Methods of speech input and output from the system.
(f) The maximum noise levels in the electroacoustical system due to the utilization of typical complementary equipment; the optimum, the maximum, and the minimum speech levels which are anticipated in the application of the electroacoustical system.
(g) The output noise levels of the system for the maximum and minimum speech levels which are anticipated in the system under consideration.
(h) Any other factors or parameters which noticeably influence speech quality.
(i) The typical and the maximum noise levels to be expected in the talker and/or listener environments on an octave-band basis if applicable.
(j) The users of the system; their anticipated training, auditory and intellectual capabilities; possibilities for the screening of listeners, etc.

APPENDIX B

Bibliography

P. W. Blye, O. H. Coolidge, and H. R. Huntley, "A revised telephone transmission rating plan," *Bell Sys. Tech. J.*, vol. 34, pp. 543–572, 1955.

The proposed method illustrates the use of a special-purpose transmission quality rating method. It is assumed that in a large proportion of the Bell System telephones, frequency response is adequate and nonlinear distortion sufficiently small. Therefore, future designs will not improve these factors, and one may simply use level of the sound pressure at the listener's ear as a measure of quality. Specifically, the rating of a connection is $-20 \log_{10} (S_L/S_T)$, where S_L is the output sound pressure and S_T is the sound pressure at the talking end. Losses are defined by positive numbers, gains by negative numbers. In practice, an artificial mouth is used as the sound source.

Although the units are decibels, the rating is termed "loudness" loss. Special-situation ratings are termed "subjective" loss and would have to be determined by subjective tests. [Utilitarian]

A. Boeryd, "Some reactions of telephone users during conversation," *Ericsson Rev.* (English ed.), vol. 2, pp. 51–58, 1964.

This article deals with some results of tests based on the Richards and Swaffield technique. The rating categories were not based on effort, but were the usual Excellent, Good, Fair, Poor, Bad. The variables were received level, room noise, line noise. The results were also evaluated in terms of the CCITT reference equivalent. Regression lines were fitted to the results. [Utilitarian]

M. H. L. Hecker and C. E. Williams, "Choice of reference conditions for speech preference tests," *J. Acoust. Soc. Am.*, vol. 39, pp. 946–952, 1966.

A dichotomy in speech preference testing methods is set up as follows. a) In paired-comparison paradigm, the test system is compared to several levels of degradation (e.g., signal-to-noise ratios) of one type of reference system. b) Again in paired-comparison paradigm, the test system is compared to several distinct types of reference systems.

It is argued that method a) is insufficiently precise because a listener is "gradually diverted from his task of making true preference judgments. He may surmise that he is expected to match a test system with one of several reference conditions, and in an effort to be consistent he may concentrate on following the changes in the reference system from test item to test item. With the listener's attention thus diverted, his actual selection of the reference condition that he considers the best match for the test system may be quite arbitrary."

On the other hand, with method b), "each comparison between the test system and a reference system appears to the listeners as if it were unrelated to any other comparison. Especially in a balanced test design, . . . the listener has no way of knowing which system is the test system and which is a reference system. He has no alternative but to listen and to express an independent preference for each pair of systems."

The experimental hypothesis: Listener variance with method b) is less than with method a).

In the experiment, the materials consisted of modified Harvard Test Sentences spoken by six speakers.

Method b) Procedure: Five reference systems of speech transmission defined.

A: "Optimal"; 50–10 000 hertz, S/N ratio = 45 decibels.
B: Bandpass filtering 800–3000 hertz, 24 decibels per octave attenuation rate.
C: Low-pass filtering hertz (24 decibels per octave) + white noise low-pass filtered 500 hertz (9 decibels per octave); peak S/N ratio = 10 decibels.
D: Reverberant echo; delay of first echo 150 microseconds, echo attenuation = 6 decibels.
E: Peak-clipping (approximately 30 decibels) followed by bandpass filtering 300–2000 hertz (24 decibels per octave).

Parameters of these systems were adjusted in pilot-experiment trials such that each was reliably preferred to next in order.

Scoring consists essentially of counting the number of times the test system is preferred. Since in this experiment there were five reference systems and a replication, the maximum number of votes was 10. A conversion to a "preference rating scale" is given to adjust for vote-splitting in isopreferent situations. This scale is given by

$$\begin{aligned} N &= 1.25\,(n-1), &\quad 0 \le n \le 9 \\ &= 10, & n = 10, \end{aligned}$$

where n is the number of votes for the test system. (Since over most of the range the transformation is linear, it is not clear that the transformation is essential.)

Two listening sessions were run. The scores of 14 of 36 listeners were retained as sufficiently consistent.

Method a) Procedure: Reference systems were wide-band speech plus low-pass filtered white noise (500 hertz, 9 decibels per octave) at signal-to-noise ratios of 45, 25, 15, 10, and 5 decibels. Pilot experimentation indicated that the extreme conditions were approximately isopreferent to systems A and E of method b); the other ratios were chosen to produce preference roughly equal to B, C, and D. Scoring as in method b), 16 of 34 listeners provided sufficiently consistent scores.

Principal Results: Standard deviation of scores for the test system with method a) were approximately 2.0 and with method b) were approximately 1.5 (in both cases in a range of 11.25). These results were taken to confirm the experimental hypothesis. [Utilitarian]

M. H. L. Hecker and N. Guttman, "A survey of methods for measuring speech quality," *J. Audio Eng. Soc.*, vol. 15, pp. 400–403, 1967.

Hecker and Guttman's review divides studies concerned with measuring speech quality into two types: utilitarian and analytic. Utilitarian methods are concerned with determining a measure of speech quality suitable for the evaluation of speech transmission systems. Included here are pair-comparison methods which, typically, have yielded isopreference contours, and rating scale methods, used to determine the amount of effort involved in conducting a conversation. These methods are used when the degree of speech distortion is not too large. When considerable distortion is present, latencies have been employed to evaluate speech transmission systems.

The analytic methods attempt analyses of speech quality to determine its psychological components. These studies use multivariate techniques, most frequently factor analyses of semantic differential datum and of similarity and preference data. Results have been variable. McGee found two dimensions for filtered speech, and McDermott found three dimensions using distortion-producing circuits. These results may be somewhat determined by the type of distortion used.

Instead of using speech transmission distortions, Voiers has analyzed individual differences among speakers, and has obtained four dimensions of speaker variability: magnitude, roughness, clarity, and animation.

The reviewers point out that the utilitarian attempts to obtain a unidimensional solution probably is not wholly valid since factor analysis yields multidimensional solutions for speech quality. Comparability of results across studies is sometimes difficult since there is system variability as well as considerable speaker and listener variability.

D. A. Lewinski, "A new objective for message circuit noise," *Bell Sys Tech. J.*, vol. 48, pp. 719–740, 1964.

This article presents a review of the absolute-judgment technique, here used to evaluate noise in transmission systems. A method is presented by which judgments from "excellent" to "poor" are normalized and plotted in terms of cumulative distributions. These show the percentage of observers making the response poor or better, fair or better, good or better, and excellent. Plotted against the variable under investigation these may then be used to set criteria. A method is also given to resolve response uncertainty during boundary conditions. The author recommends the use of anchor stimuli, since he found the results of the first exposure to the conditions under investigation to be extremely variable. [Utilitarian]

V. E. McGee, "Semantic components of the quality of processed speech," *J. Speech Hearing Res.*, vol. 7, pp. 310–323, 1964.

The purpose of this article is to derive psychological components of speech quality by multidimensional methods of analysis.

The method is as follows. One male speaker recorded 105 phonetically balanced sentences and one common test sentence "Joe took . . . ." Fifteen distortions produced, including four low-pass, four high-pass, six band-pass, and one all-pass (zero) filter distortions. Listeners heard a) complete paired-comparison schedule of test sentences, scoring in terms of "better" or "worse"; b) 15 distortion sets of six sentences each, scoring each set on 30 semantic differential scales.

Analysis of a) by Thurstone Case V yielded a quality-judgment scale. As a by-product, an order effect in favor of the second member of a pair was observed.

Analysis of b) by the Eckhart–Young method showed that 18 of the 30 scales could be represented by one "point of view," a stereotype of listener behavior. For 11 scales, differences of opinion involved two distortions, and for one scale ("rich–poor"), three distortions made significant contributions.

Among the listeners, nine of 108 were eliminated as being insufficiently consistent. The source of the different "points of view" was consistently traced to a group of 12 of the remaining 99 listeners.

Finally, 15 of the semantic differential scales, an articulation index scale, and the quality scale derived from a) were factor analyzed. Two roots were significant. One factor, the Intelligibility factor, was placed through the articulation index location and the other through a nearly orthogonal cluster identified by the big–small, wide–narrow, full–thin, smooth–rough, pleasing–annoying, natural–unnatural scales. This second factor was termed the Naturalness factor. Justification for these placements came from analysis showing that the scale values of the two factors predicted the quality scale with a multiple correlation coefficient of 0.95. Individually, the factors correlated at about 0.82 with the quality scale. [Analytic]

V. E. McGee, "Determining perceptional spaces for the quality of filtered speech," *J. Speech Hearing Res.*, vol. 8, pp. 23–38, 1965.

This paper is a complement of V. E. McGee (*J. Speech Hearing Res.*, vol. 7, pp. 310–323, 1964). The 15 frequency distortions described there were arranged into three tape recordings: T_1 was a complete paired-comparison schedule of the 15 distortions of a single one-sentence recording; T_2 was a complete paired-comparison recording using different sentences in each pair. T_3 contained six sentences recorded under each distortion. For T_1 and T_2, listeners judged a) which member had better quality and b) how similar they were (on an 11-point scale).

Results involving 100 judges showed a strong order effect—the second member of a pair was favored. In a second check of reliability, judgments on T_1 and T_2 of a selected group (high internal consistency) of 31 male and 31 female judges (reduced to make analysis manageable) were analyzed by the Eckhart–Young procedure. The four plots in the first two axes were superimposed. Similarity of coordinates was good.

On further analysis to determine perceptional spaces, four "points of view" of listeners were identified, one due to the "mean" point of view, and three due to "ideal" listeners (centroids of clusters of consistent listeners). The perceptual spaces corresponding to each "view" were found to be predominantly one-dimensional, and the two dimensions in each correlated highly with each other. Therefore, the 15 distortions were found to be in similar positions in each space. A paired-comparison quality scale was placed in these spaces and it was observed that the low-pass distortions tended to cluster on one side of the pair comparison scale, and the high- and band-pass distortions clustered on the other side.

It is concluded that "for the average American listener the motion of speech quality is a one-dimensional attribute operating at two distinct levels, or under two distinct response sets, which have been named 'lowest harmonics—YES' and 'lowest harmonics—NO.' " [Analytic]

W. A. Munson and J. E. Karlin, "Isopreference method for evaluating speech transmission circuits," *J. Acoust. Soc. Am.*, vol. 34, pp. 762–774, 1962.

This is an exploratory paper describing a modification of the paired-comparison technique for deriving a one-dimensional scale for rating speech transmission systems on the basis of listener preferences. It was intended to cover any speech transmission system, as long as it was worse than the reference system, which in this study was natural speech. The paper initially reviews and classifies other transmission rating systems as direct and indirect comparisons. Direct comparisons are then subclassified into systems with fixed reference, variable test; variable reference, fixed test; and both test and reference systems variable. The isopreference method is an example of the last type.

The method is as follows. Subjects are presented with a number of test systems, such as high-pass and low-pass filtered speech, and high-fidelity speech. In one series the speech level is held constant, for example, and the noise level is varied in five steps. From the subjects' preferences of the test system over the reference system a psychometric curve can be plotted as a function of the noise level. The 50 percent point is then taken to be the isopreferent point and determines one point on the isopreference contour for the test system as a function of the particular speech and noise level. A considerable number of points obtained in this manner for different speech and noise levels then allow the plotting of the isopreference contours for the test system. The entire procedure then has to be repeated for each additional test system. The paper shows isopreference plots for 1500 hertz, 3000 hertz low pass, 500 hertz high pass, and monophonic high-quality loudspeaker systems.

The next step was to check on transitivity within each set of isopreference contours by having the subjects make direct comparisons. The transitivity conditions were generally satisfied. A check was also made on transitivity between isopreference planes. This test was also successful in spite of the fact that the variable in one system was noise level and in the other system was speech level.

A possibility then exists of assigning numbers to the isopreference contours which would rank them in terms of absolute preference by the observers. This was done by plotting the isopreference points, speech level being held constant, of the test system noise level against the reference system noise level. An arbitrary linear Transmission Preference Level (TPL) was also drawn on the ordinate. The TPL scale is in decibels and is equivalent to the spectrum level of noise added to live speech. TPL values of other systems are determined from tests equating noise levels of the reference and test systems.

Following this an attempt was made to measure the distances between the contours to arrive at a preference difference-limen scale, since the TPL numbers contain no information on the strength of preference. To arrive at the new scale, estimates of the standard deviation of the psychometric curves were obtained. It was found that the standard deviation scale tended to increase with higher TPL, i.e., higher quality. The new scale, the Transmission Preference Unit scale, was calculated from the integral of the inverse of the standard deviation versus TPL function. On the assumption of normal distribution, e.g., that one standard deviation represents preference by about 85 percent of listeners, differences on the TPU scale can be used to estimate the percentage of users who would prefer the circuit having the higher value.

In their discussion the authors point out that the tests were devised in order to improve measurements with only a small number of observers. The test is not limited by any assumption as to which factor, such as loudness, effort, articulation, etc., contributes most to the satisfaction of the user. Additional validation is necessary before the TPU scale can be used for rating purposes in a practical sense, and also the work should be extended to include two-way conversations. Ideally, after a sufficient number of valid isopreference contours have been obtained, it would only be a matter of making the appropriate physical measurements, and a rating could be obtained from the charts. A new system would be evaluated in two stages. First, a self contained set of isopreferent contours would be determined. Then, in order to place the contours in the TPL and TPU scales, comparisons with a known system would be made; having made these comparisons, speech volumes of the known and test systems are equated.

The present tests were limited in scope, and subsequent investigators are urged to continue the work of extension and validation. [Utilitarian]

Y. Ochiai and T. Fukumura, "Timbre study of vocalic voices," *Mem. Fac. Engrg. Nagoya,* Nagoya University, vol. 5, pp. 253–280, 1953. "Timbre study of vocalic voices viewed from subjective phonal aspect, *Ibid.*, vol. 8, pt. 1, pp. 1–18; pt. 2, pp. 203–221; pt. 3, pp. 222–239; 1956.

These studies have been well summarized by V. McGee (*J. Speech Hearing Res.*, vol. 8, pp. 23–38, 1965). His summary follows.

> The theory . . . of Ochiai and Fukumura . . . concerns a detailed study of "vocalic voices" and will be referred to as the VV theory. The basic data for this theory derive from an extremely thorough analysis of five Japanese vowels spoken by five "callers with no defect of speech" and listened to by four male listeners. Originally the work was conducted at a purely physical level. Then the series of papers entitled "Timbre study of vocalic voices from subjective phonal aspect" appeared and their concern was with "a subjective study which is the other side of a perfect timbre interpretation." It is interesting, by way of comparison between the two theories, to include this quotation (Ochiai and Fukumura, 1956, p. 219):
>
> "The usual way to define the so-called importance is by describing the quality-density distributed per band by the aid of mathematically differentiating the curves of quality characteristics (vs. BED)* with regard to frequency band. This consideration is seemingly without merit and the process seems purposeless. We must reflect here that the process useful for physical quantity has nothing to do with the psychological quantity in that the magnitude of quality does not belong to physical quantity." On the surface this seems like a categorical denouncement of the AI theory which clearly fits the description of "the usual way." However, the aim of the AI theory is unambiguous and it is possible to agree with Ochiai and Fukumura without abandoning the AI theory of intelligibility.
>
> For present purposes note that Ochiai and Fukumura considered both an intelligibility factor ("phonal quality") and a naturalness factor ("vocal quality"). Initially their studies involved purely physical manipulation of a sample of vowels and voices. The phonal pattern for a particular vowel was obtained simply by averaging the components of the several utterances of that vowel by the five talkers. The vocal pattern for a particular voice was obtained simply by averaging the components of the several utterances by that voice of the five vowels. No human judgments were involved until the confusion studies in the 1956 papers. Four trained listeners made naturalness and articulation judgments—of a nature not reported in the articles—for 25 different stimuli (five voices each saying five vowels) and over eight different band-eliminating-distortion conditions. The results were summarized in the form of confusion diagrams and can be studied on page 211 of Ochiai and Fukumura (1956).
>
> Three important results emerge from these confusion studies. First, a distinct difference in the confusion patterns of phonemes and voices as a function of band-eliminating-distortion was observed. For the phonemes—referring to the five vowels of the VV study—the patterns of confusion for different conditions of frequency distortion were marked by abrupt changes in both direction and quantity. On the other hand, the pattern of confusion among the five voices was relatively stable no matter what type of distortion was involved. This was explained by a second important concept, namely that of two types of quality distribution. The phoneme or phonal aspect of timbre quality was said to have a concentrated distribution of quality, whereas the vocal aspect was typified by a dispersed distribution of quality. Stated somewhat differently, it would appear that the perception of articulation quality is more directly tied in with select portions of the speech spectrum and that the perception of naturalness quality has a more subtle association with the total spectrum. A third result of importance is the distinction between "quality-formative" and "quality-ruinous" trends due to band-eliminating distortions. By this is meant the possibility of both positive and negative importances being attached to bands of frequencies, so that the importance curve over the frequency range for, say, naturalness, might be characterized by both positive and negative values. [Analytic]

Y. Ochiai, "Phoneme and voice identification studies using Japanese vowels," *Language and Speech,* vol. 2, pp. 132–136, 1959.

This article is a brief summary of a large number of experiments carried out by Prof. Ochiai and his collaborators.

The speech material used was steady-pitched, sustained vowels produced by five speakers, and, in one instance, by a child. Vowel intelligibility and speaker identity were determined as functions of high- and low-pass filtering. The main results are that, as functions of cutoff frequency in directions of increasing distortion, the gradient of the intelligibility curves is flat to steep, whereas for identity, it is moderate (−7 percent per octave) over a wide range. The high- and low-pass curves cross at about 1100 hertz for vowels and at about 1400 hertz for voices; the "quality balancing point" is always higher for voices. [Analytic]

* BED = band eliminating distortion.

E. H. Rothauser, G. E. Urbanek, and W. P. Pachl, "Isopreference methods for speech evaluation," *J. Acoust. Soc. Am.*, vol. 44, pp. 408–418, 1968.

This work concentrates not on the question of how well a listener understands a speech signal transmitted by a system but on the question of how well he likes the signal. The latter question deals with preference.

"Speech quality includes various factors such as optimum loudness, timbre and rhythmic character, annoyance, a possible fatigue of listener, speaker identifiability, naturalness, clarity, systematic amplitude or time distortions, and many others." Quantitative evaluation of these factors may be difficult or seemingly impossible.

Preference, as a parameter of speech quality, is viewed as an expression of the "average attitude of a listener" towards a test speech signal when he compares it with a reference speech signal. "Preference is thus a relative measure of quality."

Attempts to replicate the results of Munson and Karlin's isopreference experiments were not completely successful. Listener responses were consistent enough if there was a clear difference between the conditions, e.g., high versus low levels of test signal. "This means that the decisions of the listeners are influenced by the loudness levels of the previously presented signal pairs."

In the method tested, the loudness of both test and reference signals was kept constant to minimize effects of conditions at the optimum loudness of the test signal. Only the signal-to-noise ratio of the reference signal was varied. The test and reference (high fidelity) signals were recorded on separate tracks of a tape. A listening exposure consisted of the sequence A B A B, where A is the test signal and B is the reference signal degraded by the addition of noise. Each A and B lasts five seconds. The listener registers his preference for A or B.

Even with simplification, there are many parameters of preference testing that need evaluation. These include a) with respect to the reference signal, type of test material (continuous speech or word lists, level of loudness, mode of presentation) live or recorded, type of degradation and distortion; b) with respect to the test signal, continuous text or not, loudness, type of reproduction; c) with respect to mode of presentation, the type of acoustic transducer, type of ambient noise; d) with respect to presentation format, size of listening group, degree of listener training, test repetition.

Desirable properties of the reference signal are a) wide range from high fidelity to any level of degradation; b) similarity to anticipated test signals in order to improve reliability of subjective comparison; c) simple generation and physical definition; d) simple interpretation.

The reference signal was defined as

$$r(t) = s(t) + kd(t),$$

where $s(t)$ is a high-fidelity speech signal and $d(t)$ is a degradation added in proportion k, $0 < k < 1$.

The use of $d(t) = n_0(t)$, i.e., additive noise, is not entirely satisfactory because a) the listener tends to separate speech and noise; b) noise is present even when speech is absent (this might be a factor if word lists are used); c) $r(t)$ becomes dissimilar to the test signals; d) the speech level is not defined accurately.

Worth investigating is $d(t) = s(t)n_0'(t)$. Then the multiplicative reference is $r'(t) = s(t)[1 + kn_0'(t)]$, and $S/N = 20 \log \{s(t) \mid [ks(t)n_0'(t)]\} = 20 \log 1/kn_0'(t)$. Thus S/N is independent of signal level. On contrast with additive noise, $S/N = 20 \log s(t) [kn_0'(t)] - 1$. Experimentally, standard deviation of subjective test results are smaller with multiplicative reference $r'(t)$ than with $r(t)$. Undergoing investigation is another degradation consisting of $ks(t)n^*(t)$, when $n^*(t)$ is random inversion of signal polarity.

Three groups of listeners were used in these tests: a) college students untrained in listening; b) about 20 trained students; c) staff. All listening was done with earphones. Data collection was automated. Each listener was exposed to the test signal and to the reference signal at signal-to-noise ratios expected to produce a range of preferences from low to high. For each listener, isopreference point M is the median of "displaced" preferences. Thus, e.g., in order of increasing signal-to-noise ratio, in the run of votes AABABAAA;AABBBB, M is placed at the semicolon. Individual and group distributions accumulated.

The results of the measurements were as follows.

(a) Effect of six reported tests. Test signal was "normal vocoder." Multiplicative reference $r'(t)$. Standard deviation of means of 10 listeners (the group standard deviation) was 1.3 decibels. Average standard deviation of listeners was 1.9 decibels.

(b) In a test involving high pass, low pass, and two vocoder systems, both $r(t)$ and $r'(t)$ correlated well with rank orders provided by subjects. The range of mean reference levels was 11.7 decibels for $r'(t)$ and 6.3 decibels for $r(t)$. Group standard deviations were similar (3–4 decibels for vocoders to 6–7 decibels for low pass).

Note: word intelligibilities (German monosyllabic words) were 89, 74, 65, and 69 percent for high pass, low pass, and the two vocoders.

(c) Direct comparison of $r(t)$ and $r'(t)$: These pairs of values $[r(t), r'(t)]$ roughly define equivalences: (−3, −20), (−2, −10), (3, 0), (6, 5), (13, 10). Standard deviations of $r(t)$ tend to increase with signal-to-noise, i.e., from approximately 2 decibels at low signal-to-noise to 4 decibels at high. Standard deviations for $r'(t)$ are approximately 4 decibels throughout.

(d) Intelligibilities of $r(t)$ and $r'(t)$: for $r(t)$, 25 percent at signal-to-noise of −9 decibels, 50 percent at −6 decibels, 75 percent at −2 decibels, 95 percent at −1 decibel. For $r'(t)$, 88 percent at $-\infty$, 89 percent at −10 decibels, 98 percent at 0 decibel.

(e) Loudness: Optimum speech levels were found to be 71 decibel-amperes, where decibel-ampere is the high-fidelity speech signal weighted by the *A* sound-level-meter filter. Approximately equally favored levels were 69–74 decibels. Optimum levels for high pass, low pass and normal vocoder were within 5 decibels.

(f) Other loudness measurements: Direct loudness switching between high-fidelity speech pink noise (−3 decibels per octave). For additive noise, signal-to-noise (equated in loudness) = signal-to-noise (physically measured ratio) −5 decibels. For multiplicative noise, signal-to-noise (loudness definition) = signal-to-noise (speech level definition) + 3 decibels.

(g) Tests of order of presentation: Presentations AB and BA in the manner of test presentation. When the listeners have no chance to form a "stationary concept of the test speech quality," they favor the latter member of the pair.

(h) Fatigue: no systematic effects noted.

(i) Variability of single listeners: Some instability was noted, e.g., 8-decibel spread of signal-to-noise reference (group average) at beginning of tests narrowed to 4 decibels after seven test sessions. "This reduction . . . seems to be an effect of overtraining." [Utilitarian]

M. A. Sapozhkov, *The Speech Signal in Cybernetics and Communications.* Moscow: State Communications and Radio Publishing House, 1963 (translated by U. S. Dept. Commerce JPRS:28, 117, TT:65-30045).

This extensive review does not touch directly on the problem of measurement of the quality of high-intelligibility transmission systems. It is stated that analysis of speech transmission conditions may be made from the standpoint of intelligibility, fidelity, and, as an intermediate case, voice recognizability. "There is of course a connection between these factors, and in the case of untransformed speech, this connection is almost totally single-valued. In the case of transformed speech this connection cannot be said to be single-valued, but nevertheless attempts are often made to evaluate speech transmission quality solely on the basis of intelligibility" (pp. 24–25).

Analysis by hundreds of listeners of Russian speech connects word intelligibility and quality as follows: ideal, 100–99 percent; excellent, 99–98 percent; good, 98–93 percent; satisfactory, 93–87 percent; minimally adequate, 87–77 percent; communications breakdown, 77–60 percent.

J. Swaffield and D. L. Richards, "Rating of speech links and performance of telephone networks," *Proc. IEE* (London), vol. 106, pp. 65–76, 1959.

D. L. Richards and J. Swaffield, "Assessment of speech communication links," *ibid.*, pp. 77–92.

This pair of papers, which are complementary, treat the "Immediate Appreciation Method" and how it may be used in the overall rating of a telephone system. An informational analysis is used to define the five judgment categories: poor, fair, good, excellent, and perfect. These classes are also defined by the amount by which a) the extent to which the link is distinguishable from direct air path, b) the talking/listening effort, c) the conversational effort, and d) the message rate. In the one-way test, listeners rate the received speech by judging according to an opinion scale based on effort, the best: "complete relaxation possible, no effort required," to the worst: "no meaning understood with any feasible effort." In two-way tests, conversation is generated by having the subjects solve a puzzle by means of the circuit. Afterwards ratings are made on the effort scale. The ratings are averaged over listeners to arrive at a Mean Opinion Score, which is then plotted against the transmission conditions.

The assessment methods as described in Richards and Swaffield are used to arrive at ratings of telephone networks. Additionally, the papers review historical and other current methods for assessing transmission performance. The development of a Standard Speech Link to serve as a basis of comparison is recommended, and the composition of such a system in the British Post Office is described. [Utilitarian]

W. H. Tedford, Jr., and T. V. Frazier, "Further study of the isopreference method of circuit evaluation," *J. Acoust. Soc. Am.*, vol. 39, pp. 645–650, 1966.

This article presents evidence in support of the Munson–Karlin method. The conditions used were generally poorer than those of Munson and Karlin, and it was found that the method remained applicable. For identical conditions the previous results were not replicated; this was attributed to equipment and environmental differences. The authors consider the trend of the results sufficient to recommend the isopreference method for large scale testing, and feel that a reliable transmission preference unit scale should be established. [Utilitarian]

W. D. Voiers, M. F. Cohen, and J. Michunas, "Evaluation of speech processing devices I, intelligibility, quality, speaker recognizability," AFCRL-65-826, 1965.

The quality of transmitted speech, which the authors also call esthetic acceptability, is treated by means of the standard unit-variance method. Speech was processed by four representative vocoder systems to provide standards with which experimentally processed speech is compared by listeners. The listener response data are analyzed to yield a value representing the position of the experimental system on a standard unit variance scale of esthetic acceptability.

The fundamental principle of the standard unit-variance method is one whereby inter-individual differences in preference response distribute normally with respect to an unidimensional continuum which has the properties of an equal-interval scale. The method is thus a pair-comparison method based on a principle underlying Thurstone's Law of Comparative Judgment. The name of the method is derived from the fact that the estimated true variance among the listeners in their evaluation of quality differences provides the basis for the establishment of a scale unit.

The steps required to apply the method are the following.

(a) Convert observed frequencies into proportions for each individual speaker and for each pair of comparisons.
(b) Convert proportions into arcsine values.
(c) Use analysis of variance for each of the speakers and for each pair of compared systems to obtain the true inter-listener variance.
(d) Obtain the individual Z scores between the directly compared systems and for each speaker individually.
(e) Obtain a mean Z score for each speaker using those pairs of systems which have one in common.
(f) Obtain a mean Z score of five speakers for a given system. The scale value for the system is then the sum of the three Z scores divided by four.

From these a standard unit-variance scale may be derived to provide maximum precision in the prediction of relative preference frequencies for any two systems which have been scaled by the unit-variance method. It is obtained by simple linear transformation of the unit-variance scale by means of an experimentally determined factor that is determined on the basis of observed scale values of the standard comparison systems (here vocoders). The scale values represent points on the normal distribution where the scale unit is based on the true standard deviation of the scores that represent measures of an individual preference for a given condition or system.

The other sections of the paper treat intelligibility, and a speaker-recognition method based on the semantic differential. [Utilitarian]

APPENDIX C

1965 Revised List of Phonetically Balanced Sentences (Harvard Sentences)

List 1

1. The birch canoe slid on the smooth planks.
2. Glue the sheet to the dark blue background.
3. It's easy to tell the depth of a well.
4. These days a chicken leg is a rare dish.
5. Rice is often served in round bowls.
6. The juice of lemons makes fine punch.
7. The box was thrown beside the parked truck.
8. The hogs were fed chopped corn and garbage.
9. Four hours of steady work faced us.
10. A large size in stockings is hard to sell.

List 2

1. The boy was there when the sun rose.
2. A rod is used to catch pink salmon.
3. The source of the huge river is the clear spring.
4. Kick the ball straight and follow through.
5. Help the woman get back to her feet.
6. A pot of tea helps to pass the evening.
7. Smoky fires lack flame and heat.
8. The soft cushion broke the man's fall.
9. The salt breeze came across from the sea.
10. The girl at the booth sold fifty bonds.

List 3

1. The small pup gnawed a hole in the sock.
2. The fish twisted and turned on the bent hook.
3. Press the pants and sew a button on the vest.
4. The swan dive was far short of perfect.
5. The beauty of the view stunned the young boy.
6. Two blue fish swam in the tank.
7. Her purse was full of useless trash.
8. The colt reared and threw the tall rider.
9. It snowed, rained, and hailed the same morning.
10. Read verse out loud for pleasure.

List 4

1. Hoist the load to your left shoulder.
2. Take the winding path to reach the lake.
3. Note closely the size of the gas tank.
4. Wipe the grease off his dirty face.
5. Mend the coat before you go out.
6. The wrist was badly strained and hung limp.
7. The stray cat gave birth to kittens.
8. The young girl gave no clear response.
9. The meal was cooked before the bell rang.
10. What joy there is in living.

List 5

1. A king ruled the state in the early days.
2. The ship was torn apart on the sharp reef.
3. Sickness kept him home the third week.
4. The wide road shimmered in the hot sun.
5. The lazy cow lay in the cool grass.
6. Lift the square stone over the fence.
7. The rope will bind the seven books at once.
8. Hop over the fence and plunge in.
9. The friendly gang left the drug store.
10. Mesh wire keeps chicks inside.

List 6

1. The frosty air passed through the coat.
2. The crooked maze failed to fool the mouse.
3. Adding fast leads to wrong sums.
4. The show was a flop from the very start.
5. A saw is a tool used for making boards.
6. The wagon moved on well oiled wheels.
7. March the soldiers past the next hill.
8. A cup of sugar makes sweet fudge.
9. Place a rosebush near the porch steps.
10. Both lost their lives in the raging storm.

List 7

1. We talked of the side show in the circus.
2. Use a pencil to write the first draft.

3. He ran half way to the hardware store.
4. The clock struck to mark the third period.
5. A small creek cut across the field.
6. Cars and busses stalled in snow drifts.
7. The set of china hit the floor with a crash.
8. This is a grand season for hikes on the road.
9. The dune rose from the edge of the water.
10. Those words were the cue for the actor to leave.

List 8

1. A yacht slid around the point into the bay.
2. The two met while playing on the sand.
3. The ink stain dried on the finished page.
4. The walled town was seized without a fight.
5. The lease ran out in sixteen weeks.
6. A tame squirrel makes a nice pet.
7. The horn of the car woke the sleeping cop.
8. The heart beat strongly and with firm strokes.
9. The pearl was worn in a thin silver ring.
10. The fruit peel was cut in thick slices.

List 9

1. The Navy attacked the big task force.
2. See the cat glaring at the scared mouse.
3. There are more than two factors here.
4. The hat brim was wide and too droopy.
5. The lawyer tried to lose his case.
6. The grass curled around the fence post.
7. Cut the pie into large parts.
8. Men strive but seldom get rich.
9. Always close the barn door tight.
10. He lay prone and hardly moved a limb.

List 10

1. The slush lay deep along the street.
2. A wisp of cloud hung in the blue air.
3. A pound of sugar costs more than eggs.
4. The fin was sharp and cut the clear water.
5. The play seems dull and quite stupid.
6. Bail the boat to stop it from sinking.
7. The term ended in late June that year.
8. A tusk is used to make costly gifts.
9. Ten pins were set in order.
10. The bill was paid every third week.

List 11

1. Oak is strong and also gives shade.
2. Cats and dogs each hate the other.
3. The pipe began to rust while new.
4. Open the crate but don't break the glass.
5. Add the sum to the product of these three.
6. Thieves who rob friends deserve jail.
7. The ripe taste of cheese improves with age.
8. Act on these orders with great speed.
9. The hog crawled under the high fence.
10. Move the vat over the hot fire.

List 12

1. The bark of the pine tree was shiny and dark.
2. Leaves turn brown and yellow in the fall.
3. The pennant waved when the wind blew.
4. Split the log with a quick, sharp blow.
5. Burn peat after the logs give out.
6. He ordered peach pie with ice cream.
7. Weave the carpet on the right hand side.
8. Hemp is a weed found in parts of the tropics.
9. A lame back kept his score low.
10. We find joy in the simplest things.

List 13

1. Type out three lists of orders.
2. The harder he tried the less he got done.
3. The boss ran the show with a watchful eye.
4. The cup cracked and spilled its contents.
5. Paste can cleanse the most dirty brass.
6. The slang word for raw whiskey is booze.
7. It caught its hind paw in a rusty trap.
8. The wharf could be seen at the farther shore.
9. Feel the heat of the weak dying flame.
10. The tiny girl took off her hat.

List 14

1. A cramp is no small danger on a swim.
2. He said the same phrase thirty times.
3. Pluck the bright rose without leaves.
4. Two plus seven is less than ten.
5. The glow deepened in the eyes of the sweet girl.
6. Bring your problems to the wise chief.
7. Write a fond note to the friend you cherish.
8. Clothes and lodging are free to new men.
9. We frown when events take a bad turn.
10. Port is a strong wine with a smoky taste.

List 15

1. The young kid jumped the rusty gate.
2. Guess the results from the first scores.
3. A salt pickle tastes fine with ham.
4. The just claim got the right verdict.
5. These thistles bend in a high wind.
6. Pure bred poodles have curls.
7. The tree top waved in a graceful way.
8. The spot on the blotter was made by green ink.
9. Mud was spattered on the front of his white shirt.
10. The cigar burned a hole in the desk top.

List 16

1. The empty flask stood on the tin tray.
2. A speedy man can beat this track mark.
3. He broke a new shoelace that day.
4. The coffee stand is too high for the couch.
5. The urge to write short stories is rare.
6. The pencils have all been used.

7. The pirates seized the crew of the lost ship.
8. We tried to replace the coin but failed.
9. She sewed the torn coat quite neatly.
10. The sofa cushion is red and of light weight.

List 17

1. The jacket hung on the back of the wide chair.
2. At that high level the air is pure.
3. Drop the two when you add the figures.
4. A filing case is now hard to buy.
5. An abrupt start does not win the prize.
6. Wood is best for making toys and blocks.
7. The office paint was a dull, sad tan.
8. He knew the skill of the great young actress.
9. A rag will soak up spilled water.
10. A shower of dirt fell from the hot pipes.

List 18

1. Steam hissed from the broken valve.
2. The child almost hurt the small dog.
3. There was a sound of dry leaves outside.
4. The sky that morning was clear and bright blue.
5. Torn scraps littered the stone floor.
6. Sunday is the best part of the week.
7. The doctor cured him with these pills.
8. The new girl was fired today at noon.
9. They felt gay when the ship arrived in port.
10. Add the store's account to the last cent.

List 19

1. Acid burns holes in wool cloth.
2. Fairy tales should be fun to write.
3. Eight miles of woodland burned to waste.
4. The third act was dull and tired the players.
5. A young child should not suffer fright.
6. Add the column and put the sum here.
7. We admire and love a good cook.
8. There the flood mark is ten inches.
9. He carved a head from the round block of marble.
10. She has a smart way of wearing clothes.

List 20

1. The fruit of a fig tree is apple-shaped.
2. Corn cobs can be used to kindle a fire.
3. Where were they when the noise started.
4. The paper box is full of thumb tacks.
5. Sell your gift to a buyer at a good gain.
6. The tongs lay beside the ice pail.
7. The petals fall with the next puff of wind.
8. Bring your best compass to the third class.
9. They could laugh although they were sad.
10. Farmers came in to thresh the oat crop.

List 21

1. The brown house was on fire to the attic.
2. The lure is used to catch trout and flounder.
3. Float the soap on top of the bath water.
4. A blue crane is a tall wading bird.
5. A fresh start will work such wonders.
6. The club rented the rink for the fifth night.
7. After the dance, they went straight home.
8. The hostess taught the new maid to serve.
9. He wrote his last novel there at the inn.
10. Even the worst will beat his low score.

List 22

1. The cement had dried when he moved it.
2. The loss of the second ship was hard to take.
3. The fly made its way along the wall.
4. Do that with a wooden stick.
5. Live wires should be kept covered.
6. The large house had hot water taps.
7. It is hard to erase blue or red ink.
8. Write at once or you may forget it.
9. The doorknob was made of bright clean brass.
10. The wreck occurred by the bank on Main Street.

List 23

1. A pencil with black lead writes best.
2. Coax a young calf to drink from a bucket.
3. Schools for ladies teach charm and grace.
4. The lamp shone with a steady green flame.
5. They took the axe and the saw to the forest.
6. The ancient coin was quite dull and worn.
7. The shaky barn fell with a loud crash.
8. Jazz and swing fans like fast music.
9. Rake the rubbish up and then burn it.
10. Slash the gold cloth into fine ribbons.

List 24

1. Try to have the court decide the case.
2. They are pushed back each time they attack.
3. He broke his ties with groups of former friends.
4. They floated on the raft to sun their white backs.
5. The map had an X that meant nothing.
6. Whitings are small fish caught in nets.
7. Some ads serve to cheat buyers.
8. Jerk the rope and the bell rings weakly.
9. A waxed floor makes us lose balance.
10. Madam, this is the best brand of corn.

List 25

1. On the islands the sea breeze is soft and mild.
2. The play began as soon as we sat down.
3. This will lead the world to more sound and fury.
4. Add salt before you fry the egg.
5. The rush for funds reached its peak Tuesday.
6. The birch looked stark white and lonesome.
7. The box is held by a bright red snapper.
8. To make pure ice, you freeze water.
9. The first worm gets snapped early.
10. Jump the fence and hurry up the bank.

List 26

1. Yell and clap as the curtain slides back.
2. They are men who walk the middle of the road.
3. Both brothers wear the same size.
4. In some form or other we need fun.
5. The prince ordered his head chopped off.
6. The houses are built of red clay bricks.
7. Ducks fly north but lack a compass.
8. Fruit flavors are used in fizz drinks.
9. These pills do less good than others.
10. Canned pears lack full flavor.

List 27

1. The dark pot hung in the front closet.
2. Carry the pail to the wall and spill it there.
3. The train brought our hero to the big town.
4. We are sure that one war is enough.
5. Gray paint stretched for miles around.
6. The rude laugh filled the empty room.
7. High seats are best for football fans.
8. Tea served from the brown jug is tasty.
9. A dash of pepper spoils beef stew.
10. A zestful food is the hot-cross bun.

List 28

1. The horse trotted around the field at a brisk pace.
2. Find the twin who stole the pearl necklace.
3. Cut the cord that binds the box tightly.
4. The red tape bound the smuggled food.
5. Look in the corner to find the tan shirt.
6. The cold drizzle will halt the bond drive.
7. Nine men were hired to dig the ruins.
8. The junk yard had a mouldy smell.
9. The flint sputtered and lit a pine torch.
10. Soak the cloth and drown the sharp odor.

List 29

1. The shelves were bare of both jam or crackers.
2. A joy to every child is the swan boat.
3. All sat frozen and watched the screen.
4. A cloud of dust stung his tender eyes.
5. To reach the end he needs much courage.
6. Shape the clay gently into block form.
7. A ridge on a smooth surface is a bump or flaw.
8. Hedge apples may stain your hands green.
9. Quench your thirst, then eat the crackers.
10. Tight curls get limp on rainy days.

List 30

1. The mute muffled the high tones of the horn.
2. The gold ring fits only a pierced ear.
3. The old pan was covered with hard fudge.
4. Watch the log float in the wide river.
5. The node on the stalk of wheat grew daily.
6. The heap of fallen leaves was set on fire.
7. Write fast if you want to finish early.
8. His shirt was clean but one button was gone.
9. The barrel of beer was a brew of malt and hops.
10. Tin cans are absent from store shelves.

List 31

1. Slide the box into that empty space.
2. The plant grew large and green in the window.
3. The beam dropped down on the workmen's head.
4. Pink clouds floated with the breeze.
5. She danced like a swan, tall and graceful.
6. The tube was blown and the tire flat and useless.
7. It is late morning on the old wall clock.
8. Let's all join as we sing the last chorus.
9. The last switch cannot be turned off.
10. The fight will end in just six minutes.

List 32

1. The store walls were lined with colored frocks.
2. The peace league met to discuss their plans.
3. The rise to fame of a person takes luck.
4. Paper is scarce, so write with much care.
5. The quick fox jumped on the sleeping cat.
6. The nozzle of the fire hose was bright brass.
7. Screw the round cap on as tight as needed.
8. Time brings us many changes.
9. The purple tie was ten years old.
10. Men think and plan and sometimes act.

List 33

1. Fill the ink jar with sticky glue.
2. He smoke a big pipe with strong contents.
3. We need grain to keep our mules healthy.
4. Pack the records in a neat thin case.
5. The crunch of feet in the snow was the only sound.
6. The copper bowl shone in the sun's rays.
7. Boards will warp unless kept dry.
8. The plush chair leaned against the wall.
9. Glass will clink when struck by metal.
10. Bathe and relax in the cool green grass.

List 34

1. Nine rows of soldiers stood in line.
2. The beach is dry and shallow at low tide.
3. The idea is to sew both edges straight.
4. The kitten chased the dog down the street.
5. Pages bound in cloth make a book.
6. Try to trace the fine lines of the painting.
7. Women form less than half of the group.
8. The zones merge in the central part of town.
9. A gem in the rough needs work to polish.
10. Code is used when secrets are sent.

List 35

1. Most of the news is easy for us to hear.
2. He used the lathe to make brass objects.
3. The vane on top of the pole revolved in the wind.
4. Mince pie is a dish served to children.
5. The clan gathered on each dull night.

6. Let it burn, it gives us warmth and comfort.
7. A castle built from sand fails to endure.
8. A child's wit saved the day for us.
9. Tack the strip of carpet to the worn floor.
10. Next Tuesday we must vote.

List 36

1. Pour the stew from the pot into the plate.
2. Each penny shone like new.
3. The man went to the woods to gather sticks.
4. The dirt piles were lines along the road.
5. The logs fell and tumbled into the clear stream.
6. Just hoist it up and take it away.
7. A ripe plum is fit for a king's palate.
8. Our plans right now are hazy.
9. Brass rings are sold by these natives.
10. It takes a good trap to capture a bear.

List 37

1. Feed the white mouse some flower seeds.
2. The thaw came early and freed the stream.
3. He took the lead and kept it the whole distance.
4. The key you designed will fit the lock.
5. Plead to the council to free the poor thief.
6. Better hash is made of rare beef.
7. This plank was made for walking on.
8. The lake sparkled in the red hot sun.
9. He crawled with care along the ledge.
10. Tend the sheep while the dog wanders.

List 38

1. It takes a lot of help to finish these.
2. Mark the spot with a sign painted red.
3. Take two shares as a fair profit.
4. The fur of cats goes by many names.
5. North winds bring colds and fevers.
6. He asks no person to vouch for him.
7. Go now and come here later.
8. A sash of gold silk will trim her dress.
9. Soap can wash most dirt away.
10. That move means the game is over.

List 39

1. He wrote down a long list of items.
2. A siege will crack the strong defense.
3. Grape juice and water mix well.
4. Roads are paved with sticky tar.
5. Fake stones shine but cost little.
6. The drip of the rain made a pleasant sound.
7. Smoke poured out of every crack.
8. Serve the hot rum to the tired heroes.
9. Much of the story makes good sense.
10. The sun came up to light the eastern sky.

List 40

1. Heave the line over the port side.
2. A lathe cuts and trims any wood.
3. It's a dense crowd in two distinct ways.
4. His hip struck the knee of the next player.
5. The stale smell of old beer lingers.
6. The desk was firm on the shaky floor.
7. It takes heat to bring out the odor.
8. Beef is scarcer than some lamb.
9. Raise the sail and steer the ship northward.
10. A cone costs five cents on Mondays.

List 41

1. A pod is what peas always grow in.
2. Jerk the dart from the cork target.
3. No cement will hold hard wood.
4. We now have a new base for shipping.
5. A list of names is carved around the base.
6. The sheep were led home by a dog.
7. Three for a dime, the young peddler cried.
8. The sense of smell is better than that of touch.
9. No hardship seemed to keep him sad.
10. Grace makes up for lack of beauty.

List 42

1. Nudge gently but wake her now.
2. The news struck doubt into restless minds.
3. Once we stood beside the shore.
4. A chink in the wall allowed a draft to blow.
5. Fasten two pins on each side.
6. A cold dip restores health and zest.
7. He takes the oath of office each March.
8. The sand drifts over the sill of the old house.
9. The point of the steel pen was bent and twisted.
10. There is a lag between thought and act.

List 43

1. Seed is needed to plant the spring corn.
2. Draw the chart with heavy black lines.
3. The boy owed his pal thirty cents.
4. The chap slipped into the crowd and was lost.
5. Hats are worn to tea and not to dinner.
6. The ramp led up to the wide highway.
7. Beat the dust from the rug onto the lawn.
8. Say it slowly but make it ring clear.
9. The straw nest housed five robins.
10. Screen the porch with woven straw mats.

List 44

1. This horse will nose his way to the finish.
2. The dry wax protects the deep scratch.
3. He picked up the dice for a second roll.
4. These coins will be needed to pay his debt.
5. The nag pulled the frail cart along.
6. Twist the valve and release hot steam.
7. The vamp of the shoe had a gold buckle.
8. The smell of burned rags itches my nose.
9. New pants lack cuffs and pockets.
10. The marsh will freeze when cold enough.

List 45

1. They slice the sausage thin with a knife.
2. The bloom of the rose lasts a few days.
3. A gray mare walked before the colt.
4. Breakfast buns are fine with a hot drink.
5. Bottles hold four kinds of rum.
6. The man wore a feather in his felt hat.
7. He wheeled the bike past the winding road.
8. Drop the ashes on the worn old rug.
9. The desk and both chairs were painted tan.
10. Throw out the used paper cup and plate.

List 46

1. A clean neck means a neat collar.
2. The couch cover and hall drapes were blue.
3. The stems of the tall glasses cracked and broke.
4. The wall phone rang loud and often.
5. The clothes dried on a thin wooden rack.
6. Turn on the lantern which gives us light.
7. The cleat sank deeply into the soft turf.
8. The bills were mailed promptly on the tenth of the month.
9. To have is better than to wait and hope.
10. The price is fair for a good antique clock.

List 47

1. The music played on while they talked.
2. Dispense with a vest on a day like this.
3. The bunch of grapes was pressed into wine.
4. He sent the figs, but kept the ripe cherries.
5. The hinge on the door creaked with old age.
6. The screen before the fire kept in the sparks.
7. Fly by night, and you waste little time.
8. Thick glasses helped him read the print.
9. Birth and death mark the limits of life.
10. The chair looked strong but had no bottom.

List 48

1. The kite flew wildly in the high wind.
2. A fur muff is stylish once more.
3. The tin box held priceless stones.
4. We need an end of all such matter.
5. The case was puzzling to the old and wise.
6. The bright lanterns were gay on the dark lawn.
7. We don't get much money but we have fun.
8. The youth drove with zest, but little skill.
9. Five years he lived with a shaggy dog.
10. A fence cuts through the corner lot.

List 49

1. The way to save money is not to spend much.
2. Shut the hatch before the waves push it in.
3. The odor of spring makes young hearts jump.
4. Crack the walnut with your sharp side teeth.
5. He offered proof in the form of a large chart.
6. Send the stuff in a thick paper bag.
7. A quart of milk is water for the most part.
8. They told wild tales to frighten him.
9. The three story house was built of stone.
10. In the rear of the ground floor was a large passage.

List 50

1. A man in a blue sweater sat at the desk.
2. Oats are a food eaten by horse and man.
3. Their eyelids droop for want of sleep.
4. A sip of tea revives his tired friend.
5. There are many ways to do these things.
6. Tuck the sheet under the edge of the mat.
7. A force equal to that would move the earth.
8. We like to see clear weather.
9. The work of the tailor is seen on each side.
10. Take a chance and win a china doll.

List 51

1. Shake the dust from your shoes, stranger.
2. She was kind to sick old people.
3. The square wooden crate was packed to be shipped.
4. The dusty bench stood by the stone wall.
5. We dress to suit the weather of most days.
6. Smile when you say nasty words.
7. A bowl of rice is free with chicken stew.
8. The water in this well is a source of good health.
9. Take shelter in this tent, but keep still.
10. That guy is the writer of a few banned books.

List 52

1. The little tales they tell are false.
2. The door was barred, locked, and bolted as well.
3. Ripe pears are fit for a queen's table.
4. A big wet stain was on the round carpet.
5. The kite dipped and swayed, but stayed aloft.
6. The pleasant hours fly by much too soon.
7. The room was crowded with a wild mob.
8. This strong arm shall shield your honor.
9. She blushed when he gave her a white orchid.
10. The beetle droned in the hot June sun.

List 53

1. Press the pedal with your left foot.
2. Neat plans fail without luck.
3. The black trunk fell from the landing.
4. The bank pressed for payment of the debt.
5. The theft of the pearl pin was kept secret.
6. Shake hands with this friendly child.
7. The vast space stretched into the far distance.
8. A rich farm is rare in this sandy waste.
9. His wide grin earned many friends.
10. Flax makes a fine brand of paper.

List 54

1. Hurdle the pit with the aid of a long pole.
2. A strong bid may scare your partner stiff.
3. Even a just cause needs power to win.
4. Peep under the tent and see the clowns.

5. The leaf drifts along with a slow spin.
6. Cheap clothes are flashy but don't last.
7. A thing of small note can cause despair.
8. Flood the mails with requests for this book.
9. A thick coat of black paint covered all.
10. The pencil was cut to be sharp at both ends.

List 55

1. Those last words were a strong statement.
2. He wrote his name boldly at the top of the sheet.
3. Dill pickles are sour but taste fine.
4. Down that road is the way to the grain farmer.
5. Either mud or dust are found at all times.
6. The best method is to fix it in place with clips.
7. If you mumble your speech will be lost.
8. At night the alarm roused him from a deep sleep.
9. Read just what the meter says.
10. Fill your pack with bright trinkets for the poor.

List 56

1. The small red neon lamp went out.
2. Clams are small, round, soft, and tasty.
3. The fan whirled its round blades softly.
4. The line where the edges join was clean.
5. Breathe deep and smell the piny air.
6. It matters not if he reads these words or those.
7. A brown leather bag hung from its strap.
8. A toad and a frog are hard to tell apart.
9. A white silk jacket goes with any shoes.
10. A break in the dam almost caused a flood.

List 57

1. Paint the sockets in the wall dull green.
2. The child crawled into the dense grass.
3. Bribes fail where honest men work.
4. Trample the spark, else the flames will spread.
5. The hilt of the sword was carved with fine designs.
6. A round hole was drilled through the thin board.
7. Footprints showed the path he took up the beach.
8. She was waiting at my front lawn.
9. A vent near the edge brought in fresh air.
10. Prod the old mule with a crooked stick.

List 58

1. It is a band of steel three inches wide.
2. The pipe ran almost the length of the ditch.
3. It was hidden from sight by a mass of leaves and shrubs.
4. The weight of the package was seen on the high scale.
5. Wake and rise, and step into the green outdoors.
6. The green light in the brown box flickered.
7. The brass tube circled the high wall.
8. The lobes of her ears were pierced to hold rings.
9. Hold the hammer near the end to drive the nail.
10. Next Sunday is the twelfth of the month.

List 59

1. Every word and phrase he speaks is true.
2. He put his last cartridge into the gun and fired.
3. They took their kids from the public school.
4. Drive the screw straight into the wood.
5. Keep the hatch tight and the watch constant.
6. Sever the twine with a quick snip of the knife.
7. Paper will dry out when wet.
8. Slide the catch back and open the desk.
9. Help the weak to preserve their strength.
10. A sullen smile gets few friends.

List 60

1. Stop whistling and watch the boys march.
2. Jerk the cord, and out tumbles the gold.
3. Slide the tray across the glass top.
4. The cloud moved in a stately way and was gone.
5. Light maple makes for a swell room.
6. Set the piece here and say nothing.
7. Dull stories make her laugh.
8. A stiff cord will do to fasten your shoe.
9. Get the trust fund to the bank early.
10. Choose between the high road and the low.

List 61

1. A plea for funds seems to come again.
2. He lent his coat to the tall gaunt stranger.
3. There is a strong chance it will happen once more.
4. The duke left the park in a silver coach.
5. Greet the new guests and leave quickly.
6. When the frost has come it is time for turkey.
7. Sweet words work better than fierce.
8. A thin stripe runs down the middle.
9. A six comes up more often than a ten.
10. Lush fern grow on the lofty rocks.

List 62

1. The ram scared the school children off.
2. The team with the best timing looks good.
3. The farmer swapped his horse for a brown ox.
4. Sit on the perch and tell the others what to do.
5. A steep trail is painful for our feet.
6. The early phase of life moves fast.
7. Green moss grows on the northern side.
8. Tea in thin china has a sweet taste.
9. Pitch the straw through the door of the stable.
10. The latch on the back gate needed a nail.

List 63

1. The goose was brought straight from the old market.
2. The sink is the thing in which we pile dishes.
3. A whiff of it will cure the most stubborn cold.
4. The facts don't always show who is right.
5. She flaps her cape as she parades the street.
6. The loss of the cruiser was a blow to the fleet.
7. Loop the braid to the left and then over.
8. Plead with the lawyer to drop the lost cause.

9. Calves thrive on tender spring grass.
10. Post no bills on this office wall.

List 64

1. Tear a thin sheet from the yellow pad.
2. A cruise in warm waters in a sleek yacht is fun.
3. A streak of color ran down the left edge.
4. It was done before the boy could see it.
5. Crouch before you jump or miss the mark.
6. Pack the kits and don't forget the salt.
7. The square peg will settle in the round hole.
8. Fine soap saves tender skin.
9. Poached eggs and tea must suffice.
10. Bad nerves are jangled by a door slam.

List 65

1. Ship maps are different from those for planes.
2. Dimes showered down from all sides.
3. They sang the same tunes at each party.
4. The sky in the west is tinged with orange red.
5. The pods of peas ferment in bare fields.
6. The horse balked and threw the tall rider.
7. The hitch between the horse and cart broke.
8. Pile the coal high in the shed corner.
9. A gold vase is both rare and costly.
10. The knife was hung inside its bright sheath.

List 66

1. The rarest spice comes from the far East.
2. The roof should be tilted at a sharp slant.
3. A smatter of French is worse than none.
4. The mule trod the treadmill day and night.
5. The aim of the contest is to raise a great fund.
6. To send it now in large amounts is bad.
7. There is a fine hard tang in salty air.
8. Cod is the main business of the north shore.
9. The slab was hewn from heavy blocks of slate.
10. Dunk the stale biscuits into strong drink.

List 67

1. Hang tinsel from both branches.
2. Cap the jar with a tight brass cover.
3. The poor boy missed the boat again.
4. Be sure to set the lamp firmly in the hole.
5. Pick a card and slip it under the pack.
6. A round mat will cover the dull spot.
7. The first part of the plan needs changing.
8. A good book informs of what we ought to know.
9. The mail comes in three batches per day.
10. You cannot brew tea in a cold pot.

List 68

1. Dots of light betrayed the black cat.
2. Put the chart on the mantel and tack it down.
3. The night shift men rate extra pay.
4. The red paper brightened the dim stage.
5. See the player scoot to third base.
6. Slide the bill between the two leaves.
7. Many hands help get the job done.
8. We don't like to admit our small faults.
9. No doubt about the way the wind blows.
10. Dig deep in the earth for pirate's gold.

List 69

1. The steady drip is worse than a drenching rain.
2. A flat pack takes less luggage space.
3. Green ice frosted the punch bowl.
4. A stuffed chair slipped from the moving van.
5. The stitch will serve but needs to be shortened.
6. A thin book fits in the side pocket.
7. The gloss on top made it unfit to read.
8. The hail pattered on the burnt brown grass.
9. Seven seals were stamped on great sheets.
10. Our troops are set to strike heavy blows

List 70

1. The store was jammed before the sale could start.
2. It was a bad error on the part of the new judge.
3. One step more and the board will collapse.
4. Take the match and strike it against your shoe.
5. The pot boiled, but the contents failed to jell.
6. The baby puts his right foot in his mouth.
7. The bombs left most of the town in ruins.
8. Stop and stare at the hard working man.
9. The streets are narrow and full of sharp turns.
10. The pup jerked the leash as he saw a feline shape.

List 71

1. Open your book to the first page.
2. Fish evade the net and swim off.
3. Dip the pail once and let it settle.
4. Will you please answer that phone.
5. The big red apple fell to the ground.
6. The curtain rose and the show was on.
7. The young prince became heir to the throne.
8. He sent the boy on a short errand.
9. Leave now and you will arrive on time.
10. The corner store was robbed last night.

List 72

1. A gold ring will please most any girl.
2. The long journey home took a year.
3. She saw a cat in the neighbor's house.
4. A pink shell was found on the sandy beach.
5. Small children came to see him.
6. The grass and bushes were wet with dew.
7. The blind man counted his old coins.
8. A severe storm tore down the barn.
9. She called his name many times.
10. When you hear the bell, come quickly.