Paradigms for evaluation of speech quality impact on users' behaviour

Virginie Durin, Laetitia Gros

France Télécom Research & Development, 2, av. Pierre Marzin, 22307 Lannion Cedex, France, (+33 2 96 05 94 56, +33 2 96 05 07 20), {virginie.durin, laetitia.gros}@orange-ftgroup.com

Abstract

After a short review on speech quality assessment methodologies and their drawbacks, another approach of quality is introduced. In this approach, speech quality is considered as a means of impacting the efficiency of communication (*i.e.* reaching a goal regarding to consumption of cognitive resources) and some evidence is given to study it through reaction times and performances. Conclusions of two previous experiments lead to verify that impaired speech quality required more cognitive resources to be processed by the human brain, in a simple task involving only cognitive speech processes. However the results do not show significant differences. Therefore, a dual task situation with a digit recognition memory task and a letter recognition task is proposed. Letter and digit recognition reaction times and letter, digit recognition and digits recall performances are collected. Results show a quality effect on both reaction times and on letter recognition and digits recall performances in spite of strong training effect. Suggestions are given to try to reduce the training effect and to improve the quality effect.

Keywords

Speech quality, assessment, methodology, dual task paradigm, memory recognition task, search task

1 Introduction

To assess speech quality of a telecommunication system, operators use subjective tests among others. The subjective test methodologies are described in ITU-T P. serie, especially in the P.800 [1]. 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, for example the most famous one: “ Excellent ”, “Good ”, “ Fair ”, “ Poor ” and “ Bad ”. The experimenter allocates the following values to the scores: Excellent = 5, Good = 4, Fair = 3, Poor = 2, Bad = 1. Thus, a Mean Opinion Score (MOS) is obtained by averaging the individual scores.

However, these subjective tests present many major drawbacks [2]. First of all, MOS's depend on test corpus [3]: 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. Therefore, results of different tests are not always comparable, and a MOS is theoretically valid within a test. In addition, Jekosch [4] developed an interesting point of view about quality: assessing quality is more a process that consists in comparing one's perception with one's expectations, referents, knowledge etc. This implies that quality is not a simple sound attribute but 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. But in everyday life, users almost never think about speech quality. Except in cases in which quality reaches down to such a level that communication becomes impossible, speech quality is generally not a conscious object. Merleau-Ponty [6] explains that this intellectualization biases the judgement because "we consider quality as an object *of* our consciousness whereas it is an object *for* consciousness".

In the aim to study speech quality really experienced by users in ecological situations, it can be argued to not directly ask users about speech quality but rather to study the impact of quality on their behaviour in communication tasks. Based on this idea, the general hypothesis is the following: impairments introduced into the speech signal by the telecommunication system involve additional resources to cognitively process the speech. These additional resources could be to the detriment of other activities, 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). It is interesting to note that the concept of efficiency of communication is already used in aircraft industry and related to flight performances [7, 8]. Here efficiency makes reference to pattern of communication characterized by used words, presence of feedback, etc. and is associated with technical proficiency of pilots. In other words, it was shown that pattern of communication determined performance. Thus it is reasonable to assume that the performance measures are a good way to objectivise the good running of a communication. In our case, speech impairments that could deteriorate the running of communication could be measured through performances.

In laboratory tests, we propose to study speech quality, not directly asking subjects, but observing their behaviour through criteria, such as reaction times and performances, when they achieve different tasks more or less complex, serial or parallel, and involving cognitive processes close to those involved in the real situations of communications.

Some studies give first elements: Sonntag et al. [9] measured the understanding of speech sequences uttered by six speech synthesis systems and with a natural voice, coded or not by the GSM network (so likely to be impaired) in a dual task. The primary and the secondary tasks both showed significant differences in reaction time for the different types of voices. In addition, for two of the seven voices, the differences between the coded and the non coded differences were significant. Campana et al. [10] use the dual-task paradigm, for evaluating the cognitive resource demands of different speech interfaces. Participants follow simple instructions generated by a system, while simultaneously monitoring for a simple visual probe. Performance on the monitoring task is used as a measure of cognitive resource demands; whenever language understanding is more demanding, performance on the monitoring task suffers.

Another possibility is proposed by Wilson and Sasse [11] who explored the electrophysiological way to measure the audio and the video quality. The considered criteria were the Galvanic Skin Response (GSR), the Heart Rate (HR) and the Blood Volume Response (BVP). Results show that different electrophysiological responses could be obtained for different degradations, depending on the task.

Our assumption was tested in previous experiments [12, 13] and results show the validity of the general hypothesis and the relevance of reaction times, percentage of correct responses to study quality: in some conditions, reaction times increase and percentage of correct responses decreases when speech quality impairs. However, the primary task used in [13] (computation task) resulted from complex and resource-demanding cognitive processes (mental calculation). The response variability was probably more the consequence of the variability of the various calculations complexity subjects had to achieve (e.g., 21-12 being more difficult than 21+1) than the consequence of quality impairments. Therefore, the potential influence of speech quality of audio instructions on reaction times may have been hidden by such large variations. Moreover, an observation of subjects' behaviour suggests that tasks were not treated in parallel but sequentially. The aimed concurrence between the two tasks was not guaranteed. The dichotic situation tested in [13] was not realistic and not representative of real telecommunication applications. The overlapping between presentations on left and right ears was not controlled besides. A stereo image could result and disturb the subject.

Notice that the electrophysiological measures carried out in these two first experiments do not lead to relevant conclusions.

In spite of the drawbacks of these experiments, these studies make the validation of the assumption possible. Now the new goal is to find a task easy to control, involving cognitive processes close to those of a telecommunication applications and effective in terms of methodology (*i.e.* less time consuming for a test condition, reliable, sensitive etc.). A dual task paradigm may be a solution to increase the cognitive load and make the cognitive resources demand due to the bad quality visible. However, dual task paradigm is complex for many reasons. Therefore, a very simple task involving impaired speech signals is first tested in order to avoid dual task issues.

2 Impact of speech quality on cognitive speech processes

This first experiment aimed to verify that impaired speech quality required more cognitive resources to be processed by the human brain, in a simple task involving only cognitive speech processes (and no attention sharing, decision or other high level cognitive activities). However in a simple communication or understanding task, the required additional resources (because of the degraded speech) are easily fulfilled and therefore invisible in the behaviour. Another way to load the cognitive system is to remove the meaning of sentences and makes more difficult the phonemic restoration effect [14]. Two kinds of sentences were used: "predictable" (*e.g.* the cat catches the mouse)

versus "unpredictable" (*e.g.* the kitchen catches the air) sentences. Unpredictable sentences can be defined as sentences whose words do not belong to the same lexical field. If the hypothesis is right, the intelligibility is close to 100% with typical quality levels (even with very bad condition such as MNRU at Q=5 dB) when sentences are "predictable". On the contrary, we assume the demand becomes visible and measurable with "unpredictable" sentences: the more impaired the sentences would be, the lower the intelligibility score would be. Thus the impact of speech quality on cognitive resources could be artificially studied through intelligibility test with unpredictable sentences. To test this assumption, we carry out two intelligibility tests, with quality levels known to not impair so much the intelligibility. For the first intelligibility test, sentences were extracted from a French phonetically balanced corpus (Combescure's corpus) [15]. For the second intelligibility test, unpredictable sentences were built from the Combescure's corpus. This corpus consists of ten sentence lists, each list being phonetically balanced. Therefore, phonetically balanced but not predictable set of ten sentences is obtained by mixing the words of an original ten sentence list (respecting syntaxic rules). Thus, the same content (in terms of words) is used in the two tests. In each test, two quality levels were presented: high quality (HQ) (without any degradation, bandwidth filter, etc.), and quality impaired with Modulated Noise Reference Unit 5 [16] (MNRU 5). Five ten sentence lists were presented for each quality level. Subjects listen to the sentences and write down what they understood on a paper sheet. Ten subjects performed the "predictable sentences" test and ten other subjects performed the "unpredictable sentences" test. The intelligibility result for a quality level is scored by counting all the correct words in the fifty sentences presented for a quality level and for each subject. The averaged percentages are given in the Table 1, with standard deviations in brackets.

Quality Sentences\	MNRU 5	HQ
Predictable	94.80 % (4 %)	99.87 % (0.4 %)
Unpredictable	86.79 % (7 %)	98.45 % (2%)

Table 1: percentages of correct words

For predictable sentences, the intelligibility rate is very high for the two studied quality levels. In return, for unpredictable, a decrease of intelligibility can be observed for the impaired quality level. However, the difference between the two qualities is not significant. Even removing the meaning of sentence, the cognitive load was too low to measure differences in speech quality. It seems that only dual task paradigm could overload the cognitive system enough to observe behaviour changes when quality is impaired.

3 A new dual-task paradigm

The dual task paradigm adds another task to the communication task and therefore increases the cognitive load. The dual task situation is not unrealistic since today's services allow users to do several things in the same time, especially with mobile phone (walking and crossing the street or doing one's shopping during a call). Telecommunication services combining audio and visual modalities are fast-expanding. Therefore, today, it is not unusual to have motor, visual and auditory modalities involved in a same communication situation.

3.1 Communication task

Actually, by communication task, we do not mean a communication between two people (as conversation-opinion test) but a task with vocal instructions that subjects need to understand to achieve a certain action. This pseudo-communication task may have certain similarities with a vocal server.

Our previous experiment [13] showed that the (pseudo) communication task should not be too complex like mental calculation task. The task should involve simple cognitive process such as detection or recognition. Furthermore, it does not bring into play culture level of the subjects.

In the experiment of Campana et al. [10], subjects were asked to follow simple instructions to click on the right object among many objects displayed on the screen (for example: click on the small red candle). This primary task is a good simulation for communication because it requires understanding the sentence and acting as a consequence; it is listed in the speech acts (as "a assertive act") of the theory developed by Austin and Searle [17, 18] and which formalizes language. At last, this task is convenient to measure reaction times and performances, since they do not involve too complex cognitive activities or culture issues. However authors do not observe any effect of speech interfaces on reaction times in the pseudo communication task but only in the monitoring task. No explication is given in the paper but we think that this task involves variability in the search between all of the objects. The choice of our pseudo communication task is carried out in such a way that all of others variabilities than those of speech quality are reduced.

The proposed communication task is very close to the primary task of [10]: subjects listen to a sentence in French language with the form "Is it a (colour) (case)

(letter)?". The five colours are: blue, red, green, yellow, and black. Twenty letters are chosen among the monosyllabic letters of alphabet. The two cases are: lower case and upper case. All in all, there are two hundred sentences corresponding to all the possible combinations. At the end of the sentence, a letter appears on the screen. The subject has to say as quickly as possible if the displayed letter (test letter) matches the verbally described letter (target letter).

3.2 Additional task

Previous experimentation showed that the so-called parallel tasks yield strategies hard to control. Consequently to avoid parallel treatment issues, the two tasks can be separated and it could be convenient to overload the subject before the communication task. A memory task is suitable for this protocol. In addition, memory processes are often involved in real situations of communication (for example remembering a phone number while communicating).

Sternberg [19] worked on a memory recognition task: a set of digits (called the "positive set"; remaining digits constitute the "negative set") was sequentially presented on a screen for a fixed time. Two seconds after the last digit in the set was displayed, a warning signal appears, followed by a visually-presented test digit. The subject has to say if the test digit belongs to the positive set or not. Then, the subject has to recall the positive set in the order of appearance. This task is all the more convenient because it allows overloading the cognitive system during the communication task without issues of parallel tasks strategies.

The proposed additional task is inspired from the Sternberg's recognition memory task [19]. The size of positive set is fixed to five digits. Thus positive and negative sets have the same size and subjects are not tempted to work on the negative set instead of the positive set. In addition, the pseudo communication task comes in between the presentation of the positive set and the presentation of the test digit.

The two proposed tasks allow to collect reaction times and performances, thus maximizing the chance to observe effects of speech impairments on one or the other task. Moreover, the two tasks are fast to achieve, that matches with an effective methodology.

3.3 Stimuli

The sentences ("Is it a (colour) (case) (letter)?") are uttered by a French female speaker. The choice of letters and colours is made to minimize temporal variabilities of the speech signal. Only monosyllabic letters and colours are kept. For each cue, speech signal's length is equalized. Therefore sentences have the same length (2 s).

Four quality levels are applied to these recorded sentences:

- Q1: High quality (HQ) not impaired.
- Q2: G. 729.1 coder (rate: 32 kbps).
- Q3: Narrow band AMR coder (rate: 4.75 kbps).
- Q4: Modulated Noise Reference Unit 5 (MNRU 5). MNRU 5 is used as a low reference.

The resulting sentences are up-sampled at 48 kHz to be diotically presented to subjects through headphones at a comfortable listening level of 73 dB SLP.

3.4 Procedure

For each trial, a cross appears at the centre of the screen to warn the subject of the digits presentation time and place. Then, the positive set is visually and sequentially presented at a rate of 1.2 s per digit. After this presentation, the vocal description of the target letter is played through headphones. During the audio signal presentation the cross in the centre of the screen appears again allowing to fix eyes. At the end of the sentence, a test letter appears on the screen. Subjects have to say as quickly as possible if the test letter matches the target letter within two seconds. After this communication task, a test digit is then displayed. Subjects have to decide if the test digit belongs or not to the positive set within 2 seconds. The subject presses on the "q" key with left hand or "m" key with right hand to answer true or false. For half the subjects, "q" key (resp. "m" key) is the true key (resp. false key) and for the remaining half the subjects, "q" (resp. "m") is the false key (resp. true key). Finally, subjects have to recall the positive set in the order of appearance without time limitation.

In order to motivate participants, a score calculated according to the RTs of the two tasks and errors is given at the end of the trial. In addition, for the two tasks, a green (resp. red) feedback is given for a correct (resp. false) response to subjects.

There are one hundred trials per condition: all of the possible combinations of letters and colours are used. Upper case and lower case are equally shared over the one hundred trials. The presentation of the one hundred sentences is made in a random way. Moreover for each quality condition and for the two tasks, there are as many true responses as false responses.

The test is divided into five sessions: a "control" session involving the additional task alone followed by four "quality" sessions, one for each quality level, involving the communication task in the dual-task paradigm and then alone. In order to prevent order effect on quality, the four "quality" sessions are carried out in one of the four following order (corresponding to a Latin square) : [Q3 Q4 Q1 Q2], [Q2 Q3 Q4 Q1], [Q1 Q2 Q3 Q4], and [Q4 Q1 Q2 Q3].

The "control" session is made up of one hundred trials involving the additional task alone. Subjects wait for 4 s (that corresponds to the length of the communication task in the dual-task). Subjects are invited to make a pause each twenty trials. A trial can be evaluated at 14 s ([1.2s x 5 for the presentation of the five digits of the positive set] + 4s for the pause corresponding to the communication task + 2s for the response to the test digit + recall). The "control" session can be estimated to 23 min.

Each "quality" session is divided into two parts: first the communication task is carried out with the additional task and then alone. Each part is made up of one hundred trials. A quality session lasts about 30 min (23 min for dual task + 7 min for the communication task alone).

Subjects have several training sessions before beginning the test using the same procedure that the test's. All in all, subjects practise the dual task and the two tasks alone in all conditions more than three hours.

3.5 Dependent Variables

Five dependent variables are measured:

- Letter Recognition Reaction Time (LR RT)
- Digit Recognition Reaction Time (DR RT)
- Letter Recognition Performance (LR P)
- Digit Recognition Performance (DR P)
- Digit Recall Performance (DRe P)

The LR RT (resp. DR RT) is the time from the display of the test letter (resp. digit) and the onset of response. Besides, in order to favour the observation of a possible quality effect on reaction times of letter recognition, the test letter is presented 150 ms before the end of the sentence.

3.6 Subjects

Eight subjects achieve the test, with two subjects per order of quality sessions. For a given order, one subject has the "q" key for the true response and the other has the "q" key for the false response. Subjects were told to answer as quickly as possible and make as little errors as possible. They were motivated by the score at the end of each trial. They were paid for their services.

3.7 Technical specifications

Nowadays most of widespread operating systems are multitasks and they give illusion to run many programs at once. They actually allocate resources to all the programs sequentially. There may be other applications running that have a higher priority, and hence, the operating system will not always carry out an operation whenever it is requested. By consequence reaction times measures can be biased and there is no means to know the error of the measure. To prevent time inaccuracies, we developed a dll working with Matlab to catch keyboard event. Tests were done in which the key stroke (for letter and digit recognition) is simulated by the PC at 875ms. Therefore RTs (for the two tasks) should be 875ms in theory. The results on six thousands data (one hundred trials thirty times) give a RT mean of 875ms and a standard deviation of 0.082ms.

4 Results

4.1 Data preparation

Reaction times (RTs) corresponding to non responses (*i.e.* RT = 2 s, that is to say 0.03% of the whole of the RTs collected in the two tasks) are substituted for the mean of the subject for the considered condition.

Successes number of each subject in each condition is added on the one hundred trials. Let's note that subjects did the task very well: the worst score is 81/100 and 80.6% of individual scores are above 90. These scores are transformed in RAU scores (Rationalized Arcsine Units) [20]: the success scores are not adequate for statistical analysis since they fit a binomial probability distribution. Therefore data are not normally distributed around the mean and scale values are not linear in relation to test variability. The rationalized arcsine transform is applied to have statistically valid results.

4.2 Subject variability

Figure 1 and Figure 2 show the RTs and performances for each subject, all quality levels considered.

It appears that there are around 200 ms of difference between the fastest and the slowest subject. In a same

way, there are non negligible performances differences between subjects.

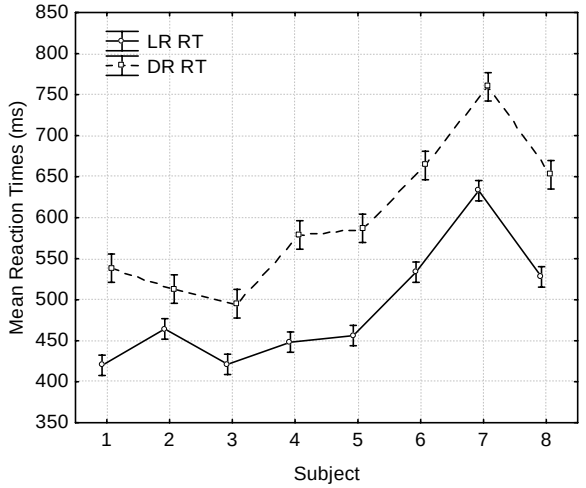

Figure 1: Mean reaction times and associated confident intervals for each subject, for letters and digit recognition.

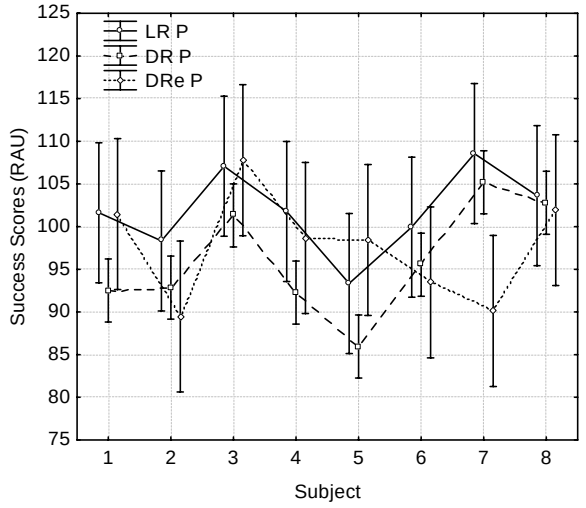

Figure 2: Success scores and associated confident intervals for each subject, for letters and digit recognition and digits recall.

Two MANOVAs, conducted on one hand on the two RTs and on the other hand on the three performances, confirm a significant effect of the factor "Subject" on both RTs ($F(14, 6382) = 90.95$, $p<0.0001$) and performances ($F(21, 63{,}722) = 3.83$, $p<0.001$).

In order to remove the inter subject variability that could hide a possible quality effect, the individual RTs and performances are centered and reduced.

4.3 Training effect

Figure 3 shows the mean reaction times for the digits and letter recognition, according to the position of sessions in the order of the test.

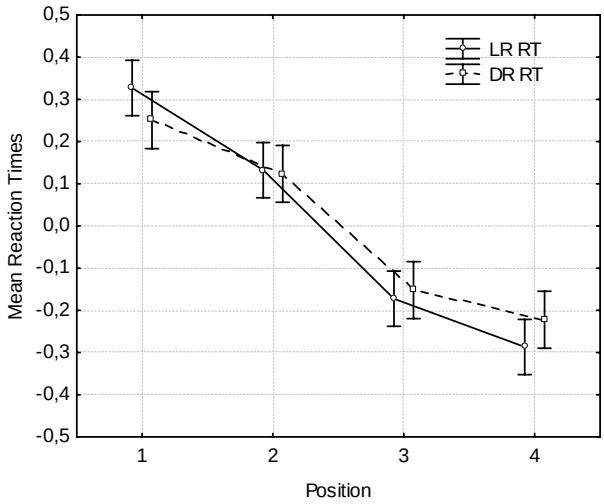

Figure 3: Mean reaction times and associated confident intervals for letters and digit recognition according to the session position.

Figure 4 shows the success scores for the digits and letter recognition and digits recall, according to the position of sessions in the order of the test.

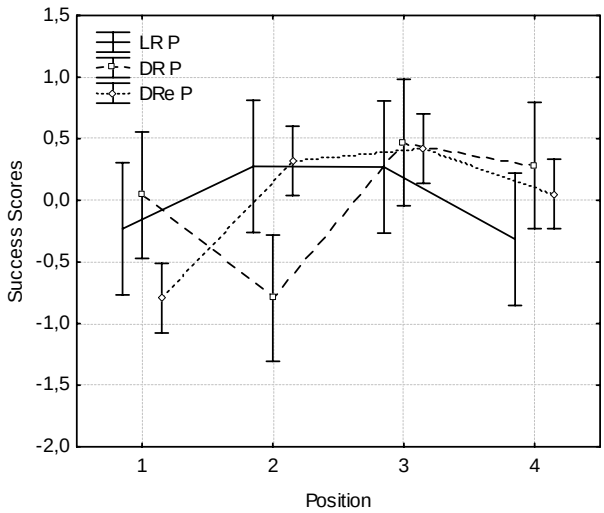

Figure 4: Success scores and associated confident intervals for letters and digit recognition and digits recall according to the session position.

Two ANOVAs are run on the reaction times, one for the letter recognition and the other one for the digit recognition, considering the two factors "Position" (with four levels: P1, P2, P3, P4) and "Quality" (with four levels: Q1 for HQ, Q2 for G.729.1 coder, Q3 for AMR coder, Q4 for MNRU 5). Three analogous ANOVA are

conducted on the success scores, the first one for the letter recognition, the second one for the digit recognition, and the last one for the digits recall. Results of these five ANOVAs are given in the Appendix 6.1 and 6.2.

These ANOVAs confirm the strong effect of position on the reaction times, either for the letter ($F(3, 3184)=70.73$, $p<0.0001$) and the digit recognition ($F(3, 3168)= 42.91$, $p<0.0001$). Nevertheless, this training effect seems to be weaker for the performances: no effect on the letter recognition ($F(3,16)=1.57$, $p=0.23$), weak effect for digit recognition ($F(3, 16)=5.30$, $p<0.01$) and effect for digit recall $F(3, 16)=17.21$, $p<0.0001$). For digit recognition case, the significant effect is more explained by the low and surprising performance for position 2 than by a training effect. In return, the pattern obtained for the digit recall looks like a training pattern (increase of the success scores with time).

The absence of training effect in performances for digit and letter recognition is not surprising since this task is rather easy, the difficulty being more in a fast achievement (of course without error). Therefore, the training effect mainly appears on the reaction times that shorten with experience. However, there is a slight training effect on digits recall, perhaps because of a specific and more complex strategy needed to encode the five digits in their order of appearance. This strategy probably improves along the test.

4.4 Quality effect

Figure 5 shows the mean reaction times for the digits and letter recognition, according to the quality level.

Figure 6 shows the success scores for the digit and letter recognition and digits recall, according to quality level.

It seems that for all dependent variables, except digit recognition performance, the worse the quality, the worse the performances and the longer the reaction times.

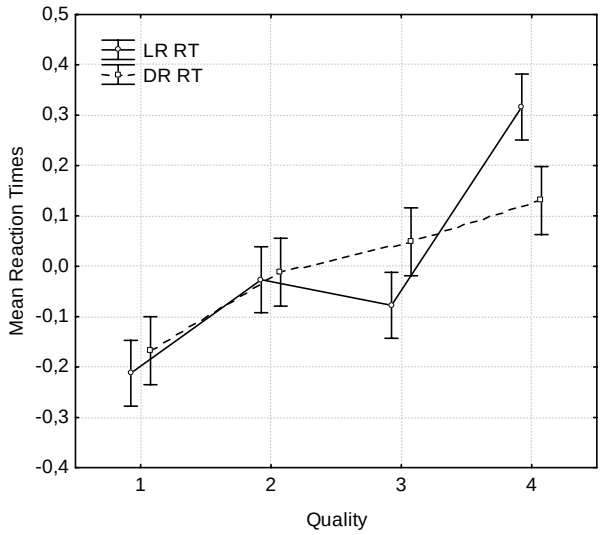

Figure 5: Mean reaction times and associated confident intervals for letters and digit recognition according to the quality level.

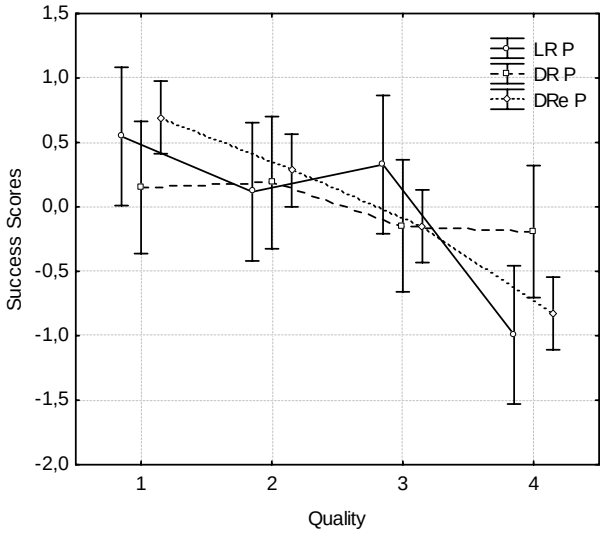

Figure 6: Success scores and associated confident intervals for letter and digit recognition and digits recall according to the quality level.

The five ANOVAs conducted for each dependent variable (cf. Appendix 6.1 and 6.2) confirm a significant quality effect on letter recognition reaction times ($F(3, 3184)=45.43$, $p<0.0001$), on digit recognition reaction times ($F(3, 3184)=13.42$, $p<0.0001$), on letter recognition performances ($F(3, 16)=7.32$, $p<0.01$) and on digits recall performances ($F(3, 16)=23.93$, $p<0.0001$). No effect is found on the digit recognition performances ($F(3, 16)=0.62$, $p=0.59$).

Table 2 gives the level of significance of differences between quality levels, for each of five dependent variables.

	LR RT	DR RT	LR P	DR P	DRe P
Q1-Q2	***p<0.05***	***p<0.05***	*p=0.64*	*p=1*	*p=0.17*
Q1-Q3	***p<0.05***	***p<0.05***	*p=0.93*	*p=0.82*	***p<0.05***
Q1-Q4	***p<0.05***	***p<0.05***	***p<0.05***	*p=0.75*	***p<0.05***
Q2-Q3	*p=0.7*	*p=0.6*	*p=0.93*	*p=0.76*	*p=0.14*
Q2-Q4	***p<0.05***	***p<0.05***	***p<0.05***	*p=0.69*	***p<0.05***
Q3-Q4	***p<0.05***	*p=0.33*	***p<0.05***	*p=1*	***p<0.05***

Table 2: p values from Honestly Significant Difference Tukey test for the five variables and all comparisons of quality levels.

The most discriminative variables are reaction times and especially the letter recognition reaction times which discriminate not only extreme quality levels (MNRU versus High quality) but also intermediate quality levels (for example High quality versus G729.1 at 32 kbps). Performances especially digit recognition performances are less sensitive to quality differences. In return, digits recall performances seem to be promising.

LR RT would be a very good candidate without the slight inflexion of the mean RT for the quality level 3. Figure 7 shows mean reaction times for each quality level and position, for LR RT. It illustrates the significant interaction between the two factors (F(9, 3184)=5.25, *p<0.0001*). It appears on this figure that the inflexion on quality 3 specifically occurs for position 1 and has no evident explanation. Notice that for one position and one quality, only two subjects are considered.

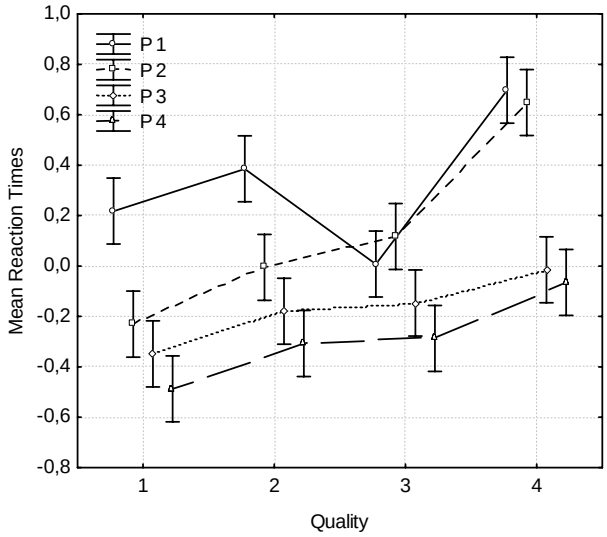

Figure 7: Mean RTs and associated confident intervals for letter recognition according to the quality level and the position.

Now one wonders if the letter recognition task could be sufficient for our purpose. In the next section, the comparison of results between the reaction times obtained with the communication task alone and those obtained with the dual-task is studied.

4.5 Dual-task effect

Figure 8 shows mean reaction times of letter recognition, obtained with the dual-task and the letter recognition alone, for each quality level.

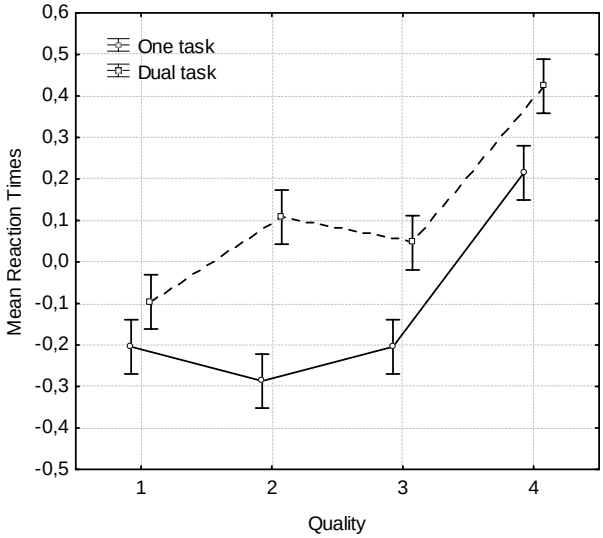

Figure 8: Mean RTs and associated confident intervals for letter recognition task alone and in dual task according to the quality level.

An ANOVA on these reaction times considering the factors "Position", "Situation" (with two levels: dual-task, one task) and "Quality" confirms that letters RTs are significantly longer in the dual-task situation whatever the quality level (F(3, 6368)=6.39, *p<0.001*). Details of ANOVA are given in Appendix 6.3. In the same way, DR RTs are shorter and digits recall performances are better (*p<0.05*) in control situation than in dual task situation, that means that the task is more complex.

Moreover, a HSD Tukey test conducted on letter recognition reaction times points out that only the difference between the quality level 4 and the three other quality levels (*p<0,05*) is significant in the one-task situation. Contrary to the dual-task situation, quality levels 1, 2 and 3 are not significantly different (*p>0.65* for all comparisons). Therefore, dual-task situation is more sensitive and relevant to a quality effect.

4.6 Conclusion

The interest of using the dual task situation is demonstrated in comparison to a single task in the sense that the dual task situation is more relevant and sensitive to quality effect. Letter and digit recognition RTs, letter recognition and digits recall performances make the differences between extreme quality levels and others levels possible. For example, letter recognition RTs makes the difference between the G.729.1 coder at 32 kbps and the High Quality. These results are encouraging because they provide more differences with more accuracy in comparison to our last studies [12, 13], by avoiding parallel tasks issues. However, no significant differences between G.729.1 and AMR coder was found, whatever the criterion considered. Differences could have been perhaps significant if there had been no training effect. The training effect observed is indeed very strong and likely weakens the quality effect. Therefore, the first improvement for further experiments would be to control this training effect. We suggest two solutions: first, it is possible to extend the training session in order to have very trained subject. However this solution makes the test very time consuming. The other alternative could be to have untrained subjects whose each group achieves only one quality condition. This alternative requires more subjects.

Some other improvements should be brought to the protocol to make it more efficient. The quality effect is less strong on digit recognition RTs than on letter recognition reaction times. However there is perhaps a way to strengthen the quality effect on digit recognition reaction times. The longest allotted time to answer to the letter recognition is 2s. Because subjects do this task quickly they can have enough time to recall in their mind the five digits before the test digit appears. This strategy can help them to keep digits in mind until the recall procedure but also to answer faster to the test digit. Therefore it could be interesting to do the experiment again with a shorter allotted time for letters. Moreover it could be interesting to use a letter cues recall procedure in place of digits recall because it implies a dual memorization of digits and letter cues which could interfere according quality. It would be indeed possible that recall of vocal items is dependent of the speech quality since it is shown that it is dependent of the hearing losses [21].

Finally, it would be also interesting to collect the subject assessment not on quality, but on theirs own performances. Other criteria like measure of cognitive load (e.g. NASA TLX) could be interesting too: for example, the NASA Task Load Index [22] is a multi-dimensional rating procedure that provides an overall workload score based on a weighted average of ratings on six subscales: Mental Demands, Physical Demands, Temporal Demands, Own Performance, Effort, and Frustration. This index is widely used in aviation area but also in other areas like cognitive psychology as a mental workload measure [23, 24].

5 References

[1] "UIT-T P.800: Methods for subjective determination of transmission quality," 1996.

[2] N. Chateau, Gros, L., Durin, V., Macé, A., "Redrawing the link between customer satisfaction and speech quality," presented at 2nd ISCA/DEGA Tutorial & Research Workshop on Perceptual Quality of Systems, Berlin, Germany, 4th-6th September 2006.

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

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

[5] L. Gros, Chateau, N., Durin, V., "Speech quality: beyond the MOS score," presented at Measurement of Audio and Video Quality in Networks (MESAQIN), Prague, 5th-6th June 2006.

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

[7] K. Mjos, "Communication and operational failures in the cockpit," *Human Factors and Aerospace Safety*, vol. 1, pp. 323-340, 2001.

[8] J. B. Sexton, Helmreich, R.L., "Analyzing cockpit communications: the links between language, performance, error, and workload," *Human Performance in Extreme Environments*, vol. 5, pp. 63-68, October 2000.

[9] G. P. Sonntag, Portele, T., Haas, F., "Comparing the comprehensibility of different synthetic voices in dual task experiment," presented at Proceedings of the 3rd ESCA Workshop on Speech Synthesis, Jenolan Caves, Blue Mountains, Australia, 26th-29th November 1998.

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

[11] G. M. Wilson, Sasse, M. A., "Straight from the heart: Using physiological measurements in the evaluation of media quality," presented at Proceedings of the Society for the Study of

Artificial Intelligence and the Simulation of Behaviour Convention 2001, Symposium on Emotion, Cognition and Affective Computing, York, UK, 21st - 24th March 2001.

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

[13] V. Durin, Gros, L., Chateau, N., "Evaluation indirecte de la qualité vocale perçue," presented at Congrès Français d'Acoustique (CFA), Tours, 21-24 Avril 2006.

[14] R. M. Warren, "Perceptual restoration of missing speech sounds," *Science*, vol. 167, pp. 392-393, 1970.

[15] P. Combescure, "20 listes de dix phrases phonétiquement équilibrées," *Revue d'acoustique*, vol. 56, pp. 34-38, 1981.

[16] ITU-T, "P.810: Modulated Noise Reference Unit," 1996.

[17] J. R. Searle, *Speech acts: An essay in the philosophy of language*. Cambrige: Cambrige University Press, 1969.

[18] Austin, *How to do things with words*. Oxford: Oxford University Press, 1962.

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

[20] G. A. Studebaker, "A "Rationalized" Arcsine Transform," *Journal of Speech, Language and Hearing Research*, vol. 28, pp. 455-462, 1985.

[21] S. L. McCoy, Tun, P.A., Cox, L.C., Colangelo, M., Stewart, R.A., Wingfield, A., "Hearing loss and perceptual effort: Downstream effects on older adults' memory for speech," *The Quarterly Journal of Experimental Psychology*, vol. 58A, pp. 22-33, 2005.

[22] S. G. Hart, Staveland, L.E., *Development of NASA-TLX (Task Load Index): Results of empirical and theoretical research*: In P.A. Hancock and N. Meshkati (Eds.). Amsterdam, The Netherlands: North Holland Press, 1988.

[23] Y. H. Lee, Liu, B.S., "Inflight workload assessment: comparison of subjective and physiological measurements," *Aviation, Space, and Environmental Medicine*, vol. 74, pp. 1078-1084, 2003.

[24] S. Haga, Shinoda, H., Kokubun, M., "Effects of task difficulty and time-on-task on mental workload," *Japanese Psychological Research*, vol. 44, pp. 134-143, 2002.

6 Appendix

6.1 Letter and digit recognition reaction times

An ANOVA conducted on the letter recognition reaction times with the "Position" and "Quality" factors gives the results in the tTable 3:

Effect	SS	df	MS	F	p
Position (1)	189.2	3	63.07	70.73	**0.000***
Quality (2)	121.5	3	40.51	45.43	**0.000***
(1)* (2)	42.2	9	4.68	5.25	**.000***

Table 3: Summary of the ANOVA conducted on the letter recognition reaction times with the "Position" and "Quality" factors.

An ANOVA is conducted on digit recognition reaction times with the "Position" and "Quality" factors gives the results in the Table 4:

Effect	SS	df	MS	F	p
Position (1)	120.4	3	40.14	42.38	**0.000***
Quality (2)	38.1	3	12.72	13.42	**.000***
(1)* (2)	17.5	9	1.95	2.06	**.030***

Table 4: Summary of the ANOVA conducted on the digit recognition reaction times with the "Position" and "Quality" factors.

6.2 Letter and digit recognition and digits recall performances

An ANOVA conducted on letters performances with the "Position" and "Quality" factors give the results in the Table 5.

Effect	SS	df	MS	F	p
Position (1)	2.42	3	.808	1.577	.234
Quality (2)	11.25	3	3.751	7.320	**.003***
(1)* (2)	2.12	9	.236	.461	.880

Table 5: Summary of the ANOVA conducted on the letter recognition performances with the "Position" and "Quality" factors.

An ANOVA conducted on digits recognition performances with the "Position" and "Quality" factors give the results in Table 6:

Effect	SS	df	MS	F	p
Position (1)	7.452	3	2.484	5.306	**.010***
Quality (2)	.930	3	.310	.662	.587
(1)* (2)	8.127	9	.903	1.929	.121

Table 6: Summary of the ANOVA conducted on the digit recognition performances with the "Position" and "Quality" factors.

An ANOVA conducted on digits recall performances with the "Position" and "Quality" factors give the results in the Table 7:

Effect	SS	df	MS	F	p
Position (1)	7.30	3	2.432	17.21	**.000***
Quality (2)	10.14	3	3.381	23.93	**.000***
(1)* (2)	4.30	9	.478	3.38	**.016***

Table 7: Summary of the ANOVA conducted on the digits recall performances with the "Position" and "Quality" factors.

6.3 Dual Task effect

An ANOVA conducted on LR RT with the "Situation" (Position, Situation (with two levels: dual-task, one task) and "Quality" factors give the results in the Table 8.

Effet	SS	df	MS	F	p
Situation (1)	92.7	1	92.7	104.6	**0.000***
Position (2)	324.5	3	108.2	122.1	**0.000***
Qualité (3)	221.9	3	74.0	83.5	**0.000***
(1)*(2)	9.9	3	3.3	3.7	**.011***
(1)* (3)	17.0	3	5.7	6.4	**.000***
(2)* (3)	48.5	9	5.4	6.1	**.000***
(1)*(2)*(3)	36.4	9	4.0	4.6	**.000***

Table 8: Summary of the ANOVA conducted on the LT RT with the "Situation", "Position" and "Quality" factors.