SECTION 8G: AVAILABILITY, PERFORMANCE OBJECTIVES AND INTERWORKING WITH TERRESTRIAL NETWORKS

REPORT 751 *

METHODS FOR THE SUBJECTIVE ASSESSMENT OF SPEECH QUALITY IN THE MARITIME MOBILE-SATELLITE SERVICE

(Study Programme 17A/8)

(1978)

1. Introduction

The assessment of the performance of a telephone connection involving a maritime mobile satellite system, which may be connected to switched national or international circuits, presents some complications in view of the diversity of conditions and degradations likely to be encountered in a typical connection. At least one of the links will always introduce a long propagation time of the order of 270 ms (mean one-way); echo suppressors and possibly other voice-switched equipment will be necessary. Other forms of degradation such as loss, circuit noise, attenuation/frequency distortion and, possibly, quantizing distortion will be contributed by the various circuit elements that form the connection containing the maritime system. It must also be taken into account that the mean speech and noise levels over the satellite link may be different in the two directions of transmission.

Ideally, the ultimate criterion of performance of such a connection is its performance in service when providing communication between customers drawn from the population of potential users. Under such conditions, all talking, listening and conversation degradations will be present. Such an approach, however, may not be practicable for various reasons, and the communication efficiency of the system must be tested by alternative methods. Laboratory simulation of both-way connections for conversation tests is feasible, provided the connections are truly representative and the tests are carried out in a methodical manner. Listening-only tests have a place in the assessment process, but it would be imprudent to attempt to predict entirely on the basis of results of listening-only tests the performance of maritime satellite systems under conversational conditions.

The purpose of this Report is to summarize various methods that could be used for listening-only and for conversational tests. The Report suggests which methods are most appropriate to study the effects of various types of degradation that may be encountered in establishing a maritime mobile-satellite service and which are most appropriate to establish the various performance requirements.

2. Transmission objectives

A major aim in establishing a maritime mobile-satellite service is to enable connections to be made between ships and terrestrial subscribers that are as good as, or nearly as good as, normal terrestrial connections. Several decisions have to be made in developing the service and the following is a list of areas where assessment methods play a part in making the decisions:

- determination of the most suitable speech modulation system;
- for modulation systems that produce speech correlated noise, the determination of the level of white noise that is subjectively equivalent to the speech correlated noise (see Report 750, Kyoto, 1978);
- the determination of the minimum performance standards for distress communication;
- the determination of the effect of delay, echo, and voice activated carrier switching;
- the determination of the required performance of the overall system;
- the in-service investigation whether the service meets the required level of performance and where particular areas of difficulty may exist.

3. Listening-only tests

There are many types of listening-only tests, but they can be divided into two broad categories: those that measure the intelligibility of speech received over the channel under test and those that measure the overall quality of the received speech.

Articulation tests are used for the former purpose. These are relatively simple to perform, although requiring trained subjects, and the analysis is straightforward.

Opinion scales and isopreference tests are amongst those used for the latter. In general, these methods require untrained subjects.

* The Director, CCIR, is requested to bring this Report to the attention of the CCITT.

3.1 *Articulation tests*

3.1.1 *General principles*

The purpose of articulation tests is to obtain a measure of the intelligibility of the speech transmitted over the channel under test.

Generally, the tests are conducted by reading standardized speech material over the channel under test. The percentage of speech sounds received correctly by a team of listeners is taken as a measure of the intelligibility of the channel. This may be expressed as word, logatom, sound, consonant, or vowel articulation.

A great variety of practical methods exists, differing mainly on the speech material used and the general experimental conditions, and are well suited for the determination of the minimum performance standards for distress communications.

The task presented to the subjects is simple and tests of this kind are relatively easy to perform. The analysis of the results is also straightforward. The interpretation of results will depend upon the specific experimental method used.

3.1.2 *Logatom articulation method*

Logatoms are meaningless syllables consisting of three components as follows:

consonant(s)-vowel-consonant(s).

Each of the sounds constituting the logatom is chosen at random from a list of phonemes which are representative of different languages.

The method of forming logatoms is described in detail in CCITT Recommendation P.45, Red Book, Volume V, page 74.

The method consists of having at the receive end of the channel under test a trained subject listen to a logatom list. At the other end, the logatoms (without a carrier phrase) are enunciated either directly by speakers or by means of a previously recorded magnetic tape. (A complete test may comprise 500 logatoms, pronounced by several speakers.) The subject writes down the sounds he hears, thus making a list of "received" logatoms, which are then compared with those of the list read out.

The percentage of correctly received logatoms is called "logatom articulation".

3.1.3 *Sound articulation method*

A complete test comprises 100 syllables (each consisting of one or more sounds), which are chosen at random from a given language (e.g. Japanese).

A trained speaker pronounces these syllables at a constant speaking level and a subject writes down the syllables he hears. The percentage of correctly received sounds is called "sound articulation".

3.1.4 *CCITT articulation method*

Attention is also drawn to the method specified in (CCITT Recommendation P.45, Red Book, Volume V, pages 69-114).

This method was originally designed for the measurement of Articulation Reference Equivalents (AEN), but may be used for other purposes.

3.1.5 *Phonetically balanced word method (PB)*

Attention is further drawn to voice intelligibility measurements using the PB-50 word test of the American National Standard S3.2-1960: Method for the Measurement of Monosyllabic Word Intelligibility.

Experiments have shown that the PB-50 word tests can be modified from the standard procedures without significantly affecting measurement reliability and accuracy [Milner, 1973; Milner and Golab, 1975]. In order to conduct a comprehensive intelligibility test programme within a limited time in both maritime and aeronautical experiments over ATS-6, the reading rate was increased to one word per 2.5 s, the carrier phrase was dropped, the tests were limited to 400 words per test, and 68 scrambled orders of the word lists were created. These modifications in the American National Standard PB word test procedures did not comprise listener performance. Listener memorization was precluded and a wide range of parametric conditions were accommodated while maximizing the efficiency of use of satellite and flight test time (see Report 599).

3.2 *Opinion scales*

The opinion-scale methods seek to characterize speech transmission channels by using a limited set of verbal descriptions (opinion scale) of the channel quality. After listening to speech transmitted through the channel under test, untrained subjects are asked to indicate which quality description is the most appropriate for this channel.

The results from tests of this kind may be expressed as the percentage of responses in each of the various categories adopted.

In addition, certain categories may be grouped together – for example, if a seven-point scale is used, adding together groups 4, 5 and 6 will give the percentage of satisfied users and adding together groups 3, 4, 5 and 6 will give the percentage of fairly satisfied users. If a five-point scale is used, the corresponding groups are A and B and A, B and C.

A mean opinion score may also be computed if numerical values are allocated to the categories.

In general, opinion tests are easy to administer and the analysis of the results is straightforward. In this method, however, the task of the listeners is difficult in the sense that they must express their opinion by choosing one of a set of quality descriptions. The results are dependent upon the scale description which must adequately reflect the quality range covered in the test. When adequate formulations are used, the interpretation of the results is then straightforward.

Furthermore the results can be used to calculate by interpolation the level of white Gaussian noise that would be subjectively equivalent to the level of speech correlated noise. The subjects would listen to channels containing speech correlated noise and to channels containing white Gaussian noise (reference channels). The noise levels introduced into the reference channels should cover a range of Mean Opinion Score (MOS) values larger than the channels containing speech correlated noise.

Two typical opinion scales are given below:

3.2.1 7-point scale

Score	*Quality description*
6	Ideal circuit.
5	Excellent circuit. Possible to relax completely during call. Very agreeable.
4	Good circuit. Necessary to pay attention, but not necessary to make a special effort. Agreeable circuit.
3	Fair circuit. A moderate, but not too great, effort is necessary. Not a very agreeable circuit.
2	Poor circuit. Listening is possible, but somewhat difficult. Listening disagreeable.
1	Bad circuit. Can be used only with great difficulty. Listening very disagreeable.
0	Very bad circuit. Practically unusable.

3.2.2 5-point (listening effort) scale (Reference: Annex 5 to Question 1/XII, Part C, CCITT Green Book, Volume V)

A	Complete relaxation possible: no effort required.
B	Attention necessary: no appreciable effort required.
C	Moderate effort required.
D	Considerable effort required.
E	No meaning understood, even with considerable effort.

One of the problems of the above scales is that they are not easily translated into other languages, causing difficulties to subjects in their understanding of the difference between scale points. A possible consequence is that tests conducted in different languages but otherwise with the same parameters may not be comparable. To overcome this language problem a simple scale, easy to translate into different languages – such as very good, good, fair, poor, very poor – might be used. It is, however, very important in using such a scale that the correct question is presented to the subject; otherwise it may not be clear whether the speech or the channel is being judged, the results being as a consequence inconsistent.

Opinion scale methods are suitable for assessing the effect of most types of degradation, especially when the transmission aim is for good quality rather than mere intelligibility. They are not able to assess the effects of delay and echos.

3.3 *Comparison methods for determination of equivalent noise*

3.3.1 *General*

The use of equivalent noise is motivated by the need for assessing the effect on overall transmission quality of circuits introducing speech correlated types of distortion. For this purpose, the equivalent noise of the "unknown" system can be determined for this system alone without any additional degradation introduced (Report 750, Kyoto, 1978). The overall quality of various connections may then be calculated using ordinary transmission performance calculations. (Reference: CCITT Orange Book, annexes to Question 7/XII).

Being concerned with basically ordinary telephone connections, the subjective assessment methods used to determine equivalent noise should be based on judgement of overall quality. Also, if consistent results are to be obtained, this judgement process should obey a transitivity law, i.e., if A = B and B = C then C = A.

The tasks presented to the subject in the methods described below are somewhat different, asking for preference, similarity and rank order respectively. It is believed however that the methods are all based on overall quality judgements.

3.3.2 *Isopreference method*

In the isopreference method, the channel under test is assessed by comparison with a reference channel which carries varying amounts of noise. For each pair of reference and test channels, the subject is asked to indicate which of the two speech samples he or she would prefer to listen to. The isopreference (noise) level is then the value where 50% of the listeners prefer one channel and 50% the other channel. Since, in practice, the isopreference level is unlikely to be obtained, an interpolation procedure is used to arrive at the isopreference level and tests have shown that the transitivity requirement is met by this method.

3.3.3 *Adjustment method*

This method is also called the method of average error. In this method, the subjects control the noise power in the reference channel. They are asked to adjust the reference channel by selecting the amount of noise required so that this channel and the channel under test are equivalent with respect to overall quality. In this method, the equivalent noise level is obtained directly by the setting of the noise level.

Tests have shown that the transitivity requirement is also met by this method.

3.3.4 *Rank order method*

The method consists of rank ordering a mixture of known and unknown processing techniques in order of performance and is known in statistical experimentation as the Youden Squares (YS) design. The term "known" transmission technique refers to one where the subjective quality is known in terms of an analogue signal with Gaussian noise. An "unknown" transmission processing technique is any technique for which the quality is to be determined.

The mixture of known and unknown transmission techniques is composed such that, by careful preselection, the quality range of the known techniques embraces those of the unknown techniques. Then the equivalent noise values for the unknown techniques are obtained by interpolation.

The Youden Squares design allows the subjective rank ordering of many stimuli such as speech samples processed by different techniques by presenting only a small fraction at one time to test subjects. This fraction is called a block. For example, one can rank order thirteen stimuli by presenting only blocks of four at a time and the design is called a 4 × 13 YS. The stimuli, as used here and designated S_i, are recorded speech samples processed by different techniques under specified conditions.

The Youden Square assures that the division into blocks for ranking maintains a certain balance, in particular, that each stimulus S_i is listened to an equal number of times in the experiment and each possible pair of stimuli occurs equally often in an equal number of blocks.

Because it is generally too difficult for most people to compare more than 4 samples at a time, the 13 × 4 Youden Square design is most often used. In this design 6 "unknown" systems are mixed with 7 systems of known analogue signal to noise ratio.

For each block the subject is asked to rank the 4 stimuli by assigning numbers from 1 (for the best) to 4 (for the worst) to the speech samples in accordance with his own preference.

The coefficient concordance test should be performed to ensure a high probability that the observed rank order exists. The analysis consists of computing the mean rank for each system. For the conversion from rank order to equivalent noise, a regression curve mean rank versus S/N, for the analogue treatments is determined. The mean ranks of the unknown treatments are transferred to the S/N domain with this regression curve.

Example of the use of this method may be found in Report 752, Annex II.

4. Conversational tests

4.1 *General principles*

Conversational tests are primarily concerned with the overall quality of telephone connections. The effects arising in a conversation are taken into account in as realistic a way as possible.

The tests are performed by having pairs of untrained subjects conduct conversations over the connection under test. Immediately after having completed the conversation task without any disturbance, the subjects are examined on their opinion of the circuit. The method might therefore be used on either real or simulated connections.

The questionnaires used may vary, but generally an opinion scale is included. The subjects are asked to indicate which quality description is in accordance with his opinion of the circuit quality. In addition, subjects may be asked whether they experienced difficulty or not. If they did, further questions may be asked to probe the nature of difficulty, but without suggesting possible types of difficulty. The subjects are interrogated independently and they are not allowed to discuss their opinions (Reference: Appendix 2 to Annex 2 to Question 4/XII, CCITT Green Book, Volume V).

The conduct of conversation tests is very time-consuming, since the subjects are allowed to perform their task at their own speed. The generation of a realistic conversation between them is one of the main problems posed in experiments of this kind.

The method of analysis of the results is similar to that used in opinion scale testing. The mean opinion score and the percentage score in each category and in groups of categories are computed. In addition, the percentage of subjects who experienced either difficulty or any of the special kinds of distortions listed in the questionnaire is also given.

4.2 *Practical methods*

A particular method of carrying out conversational tests is described in Annex 2 to Question 4/XII, CCITT Green Book, Volume V. A five-point rating scale is generally used and contains, for example, the quality descriptions "Excellent", "Good", "Fair", "Poor" and "Bad". The mean Opinion Score (MOS) is computed as follows:

$$\text{MOS} = (4E + 3G + 2F + P)/N$$

where,

E: number of calls rated excellent,
G: number of calls rated good,
F: number of calls rated fair,
P: number of calls rated poor,
N: total number of calls.

(The number of calls rated "Bad" is given a zero coefficient and affects the MOS by being included in N.)

5. Experimental set-up

Figure 1 shows a laboratory simulation of a typical subscriber-to-subscriber connection which includes degradations that might be experienced on a maritime satellite call. When listening-only tests are performed, some of these degradations cannot be evaluated and equipment such as propagation delay equipment and echo suppressors can be removed from the simulated connection.

For practical reasons, it may be desirable to carry out listening-only tests over the satellite link taken in isolation. However, care must be taker. in drawing conclusions from such tests, since, ideally, these tests should take into account factors such as loss, distortion and circuit noise which may influence the speech signal applied to the satellite system modulators, both at the ground station and at the ship terminal; also the presence and characteristics of any ambient noise at the listening locations must be considered. Tests should be conducted, separately, in both directions of transmission.

The noise arising in the terrestrial extension and which is to be injected into the simulated connection needs to be determined and will have to take into account the fact that terrestrial extensions will vary considerably in length.

For comparison purposes during tests, a reference connection is needed; this should be similar to the connection under test, but the maritime link should be replaced by a fixed service satellite link.

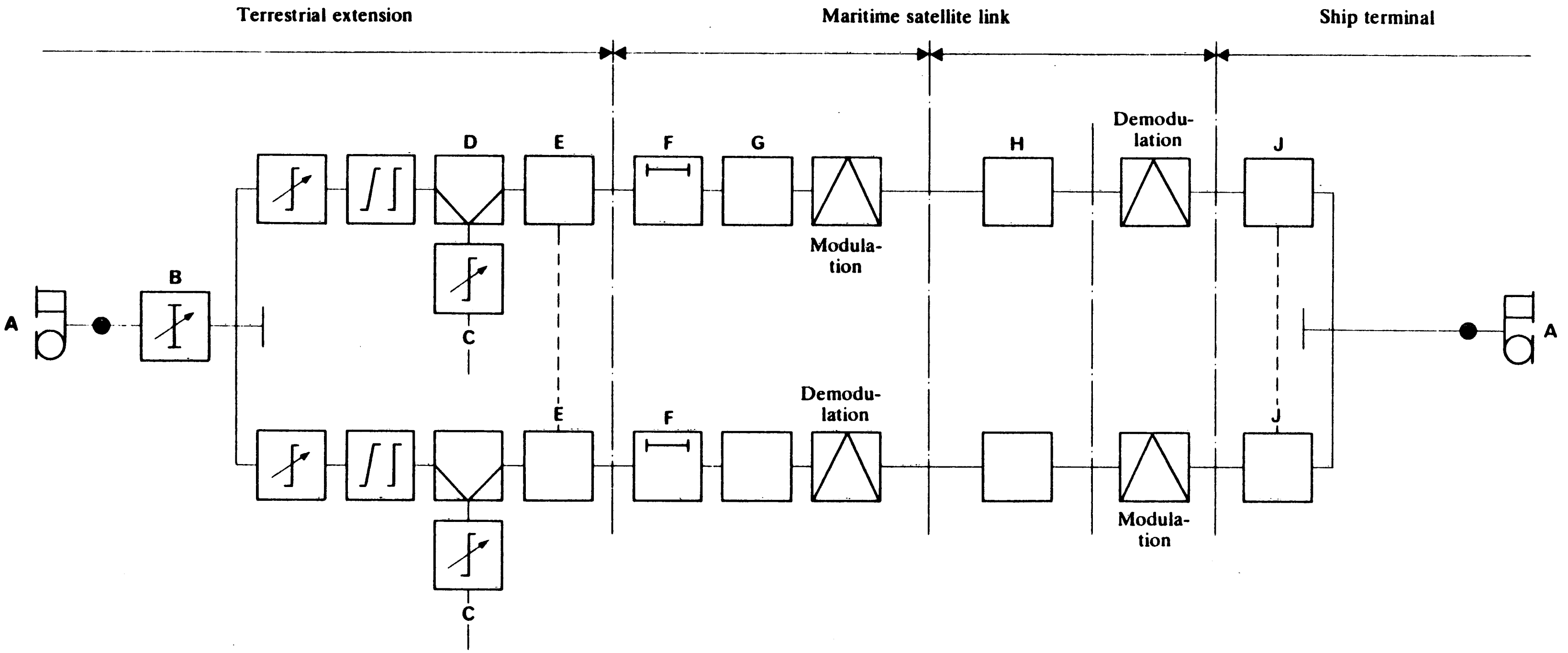

FIGURE 1 – *Laboratory simulation of typical subscriber to subscriber connection*

A: ambient noise	D: noise inject pad	G: voice activator
B: artificial line of variable length	E: half-echo suppressor	H: satellite simulator
C: noise	F: delay	J : half-echo suppressor

Note. — With four-wire telephone system on board ship the four-to-two-wire terminating set will not be required. See also Recommendation 550 as regards the half-echo suppressor at the ship end of the connection.

6. Conclusions

Subscriber-to-subscriber conversation tests are the only satisfactory method of evaluating the performance of a telephone connection involving a system in the maritime mobile-satellite service. Such tests are the only ones that can take into account the degradations introduced by propagation delay through the satellite. Unfortunately, conversation tests are difficult and time consuming to carry out; listening-only tests are simpler to carry out and have a place in certain aspects of evaluation, but their limitations must be borne in mind when making decisions based on results from such tests.

Especially, it should be recognized that articulation tests are not very suitable for the assessment of ordinary telephone connections. These methods, however, may be applied for special purposes, e.g. determination of the effect of voice activated carrier switching (trunkation of initial syllables) and performance standards for distress communication.

For the determination of the most suitable speech modulation method and assessment of equivalent noise levels of the satellite link, listening-only methods assessing overall quality should be used.

The following recommendations are submitted:

- listening-only tests may be used in preliminary tests for the comparison of modulation methods; ideally they should be carried out over a complete connection;
- listening-only tests assessing the equivalent noise of the modulation methods may be carried out to enable transmission performance calculations to be performed on a variety of representative complete connections;
- conversational laboratory tests over a simulated connection should be carried out, preferably under the auspices of an international forum;
- eventually, in-service conversational tests should be performed.

REFERENCES

MILNER, P. [June, 1973] Advantages of experienced listeners in intelligibility testing. *IEEE Trans. Audio and Electroacoustics*, Vol. AU-21, 161-165.

MILNER, P. and GOLAB, J. [April, 1975] Intelligibility of voice transmission through a satellite relay system. *J. Acous. Soc. of Amer.*, Vol. 57-Supp. 1, 23.

BIBLIOGRAPHY

KENDALL, M. G. [1974] *Rank correlation methods*, Chapters 9 and 10. Charles Griffin & Co., Ltd.