A New Method for Testing Communication Efficiency and User Acceptability of Speech Communication Channels

Sander J. van Wijngaarden, Paula M.T. Smeele, Herman J.M. Steeneken

TNO Human Factors
PO Box 23
3769 ZG Soesterberg
The Netherlands
{VanWijngaarden,Smeele,Steeneken}@tm.tno.nl

Abstract

The performance of speech communication channels featuring long delay times is usually subjectively experienced as lower than similar channels without delay. Yet most conventional speech intelligibility and speech quality tests are not sensitive to the effects of delay. Moreover, these conventional test do not take the effects of human compensating strategies into account, which help cope with adverse communication conditions by adapting our speech. Test types that do incorporate such effects are sometimes known as 'speech communicability' tests. Based on the lessons learned from literature on speech communicability testing, a list of requirements for the design of a good communicability test method was composed, followed by the actual design of a new test method combining attractive features of existing communicability tests. The suitability of the test design was verified by conducting a pilot experiment. The results of this experiment show that the new method is capable of measuring efficiency and acceptability, and is sufficiently sensitive to delay and background noise.

1. Introduction

Performance of speech communication channels (or systems) is quite often measured in terms of speech intelligibility or speech quality. These are useful performance characteristics, that are known to be powerful when comparing the performance of systems, but they do not take the adaptive nature of human conversation into account. Under adverse conditions people often use compensating strategies: an increased level of ambient noise, for instance, may be countered by raising ones vocal effort (the so-called Lombard effect), or by hyper-articulating,

Straightforward intelligibility measures are therefore sometimes unable to predict performance in practice. An important characteristic of many communication channels, known to clearly influence conversations, is a (long) transmission delay. Speech intelligibility and quality are not influenced by delay; yet channels producing perfectly intelligible speech at a delay of many seconds are almost certainly unacceptable for most applications (mostly because of effects of transmission delay on the process of conversational turn-taking).

Speech communication performance tests that are designed to be sensitive to conversational effects as mentioned above, are often referred to as 'Speech Communicability' tests. We will use this term (or simply 'communicability' test) to indicate this type of method throughout this paper.

The object of the study described in this paper was to implement a new speech communicability test, optimised according to a specific set of requirements, based on principles and results reported in literature.

2. Literature on communicability testing

The awareness that performance of communication systems consists of more than just speech intelligibility or speech quality, has clearly been heightened by the introduction of systems featuring long delay times. The first speech communication performance tests that address the effects of transmission delay date back to the early 1960's, when satellites were starting to be considered as a feasible means of providing transcontinental telephone connections [1,2,3].

The problems introduced by long delays were found to be twofold. First of all, long delay times are the cause of echoes. This became of lesser importance once the development of better echo suppression techniques started to pay off [1,4]. The second problem introduced by delay is more fundamental: long delays introduce conversational difficulties, due to the effect on the turn-taking process.

Procedures for evaluating performance of systems featuring delay in real-time became known as Speech Communicability test. Communicability tests introduced by that name include the Free Conversational Test [5], the Diagnostic Communicability Test [6] and the NRL Communicability test [7]. A more recently developed communicability test in this category is the Arcon Communicability Exercise (ACE) [8]. Other (similar) recent methods, not specifically labeled 'Communicability tests', include the methods applied by Kitawaki [9,10,11].

Communicability tests are targeted at obtaining two categories of test results: results regarding the *efficiency* of speech communication, and results regarding the *acceptability* of the communication channel. Efficiency is best measured objectively, by keeping track of subjects' performance on a communication task. Acceptability is a subjective measure – it can only be obtained by recording users' opinions.

All communicability tests must somehow require subjects to interact in some form of conversation. This is often a structured conversation, which restricts the subjects to predefined conversational patterns: the stronger the restrictions, the more predictable the course of the conversation. This has both advantages and drawbacks; communicability tests using more structured conversations may yield more reproducible estimates of the efficiency of speech communication, but may represent real-life conversations less accurately than free conversations. With free conversations (eg. [2,3,5])., the only feasible way to obtain any measure of communicability usually turns out to be by means of post-hoc interviews or questionnaires. Hence, communicability tests based on free conversation will only reflect the acceptability component (and not the efficiency component) of communicability. An example of a communicability test based on clearly structured conversations is the NRL Communicability test [7]. Several others [3,8] take an intermediary position by using tasks that lead to only moderate structuring. Kitawaki [9,10,11] combined structured and free conversation, by implementing several separate (consecutive) tests.

The way to 'force' the subjects to engage in conversation is generally to give them a joint task. Successful completion of this task depends on the collaboration between subjects, who must therefore communicate with each other. When designing a free conversation test, this joint task only serves to generate a topic of conversation. As long as it is sufficiently motivating to keep the subjects interested, any task is suitable.

For a structured conversation test, the need for the task to be intrinsically motivating is also present, but additional requirements must be imposed. The 'rules' should prevent subjects from reducing the conversation to a fully 'programmed' level, avoiding any kind of interruptions. If the normal pattern of interruptions within a conversation is eliminated, the effect of transmission delay will be strongly reduced. The task should also make sure that a sufficiently large vocabulary is used; by using only few words (such as, for instance, digits) certain kinds of deterioration of the speech signal may pass by the subjects unnoticed. Examples of tasks used in more structured tests are derivatives of the popular Battleships game [7,8], comparing graphs [3], stock trading [6], verifying numbers, completing words with missing letters, and verifying city names [11].

A great advantage of the more structured tests is that they allow for objective measures of efficiency (or effectiveness), such as the number of words necessary to describe a graph [3] or measures derived from (partial) task completion times [11].

3. Development of a New Test Method

3.1. Test method requirements

The lessons learned from literature are summarized in the following list of requirements for a new test method. Some dependency exists between requirements listed separately.

- *The test should incorporate objective (efficiency) and subjective (acceptability) performance.* Whereas most existing methods focus on either efficiency or acceptability, we want to be able to measure both in a single test.
- *The test should reflect all conceivable influences on real-time communication performance* This includes (at least) all effects related to speech quality, speech intelligibility, and the effects of delay on conversations.
- *The test should make use of 'semi-structured' conversations.* This is a compromise between an optimally representative test (free conversation) and a test that permits accurate estimates of efficiency.
- *The test must allow manipulation of 'context'.* Assuming that the communicability task is based on communication of certain key-phrases, the set of phrases that is used should be varied.
- *The test should show significant effects with sufficiently small test populations.* In other words: the test should be sufficiently *sensitive* to the effects that are sought after.
- *The task should be insensitive to changes in subjects' game (task) strategies.* Any task may probably be tackled using a variety of strategies: the more structured the task, the smaller the number of possible strategies. When subjects change strategies halfway through the test, this may introduce statistical spread that reduces the sensitivity to the test.
- *The task should be intrinsically motivating.* If the task is not sufficiently motivating, the subjects may loose interest; the sensitivity of the test will suffer.

3.2. Communicability test design

Based on the list of requirements given in the previous section, numerous different interpretations of a communicability test could be designed. The overall design of a communicability test may be thought of as consisting of three factors: the communication task, collection of objective efficiency data and collection of subjective acceptability data.

The only really feasible way to collect acceptability data is by handing out suitable questionnaires to the subjects (after each condition, with some control questions after completion of the whole test).

More design freedom exists for the objective efficiency data. Some possible indicators of communication efficiency are: time needed for completion of (part of) the task, number of words used and number of repetitions of key phrases. Measures related to time are by far the easiest to measure, because they can be stored automatically without intervention of a test leader. In our pilot experiment (section 4), we focused mainly on response time.

Since we require the task to be intrinsically motivating, it is attractive to shape the communication task in the form of a game. Existing games cannot easily be adopted as a communication task, unless the rules are modified to obtain a desirable structure of the conversations, and the possibility to measure well-defined response times.

For our pilot study, we derived a card game from the well-known gambling game 'Black Jack'. We introduced a bonus-system to increase motivation, especially since this really brings the 'gambling-aspect' of the game to life. In reality, the course of the games was pre-designed, making the outcome of the game (number of times the subjects win and loose), and in connection with the bonus, very predictable. The subjects were not aware of this; they were told that all cards would be drawn randomly. Hence, we were able to impose a pre-designed structure on the game, without ruining the subjects' motivation.

3.3. Brief description of the game

An implementation of the game was realised using two notebook-computers (one for each subject), connected over a network. Although the game is based on Black Jack, there are several differences. A very important difference is that the two subjects play *together* against the 'bank' (played by a computer program). They win or loose 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 induce 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.

We will omit the precise game rules; they were optimised to shape the communication task according to the requirements given in section 3.1, and may be seen as only one example of many possibilities.

4. Results of the Pilot Experiment

4.1. Conditions, subjects and stimuli

From a pool of 16 students, 8 same-gender pairs (5 male, 3 female) of subjects were formed. All pairs of subjects were unacquainted with each other. They were given instructions (written and orally), explaining the game, after which they experienced 30 minutes of hands-on training.

Eight channel conditions were tested. These conditions differed with respect to the level of background noise and the transmission delay. The roundtrip delay times were 0, 800 and 1600 ms; the background noise conditions were: no noise, 73, 67 and 61 dB(A). Noise with the same approximate long-term spectrum as speech for an average male speaker was used.

The subjects communicated through headsets. The speech level produced by these headsets were calibrated to correspond to 70 dB(A), assuming an average male voice at normal vocal effort (60 dB(A) at 1 meter distance). The actual speech levels observed during the tests varied, subject to the vocal effort applied by the subjects. All speech was bandwidth-limited to telephone quality (300-3400 Hz).

Two different sets of keywords were used. The first (more redundant) set was the NATO spelling alphabet. Subjects were familiarised with this alphabet during the training sessions. The second set of keywords consisted of nonsense CVC (consonant-vowel-consonant) rhyme words. Between the words the subjects had to choose from, each time only the initial consonant differed (eg. bil, kil, fil, ril, …). For each condition, three games (varying in length between 1-4 card-drawings) were played for both types of keywords. The time needed for each condition was approximately 10 minutes for both types of keywords (1 – 2 minutes per game).

4.2. Objective efficiency results

The average response time needed for selecting each card was used as an efficiency measure. Approximately 50 estimates of the response time were obtained for each condition and type of keywords, from a total of 8 subject pairs. The effect of delay on the average response time is shown in figure 1, for both types of keywords at a noise level of 67 dB(A). In figures 2 and 3, response times are given as a function of background noise level for conditions with and without delay.

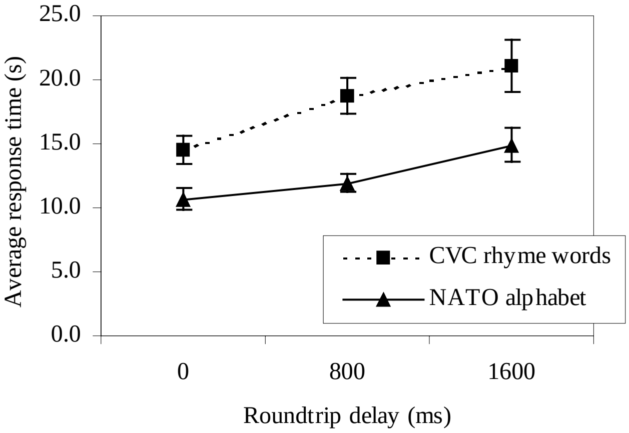

Figure 1. Average response time (NATO spelling alphabet and CVC rhyme words) as a function of roundtrip-delay at a background noise level of 67 dB(A). The error bars indicate the standard error (N=50).

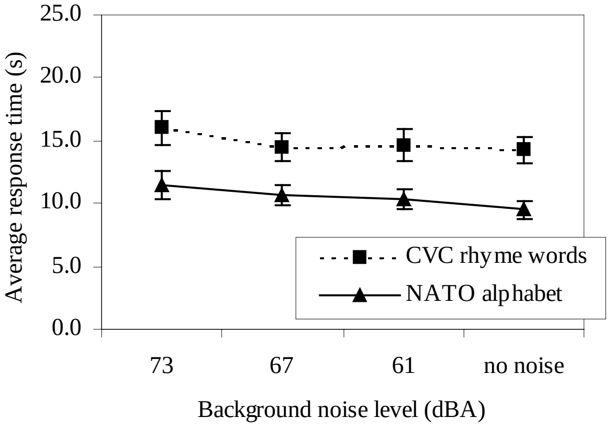

Figure 2. Average response time as a function of background noise level for the conditions without delay (N=50).

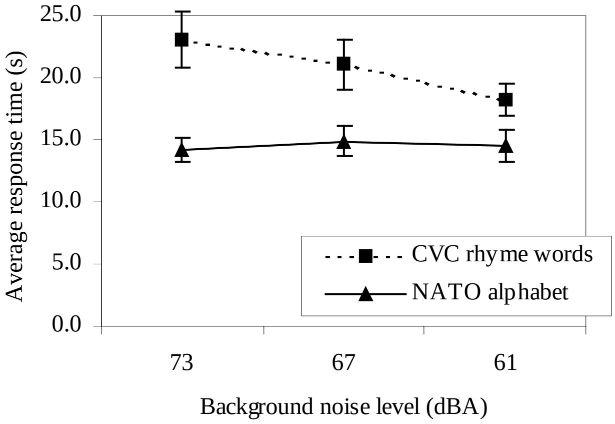

Figure 3. Average response time as a function of background noise level for conditions with a roundtrip-delay of 1600 ms (N=50).

All three figures 1-3 show statistically significant differences between both types of key words. In figure 1, there is also a statistically significant increase in average response time (decrease in efficiency) as a function of the roundtrip delay time. Figure 2 shows virtually no effect of the level of background noise on the average response time; none of the differences in average responses times are significant. Only when an appreciable delay is applied, and then only for the rhyme words, is a significant effect found for background noise level (figure 3). The effect of compensating strategies (increased vocal effort, hyper-articulating) are observed to have an equalising effect on communication efficiency.

4.3. Subjective acceptability results

The questionnaire used in the pilot experiment comprised several questions related to speech quality and acceptability. In this paper, we will only look at subject's responses to the following question: "if you were to use this communication channel during your daily activities, to which degree would you find it acceptable?" Subject responses were given on a 5-point scale. Mean values of the responses from all 16 subjects are given in figure 4.

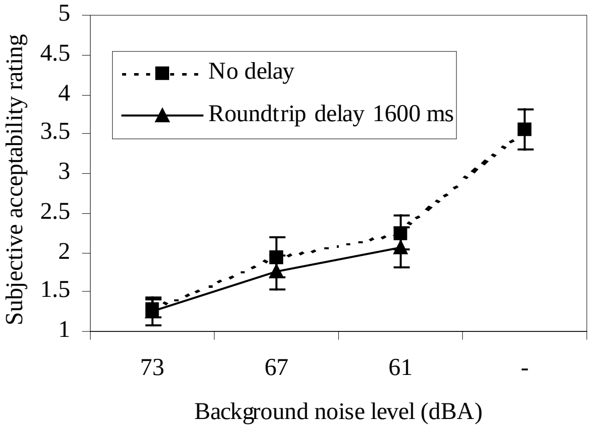

Figure 4. Mean responses (N=16) regarding acceptability for daily use (5-point scale) as a function of background noise level. The error bars indicate the standard error

While communication efficiency was found to be relatively independent of background noise level, the acceptability of the channel clearly *is* reduced by high levels of background noise.

For the effect of delay, almost the reverse is true: the effect of delay is clearly present in the efficiency results, but it did not have a great impact on acceptability. Even the longest delay time (1600 ms roundtrip) was informally reported by subjects to be only a minor inconvenience, compared to the background noise conditions that were present in the same experiment. This indicates that the outcome of the acceptability rating for a certain condition is influenced by the context of other test conditions among which it is presented. This is a valid effect, and may be considered as a normal feature of the test. This implies that the selection of test conditions from which to compose an experiment, should be carefully chosen, and presented in a properly counter-balanced design (as was the case in our pilot experiment).

5. Conclusions

The results of the pilot experiment show that the requirements for a communicability test given in section 3.1 are met by this particular implementation of a communicability test. The pilot experiment was set up to be relatively 'small': only 16 subjects, and only 3 games (approx. 5 minutes) per condition for each type of keyword. Still, the communicability test design proved effective in measuring both efficiency (objective) and acceptability (subjective) of real-time communication channels. The acceptability aspect was found to depend strongly on the level of background noise, but not on the transmission delay. For the efficiency aspect, the opposite result was found: there was a clear effect of delay, but (in the absence of delay) not of the level of background noise. At a roundtrip delay time of 1600 ms, an effect of background noise on efficiency was measurable. Compensation strategies (such as the Lombard effect and more careful articulation) may explain why the effect of noise on communication efficiency remains small.

6. References

[1] Gardner, M.B. and Nelson, J.R., "Combating echo in speech circuits with long delay," *J. Acoust. Soc. Amer.* 35(11): 1762-1772,1963.

[2] Helder, G.K., "Customer evaluation of telephone circuits with delay," *Bell Syst. Tech. J.*, September 1966: 1157-1191, 1966.

[3] Krauss, R.M. and Bricker, P.D., "Effects of transmission delay and access delay on the efficiency of verbal communication," *J. Acoust. Soc. Amer.*, 41(2): 286-292, 1966.

[4] Riesz, R.R. and Klemmer, E.T. "Subjective evaluation of delay and echo suppressors in telephone communications," *Bell Syst. Tech. J.*, November 1963, 2919-2941, 1963.

[5] 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.

[6] Voiers, W.D. *Exploratory research on the feasibility of a practical and realistic test of speech communicability*, (Final report, April 1978). Dynastat Inc, 1978

[7] Schmidt-Nielsen, A. and Everett, S.S. *A conversational test for comparing voice systems using working two-way communication links*, (NRL Report No. 8583, June 1982) Washington DC, USA: Naval Research Laboratory, 1982.

[8] Woodard Kreamer, E.. and Tardelli, J.D. "Communicability testing for voice coders," *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing* 1996.

[9] Kitawaki, N, & Itoh, K. "Pure delay effects on speech quality in telecommunications," *IEEE Journal on Selected Areas in Communication*, 9(4): 586-593, 1991

[10] Kitawaki, N, & Itoh, K. "Delay effect assessment taking into account human factors in telecommunications," *Proc. 13th Symposium on Human Factors in Telecommunication*: 555-562,1991

[11] Kitawaki, N., Kurita, T. & Itoh, K. (1991). Effect of delay on speech quality. *NTT Review* 3(5), 88-94,1991.