International Telecommunication Union

ITU-T

TELECOMMUNICATION
STANDARDIZATION SECTOR
OF ITU

P.1312

(02/2016)

SERIES P: TERMINALS AND SUBJECTIVE AND OBJECTIVE ASSESSMENT METHODS

Telemeeting assessment

Method for the measurement of the communication effectiveness of multiparty telemeetings using task performance

Recommendation ITU-T P.1312

ITU-T P-SERIES RECOMMENDATIONS

TERMINALS AND SUBJECTIVE AND OBJECTIVE ASSESSMENT METHODS

Vocabulary and effects of transmission parameters on customer opinion of transmission quality	Series	P.10
Voice terminal characteristics	Series	P.30
		P.300
Reference systems	Series	P.40
Objective measuring apparatus	Series	P.50
		P.500
Objective electro-acoustical measurements	Series	P.60
Measurements related to speech loudness	Series	P.70
Methods for objective and subjective assessment of speech quality	Series	P.80
		P.800
Audiovisual quality in multimedia services	Series	P.900
Transmission performance and QoS aspects of IP end-points	Series	P.1000
Communications involving vehicles	Series	P.1100
Models and tools for quality assessment of streamed media	Series	P.1200
Telemeeting assessment	**Series**	**P.1300**
Statistical analysis, evaluation and reporting guidelines of quality measurements	Series	P.1400
Methods for objective and subjective assessment of quality of services other than voice services	Series	P.1500

For further details, please refer to the list of ITU-T Recommendations.

Recommendation ITU-T P.1312

Method for the measurement of the communication effectiveness of multiparty telemeetings using task performance

Summary

Recommendation ITU-T P.1312 describes a subjective test method for quantifying the effectiveness of telemeeting systems in conveying information in multiparty conversation scenarios. This method utilizes a predefined set of tasks designed to provoke rapid turn-taking and concurrent talking among participants. The goal of these tasks is to stress conferencing systems in order to clearly and measurably demonstrate their limits.

The method measures the rate at which multiple participants exchange information to assess the effectiveness of communication systems compared to face-to-face communication. This constitutes an objectively-measured performance metric providing stable and repeatable results across different laboratories and that is highly discriminative across various types of system properties.

This method may be used to evaluate individual subcomponents of a telecommunication system or the system as a whole. The proposed method is especially sensitive to the ability to transmit multiple concurrent voices and maintain their independence. The test has demonstrated sensitivity to system parameters including duplex capability, spatialization, bandwidth and delay. As such, it is suitable for the assessment of multiparty telemeeting systems (see Recommendation ITU-T P.1301 for general guidelines and additional methods).

History

Edition	Recommendation	Approval	Study Group	Unique ID*
1.0	ITU-T P.1312	2016-02-29	12	11.1002/1000/12751

Keywords

Double-talk, face-to-face reference, interactive test method, system effectiveness, task performance measurement, telemeeting.

* To access the Recommendation, type the URL http://handle.itu.int/ in the address field of your web browser, followed by the Recommendation's unique ID. For example, http://handle.itu.int/11.1002/1000/11830-en.

FOREWORD

The International Telecommunication Union (ITU) is the United Nations specialized agency in the field of telecommunications, information and communication technologies (ICTs). The ITU Telecommunication Standardization Sector (ITU-T) is a permanent organ of ITU. ITU-T is responsible for studying technical, operating and tariff questions and issuing Recommendations on them with a view to standardizing telecommunications on a worldwide basis.

The World Telecommunication Standardization Assembly (WTSA), which meets every four years, establishes the topics for study by the ITU-T study groups which, in turn, produce Recommendations on these topics.

The approval of ITU-T Recommendations is covered by the procedure laid down in WTSA Resolution 1.

In some areas of information technology which fall within ITU-T's purview, the necessary standards are prepared on a collaborative basis with ISO and IEC.

NOTE

In this Recommendation, the expression "Administration" is used for conciseness to indicate both a telecommunication administration and a recognized operating agency.

Compliance with this Recommendation is voluntary. However, the Recommendation may contain certain mandatory provisions (to ensure, e.g., interoperability or applicability) and compliance with the Recommendation is achieved when all of these mandatory provisions are met. The words "shall" or some other obligatory language such as "must" and the negative equivalents are used to express requirements. The use of such words does not suggest that compliance with the Recommendation is required of any party.

INTELLECTUAL PROPERTY RIGHTS

ITU draws attention to the possibility that the practice or implementation of this Recommendation may involve the use of a claimed Intellectual Property Right. ITU takes no position concerning the evidence, validity or applicability of claimed Intellectual Property Rights, whether asserted by ITU members or others outside of the Recommendation development process.

As of the date of approval of this Recommendation, ITU had not received notice of intellectual property, protected by patents, which may be required to implement this Recommendation. However, implementers are cautioned that this may not represent the latest information and are therefore strongly urged to consult the TSB patent database at http://www.itu.int/ITU-T/ipr/.

Table of Contents

Recommendation ITU-T P.1312

Method for the measurement of the communication effectiveness of multiparty telemeetings using task performance

1 Scope

This Recommendation describes a method for quantifying the amount of information that multiple interlocutors correctly exchange over a system. The measure serves to assess the effectiveness of a voice communication system by comparing the amount of correct information exchanged over the system relative to the amount of correct information these interlocutors achieve in face-to-face communication. This metric qualitatively evaluates a multiparty conferencing system's ability to transmit multiple voices and maintain their independence. The method has demonstrated sensitivity to system parameters including duplex capability, spatialization, bandwidth and delay.

The method described herein applies only to voice communication; audiovisual communications is not covered by the scope of this Recommendation. The method does not include the measurement of technical properties of the systems.

2 References

The following ITU-T Recommendations and other references contain provisions which, through reference in this text, constitute provisions of this Recommendation. At the time of publication, the editions indicated were valid. All Recommendations and other references are subject to revision; users of this Recommendation are therefore encouraged to investigate the possibility of applying the most recent edition of the Recommendations and other references listed below. A list of the currently valid ITU-T Recommendations is regularly published. The reference to a document within this Recommendation does not give it, as a stand-alone document, the status of a Recommendation.

[ITU-T P.800] Recommendation ITU-T P.800 (1996), *Methods for subjective determination of transmission quality*.

[ITU-T P.805] Recommendation ITU-T P.805 (2007), *Subjective evaluation of conversational quality*.

[ITU-R BS.1116] Recommendation ITU-T BS.1116-3 (2015), *Methods for the subjective assessment of small impairments in audio systems*.

3 Definitions

3.1 Terms defined elsewhere

This Recommendation uses the following terms defined elsewhere:

3.1.1 concurrent talk [b-ITU-T P.1311]: When far-end speech of two or more talkers occurs simultaneously at a given point, typically the near end terminal.

3.1.2 multiparty [b-ITU-T P.1301]: More than two persons. Example: More than two persons are participating in a telemeeting, having a conversation, performing a test task together, etc. The term multiparty does not specify if the persons are distributed across two or more locations. If not explicitly stated differently, multiparty implicates that the persons are at two or more than two locations. When further specification is necessary, additional terms will be used or the number of locations will be explicitly stated.

3.1.3 one per site [b-ITU-T P.1301]: One person per connected location. Example: In a multiparty one-per-site telemeeting, more than two sites are connected with only one person present at each site.

3.2 Terms defined in this Recommendation

This Recommendation defines the following terms:

3.2.1 performance score: The number of correct sequences conveyed by an individual, team or group.

3.2.2 system effectiveness: The ratio, expressed as a percentage, between the performance score achieved on a given system and the performance score obtained in face-to-face communication.

3.2.3 reader: A participant for whom the role is to read out loud a sequence of letters or words to his or her partner(s).

3.2.4 responder: A participant for whom the role is to register the sequence of letters or words his or her partner communicates.

3.2.5 face-to-face meeting: A meeting for which all participants are collocated.

3.2.6 duplex: A communication channel for which signals can flow in both directions.

3.2.7 spatialization: The process of rendering audio sources in a three-dimensional environment.

4 Abbreviations and acronyms

This Recommendation uses the following abbreviations and acronyms:

SuT System-under-Test

5 Conventions

The keywords "is required to" indicate a requirement which must be strictly followed and from which no deviation is permitted if conformance to this Recommendation is to be claimed.

The keywords "is recommended" indicate a requirement which is recommended but which is not absolutely required. Thus this requirement need not be present to claim conformance.

The keywords "is prohibited from" indicate a requirement which must be strictly followed and from which no deviation is permitted if conformance to this Recommendation is to be claimed.

The keywords "can optionally" indicate an optional requirement which is permissible, without implying any sense of being recommended. This term is not intended to imply that the vendor's implementation must provide the option and the feature can be optionally enabled by the network operator/service provider. Rather, it means the vendor may optionally provide the feature and still claim conformance with this Recommendation.

6 Task effectiveness test method

6.1 Purpose

The purpose of the Recommendation is to enable multiparty conversational tests of voice-only communication systems that are sensitive to subtle changes of system parameters, to produce results that are repeatable in different laboratories with predictable test duration.

The test described in this Recommendation is based on the objectively-measured performance of participants when solving tasks over the communication system-under-test (SuT). This is unlike the conversational tests of [ITU-T P.805], which are based on subjective impressions reported by

participants. The method provides an objective performance measure that reduces the mutual influence of test conditions between experiments.

Another important aspect of this test is that it relates the performance achieved using the SuT to the performance achieved through (in-person) face-to-face conversation. Using in-person meetings as the reference not only minimizes individual biases, but also leads to a metric that is directly relevant to a predominant use case of telecommunication systems, which is to substitute for in-person meetings. By measuring the rate of information exchange attained when using the SuT and relating it to the information exchange rate attained through in-person meetings, a measure of relative communication effectiveness is obtained that is easily interpretable and directly relevant to the systems' intended use.

The tasks used in the test are highly structured and designed to be challenging, which serves two purposes.

Firstly, it emphasizes the aspects of a communication system that can only be assessed in conversational tests (as opposed to listening-only tests), thereby maximizing the information gained from the investment in a test. By encouraging concurrent talking, the systems' duplex performance and ability to transmit multiple voices has a stronger effect on results than it would in less-demanding conversational tasks. Furthermore, by emphasizing rapid turn-taking, this method is sensitive to system delay.

Secondly, the structured tests generate highly reproducible results, which necessarily should be the objective of any standardized test. The tests place low demands on social or verbal skills and are largely unaffected by listener training and experience.

It is obvious that tasks specified in this Recommendation do not resemble a typical conversation. However, it covers important aspects of conversation such as: single talk, concurrent talk, turn taking, listening and understanding, back channels. Therefore, it is expected that systems leading to superior performance in the tasks of this Recommendation also exhibit superior performance in real-world meetings, although the magnitude of the difference may be different.

6.2 Test principle

This subjective test paradigm involves the participation of four or more subjects grouped into teams to complete tasks concurrently and on the same communication link. Depending on the conditions under test, subjects may be seated in the same or separate rooms to perform a series of tasks. Acoustic noise environments may be simulated in one or more booths.

The task consists of one member of each team, the "reader", communicating a sequence of letters or words to their partner(s), the "responder(s)". The objective is to convey the highest number of sequences within a fixed amount of time. Each team progresses at its own pace during this period so the number of iterations achieved by each team will be different. The teams are free to use any strategy they wish and can freely interact with the group to perform their tasks. An example of observed strategies is provided in Appendix II.

During a test, the designated readers pronounce a sequence of six target letters or words. Their partners enter the sequence they hear via a simple graphical interface by selecting each letter or word from a set of possible responses composed of the target and foil alternatives. The task for a given team continues with a new sequence if the submitted response is correct. If incorrect, the responder is visually alerted and prompted to submit a corrected sequence. The reader is not provided with the next test item and the test will not progress until the responder has entered the correct response.

In the letter task, the targets are randomly drawn from the alphabet and letters can be repeated within a sequence. The responders select their responses from the full alphabet.

The sequences of target words are randomly drawn from a large pool of words. Each of the targets and associated sets of five foil alternatives drawn from phonological neighbour words are presented to the responders as possible answers. The pool of words should be large enough so that target words are not repeated within an experiment. A recommendation on how to construct databases for different languages can be found in Annex A.

Each task within a trial is time-limited by an arbitrary length of 60 s. The task starts and ends at the same time for both teams. The role (reader/responders) switches once the time has elapsed and the participants perform both letter and word tasks as reader and responder on every system under test. With a group of four participants (i.e., two teams), a total of four tasks is performed per session on a given system, resulting in sessions of four minutes. Each group completes several sessions on each of the systems.

An example of subject instructions and captures of participant screens is provided in Appendix I.

6.3 Reference condition and test scenarios

The test method allows for the evaluation of different scenarios. The noise-free face-to-face scenario is mandatory and should be included in all experiments as it serves as a reference condition in this method. Members of the same team should be facing each other at a distance d of about 1.3 metres. Teams should be seated side by side at about one metre from each other as depicted in Figure 1.

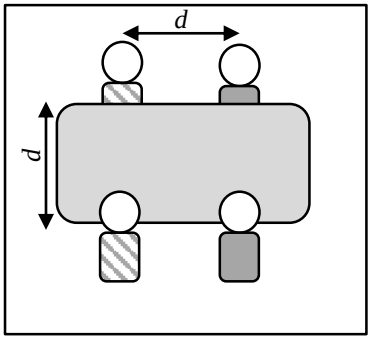


Figure 1 – Reference face-to-face scenario

6.3.1 Example of test scenarios

The allocation of talker and responder roles as well as the position of the subjects are very important and should be documented in the test design and report. When evaluating the performance of hands-free devices, results can vary depending on the location of the readers and responders as different signal processing behaviors are invoked. Devices may behave differently when both readers are around the endpoint device, or when one of them is at the far-end. When both cases are evaluated in an experiment, it is recommended to consider these cases as separate conditions.

6.3.1.1 Four individual participants (one-per-site)

Figure 2 shows a test scenario with four individual participants (one-per-site).

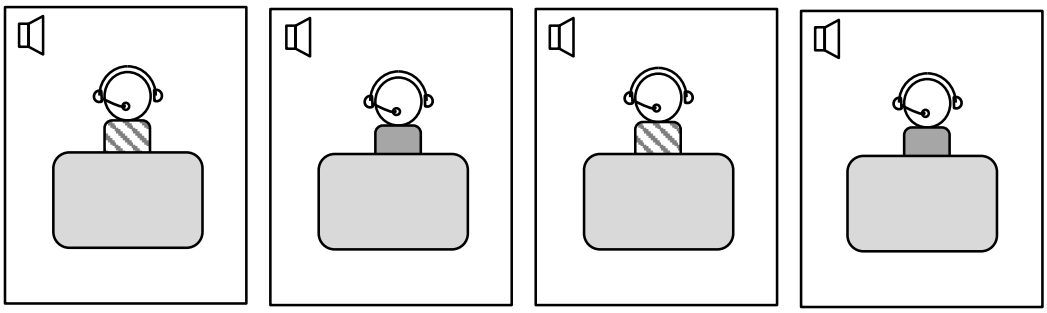


Figure 2 – Four individual participants

Each participant sits in a separate room joining the conference from the individual endpoints.

Background noise can be played in one or more individual rooms to simulate specific conditions.

6.3.1.2 Two participants sit around a hands-free device, two individual participants

Figure 3 shows a test scenario with two participants sitting around a hands-free device, two individual participants.

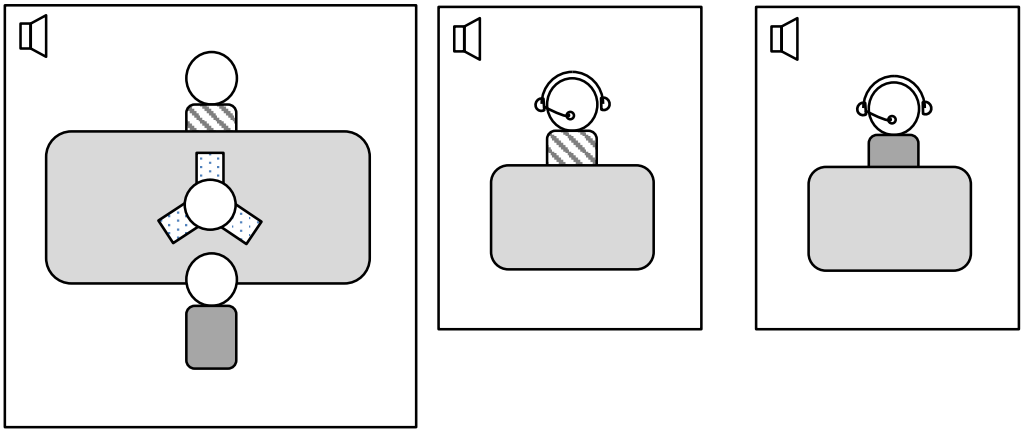


Figure 3 – Two participants around a hands-free device, two individual participants

One member of each team sits around a conference device while the other two sit in separate rooms joining the conference from the individual endpoints.

6.3.2 Two hands-free devices and two participants around each device

Figure 4 shows a test scenario with two hands-free devices and two participants around each device.

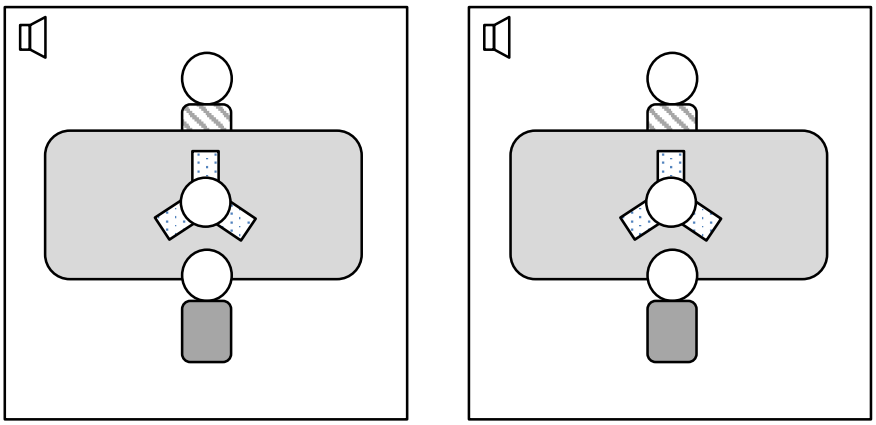


Figure 4 – Two hands-free devices and two participants around each device

The teams are split into two rooms. One member of each team sits around a hands-free device.

6.4 Test facilities and equipment

The test should be conducted in a realistic communication environment and all processes in the communication link are required to be real time.

Switching between conditions that involve different coders, network parameters and/or hands-free devices must be transparent to the subjects. This may require specialized instrumentation and procedures. The same applies when switching between hands-free devices. Devices should ideally be hidden behind an opaque screen that does not affect its acoustic performance.

Communication link asymmetry between participants is typical for many live communication scenarios. An asymmetric scenario may be defined by different acoustic noise environments, different transmission conditions, or different end-points.

Depending on the scenario, subjects may sit in the same room for face-to-face communication, in two or more sound-proof rooms when some participants sit around a device, or in individual rooms when participants use individual end-points. Acoustic noise environments can be simulated and can be the same or different in each room. Various environments that might be used for such an experiment include: a quiet room, an office, a vehicle, a railway station, a train, a cafeteria. A quiet room might be simulated by the introduction of a suitable level of Hoth noise for example. Room reverberation may also be considered as an experimental variable.

A description of sound-proof rooms can be found in [ITU-T P.800].

6.5 Test administrators

Test administrators play an important role throughout the execution of the tests. During the preliminary conditions, they ensure that the participants understand the task and the interface. For the remainder of the test, the administrator acts as a moderator and observer.

Conversational tests that evaluate live systems are prone to connection issues and the test administrator should ensure that the properties of the connection link function as expected for the condition under test. Any deviation from these properties may introduce bias in the results and must be documented. The test administrator should also ensure that the participants execute the tasks appropriately, making notes of any anomalies or disruption in the execution of the task. This proper documentation enables the experimenter to account for any disrupted sessions during the analysis.

6.6 Subjects

The definitions of naive, experienced or expert subjects are provided in [ITU-T P.805]. It is expected that the experience of the subjects would not affect the performance measurements provided by this method, nevertheless it is recommended not to mix subjects with different experience levels. This is particularly important if the experimenter also gathers subjective impressions during the test. The required subject experience may depend on the questions and the desired degree of precision in the results. It is recognized that expert subjects provide more critical assessments than naive subjects.

Formal tests should represent a random cross-section of the population unless the design factors of the test require specific demographics, such as gender, age or other socio-economic characteristics.

The test administrator should take special care to balance gender pairing and distribution across all sessions. Unless specified otherwise, the experiment should be conducted with unisex as well as mixed-gender groups. Care should be taken when pairing subjects into teams to achieve the desired gender balance. Voice characteristics should also be considered as the tasks are less challenging with spectrally separated voices.

The number of subjects required for an experiment depends on the desired statistical significance for which a significance threshold of 0.05 is recommended. The required number of subjects depends on how different the test conditions are and the number of participants per group. As an example, all test results from Appendix III are derived from ten groups of four participants.

6.7 Experiment design

The test design should conform to the same rules as listening-only tests, including the need for preliminary conditions, reference conditions and the order of the conditions.

Preliminary conditions are essential for the participants to get acquainted with the SuTs, the tasks and the user interface, which require more practice than traditional listening tests. The experimenter should ensure that an appropriate number of preliminary trials are included at the beginning of each individual test. Experimental trials resulted in a consensus that four trials, including the face-to-face condition, is a minimum number of preliminary trials.

Due to the duration of each individual task, or trial, only a limited number of conditions can be evaluated in each experiment. Additionally, repeating each trial several times ensures consistency and allows for robust task performance measurements. As such, a minimum of three rounds is recommended.

The test structure should be broken into blocks of trials where each block is separated by a break. Table 1 provides an example test design showing the block structure. A block typically contains 4 to 8 conditions which would result in 20 to 40 minutes to complete a block. Because subjects have to change rooms to attend the face-to-face trials, it is practical to always start a block with the face-to-face condition, then to split the group for the remaining conditions. It also serves the purpose of re-exposing the participants to the reference at the beginning of each block. The order of other conditions should be randomized for each block and group of subjects.

The typical trial duration is four minutes as each participant acts one minute as reader and one minute as responder for both letters and words. When estimating the total length of the test, additional inter-trial periods should be factored into account for:

– the initialization of the system settings for the next condition;

– the time needed to collect any subjective opinions if required;

– the interaction between the test administrator and the participants. After each change of condition, especially when switching systems, the test coordinator should communicate with each participant to ensure that the connection was initialized correctly. This step also allows the participants to recognize and locate one another's voice.

Table 1 – Example test design

Block	Time (min)	Description
Preliminaries	10	Instructions
	4	Trial – Face-to-face
	2..5	inter-trial period
	5	Split into rooms
	4	Trial – Condition $x \in [1..N]$
	2..5	inter-trial period
	4	Trial – Condition $x \in [1..N]$
	2..5	inter-trial period
	4	Trial – Condition $x \in [1..N]$
	2..5	inter-trial period
Break	5	
Block 1	4	Trial – Face-to-face
	2..5	inter-trial period
N times	4	Trial – Condition $x \in [1..N]$
	2..5	inter-trial period
Break	5	
Break	5	
Block M	4	Trial – Face-to-face
	2..5	inter-trial period
N times	4	Trial – Condition $x \in [1..N]$
	2..5	inter-trial period

6.8 Task performance measurement

Examples of task performance and system effectiveness data are provided in Appendix III.

The method proposes the use of both a letter and word task to account for different aspects of intelligibility. One differentiating aspect is the task difficulty as the word task draws its targets and foil alternatives from a large pool of words while the letter task relies on a much more limited and known set, the alphabet. Therefore, each group is expected to successfully communicate more sequences in the letter task than in the word task.

Due to differences in approaches to task completion, voice characteristics and subject personalities, the number of correct sequences communicated on the SuTs may also vary from group to group while maintaining the SuTs' ranking consistency. It is recommended to perform an inter-group normalization of the performance score to maximize the discrimination power of the test results. An inter-team or inter-subject normalization may be performed instead depending on the desired analysis.

6.8.1 Example normalization of the group performance score

The following procedure normalizes the group performance score distributions in terms of mean and standard deviations for each task. The normalization first standardizes the performance score

sample distribution of each group, resulting in a standard deviation of one and a mean of zero. Then the distribution of each group is scaled back to the overall standard deviation and mean. A similar normalization is proposed in clause 4 of [ITU-R BS.1116].

Consider a sub-trial i representing one round of the task t, on a system or condition c, where each participant of a group g acts as reader and responder. The task performance score m_{gtci} for a given trial is defined as the sum of sequences the group successfully conveyed during the trial.

Depending on the groups and the approaches used to complete the tasks, the value of measurements m_{gtci} may vary considerably from one group to another. To account for these variations, a normalized performance score z_{gtci} can be expressed by:

$$z_{gtci} = m + \frac{\left(m_{gtci} - m_{gt} \right)}{s_{gt}} . s$$

where:

z_{gtci} is the normalized score of performance measure m_{gtci}

m_{gt} is the mean performance measure of the group g on task t

s_{gt} is the performance measure standard deviation of the group g on task t

m is the overall mean performance score across all groups

s is the overall performance score standard deviation across all groups.

6.8.2 System effectiveness metric

The system effectiveness metric is a useful interpretation of the test results. It represents the performance of the systems under test compared to the performance achieved in in-person, face-to-face conversation. The system effectiveness is expressed as the ratio between the performance score achieved on a given system and the score obtained in face-to-face communication. The effectiveness, represented as a percentage, where face-to-face is defined as 100%, is easily interpretable and directly relevant to the systems' intended use.

6.8.3 Performance analysis

System effectiveness can be reported per trial, task or conditions depending on the analysis needs.

When the number of groups is small, it is recommended to have balanced gender distribution within each group and to ensure that readers are of the same gender. In the case where this cannot be achieved, it is recommended to analyse the performance of each group individually due to the potential impact of voice characteristics on task difficulty.

When performing per-participant analysis, the composition of the team and voice characteristics should be considered.

Annex A

Construction of the pool of words for the word task

(This annex forms an integral part of this Recommendation.)

The word task described in this Recommendation requires a database of words from which the target and phonological neighbour foil alternatives are drawn. The database should be constructed from a phonological neighbourhood densities corpus. Such corpora can be found in the literature, like the CLEARPOND database that provides lexicons (a list of words composing a language) in English, Dutch, French, German and Spanish [b-CLEARPOND].

Phonological neighbourhood densities databases include a lexicon with 'neighbourhood density' and 'word frequency' metrics for each of the words. The 'neighbourhood density' (or neighbourhood confusability) refers to the number of phonologically similar words in the lexicon, corresponding to the number of words derived from the target word by addition, deletion or substitution of a single phoneme. The 'word frequency' relates to the number of times a word is used in a given language.

Creation of the target word list

The list of target words for the word task can be constructed by extracting from the neighbourhood density database words having at least five phonological neighbours and a significant 'word frequency', one per million for example. At least five of the phonological neighbour words should also have a high 'word frequency'.

Selecting words having a high 'word frequency' increases the probability of being part of the participant's vocabulary and reduces the probability of incorrect pronunciation and undesired slowdown of task completion.

Creation of the list of foil alternatives

Each target word must have a list of foil alternatives corresponding to the phonological neighbour words provided by the neighbourhood density database. This list should be composed of a minimum of five words with significant 'word frequency'. If more than five neighbours are available, then five foil alternatives should be randomly drawn when word sequences are generated.

Words from the target word list may be part of the possible foil alternatives of other target words.

Pruning of the database

The experimenter should verify that the constructed lists of target and alternative words only contain words appropriate for presentation to subjects. In particular words that may offend or cause discomfort to participants should be excluded.

Appendix I

Example of instructions

(This appendix does not form an integral part of this Recommendation.)

Instructions to participants

In this experiment systems that might be used for telecommunication services are evaluated.

You are going to complete tasks by conversing with other participants. In some situations, you will complete the tasks face-to-face, in other situations the group will split into separate rooms to complete the tasks using audio conferencing systems of varying quality.

You will be paired into teams to perform the tasks which consist of conveying sequences of letters or words to your partner. In turns, one participant of each team will be the **Reader** while the other team member(s) will be the **Responders**.

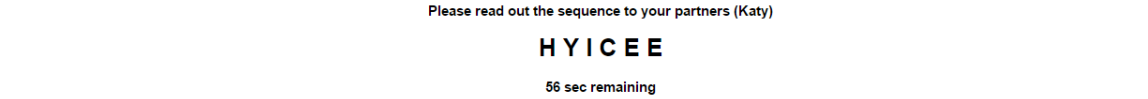
The Reader's screen displays the sequence to communicate to their partner(s) by reading out loud. A new sequence will appear as soon as their partner(s) have entered the correct sequence on their screen. The Responders enter one letter or word per row. Example screens are provided on the next page.

The objective is to read and enter **as many letter or word sequences as possible** within the allotted time. You are free to use any approach to complete the tasks. Crosstalk and interruptions are expected as both teams complete the tasks concurrently.

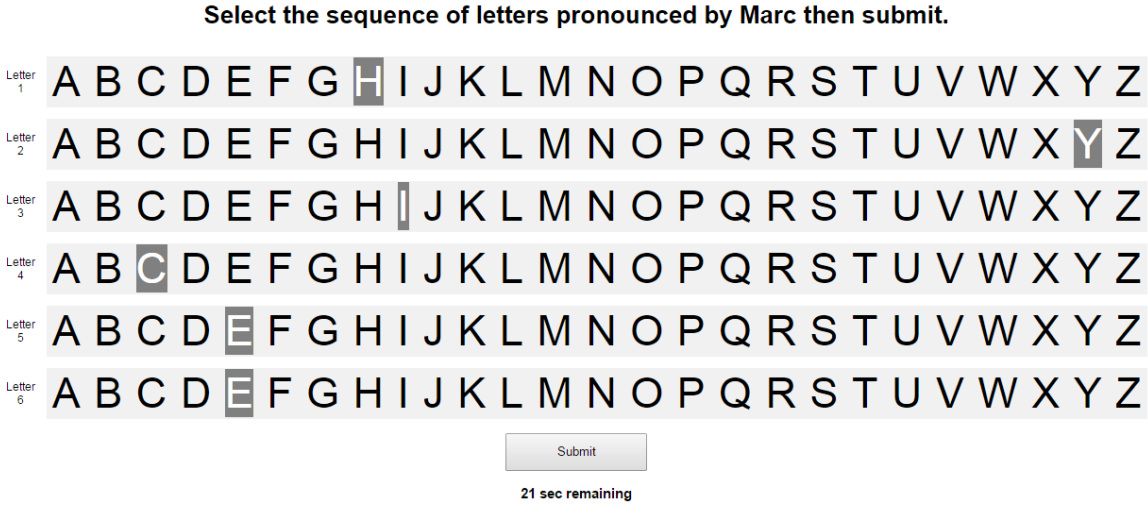
Be as fast as possible: your goal when using the conferencing systems is to reach similar scores to those achieved in face-to-face situations.

Thank you very much for participating in this study.

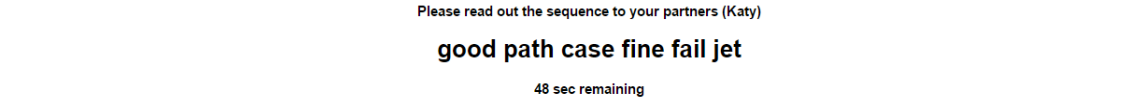
EXAMPLE OF THE READER'S SCREEN FOR THE LETTER SEQUENCE



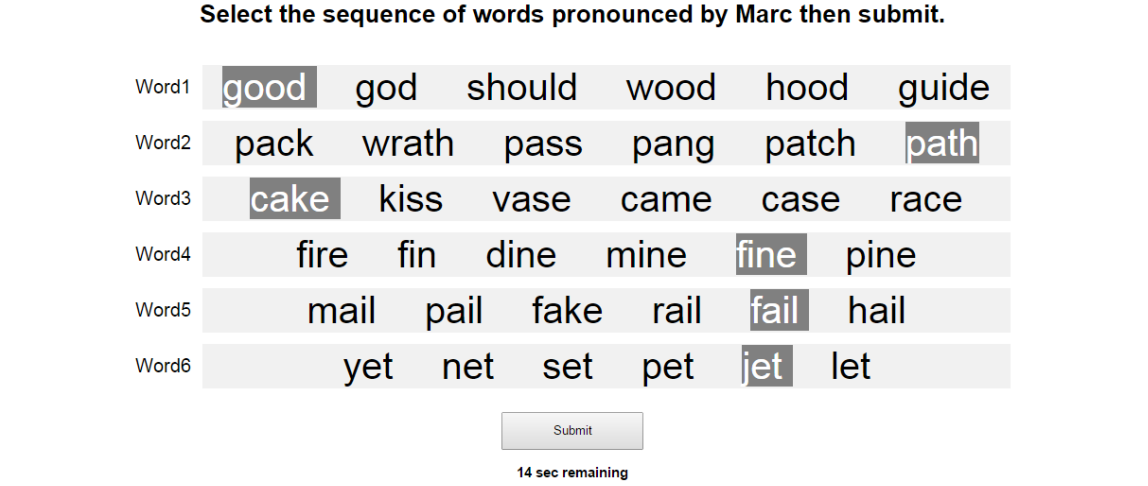
EXAMPLE RESPONDER'S SCREEN FOR THE CORRESPONDING LETTER SEQUENCE



EXAMPLE READER'S SCREEN FOR THE WORK SEQUENCE



EXAMPLE RESPONDER'S SCREEN FOR THE CORRESPONDING WORD SEQUENCE



Appendix II

Example approaches subjects may use to complete the tasks

(This appendix does not form an integral part of this Recommendation.)

Participants are free to choose the most appropriate approach to complete the tasks. The most encountered approaches are:

- **NATO-style phonetic alphabet** – When conveying letters, talker uses words to describe the letters ("A" as "Alpha", "B" as "Boy").
- **Crosstalk** – With each talker/listener team focused on each other's voice, disregarding the others. This seems especially effective in groups with unbalanced genders.
- **Staggered** – No crosstalk. One team's talker waits for the other team's talker to speak to finish.
- **Spelling bee** – When reciting words the talker reads the word, spells it out and uses it in a sentence.
- **Ongoing repetition** – The talker repeats everything on the list over and over without stopping, so the responder can "catch" the letters/words they did not hear before.
- **Targeted repetition** – The responder asks for the specific letter or word he/she feels is incorrect. (For example, the responder requests "Say the 4th one again". The talker says the fourth letter/word listed on their screen).
- **Progressive amplitude** – The talker develops increased volume (shouting), while leaning "into" or closer to the speaker/phone at a decreased angle. Instead of sitting straight and speaking at a normal volume, the talker shouts as close to the speaker as he/she can without distortion. For speakers on headphones, this is achieved by moving the microphone closer to their mouth while shouting. This is perceived by the talker to provide better clarity, while the responders find this causes distortion. Responders then ask for letters/words to be repeated due to the perceived distortion.

Appendix III

Example test results

(This appendix does not form an integral part of this Recommendation.)

This appendix provides examples of effectiveness metrics generated using the method described in this Recommendation. All experiments used four participants for each of the ten sessions.

An experiment evaluating a wideband spatial, a wideband mono and a narrowband mono system for which all participants joined from individual booths using stereo headsets is shown in Figure III.1:

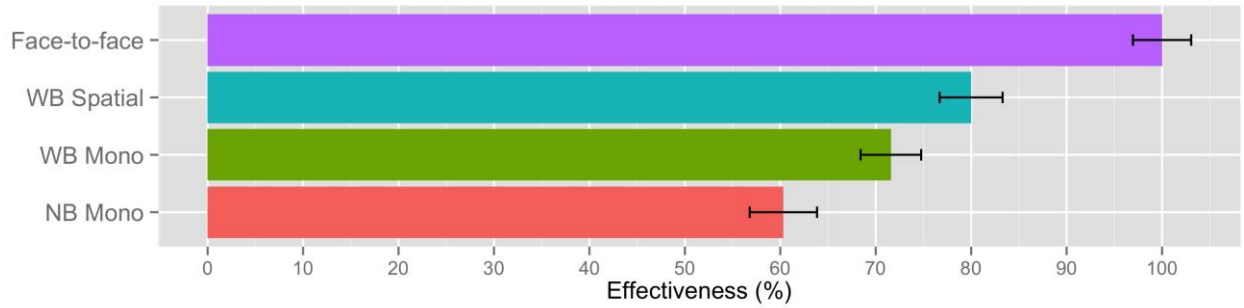


Figure III.1 – Evaluating wideband spatial, a wideband mono and narrowband mono systems

The same systems as above evaluated with headsets featuring limited full-duplex capabilities, resulting in a reduction of performance is shown in Figure III.2:

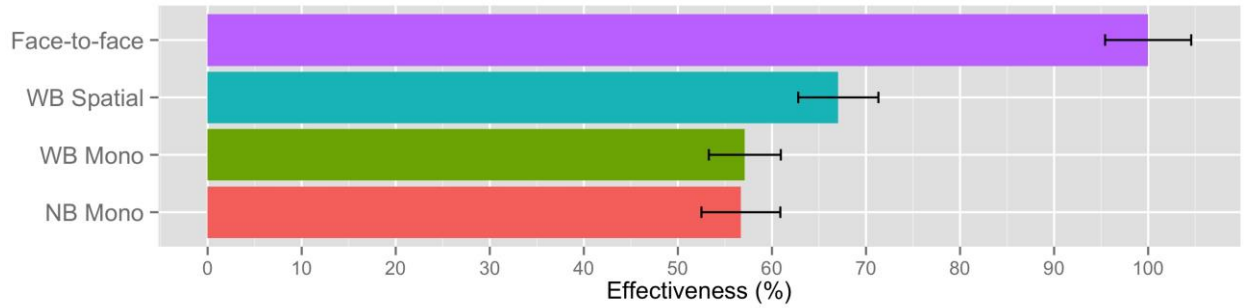


Figure III.2 – Evaluating systems using headsets featuring limited full-duplex capabilities

An experiment evaluating the performance of a wideband spatial, a wideband mono and narrowband mono conference phones is shown in Figure III.3. Two participants were sitting around the device while the two other participants joined from individual booths using headsets:

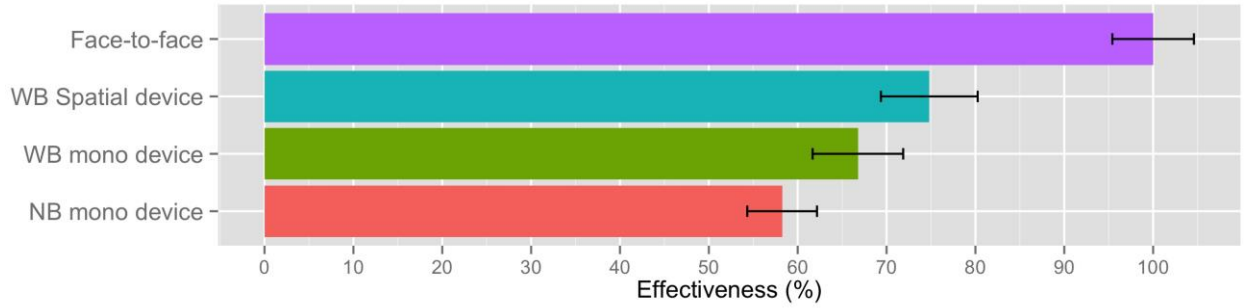


Figure III.3 – Evaluating systems using conference phones

An experiment evaluating the impact of added roundtrip transmission delay to a wideband system where all participants joined from individual booths using headsets is shown in Figure III.4. The conditions also included a narrowband mono system with modified IRS receive filter (MIRS-RX):

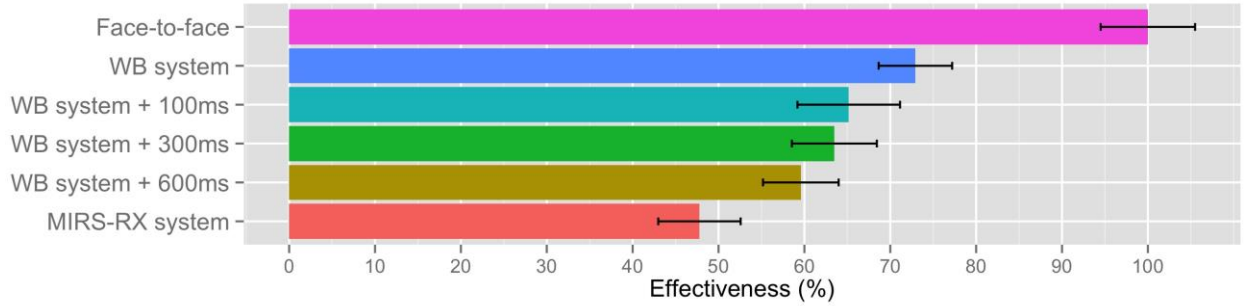


Figure III.4 – Evaluating a wideband system with added roundtrip transmission delay

Bibliography

[b-ITU-T P.800.1] Recommendation ITU-T P.800.1 (2016), *Mean Opinion Score (MOS) terminology*.

[b-ITU-T P.831] Recommendation ITU-T P.831 (1998), *Subjective performance evaluation of network echo cancellers*.

[b-ITU-T P.832] Recommendation ITU-T P.832 (2000), *Subjective performance evaluation of hands-free terminals*.

[b-ITU-T P.1301] ITU-T Recommendation P.1301 (2012), *Subjective quality evaluation of audio and audiovisual multiparty telemeetings*.

[b-ITU-T P.1302] Recommendation ITU-T P.1302 (2014), *Subjective method for simulated conversation tests addressing speech and audiovisual call quality*.

[b-ITU-T P.1311] Recommendation ITU-T P.1311 (2014), *Method for determining the intelligibility of multiple concurrent talkers*.

[b-CLEARPOND] Marian, V., Bartolotti, J., Chabal, S., Shook, A. (2012), *CLEARPOND: Cross-Linguistic Easy-Access Resource for Phonological and Orthographic Neighborhood Densities*, PLoS ONE, August. <http://journals.plos.org/plosone/article?id=10.1371%2Fjournal.pone.0043230>

SERIES OF ITU-T RECOMMENDATIONS

Series A	Organization of the work of ITU-T
Series D	General tariff principles
Series E	Overall network operation, telephone service, service operation and human factors
Series F	Non-telephone telecommunication services
Series G	Transmission systems and media, digital systems and networks
Series H	Audiovisual and multimedia systems
Series I	Integrated services digital network
Series J	Cable networks and transmission of television, sound programme and other multimedia signals
Series K	Protection against interference
Series L	Environment and ICTs, climate change, e-waste, energy efficiency; construction, installation and protection of cables and other elements of outside plant
Series M	Telecommunication management, including TMN and network maintenance
Series N	Maintenance: international sound programme and television transmission circuits
Series O	Specifications of measuring equipment
Series P	**Terminals and subjective and objective assessment methods**
Series Q	Switching and signalling
Series R	Telegraph transmission
Series S	Telegraph services terminal equipment
Series T	Terminals for telematic services
Series U	Telegraph switching
Series V	Data communication over the telephone network
Series X	Data networks, open system communications and security
Series Y	Global information infrastructure, Internet protocol aspects, next-generation networks, Internet of Things and smart cities
Series Z	Languages and general software aspects for telecommunication systems

Printed in Switzerland
Geneva, 2016