interspeech 2008 incorporating SST 08

Measuring Speech Quality Impact on Tasks Performance

Virginie Durin[1], *Laetitia Gros*[1]

[1]Orange Labs, Lannion, 22300, France

virginie.durin@orange-ftgroup.com, laetitia.gros@orange-ftgroup.com

Abstract

This paper deals with perceptual test methodologies to assess speech quality of telecommunication systems. Faced with the tremendous evolution in the telecommunications industry and with drawbacks of typical methodologies recommended by ITU-T, a new way to assess speech quality is investigated. It consists in measuring performance (*i.e.* reaction times and error rates) when subjects are performing two overlapped tasks involving degraded speech signals. This paper presents three tests based on theses two tasks. These tests are different in their protocols. Since results depends on the considered protocol, they are compared to define the most sensitive protocol to quality effect on performance.

Index Terms: speech quality, assessment, methodology, performance

10.21437/Interspeech.2008-538

1. Introduction

Telecommunications industry has experienced two tremendous evolutions for the last twenty years [1]. Firstly, the time of the land-line phone working on analog networks is over: technological improvements such as mobile networks and packet voice have created a strong diversity of terminals (analog land-line, Internet Protocol land-line, mobile etc.) and qualities (presence of distortions, delays, echoes, packet losses, noises etc.), services, use and pricing. These new technologies have also introduced new uncontrolled speech impairments. Consequently, operators need to assess this diversity of qualities corresponding to a new diversity of telecommunication systems. Secondly, nowadays the market is highly competitive meaning that price and quality of services are two key issues that telecommunications operators have to control. Research labs are asked to produce data that reflect customers' opinions on quality and their subsequent satisfaction in their daily use of services. Not only these phenomena intensify the necessity to assess speech quality but also typical methodologies are consequently challenged and subject to evolutions.

Subjective tests can be used by operators to assess speech quality of a telecommunication system. Subjective test methodologies are described in ITU-T P. serie, especially in P.800 [2]. In listening tests, subjects listen to speech samples processed by the system under study, and are asked to assess their quality by giving a score on a five-level scale such as "Excellent", "Good", "Fair", "Poor" and "Bad". The experimenter respectively allocates the scores 5, 4, 3, 2 and 1 to these categories. Thus, a Mean Opinion Score (MOS) is computed by averaging the individual scores. However, these subjective tests present many major drawbacks [1]. First of all, MOS depend on the distribution of qualities within the test corpus: the bias comes from the fact that whatever the corpus is, subjects tend to assign stimuli to categories in such a way that all categories are used about equally often [3, p.108]. Therefore, results of different tests are not always comparable, and therefore a MOS score is valid only for a given corpus. According to Jekosch [4] sound quality should not be studied as an object or a simple sound attribute, as assumed in typical listening tests, but as a process that consists in comparing perception with expectations, referents, knowledge etc. This implies that quality is experienced, hence context dependent [5]. However, the current methodologies do not consider the diversity of services and contexts of use (environment, other activities, aims) that results in various expectations and internal references. Finally these methodologies are not realistic since they are based on an explicit judgement that biases the speech quality percept [6]. In everyday life, speech quality is indeed generally not a conscious object, except in cases in which quality is so degraded that communication becomes impossible.

In order to study the speech quality really experienced by users in ecological situations, it can be argued that we should not directly ask users about speech quality but rather study the impact of quality on their behaviour in communication tasks. Based on this principle, the general hypothesis is the following: impairments introduced into the speech signal by the telecommunication system require additional resources to cognitively process the speech. These additional resources could be to the detriment of other activities and could impact the human behaviour and likely the user satisfaction. Therefore, quality is considered as a means of impacting the efficiency of communication (*i.e.* reaching a goal regarding to consumption of cognitive resources). We assume that performance measure is a good way to objectivise the good running of a communication. In our case, speech impairments that could deteriorate the progress of communication could be measured through performance. In laboratory tests, we propose to study speech quality by observing subject's behaviour through performance criteria (such as reaction times and error rates) when they achieve different tasks more or less complex, serial or parallel, requiring comprehension of degraded speech signals. These tasks are also supposed to involve cognitive processes close to those of real situations of communications.

Several previous studies [7–10] based on dual tasks and attention sharing encourage this approach and make the assumption correct. Three tests had been achieved to observe and optimize the impact of quality on these criteria since then. The first two tests, called Test 1 and Test 2 in this paper, are respectively detailed in [12, 13]. The third test, Test 3, is a new one. In this paper, main results of the three tests are described and compared in order to select, if possible, the most sensitive protocol to quality effect on the considered criteria (*i.e.* reaction times and error rates). The three tests are based on two over-

Accepted after peer review of full paper

lapped tasks: a memory recognition task with digits and a letter recognition task.

2. Methodology

2.1. Tasks

The first task is inspired from Sternberg's memory recognition task [11]: a set of five digits is sequentially displayed on a screen for a fixed duration. Four seconds after the last digit of the set is displayed, a test digit is presented visually. The subject has to decide if the test digit belongs or not to the five digit set. Then, the subject has to recall the five digits in the order of appearance without time limit. For the three tests described in this paper, it is chosen to present visually or auditorily the five digit set depending on the test (see table 1). However, the test digit is always presented visually. For more details about the tasks and the procedure, see [12, 13].

The second task is a letter recognition task. The subject listens to a sentence describing a target letter. At the end of the sentence, a test letter appears on the screen. The subject has to decide if the test letter matches the target letter. This task is used in all three experiments. This task comes in between the presentation of the digit set and the presentation of the test digit of the memory recognition task as illustrated on figure 1.

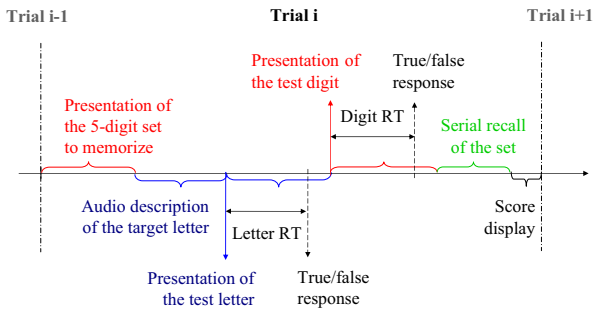

Figure 1: *Schema of a trial performed in the three tests.*

2.2. Stimuli

Different quality levels are applied to the sentences describing the target letter according to the test (see table 1). When digits are presented auditorily, the quality level of the digits and sentences is the same.

- Q1: High Quality (HQ) not impaired
- Q2: G. 729.1 codec (rate: 32 kbps)
- Q3: Narrow band AMR codec (rate: 4.75 kbps)
- Q4: Modulated Noise Referenced Unit 5 dB, MNRU 5

Test 1 involves four quality levels covering a wide range of quality. MNRU 5 is used as a low reference since it is perceptually a very impaired quality level [14]. Therefore a strong quality effect on performance between the worst level (MNRU 5) and others is expected. Thus, it enables to calibrate the scale of reaction times as well as the number of errors one. Unlike MNRU 5 quality level, impairments introduced by other codecs (G.729.1 and AMR) are slighter and their corresponding quality levels are perceptually closer. Tests 2 and 3 focus on a quality effect between HQ, G.729.1 and AMR quality levels.

2.3. Dependent Variables

For each test, five dependent variables are measured:

- Letter recognition Reaction Time (Letter RT)
- Digit recognition Reaction Time (Digit RT)
- Letter recognition Errors (L E)
- Digit recognition Errors (D E)
- Digit Recall Errors (DR E)

The letter RT (resp. digit RT) is the time from the display of the test letter (resp. digit) to the onset of the subject's response.

2.4. Design

In addition to the number of quality levels and the modality of presentation of the digits set, two modes of conditions presentation can distinguish the three tests (see table 1). Tests involve quality levels assessed over one hundred trials each. In Test 1 and 2, one quality level corresponds to one session (*i.e.* "ordered design" in table 1). In Test 3, quality levels are interlaced: they are uniformly and randomly mixed all over the sessions (*i.e.* "random design" in table 1). Then, in this test, a session is arbitrary made of one hundred trials of different quality levels. With this design, it can be expected that it improves the quality effect on performance: the training focused on the first sessions is indeed no more related to specific quality levels depending on the subjects (as in ordered design) but on all quality levels.

Test	Quality level	Subjects number	Digits set presentation modality	Design
Test 1	HQ, G.729.1, AMR, MNRU 5	16	visual	ordered
Test 2	HQ, G.729.1, AMR	18	auditory	ordered
Test 3	HQ, G.729.1, AMR	20	auditory	random

Table 1: *Characteristics of three tests.*

3. Results

The three tests show similar effects from test to the next: a strong between subject variability, a training effect and a quality effect. The training effect corresponds to the improvement of performance (numbers of errors decrease and reaction times shorten) as subjects advanced in the test. The quality effect is the decrease of performance (numbers of errors increase and reaction times lengthen) when quality is impaired. Over the three tests, the between subject variability and the training effect are always observed, especially on reaction times (RT). However, the quality effect seems to depend on the presentation modality and the test design. This paper focuses on this effect. RT and error rates on digit recognition are not depicted in this paper since they are not the most sensitive variables to a quality effect. They are described in [12, 13].

3.1. Error Rates

Figure 2 shows the quality effect on the number of errors of letter recognition and digits recall for the three tests. In the three tests, chi^2 tests are conducted on the errors of letter recognition and digits recall. They confirm a significant quality effect on error rates (see section 6 Appendix): numbers of errors increase when quality is impaired. In Test 1, the quality effect for letter recognition and digits recall comes from the difference between the number of errors obtained for HQ, G.729.1 and AMR versus MNRU 5 as expected (see section 2.4 Design). Therefore Test 1 fails to discriminate all quality levels. In Test 2, numbers of errors of letter recognition and digits recall for each quality levels are significantly different. In Test 3, no quality effect is observed on digits recall errors. There is a quality effect on letter recognition errors, but only between the number of errors obtained for G.729.1 and AMR.

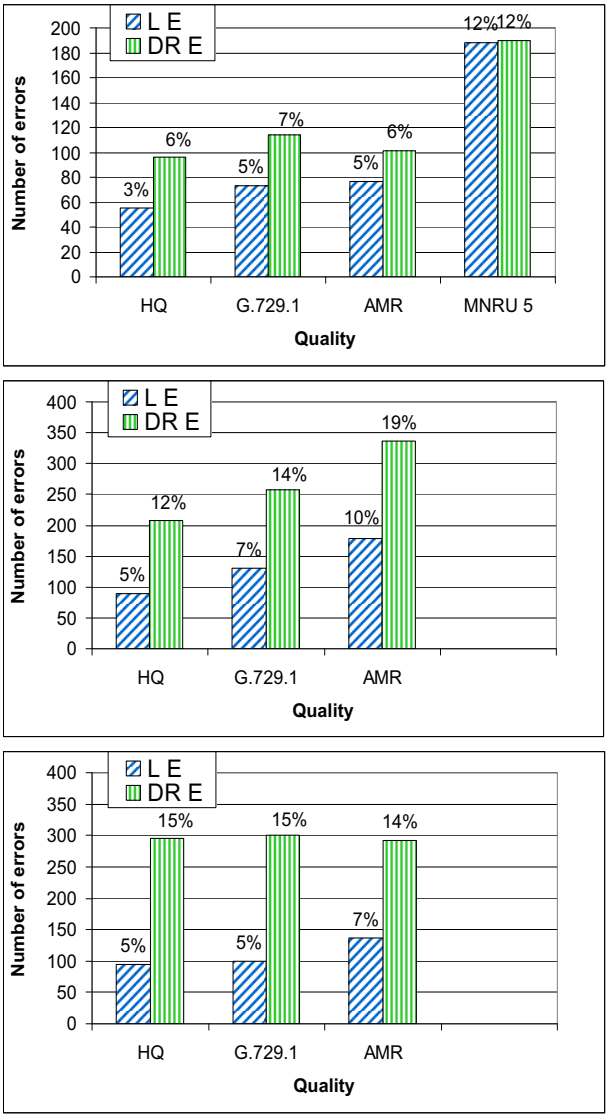

Figure 2: *Number of errors and equivalent percentages per quality for letter recognition (L E) and digits recall (DR E) in test 1 (up), test 2 (middle) and test 3 (bottom).*

To conclude, it seems that the procedure of Test 2 is the most sensitive to quality effect: it enables to discriminate all quality levels although they are perceptually close.

3.2. Reaction Times

For each test, analyses of variance (ANOVA) run on individual letter recognition RT, considering the following factors: "Subject", "Session" and "Quality". They show a quality effect on RT whatever the test (see section 6 Appendix): RT lengthen when quality impairs. Figure 3 shows the mean RT of letter recognition for the three tests. In Test 1, RT obtained for each quality level are significantly different except between G.729.1 and AMR codecs. In Test 2, RT obtained for each quality levels are significantly different. In Test 3, RT for HQ and G.729.1 are not significantly different.

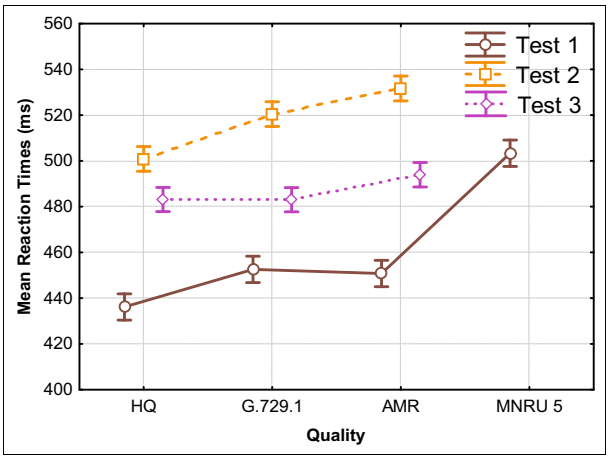

Figure 3: *Mean Reaction Times per quality for letter recognition in Test 1, Test 2 and Test 3.*

Once more, it seems that the protocol of Test 2 is more appropriate to show quality differences. It enables to discriminate RT for each quality level although they are perceptually close.

4. Conclusion

The main results from the three tests presented in this paper show that the quality effect on general performance (reaction times and errors rates) depends on the modality of digits set presentation (visual in Test 1 versus auditory in Test 2) and the mode of conditions presentation ("ordered design" in Test 2 versus "random design" Test 3). Firstly, visual presentation of the digit set is less efficient than auditory presentation in terms of significant differences between quality levels. It can be noticed that MNRU 5 quality level is not assessed in Test 2. However, it can be assumed that quality effect between this quality level and others would still be significantly different since this level is a particulary poor quality. Secondly, the presentation of conditions in an ordered design seems to be more sensitive to quality effect on performance than in a random design. Contrary to what was expected, the quality effect obtained with the random design is less strong that the quality effect obtained with the ordered design. Therefore it can be supposed that switching of quality levels from a trial to the next within a session can create a surprise effect. It may disrupt the establishment of subject's strategy and thus mask the quality effect. Finally, the comparison of these three tests enables to distinguish the most sensitive

protocol to a quality effect on performance: an auditory and ordered presentation of the conditions. Further works turn towards the validity of the methodology in terms of repeatability and stability of the measure, and especially in terms of the number of subjects.

5. Appendix

Test	Quality Effect	Effect on Pairs of Qualities
Test 1	$\chi^2(3) = 119$, $p < 0.001*$	Q1-Q2: $p = 0.1$ Q2-Q3: $p = 0.7$ Q3-Q4: $p < 0.001*$
Test 2	$\chi^2(2) = 32$, $p < 0.001*$	Q1-Q2: $p < 0.05*$ Q2-Q3: $p < 0.05*$
Test 3	$\chi^2(2) = 10$, $p < 0.01*$	Q1-Q2: $p = 0.7$ Q2-Q3: $p < 0.05*$

Table 2: *Quality effect on the number of errors of letter recognition. Results of Chi^2 tests for quality effect on all quality levels and on pairs of qualities. Significant differences are asterisked.*

Test	Quality Effect	Effect on Pairs of Qualities
Test 1	$\chi^2(3) = 50$, $p < 0.001*$	Q1-Q2: $p = 0.2$ Q2-Q3: $p = 0.4$ Q3-Q4: $p < 0.001*$
Test 2	$\chi^2(2) = 38$, $p < 0.001*$	Q1-Q2: $p < 0.05*$ Q2-Q3: $p < 0.05*$
Test 3	$\chi^2(2) = 0.1$, $p = 0.9$	

Table 3: *Quality effect on the number of errors of digits recall. Results of Chi^2 tests for quality effect on all quality levels and on pairs of qualities. Significant differences are asterisked.*

Test	Quality Effect	HSD Tukey Test
Test 1	$F(3, 6378) = 152$, $p < 0,001*$	Q1-Q2: $p < 0.001*$ Q2-Q3: $p = 1.0$ Q3-Q4: $p < 0.001*$
Test 2	$F(2, 5378) = 18$, $p < 0,001*$	Q1-Q2: $p < 0,05*$ Q2-Q3: $p < 0,05*$
Test 3	$F(2, 5645) = 5$, $p < 0,01*$	Q1-Q2: $p = 1.0$ Q2-Q3: $p < 0.05*$

Table 4: *Quality effect on letter recognition reaction times. Results of ANOVA and Honestly Significant Difference (HSD) Tukey Tests. Significant differences are asterisked.*

6. References

[1] Gros, L., Durin, V. and Chateau, N., "Redrawing the link between customer satisfaction and speech quality," Acta Acustica united with Acustica, 94, pp. 32-42, 2008.

[2] ITU-T, "Recommendation P.800, Methods for subjective determination of transmission quality", Geneva, Switzerland, 1996.

[3] Gescheider, G. A., "Psychophysics: Method and Theory", Hillsdale, NJ: Lawrence Erlbaum Associates Publishers, 1976.

[4] Jekosch, U., "Voice and Speech Quality Perception": Springer, 2005.

[5] Gros, L., Chateau, N. and Durin, V., "Speech quality: beyond the MOS score", presented at MESAQIN, Prague, 2006 June 5–6.

[6] Merleau-Ponty, M., "Phénoménologie de la perception": Gallimard, 1945.

[7] Sonntag, G. P., Portele, T. and Haas, F., "Comparing the comprehensibility of different synthetic voices in dual task experiment", presented at the 3rd ESCA, Jenolan Caves, Australia, 1998 November 26–29.

[8] Campana, E., Tanenhaus, M. K., Allen, J. F. and Remington, R. W., "Evaluating Cognitive Load in Spoken Language Interfaces using a Dual-Task Paradigm", presented at the 8th ICSLP, Jeju Island, Korea, 2004 October 4–8.

[9] Gros, L., Chateau, N. and Macé, A., "Assessing speech quality: a new approach", presented at Forum Acusticum, Budapest, 2005 September 29–4.

[10] Durin, V., Gros, L. and Chateau, N., "Evaluation indirecte de la qualité vocale perçue", presented at CFA, Tours, 2006 Avril 21–24.

[11] Sternberg, S., "Memory Scanning: Mental Processes Revealed by Reaction-Time Experiments", American Scientist, 57, pp. 421–457, 1969.

[12] Durin, V., Gros, L., "Toward the evaluation of speech quality impact on users' behaviour through a recognition dual task", presented at 19th ICA, Madrid, 2007 September 2–7.

[13] Durin, V., Gros, L., "Reaction times and performances in recognition tasks to assess speech quality", presented at AES 124th Convention, Amsterdam, The Netherlands, 2008 May 17–20.

[14] ITU-T, "Recommendation P.810, Modulated Noise Reference Unit", Geneva, Switzerland, 1996.