Communicability Testing for Voice Communications

Sander J. van Wijngaarden[1]*, John D. Tardelli*[2]*, Hisham Hassanein*[3]*, Brian Ray*[4]*, and John S. Collura*[5]

[1] TNO Human Factors, PO Box 23, 3769 ZG Soesterberg, The Netherlands, VanWijngaarden@tm.tno.nl
[2] ARCON Corporation, 260 Bear Hill Road, Waltham, Massachusetts 02154,, USA, jdt@arcon.com
[3] Communications Research Centre, 3701 Carling Ave, Ottawa, On, Canada, hisham.hassanein@crc.ca
[4] Defence Science and Technology Laboratory, bray@dstl.gov.uk
[5] NSA, R224, 9800 Savage Road, Fort George Meade, MD 20755-6000, USA, jscollu@alpha.ncsc.mil

ABSTRACT

Traditionally, the performance of low bit rate coders was measured by two performance characteristics, namely quality and intelligibility. However, those characteristics are not adequate for measuring the performance over a realistic communication link, where factors such as channel conditions (errors, delay, etc.), dynamic speaker compensation (raising the voice under noisy acoustic conditions), and speakers' capability of interruption, also play a role. Communicability tests were designed for this purpose. This paper is meant to be an introduction to four communicability tests proposed for the selection of the STANAG 4591 coder. The four tests have many features in common. They differ mainly in the type of task used to exercise the communication system. Two of these tasks are variations of games battleship and blackjack, the remaining two tasks are proof reading and determining differences between photos.

1. INTRODUCTION

A voice communication channel may be described as a combination of input/output transducers, a vocoder, a transmission channel and background acoustic environment(s). Conventional methods for measuring the vocoder performance are passive, in the sense that they do not provide for speaker interaction and adaptation. In the context of the development of STANAG 4591, a voice coder selection procedure was designed which was largely based on passive vocoder performance tests. These tests included intelligibility, quality, speaker recognizability, intelligibility of whispered speech, and language dependency. However, under adverse conditions, talkers tend to use compensatory strategies which may affect the vocal effort, talking speed, and pitch..

In order to reflect the dynamic changes in a communication system, various attempts were made (and reported in literature) to design so-called "speech communicability tests": tests that measure the performance of an actual system in a conversational mode. Passive performance tests and speech communicability tests each have their specific advantages, but passive tests are currently more popular, and are applied more often. However, it is expected that communicability tests will gain in importance with the increased use of packetized voice and the need for establishing Quality of Service criteria.

In addition to the passive voice coder tests (phases I and II of the STANAG 4591 selection procedure), speech communicability will be measured in a number of realistic scenarios (phase III). Four proposed communicability tests are presented in this paper.

2. COMMUNICABILITY TEST REQUIREMENTS

Communicability testing mainly measures the *ease* of communication and the ability of the talkers to *interact* in a timely fashion, given a specific communication system. Two attributes are measured, the *efficiency* of the communication system, and the degree of acceptability to the users. Efficiency can be evaluated using an objective measure of the subjects' performance in achieving certain pre-assigned tasks, provided that these tasks allow for sufficiently reliable quantitative efficiency measures. Acceptability on the other hand can only be measured subjectively.

The various communicability tests reported in literature differ most clearly with regard to the tasks that the subjects are given. This is usually a joint collaborative task, which requires elaborate speech communication for the subjects to perform well. In order to design an efficient and accurate communicability test, the choice of a suitable task is crucial.

The following requirements are desirable in a communicability test

- The test should preferably be able to simultaneously measure the efficiency of the system (objective measure), and the acceptability to the user (subjective measure).
- The test should make use of semi-structured conversations (too 'open' conversations make it impossible to measure communication efficiency, but too structured communications do not leave room for the subjects to develop a balanced opinion on the channel)
- The task should be easily learned
- The task should be intrinsically motivating
- Repetitions of the task should be equivalent in their exercising of the communications system
- The task should allow for interruptions
- The results of the test should reflect the quality of the communication system and not the skill of the communicators in performing their task.
- The test should show significant effects with sufficiently small test populations

0-7803-7549-1/02/$17.00 ©2002 IEEE.

- The test should be insensitive to changes in subjects' task strategy

The proposed communicability tests are based on two subjects, performing a task, which requires collaborative communication. The subjects are linked by the communication system under test, and subjected to a background noise environment. The task performance may be scored for success and the subjects are asked to complete a questionnaire as to their opinion of characteristics of the communication such as quality, ease of use, acceptability.

Choices of the task for the tests include variations of games such as *Battleship* and *Black Jack*, and tasks such as *proof reading* and determining the *differences between photos*.

Communication effectiveness is obtained by asking the subjects to rate the link on a scale similar to that used for MOS (Mean Opinion Score). The scale varies from 5 to 7 point scale for a finer resolution.

3. DESCRIPTION OF THE PROPOSED TESTS

Based on the requirements listed above, the next sections describe the four games/tasks proposed for the evaluation of the STANAG 4591 coder.

3.1. The ARCON Communicability Exercise (ACE-95) [1]

The objective of the ACE-95 task is for the two participants to work cooperatively to find and destroy a computer-controlled target. In order to maintain interest and equivalence, the task is adaptive with level of difficulty changing to reflect the success rate of the communicating pair. The computer controlled target has the capability of changing direction, repairing itself when *damaged*, detecting an active search for it, and moving faster as pairs become more sophisticated with the task. Each ACE task lasts for five minutes or until a target is destroyed. If the target is destroyed within a given time, one or more "bonus" targets are available.

The communicators fill two positions that require cooperative discussion and tasking. The first is the Sensor Operator (SO). It is the responsibility of the SO to locate. and track the target. The SO must communicate target direction or coordinates to the Fire Operator (FO), who has the responsibility of selecting a *weapon* and fire point, and communicating this information to the SO. The target is located on a 20x20 grid with rows and columns labeled with rhyming DRT words. Thus, communication of locations involves using these words. All coordinates and actions require verbal verification between the communicators. The success rate of these exchanges is measured. Training on the task to insure that pairs understand the game rules and operation is straight forward. The use of field transducers, half-duplex configurations and other system functions require additional training.

For test series, a blind, random test format is used to control for "carry-over" effects where rated opinions are affected by the system or systems presented before the current one. Each communicator pair is presented with a different randomization. Since communicator pairs evaluate every system in an unknown (to them) order, this will aid in controlling error introduced by inter-subject differences in subjective rating origins and scaling. Test scenarios are normally asymmetric with regards to the acoustic environments at each end of the communication. The way that the task is presented is that there are two laptop computers, each with the programming necessary to present one role of the ACE task (Fire Operator or Sensor Operator). In order to insure each system is tested with individual communicators in each ACE task role, one is assigned to each sound isolation room. These do not vary. The communicators switch rooms (scenario environment, transducers, rooms, and roles) after each test series. The repeat administrations are designed such that the roles and rooms are reversed. The rating scale is automatically presented by the computer upon completion of the task and takes about two minutes to complete. The design of the rating scale is critical to the success of any communicability test methodology. Categorical identifiers are important as is the phrasing of the item. A brief 7-point rating scale was developed to insure it reflected the opinions of the communicators regarding the effort required, quality and overall acceptability of the communication system rather than the difficulty of the task.

The ACE-95 was used as one the tests for the selection of the MIL STD 3005 coder.

3.2. The TNO Communicability Test [2]

The TNO communicability test is based on the Black Jack card game. Two subjects play cooperatively against the "bank", which is computer-controlled. The subjects participate in the game by selecting cards on a computer terminal, and by communicating with each other over the communication channel under test.

The course of the game is always pre-designed, but in such a way that the subjects are given the illusion that all cards are drawn completely at random. The illusion of randomness is necessary to keep the subjects motivated. However, because of the pre-determined course of the game, more accurate efficiency measures can be taken.

The game features a bonus-system, which introduces time pressure (the bonus decreases if the subjects take more time to communicate), and also adds a realistic gambling aspect. This gambling aspects greatly increases subjects' motivation, although the actual bonus that the subjects can earn is also largely pre-determined.

Although the game is based on Black Jack, there are several differences. A very important difference is that the two subjects play together (instead of individually) against the 'bank'. They win or lose together, and are each given the same bonus. Also, some simplifications are introduced in comparison to the original game. For instance, by removing the possibility to 'fold' from the game, the number of good strategies to win the game is greatly reduced. Hence, the subjects' choices become very predictable, and their actual performance depends much more on their ability to communicate well than their aptness at the game. To create time pressure (necessary to motivate the subjects to communicate efficiently) the bonus decreases with time. The players communicate about their cards through 'key words' (or key phrases). The subjects are each given 5 word-card combinations. The subjects decide together which key word to choose, hence which card each player is given. Since the same key word is linked to different cards for both players, some discussion is needed to find out the 'optimal' key word to choose. A typical utterance by one of the players could be: "For 'Romeo' I have a three; what do you have?", or: "We'll take 'Echo', unless you have a word that gives you 21". The design of the game leads to discussions which are structured to a certain degree, and which may be manipulated by using different sets of key words.

The TNO communicability test was applied in a study on the relation between communication efficiency and acceptability, (trade-offs between background noise and channel delay). It is also used in a study on communication efficiency for the International Space Station.

3.3. The CRC communicability test [3]

A pair of subjects are asked to perform a proofreading task. For each trial, the two subjects are provided with slightly different versions of a page of text, and they are asked to identify the differences between the two versions. For each trial the subjects are given 90 seconds to perform the proof reading task. At the end of each trial, the subjects are asked to evaluate the voice communications system which they have just used. The subjects are given 15 seconds to indicate their responses before proceeding to the next trial. On each trial, the characteristics (e.g. transmission delay, bit errors, background noise, etc.) of the voice communications system are varied in a manner which is unpredictable to the subjects. Also, for each trial, the subjects are provided with a new page of text for the proofreading task. Each pair of subjects are exposed to the same system configuration twice: once with Subject A as the Talker and Subject B as the Listener, and vice versa. As such, each pair of subjects will generate two data points for each configuration of the communications systems. This equalizes any biases which may result in the data due to effects Talker or Listener. Each subject is allowed to change his own text to be identical to the text of his partner, regardless of who was doing the reading on that trial. Apart from equalizing the task demands per trial, this increases the amount of interaction and interruption. This is particularly useful when evaluating the effects of delay.

The CRC communicability test was used in a pilot project.

3.4. The DERA communicability test [4]

DERA proposes to use a Free Conversational Test (FCT) with a five-point score as the communicability test. The task serves no other purpose than to facilitate the use of the communication link to pass information with a low a-priory probability. The task is formed by passing both parties on the link, photographs which differ in a number of ways. The subjects are asked to determine the differences between the photos. (In some previous work the task was to identify which photograph was taken first).

4. CONCLUSION

This paper has outlined four different communicability test methods proposed for Phase III of the selection of NATO STNAG 4591. Each of these tests has its own specific character, and its own merits. Given the increasing importance of including real-time aspects in voice coder performance evaluations (specifically in relation to packetized speech transmission), communicability tests are likely to be applied more often in the future. The fact that a variety of methods was proposed for the STANAG 4591 selection procedure, indicates that among the four proposed tests, it will not be difficult to find an adequate one.

5. REFERENCES

1. E.W. Woodward and J.D. Tardelli, “Communicability Testing for Voice Coders, ”IEEE International Conference on Acoustics, Speech, and Signal Processing, pp. 1153-1156, Atlanta, GA, 1996

2. S.J. van Wijngaarden, P.M.T. Smeele, and H.J.M. Steeneken, “A New Method for Testing Communication Efficiency and User Acceptability of Speech Communication Channels,” Proc. Eurospeech, pp.1675-1678, 2001

3. G. Soulodre, M. Lavoie, L. Thibault, and T. Grusec, “Test Method for the Assessment of Transmission Delay in Military Voice Communications Systems,” CRC Report, 1998.

4. Butler, L.W. and Kiddle, L. The rating of delta sigma modulating systems with constant errors, burst errors, and tandem links in a free conversation test using the reference speech link, (Rpt. No. 69014, Feb. 1969) Signals Research and Development Establishment, Ministry of Technology, Christchurch, Hants, 1969.