Subjective assessment of time-varying speech quality

Laetitia Gros, France Télécom R&D, Lannion, France, laetitia.gros@rd.francetelecom.fr

1. Introduction

The last decade has seen a considerable evolution in the telecommunications sector with the development of real-time services on the Internet such as telephony. These services come up with several technical issues that have to be taken into account by operators. One of them is time-varying speech quality resulting from delays and random losses of packetized information during connection. In order to assess speech quality, typical methods described in recommendation ITU-T, P.800 use short stimuli (8 s) in subjective listening tests. These stimuli are speech samples impaired by the system under study. In the ACR method (Absolute Category Rating) which is a single stimulus assessment, listeners assess sample's quality giving a mark on a five-category scale: 5 for "Excellent", 4 for "Good", 3 for "Fair", 2 for "Poor" and 1 for "Bad". Subsequent analyses usually use averaged scores obtained from all listeners called MOS (for Mean Opinion Score). Although well suited for time-constant speech quality, these methods do not take into consideration integration of non-steady quality over long periods. In this context, new protocols suitable to time-varying speech quality need to be developed. In this paper we present a few results obtained for time-varying speech quality with an appropriate methodology as follow.

2. General methodology

In order to present realistic quality fluctuations, long speech sequences (between 45 seconds and 3 minutes) were presented to listeners. Then, in order to measure the instantaneous perceived quality (quality perceived at any instant of a heard sequence), a method on continuous judgement was used (SSCQE method generally used in the audiovisual field): subjects listened a long speech sequence through handset and in the same time they moved a cursor along a continuous scale including the five items "Excellent", "Good", Fair", "Poor" and "Bad" so that its position reflected their opinion on quality at any instant. The cursor was held on a box allowing an 11-cm movement. Six boxes were connected to a PC that recorded every 500 ms the cursors' position coded from 0 (scale's bottom) to 255 (scale's top) in a data file on the hard disk. For each subject, each T-sec sequence provided a data file of 2xT values (one instantaneous score every 500 ms during T s) and a scalar value (a global quality judgement). The 2xT values ranging from 0 to 255 were subsequently linearly transformed to MOS units between 1 to 5. Secondly, at the end of each sequence, subjects were asked to rate its overall quality by giving a mark on the typical ACR scale (5-categories scale with the French equivalents of Excellent, Good, Fair, Poor and Bad) written on a sheet.

3. Study of un-guaranteed QoS impairments

Figures 1 and 2 show mean instantaneous judgments obtained for long speech sequences degraded at different level with two types of impairments, MNRU and packet losses. Packet losses were introduced by the mean of an Internet simulator using the codec G 723.1 (30-msec frames, 3 frames per packet). In addition, on these figures, one can see mean overall scores obtained at the end of these long speech sequences on the typical ACR-scale, and also mean overall scores obtained in a standard protocol (ACR) with a 8-sec speech sample impaired with MNRU and packet losses at the same objective levels.

As can be seen, for MNRU, subjective quality is quite constant for a same objective level. In return, for packet losses, none of the five objective quality levels was rated as being constant over the 3 minutes. Differences of almost 1 MOS between two samples of a same objective quality level can be noticed. Moreover, MOS obtained with a short sample and with a long speech sequence are similar for standard impairments whereas they are quite different for packet losses because of quality fluctuations.

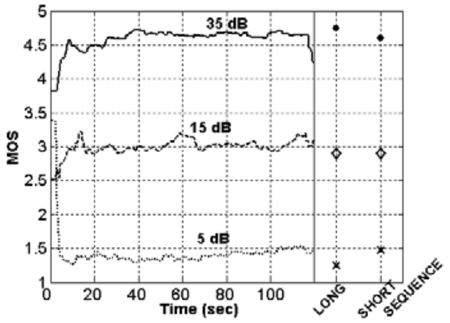


Fig. 1: Mean instantaneous judgments and mean opinion scores (with 2-min and 8 sec speech sequences) obtained for 35 dB, 15 dB, 5 dB of MNRU.

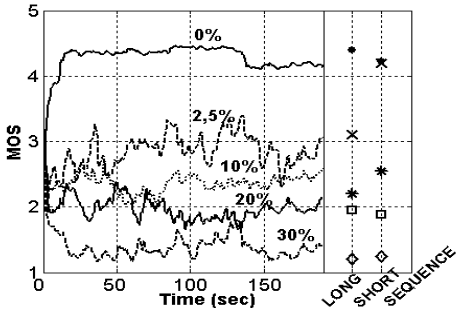


Fig. 2: Mean instantaneous judgments and mean opinion scores (with 3-min and 8 sec speech sequences) obtained for 0%, 2.5%, 10%, 20%, 30% of packet losses.

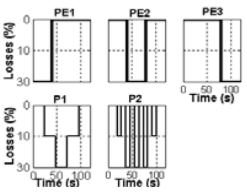
So, standard methods using short speech samples are suitable to standard impairments but not to impairments characterizing un-guaranteed QoS networks as packet losses. In order to measure perceived quality for such networks, it is advisable to use long speech sequences. Moreover, a method of continuous judgment is very useful for studying the impact of such impairments at any time.

4. Relationship between instantaneous and overall quality judgments

In addition, the double protocol was used to study the relationship between instantaneous and overall quality judgments and so to determine the impact of quality fluctuations on overall quality judgment. For example, we presented to 24 subjects a 3-min speech sequence with a period of degradation (30 % of packet losses) at the beginning, at the middle and at the end of the sequence. The duration T of this period of degradation could be 15 s, 30 s or 1 min. Subjects were asked to assess their quality continuously and at the end of each sequence, according to the double protocol (see §.2). Results showed that overall judgments were highly correlated to the averages of instantaneous judgments (r = 0.97, $p<0.05$). Of course, the longer the degradation, the worst the overall judgment. A variance analysis (ANOVA) revealed a significant effect of the duration T (F(2,46) = 28.33 $p<0.001$). However, overall judgements seemed to be influenced by a recency effect, *i.e.* subjects were more influenced by the last moments of the sequence: the more the degradation appears at the end of the sequence, the worst the overall judgement. The ANOVA confirms this observation since the location effect is significant (F(2,46) = 11.86 $p<0.001$). The overall judgment failed of about 0.6 MOS when the degradation came from the beginning to the end of the sequence, all durations considered. So the overall judgment seems to be determined by an average of the instantaneous judgments, weighted by mnesic processes.

5. Influence of semantic and verbal contents

In order to study the influence of the semantic content on quality judgments, we carried out an experiment in which speech signals in French and Greek were presented to listeners. Two-min speech signals (one in French, and its translation in Greek) were used. These two signals were impaired with three percentages of packet losses: 0%, 10% and 30%. Then five quality profiles *i.e.* five quality evolutions were realised by concatenation of extracts of the two signals at 0, 10 and 30% (see Fig.4)



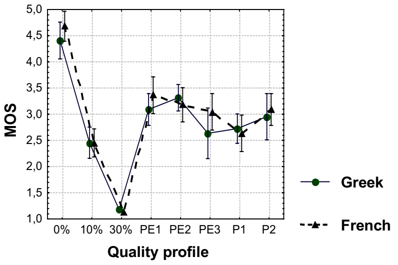


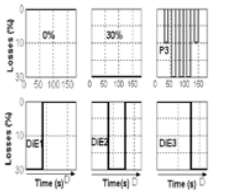
Fig. 4 *Fig.5*

For each presented sequence, subjects were asked to assess its quality continuously and at the end, according to the double protocol.

Figure 5 shows MOS obtained at the end of each sequence according to the quality profile (abscissa) and to the language (parameter). It appears that MOSs are mainly influenced by the time distribution of degradation, whatever the language considered. And MOSs obtained with the two languages are quite similar.This observation is confirmed by a variance analysis that revealed a significant effect of the factor Quality profile (F(7,147)=63.53 *p<0.0001*) and no effect of the factor Language (F(1,21)=1.58 *p=0.22*). Therefore, in a quality measurement context, quality judgements seem to be mainly elaborated on the basis of the time distribution of impairments and the semantic content has a weak impact.

6. Role of the motor task

In order to study the impact of the motor task involved by the continuous judgment with the cursor, we realised an experiment in which we presented speech sequences already assessed in previous experiments with the double protocol. These speech sequences were degraded according to a few quality profiles (fig.6). In this experiment, subjects were asked to give only an overall judgment at the end of each sequence.



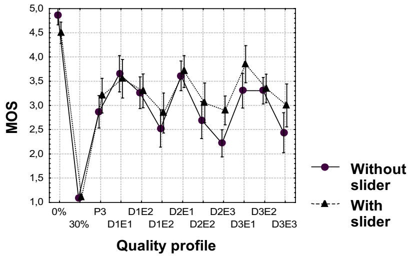


Fig. 6 *Fig. 7*

On the figure, one can see MOS obtained according to the quality profiles and to the procedure (with or without cursor).It appears that MOSs are mainly influenced by the time distribution of degradation (that is by quality profile). And MOSs obtained with and without continuous judgment are quite similar. This result is confirmed by a variance analysis that revealed a significant effect of the factor Quality profile (F(11,451)=61.88 *p<0.05)* and no effect of the factor Protocol (F(1,41) = 3.38 *p = 0.062*). The method of continuous judgment can then be used without biasing the overall judgment.

7. Influence of the environment

Finally, in order to study the influence of a real environment, we carried out an experiment in two steps. First, a listening test was run in outside conditions. Subjects were seated in a noisy and bustling place, with cars and people passing by. Subjects heard a speech sequence through a mobile. The quality could be degraded applying a decay on the mobile radio reception level. The decay was applied according to 5 quality profiles: no decay, decay during all the sequence, or decay at the beginning, the middle or at the end of the sequence. At the end of each sequence, subjects were asked to give a quality score on the standard MOS scale. Moreover, each sequence was recorded at the mobile output in order to have the sequence such as the subject heard it. In a second time, these recorded sequences were reproduced in laboratory through handsets, and were rated by other listeners with the double protocol. The figure shows MOSs obtained according to the quality profile and to the environment (real environment or laboratory).

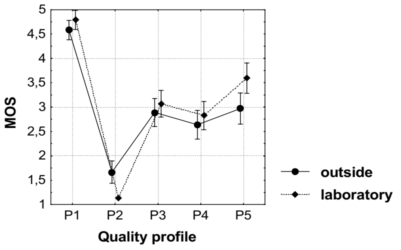


Fig. 8: MOS according to the quality profile (abscissa) and to environment (parameter).

Once again, MOSs are mainly influenced by the time distribution of degradation, that is by quality profile (F(4,140)=270.8 *p<0.001*). And MOSs obtained in the two environments are quite similar (no effect of environment F(1,35) =1 *p=0.32*)

8. Conclusion

As a conclusion, we have a methodology suited to the problem of time-varying speech quality, that allows to study the impact of new impairments both on instantaneous quality judgment and overall quality judgment. Moreover this method allows a subjective quality measurement representative of a subjective measurement that could be realized in a real environment. Additionally, what we learned from this study is that quality judgments are mainly elaborated on the basis of time distribution of acoustic manifestations of impairments. And in a context of measurement, semantic content has a weak impact. Probably, in a real situation of communication, semantic aspects become more important. However, this result validates the use of objective measure instruments. In addition, it seems that the overall judgment is determined by an average of instantaneous judgments, weighted by a recency effect. This process could be modelled and integrated in an objective measure instrument. Nevertheless, all these results were obtained in a listening context. Further studies will deal with the influence of the listening test and more generally the cognitive resources mobilization by other activities on quality judgments.

References:

ITU-T P. 800 (1996). Methods for subjective determination of transmission quality.

Gros, L., Chateau, N. (2001). Instantaneous and overall judgements for time-varying speech quality: Assessments and relationships, acta acustica Acustica, 87, 367-377.

ITU-R-BT-500-8 (1998). Methodology for the subjective assessment of the quality television.

ITU-T P.810 (1996). Modulated Noise Reference Unit (MNRU).

Acknowledgment

I would like to thank Noel Chateau (FTR&D) for his precious support during all this research, and also Sylvain Busson for his important help during the experiment described in § 7.