

IEEE
No. 297
JUNE 1969

IEEE Recommended Practice for
SPEECH QUALITY MEASUREMENTS

IEEE No. 297



PUBLISHED BY
THE INSTITUTE OF ELECTRICAL AND ELECTRONICS ENGINEERS, Inc.
345 EAST 47 STREET, NEW YORK, N. Y. 10017



Documents

Document Sharing Hub

Authorized licensed use limited to: Tsinghua University. Downloaded on February 25, 2026 at 09:17:14 UTC from IEEE Xplore. Restrictions apply.

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. Rothaus, *Chairman*
W. D. Chapman, *Vice-Chairman*

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

Consultants

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

© Copyright 1969 by The Institute of Electrical and Electronics Engineers, Inc.

This publication may be reproduced, without change, in part or in its entirety,
provided that notice of its copyright by the IEEE is included

CONTENTS

1	Introduction	3
1.1	Scope	3
1.2	Review of Past Work	3
1.3	Background of Recommendations	3
1.4	Basic Considerations in Preparing the Practices	4
2	Definitions	4
3	Isopreference Method	4
3.1	Synopsis	4
3.2	Test Procedure	5
3.3	Loudness of Speech Signals	5
3.4	Speech Material	5
3.5	Talkers	5
3.6	Listeners	5
3.7	Reference Signals	6
3.8	Analysis of Results	6
4	Relative Preference Method	6
4.1	Synopsis	6
4.2	Test Procedure	6
4.3	Loudness of Speech Signals	7
4.4	Speech Material	7
4.5	Talkers	7
4.6	Listeners	7
4.7	Reference Signals	7
4.8	Analysis of Results	7
5	Category-Judgment Method	7
5.1	Synopsis	7
5.2	Test Procedure	7
5.3	Loudness of Speech Signals	8
5.4	Speech Material	8
5.5	Talkers	8
5.6	Listeners	8
5.7	Reference Signals	8
5.8	Analysis of Results	8
Appendix A	Documentation Requirements	8
Appendix B	Bibliography	9
Appendix C	1965 Revised List of Phonetically Balanced Sentences (Harvard Sentences)	15

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 (Rothausen) 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. [Ferguson]

^{*} 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 [$\text{dB}(A)$].

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.

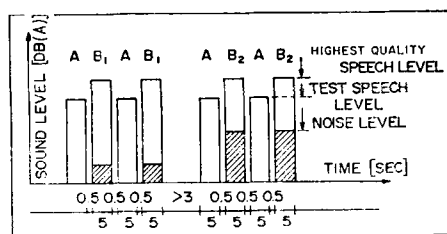


Figure 1. Example of isopreference test pair sequence.

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 $n_o(t)$ should be pink noise weighted by the A curve described in USA Standard S1.4-1961 (A -Weighted Pink Noise)

$$d(t) = n_o(t).$$

- (b) **Multiplicative Random 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.

$$d'(t) = s(t) \cdot n'_o(t).$$

Figure 2 illustrates the comparison between the two reference signals.

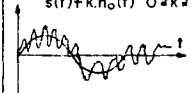
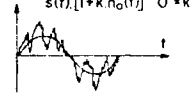
ADDITIVE REFERENCE	MULTIPLICATIVE REFERENCE
	
$s(t) + k \cdot n_o(t)$	$s(t) \cdot [1 + k \cdot n'_o(t)]$
$S/N = 20 \log \frac{S}{k N_o}$	$S/N = 20 \log \frac{S}{k S N'_o}$
$S/N = 20 \log 1/k$ - LEVEL OF NOISE SIGNAL $n_o(t)$ + LEVEL OF SPEECH SIGNAL $s(t)$	$S/N = 20 \log 1/k$ - LEVEL OF NOISE SIGNAL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

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