euronoise

Effect of degradations' distribution in a corpus test on auditory ratings

Nicolas Côté and Virginie Durin

France Télécom, 2 avenue Pierre Marzin, 22300 Lannion, France
nicolas.cote@orange-ftgroup.com

Speech quality of telecommunication systems is usually evaluated thanks to auditory tests, which must be carried out in accordance with ITU-T Recommendations. In these tests, subjects are asked to assess the quality of speech sample by giving a score on a five-level scale. The averaging of subjects' scores yields the Mean Opinion Score (MOS) which represents the speech quality for a given condition. However, MOS values can be strongly influenced by many factors. In this paper, we focus on a specific bias, the distribution of the impairments in the corpus test by considering three degradation types: bandwidth (or frequency content), continuity and noisiness. In the specific case of frequency content, the rating of a narrow-band condition may have a lower quality score in a mixed-band corpus (mixed of narrow-band and wideband conditions) than in a purely narrow-band corpus. Consequently, the validity of MOS values is theoretically limited within a test which prevents MOS comparison between different tests. Finally, two suggestions are proposed to limit or avoid this effect: first, an improvement of auditory tests methodology and then, a new approach to assess speech quality, based on the subject behaviour.

1 Introduction

1.1 Speech Quality

Speech quality perception is a complex phenomenon to define. According to the point of view of Blauert and Jekosch [1], perceived speech quality is a process that consists in comparing subject perception with its expectations, referents, knowledge etc. The result of this process is quality related to the speech event. Although all these parameters are rarely taken into account in other speech quality definition, a consensus exists between authors: speech quality is considered as a multidimensional object. For example, for Wältermann [2], speech quality can be described according to three dimensions: frequency content, noisiness and continuity.

However in the most widespread methodologies used by operators, speech quality is considered as an unidimensional object. The subjective test methodologies are described in ITU-T P. serie, especially in P.800 [3]. In listening-only tests (LOTs), subjects listen to speech samples processed by the system under study, and are asked to assess their overall quality by giving a score on a five-level discreet category scale such as 'Excellent', 'Good', 'Fair', 'Poor' and 'Bad'. The experimenter preannotated these categories with the scores 5, 4, 3, 2, 1 and assumed that each intervals occupies the same perceptual interval. Thus, a Mean Opinion Score (MOS) is computed by averaging the individual scores.

1.2 Factors influencing MOS

Many factors could influenced the MOS values. All these factors are described by Möller in [4]. However, some of these factors are briefly introduced in this part.

- The traduction of the categories' names influences the MOS values. For example, Zielinsky et al [5] showed that the semantic difference between the terms 'Poor' and 'Bad' is not equal to their translated equivalent. Consequently subjective judgments are different according to the country.
- The subject has also a personal view of the perceptual dimensions involved in the overall speech quality. For a specific condition, two subjects could rate it differently which results in a high standard deviation. However, this effect is attenuated by using a large number of subjects, set to 24 to 32 in ITU-T methodologies.
- The number of categories on the judgment scale. Even though a discrete 5-point scale seems to be the preferred scale in terms of 'ease of use', a 5-point MOS scale has a relatively low sensitivity [6]. A continuous scale may be used for rating speech quality. However, Malfait in [7] describes that subjects mostly judge the stimuli on the numbers or the name of the categories even on a continuous scale.
- There are other effects due to the scale, as the 'saturation effect'. The naïve subjects do not use the extreme categories of the scale. All these effects due to the scale are described in [8].
- The just precedent stimulus has an influence on the judgement of the actual stimulus. This is called the order effet. This effect may be attenuated using different listening order for each group of 4 or 8 subjects.
- The subjects expectations, involved during the assessment of the stimuli, correspond to its own mean overall experience in telecommunications. Consequently, judgments varied according to the personal internal reference. To avoid this effect some stimuli are listened during a training period before the experiment. These stimuli are called anchors.
- The range of degradations included in the corpus test and its distribution has an influenced on the MOS values. This is called corpus effect.

The context effect includes the last three points. It is a strong effect of many judgment scales. This paper is focused on the corpus effect on the resulting MOS values. Few studies have reported the influence of these factors on MOS values and especially the influence of degradations in the corpus test on MOS values. Möller and Raake [9] studied the effect of condition bandwidth on MOS. In other words, they studied the corpus effect by considering only frequency content of the corpus. This paper focuses on the corpus effect by considering three cited dimensions, *i.e.* the effect of bandwidth, continuity and noisiness of test conditions on MOS values.

Some biaises may be reduced by means of transformations on the MOS values, using a simple linear normalisation computed by:

$$MOS_{norm,i} = \frac{MOS_i - MOS_{min}}{MOS_{max} - MOS_{min}} * 3.5 + 1 \qquad (1)$$

where i corresponds to the current condition, MOS_{min} and MOS_{max} are the minimum and maximum MOS value of the corpus and $MOS_{norm,i}$ the resulting normalized MOS value. A second proposition is to use reference conditions in the corpus-test. ITU-T organism decided to introduce in each corpus reference conditions to simulate speech codecs degradations. The procedure to create these conditions called Modified Noise Reference Unit (MNRU) is described in [10, 11]. A parameter called Q defined in dB is used to quantify the degradation in this reference condition. This value corresponds to a signal to noise ratio. In this specific case, the noise is correlated to the speech signal. However, nowadays, MNRUs do not represent the current diversity of degradations and are not sufficient to reduce the corpus effet.

Consequently the choice of the corpora in Section 2 takes into account the points described above. These corpora are used to evaluate the biases due to the corpus effect in Section 3. Finally, suggestions are made to limit or avoid the corpus effect in Section 4.

2 Corpora

Many corpora are carried out during competitions conducted by normalization organisms (such as ITU-T or ETSI) aiming at select standard speech codecs. For our study, commun conditions evaluated in different corpora are compared. Because of the influence of criteria described above (Section 1.2), comparisons are made for corpora coming from the same standard competition involving the same methodology and same language. The selected corpora are described in Table 1. There is a comparison of corpora for each perceptual dimension under study.

For Frequency Content dimension, two corpora are compared. The first one, the narrow-band corpus, includes only Narrow-Band conditions (NB, [300-3400 Hz]). In the second one, the mixed-band corpus, WideBand conditions (WB, [50-7000 Hz]) are presented with NB conditions. In addition the corpora are especially carried out to study this specific corpus effect.

The corpora used for Noisiness dimension correspond to the ITU-T G.729 [12] competition. In this case two corpora are compared, a first one with only clean speech, a second one including clean speech, noisy speech signals (*i.e.* speech transmitted with an ambient noise at the sending side) and transmission errors. Note that 'clean' refers to non-noisy condition but also includes codec or MNRU degradations. In addition, each corpus is assessed in three languages.

For Continuity dimension, the corpora were carried out during the ETSI competition for the AMR-NB speech codec. Results come from two comparisons and three corpora: one is compared to the two others. The first corpus includes only error-free conditions and the second and third ones include error-free conditions and conditions impaired by transmission errors. There are forteen conditions in common between the first and the second corpora, and ten conditions in common between the first and the third corpora.

3 Results

In this section, results show the influence of the corpus effect. This section is divided in three parts according to the perceptual dimensions described in Section 1.1.

3.1 Frequency content

The figure 1 shows the MOS values of the narrow-band conditions in each corpus. In the purely narrow-band corpus, the non-degraded NB condition is not perceived as being degraded, which results in a mean quality of 4.41 MOS (cross point in the right hand corner). In a mixed-band corpus, where high quality wideband conditions are introduce, NB conditions are rated lower (3.88 MOS) than in the purely NB corpus. If MOS values were the same in two corpora, the points should align according the function $y = x$. However, a compression of the MOS values in mixed-band corpus is observed. A exponential relationship between the two corpora is estimated. The obtained mapping function corresponds to:

$$MOS_{NB} = 14.45 * (\exp(\frac{MOS_{MB} - 1}{13.50}) - 1) + 1 \quad (2)$$

where MOS_{MB} and MOS_{NB} correspond to the MOS values in the mixed-band corpus and the narrow-band corpus respectively. An ANalysis Of VAriance (ANOVA) with the three factors *talker*, *condition* and *corpus* reveals that the factor *corpus* is significant ($F(1, 72) = 41, 17$, $p < 0.001$). This confirms significant differences in MOS values between both corpora.

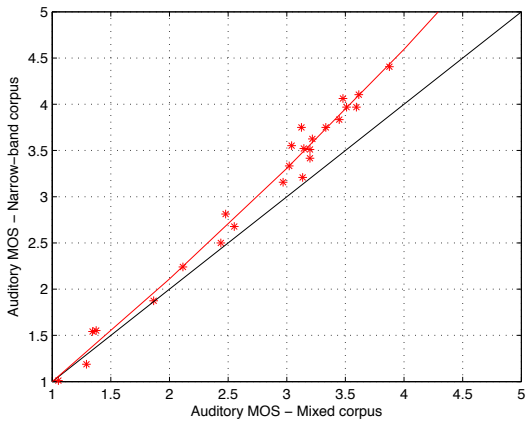

Figure 1: Results of narrow-band conditions in a purely narrow-band and mixed-band corpora.

3.2 Noisiness

The figure 2 shows the results of the clean conditions in the clean and mixed (here clean + noisy speech) corpora. It can be observed higher MOS values of clean conditions in the mixed corpus than in the clean corpus. Here, noisy conditions which have a lower quality than clean conditions are introduced in the mixed corpus. Consequently, clean conditions of the mixed corpus are judged higher than in the clean corpus. The difference between both tests corresponds to a constant

Dimension	Frequency Content	Noisiness	Continuity
Name of the competition	Q.8/12	G.729	AMR-NB
Year	2004	1995	1999
Language	French	Japan, Engish, French	German
Test localisation	FT–R&D	NTT, Nortel, FT–R&D	Berkom
Number of corpus	2	2	3
Reference corpus	NB	Clean	Error-free
Mixed corpus	NB + WB	Clean + Noisy	Error-free + Transmission errors
Common conditions	25	30	10 and 14
Degradation types of common conditions	Codec (G.729, G.711,G.726)	Codec (G.729, G.726), MNRU	Codec (AMR-NB, GSM-EFR), MNRU

Table 1: Summary of the three comparisons of auditory corpora.

offset. An ANOVA with the four factors *talker*, *corpus*, *language* and *condition* is conducted. The factor *corpus* is significant ($F(1, 54) = 46.35$, $p < 0.01$). Therefore ANOVA confirms significant differences in MOS for common conditions assessed in different corpora. Distance between each points from two corpora are measured (σ). The resulting corpus effect seems to be less important in Noisiness dimension ($\sigma = 0.055$) than in Frequency Content dimension ($\sigma = 0.069$).

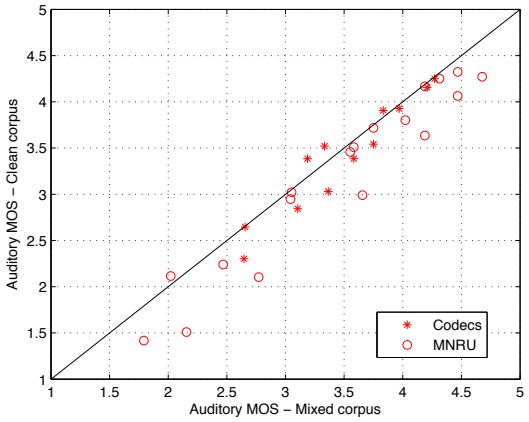

Figure 2: Results of clean conditions in a clean and mixed (clean and noisy) corpora.

3.3 Continuity

The figure 3 shows the results of the error-free conditions in error-free and mixed (error-free and transmission errors) corpora. As Noisiness dimension, it can be observed that common conditions (*i.e.* error-free) are rated higher in mixed corpus than in the error-free corpus. Two ANOVA are conducted for each comparison, with two factors *corpus* and *condition*. For the first comparison (corpus 1 and 2), *corpus* is significant ($F(1, 13) = 12.24$, $p < 0.01$). For the second comparison (corpus 1 and 3), *corpus* is also significant ($F(1, 9) = 22.14$, $p < 0.01$). Whatever the comparison, ANOVA confirms the corpus effect. Distance between each points from two corpora are measured (σ). The resulting corpus effect seems to be more important in Continuity dimension ($\sigma = 0.076$) than in Frequency Content dimension.

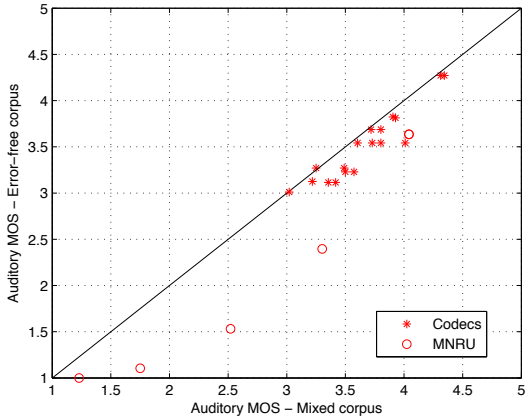

Figure 3: Results of error-free conditions in a error-free and mixed (error-free and transmission-errors) corpora.

3.4 Summary

Based on results for the three dimensions, introduction of high-quality conditions in a reference corpus leads to a 'negative' corpus effect: MOS values of conditions in the reference corpus are significantly lower (case of Frequency Content). On the contrary, introduction of low-quality conditions in a reference corpus leads to a 'positive' corpus effect: MOS values of the reference corpus are significantly higher (case of Noisiness and Continuity). For the latter case, in mixed corpora, subjects seem to pay less attention to degradations due to low bit-rate speech codecs than in the reference corpora. From a perceptually point of view, transmission errors or background noise are the dominent degradations and they likely mask the speech codec distortions. To reduce the corpus effect in Noisiness dimension, ITU-T organism decided to use the Degraded Category Rating (DCR) methodology [3] instead of using the usual method described in 1.1 to assess the speech quality of noisy speech conditions. In DCR methodology there is no more corpus effect since the stimuli are presented to the listeners by pairs A-B in which A is the reference and B the degraded sample.

In addition, the corpus effect seems to be asymmetric. The effect in Noisiness dimension seems to be weaker than the effect in Frequency Content dimension. On the contrary, for Continuity dimension, the effect is stronger than in Frequency Content dimension. How-

ever, it can be explained by the presence of MNRU conditions in Continuity corpora. In order to differentiate the 'negative' and 'positive' corpus effect a new auditory test could be carry out. A new corpus including a subset of the reference Frequency Content corpus (NB) and conditions impaired by smaller bandwidth is needed.

4 Discussion

In order to rule out this specific distortion of the auditory results two proposals are described in this part. Firstly, reference conditions could be introduced in the corpus-test. The MNRU are used to only simulate a specific coding degradation (logarithmic PCM coding). References should described the whole speech quality dimensions as expressed by [2]. In addition it is shown that Loudness is also a relevant perceptual dimension of speech quality [13]. Nine references could be used in all auditory tests corresponding to two reference points per quality dimensions and the non-degraded full-band condition. For each reference points, the degradation impacts only one perceptual dimension. These nine references are described in Table 2. These references would rule out the biases described in part 3 and thus improve the reliability of the derived MOS values.

Name	Description
Full-band	non-degraded
Noise 1	Hoth noise 20 dB SNR
Noise 2	Hoth noise 12 dB SNR
Bandwidth 1	NB, IRS
Bandwidth 2	WB, P.341
Continuity 1	Clipping 2 %
Continuity 2	Clipping 10 %
Loudness 1	Attenuation 10 dB
Loudness 2	Attenuation 20 dB

Table 2: Description of the proposed nine references.

A second possibility to avoid the dependence of MOS values to the distribution of qualities within the corpus is to use another subjective methodology. This methodology takes into account several drawbacks of current methodologies. Through the introducing of this complementary methodology a new approach of speech quality is proposed.

Of course, the MOS values dependence on the distribution of qualities in the corpus is first considered. As it is widely described in this paper, it is a strong drawback since it prevents comparisons of MOS values between different tests. Moreover, one of the principles of the new methodology is to consider that explicit judgements of subjects in current methodologies can biased the speech quality percept [14]. Furthermore, these judgements are not realistic since speech quality assessment is rarely a conscious object in real life, except in cases in which quality is so degraded that communication becomes impossible. In addition, 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. These arguments are detailed in [15]. 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 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.

The chosen tasks are two overlapped tasks: a digit memory recognition task (based on Sternberg's task [16]) and letter recognition task. Three different quality levels are applied to audio signals describing digits and letters. Reaction times and errors rates of subjects are measured. For more details on the methodology, see [15]. Results show a quality effect: the more the quality is impaired, the more the performance decreases (*i.e.* reaction times lengthen and error rates increase). This methodology enables to discriminate the three quality levels. Contrary to conscious quality judgment, it can be assumed that reaction times and errors rates do not depend on the distribution of qualities within the test corpus since there is no test corpus anymore (for example each new quality level could be measured with the non coded reference only). Then, a new quality scale giving an absolute score based on reaction times and error rates may be considered. Other relevant variables could be added to reaction times and errors rates for a more accurate measure.

References

[1] J. Blauert and U. Jekosch, "Sound-quality evaluation a multi-layered problem," *Acta Acustica united with Acustica*, vol. 83, pp. 747–753, Sep 1997.

[2] M. Wältermann, K. Scholz, A. Raake, U. Heute, and S. Möller, "Underlying quality dimensions of modern telephone connections," in *INTERSPEECH*, 2006, preprint 1089.

[3] ITU–T Rec. P.800, *Methods for Subjective Determination of Transmission Quality*, International Telecommunication Union, Geneva, 1996.

[4] S. Möller, *Assessment and Prediction of Speech Quality in Telecommunications*, Kluwer Academic Publ., Boston MA, 2000.

[5] Zieliński S., Brooks P., and Rumsey F., "On the use of graphic scales in modern listening tests," in *Proc. 123rd Audio Eng. Soc. Conv.*, New York US, Oct. 2007, preprint 7176.

[6] Preston C.C. and Colman A.M., "Optimal number of response categories in rating scales: Reliability, validity, discriminating power, and respondent preferences," *Acta Psychologica*, vol. 104, no. 1, pp. 1–15, 2000.

[7] ITU–T COM 12 Contr. C.120, *Investigating the Proposed P.OLQA Subjective Test Method*, International Telecommunication Union, Geneva, Sep 2007.

[8] E. C. Poulton, "Models for biases in judging sensory magnitude," *Psychological Bulletin*, vol. 86, no. 4, pp. 777–803, 1979.

[9] S. Möller, A. Raake, N. Kitawaki, A. Takahashi, and M. Wältermann, "Impairment factor framework for wide-band speech codecs," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 14, no. 6, pp. 1969–1976, Nov. 2006.

[10] B. H. Law and R. A. Seymour, "A reference distortion system using modulated noise," *IEE*, pp. 484–485, Nov 1962.

[11] ITU–T Rec. P.810, *Modulated noise reference unit (MNRU)*, International Telecommunication Union, Geneva, 1996.

[12] ITU–T Rec. G.729, *Coding of speech at 8 kbit/s using conjugate-structure algebraic-code-excited linear prediction (CS-ACELP)*, International Telecommunication Union, Geneva, 2007.

[13] N. Côté, V. Gautier-Turbin, and S. Möller, "Influence of Loudness Level on the Overall Quality of Transmitted Speech," in *Proc. 123rd Audio Eng. Soc. Conv.*, New York US, Oct. 2007, preprint 7175.

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

[15] V. Durin and L. Gros, "Reaction times and performances in recognition tasks to assess speech quality," in *Audio Engineering Society 124th Convention*, Amsterdam, The Netherlands, 17-20 May 2008.

[16] S. Sternberg, "Memory scanning: Mental processes revealed by reaction-time experiments," *American Scientist*, vol. 57, no. 4, pp. 421–457, 1969.