The impact of listening and conversational situations on speech perceived quality for time-varying impairments

L. Gros, N. Chateau,
France Telecom R&D
33 (0)2 96 05 13 16
laetitia.gros@rd.francetelecom.fr, noel.chateau@rd.francetelecom.fr

Summary

This paper relates an experiment conducted in order to compare the listening and conversational situations in assessment of time-varying speech quality. This experiment was realised in two stages: a conversation-opinion test followed by a listening one with recordings of conversations. In this way, the judgement on perceived quality in the two contexts has been compared. It appears that judgements are similar in a conversational context and in a listening one.

Keywords: speech quality, subjective assessment

Introduction

In order to assess perceived speech quality in telephony, there are two types of standard methods: listening and conversation-opinion tests [1]. For time-constant quality, these two types of tests give similar results. However the Internet-telephony context introduces a new constraint resulting in losses of packetized information during connection: quality can vary strongly during a same communication. Previous studies relate on subjective evaluation of time-varying speech quality, in a listening context [2]. A method of continuous judgement was used in order to study the impact of impairments variations on perceived quality at any instant. This method was associated with a standard procedure (ACR) for the assessment of the overall judgement (at the end of the sequence). With this double procedure, the overall quality score can be explained by the evolution of perceived quality during the sequence. A recency effect was found: the perceived overall quality was more influenced by a quality variation placed at the end of the sequence than by a similar quality variation placed at the beginning. So, the overall judgement is elaborated principally on the basis of time distribution of impairments weighted by memory processes. Now, perceived quality may be different in a conversational context. In effect, attention is differently shared in the two situations: in listening situation, all the attention is directed to quality and the task of rating. In a conversational context, one part of the attention is used for the action of communicating. This could deteriorate the performance of quality assessment and induce different ratings in the two different situations. In order to compare quality perceived both in listening and conversational contexts, an experiment was conducted in two stages: a conversation-opinion test followed by a listening one with recordings of conversations.

Method

Method for conversation-opinion test

According to the recommendation [1], couple of subjects[1] were seated in separate sound - proof cabinets. They were asked to communicate through an IP-telephony set, with the help of pretexts such as ordering a pizza. At the end of each conversation, they were asked to assess the quality on several five-point category scales. They were seven criterions as, for example, the global quality or the perception of defaults. Conversations were recorded at sound card input and output, in order to have for each speaker, the conversation such as he heard it *i.e.* his own voice, direct, and the interlocutor's voice through the web. The speech signal was degraded by means of the soft Netdisturb placed between the two terminals, which introduced impairments in real time during the communication according to scenarios pre-defined and programmed, named

[1] 24 naive subjects participated, in 12 groups of 2 subjects

quality profiles. In this paper, we present results obtained for six profiles that contains variations of packet losses over 2 min (see figure 1).

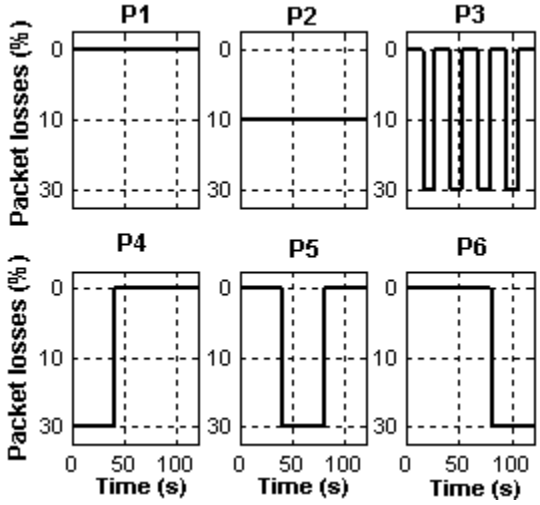

FIGURE 1: The six quality profiles used in the experiment.

Method for listening test

For each quality profile, four conversations among the twelve conversations recorded during the conversation-opinion test were retained for their natural character. For each of the 24 conversations retained, there were two sequences (A and B) corresponding to the two stereophonic recordings realised during the conversation-opinion test, one for each terminal. A third sequence (AB) was realised by mixing the two channels containing the transmitted voice (one in sequence A and one in sequence B). Thus, for each of the 24 conversations, there were three sequences (A, B and AB) of 2-min duration. Sequences A, B and AB were presented in three different listening tests, named respectively Test A, Test B and Test AB. In each test, for each sequence, the listeners[2] were asked to continuously assess the quality (see [2]), and to rate the overall quality at the end of each sequence on the recommended classic 5-items MOS scale [1] (5 =Excellent, 4=Good, 3=Fair, 2=Poor, 1=Bad).

[2] 23, 17, and 15 subjects participated respectively for test A, B, C

Results

Conversation-opinion test

An analysis of Pearson's linear correlations between mean opinion scores obtained for the seven criterions shows that they are highly correlated ($0.90 < r < 0.98$). Furthermore we will present only the results obtained with the criterion "global quality": the effects found with this criterion will be the same than those found with the other criterions, considering high correlations between the criterions.

Listening test

Sequences corresponding to a same quality profile were rated in a similar way, whatever the communication heard and whatever the test realised. The distinction between sequences seems to be made more on the basis of quality variations than on the basis of the content of communication. It is confirmed by inter-group analysis of variance (ANOVA) conducted on the overall scores. Although this ANOVA reveals an effect of the inter-group factor Test ($F(1,162) = 6.66$, $p<0.005$), the effect of intra-group factor quality profile predominates ($F(5,810) = 158.13$ $p<0.0001$), and no systematic difference between the three tests appears. So the effect of quality profile seems to be stronger than the effects of communication and test, although sequences for a same profile could be very different because of the verbal content and the time repartition of speech between interlocutors. Considering these observations, in order to compare results obtained in the two situations, overall scores were averaged for each quality profile, all communications and tests considered.

Comparison between listening and conversational tests

Figure 2 shows mean opinion scores and standard deviations obtained for the six quality profiles, in the two contexts. It can be noticed that standard deviations are systematically more important in the conversational situation than in the listening one. This observation can be explained by a more important variability of

stimuli for a same quality profile (verbal and semantic contents, time repartition of speech between the two interlocutors) in the conversation test than in the listening test. In return, mean opinion scores seem to be similar for the two contexts.

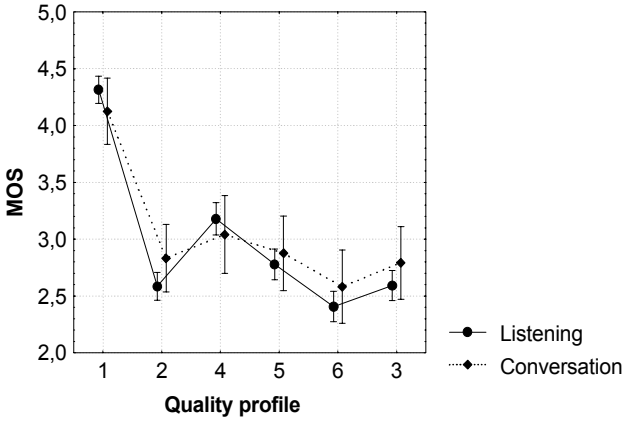

FIGURE 2: Mean overall scores for each profile, in the two experimental situations (parameter).

It is confirmed by an inter-group ANOVA: the effect of the inter-group factor Situation is not significant ($F(1,62) = 0.56$ $p = 0.4$). In return, there is a significant effect of the intra-group factor quality profile: $F(5,810)=58.027$ $p<0.0001$. And no interaction between factors Profile and Situation is revealed ($F(5,810) = 1,26$ $p = 0,28$). It can be noticed that the recency effect (the last moments of the sequence more influence subjects' opinion) seems to be stronger in listening situation than in conversational situation. When the degradation moves from the beginning to the end of the sequence, the overall MOS falls with 0.78 MOS in listening situation where as it falls with 0.44 MOS in conversational context.

Discussion and conclusion

As it has been shown in our previous studies [2], subjects take into consideration a weighted average of instantaneous judgments when giving their overall judgments, in favor of the last moments of the sequence (recency effect). With this experiment, it appeared that there is no real difference between quality judgments in a listening context and in a conversational one. Although the recency effect seems to be less important, it can not be explained by interferences between memory and attentional processes. In case of such interferences, a share of attention, as in a conversational situation, could deteriorate the storage in memory and result in a more important recency effect, contrary to our result. In return, one can observe that mean opinion scores obtained in a conversational situation have a less important dynamic than mean opinion scores obtained in a listening one. Subjects would be more critical in a listening test and their judgments would be more discriminating. It could explain a more important recency effect in a listening situation. However, this difference of strategy has a weak impact on quality judgments that are similar for the two situations. So, for IP-telephony with packet-losses only, one can imagine substituting conversation-opinion tests by listening tests which are less expensive in time and cost.

References

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

[2] L. Gros, N. Chateau. “Instantaneous and overall judgments for time-varying speech quality: Assessments and relationships". *Acta Acustica* Acustica, 2001, 87, 367-377

[3] S. Glucksberg, G.N. Cohen. "Memory for non attended auditory material". Cognitive Psychology, 1970, 1, 149-156.