Question(s):	VQEG	**Meeting, date:**	April. 25-29, 2005
Study Group:	**Working Party:**	**Intended type of document** (R-C-D-TD):	
Source:	NTT (Nippon Telegraph and Telephone Corporation), Japan		
Title:	Proposal of adjustment and aggregation procedures for subjective and objective quality assessment values in multi-laboratory comparative test		
Contact:	Takaaki KURITA NTT Japan		Tel: +81-422-59-6936 Fax: +81-422-59-5671 Email: kurita.takaaki@lab.ntt.co.jp
Contact:	Jun OKAMOTO NTT Japan		Tel: +81-422-59-6526 Fax: +81-422-59-5671 Email: okamoto.jun@lab.ntt.co.jp

Please don't change the structure of this table, just insert the necessary information.

Summary

This contribution proposes including a common set of processed video sequences (PVSs) in subjective quality assessment experiments conducted by multiple testing labs to achieve fair comparisons of the proposed objective quality models. In addition, it proposes that when there are differences in subjective quality among multi-laboratory experiments, the subjective and objective assessment values of all experiments should be adjusted before the proposed models are compared.

1. Introduction

In the performance test of the VQEG multimedia group, the number of processed video sequences (PVSs) (= no. of SRCs x no. of HRCs) will become enormous because there are many SRCs (source clips) and HRCs (test conditions). Therefore, PVSs will be divided up in the Independent Laboratory Group (ILG) and proponent laboratories, and subjective quality assessment experiments at those laboratories will be conducted. The data processing procedures for both subjective and objective assessment values will directly affect the performance comparison among proposed models. Therefore, deciding the procedures is a very important issue.

In this contribution, first, we discuss the problems with experimental results in a multi-laboratory comparative test and the problems with the current proposed aggregation procedures. Next, we propose adjustment and aggregation procedures for subjective and objective quality assessment values in order to achieve fair comparisons of the proposed objective quality models.

2. Problems with multi-laboratory comparative test

Generally, there are many cases where subjective assessment values were different even when the experiments used the same test conditions, because of differences in languages, countries, and laboratories. One of the causes is the "language dependence" of quality-scale terms [1]-[3]. The feelings for terms differ among languages or countries (Appendix 1-A). Moreover, it was reported that the Japanese and Western (Canadian English, French, and German) MOS (mean opinion scores) values were different for the speech signals with the same degradation conditions (Appendix 1-B) [4]. Also in the VQEG phase I test, differences in assessment values for the same HRCs among laboratories were indicated (Appendix 1-C) [5].

These differences in assessment values among laboratories (hereafter “between-lab effects”) may appear more conspicuously because of the individual lab experiments using different PVSs. Therefore, it is necessary to confirm these effects in experiments to maintain the consistency and reliability of experimental results. We think that VQEG should consider “between-lab effects” so that the experiments will be valuable.

3. Problems with the current proposed procedures

To date, two basic plans have been proposed for the aggregation procedures to compare the performance among objective quality models. Plan 1 processes the experimental results of laboratories individually (Individual processing method). Plan 2 integrates all experimental results of laboratories without adjustments (Integration processing method). The problems with these two plans are discussed below when there are between-lab effects.

Plan 1: Individual processing method

1) Each experiment may choose a different best model.
2) The number of PVSs assessed in each experiment is not considered. Therefore, even the best model in an experiment that has a small number of PVSs will equivalent to the best model in other large-scale experiments.
3) In any models, the accuracy of quality estimations will decrease because "between-lab effects" are not taken into account.

Plan 2: Integration processing method

1) A model of the proponent laboratory, which will assess the largest number of PVSs, will gain an advantage because of the “language dependence” (see note 1).
2) Same as Plan 1-3.

Note 1: NTT will be ready to conduct a large-scale subjective assessment experiment using about 1000 PVSs.

4. Proposal

To resolve the above problems, we propose the following two-step procedure. Step 1: a common set of PVSs will be included in the subjective assessment experiments of multiple testing labs and the differences in qualities among the experiments will be examined. Step 2: when there are differences, subjective and objective quality assessment values of all experiments will be adjusted and we will then compare objective quality models. The details are described below.

Step 1: Examination of the differences of quality characteristics among experiments

1-1) Subjective quality assessment including a common set of PVSs will be performed.

Spatially and temporally degraded processed sequences whose quality covers a wide range of MOS (1-5) should be used as a common set of PVSs.

1-2) Statistical analysis of the difference in quality characteristics of common PVSs among experiments will be performed.

When there are no statistically significant differences among experiments, experimental results will be aggregated without adjustments, and then models will be compared.

Step 2: Adjustment for the subjective and objective assessment values of experiments

2-1) Calculate reference quality characteristics.

Reference quality characteristics as the standard of adjustment will be calculated by averaging the DMOS of common PVSs per PVS (refer to appendix 2).

2-2) Adjust the subjective assessment values and aggregate to one subjective quality database.

A mapping function, such as a polynomial function, between the reference quality characteristics and the quality characteristics of common PVSs will be calculated for each experiment. Then the subjective assessment values of PVSs (PVSs other than those of the common set) will be adjusted by using each mapping function. The adjusted subjective assessment values will be aggregated to one subjective database.

2-3) Adjust the objective assessment values of proposed models.

The objective assessment values of the proposed models will be adjusted by using the above mapping function for each experiment of the proponent laboratory.

2-4) Compare the performance of proposed models.

The performance of proposed models will be compared by using an aggregated subjective assessment database and adjusted objective values.

Summary

We proposed a two-step procedure to decrease the "between-lab effects" and achieve a fair comparison of models. The first step is multi-laboratory experiments including a common set of PVSs and the second step is adjustment for subjective and objective assessment values.

References

[1] ITU-R Report BT.1082-1, “Studies toward the unification of picture assessment methodology,” 1990.

[2] B. L. Jones, P. R. McManus, “Graphic scaling of qualitative terms,” SMPTE J., pp. 1166-1171, Nov. 1986.

[3] N. Narita, “Graphic scaling and validity of Japanese descriptive terms used in subjective-evaluation tests,” SMPTE J., pp. 616-622, July 1993.

[4] ITU-T Delayed Contribution COM12-D146-E, “Performance evaluation of E-model and proposed modifications,” Sep. 2003.

[5] ITU-T Contribution COM9-80-E, “Final report from the video quality experts group on the validation of objective models of video quality assessment,” June 2000.

[6] J. P. Guilford, “Psychometric methods,” pp. 223-241, McGraw-Hill, N.Y., 1958.

Appendix 1 Difference in quality assessment values among laboratories

A: Influence of language dependence on rating terms

The perceived positions and intervals of quality-scale terms on the successive scale obtained using a graph-scaling method are shown in Fig. 1 [1]-[3]. The perceived position and interval of terms (Japan, Germany, USA, France, and Italy) are different in different countries. In particular, the intervals between “poor” and “bad” are very close in the USA, France, and Italy, whereas they are almost equally spaced in Japan and Germany.

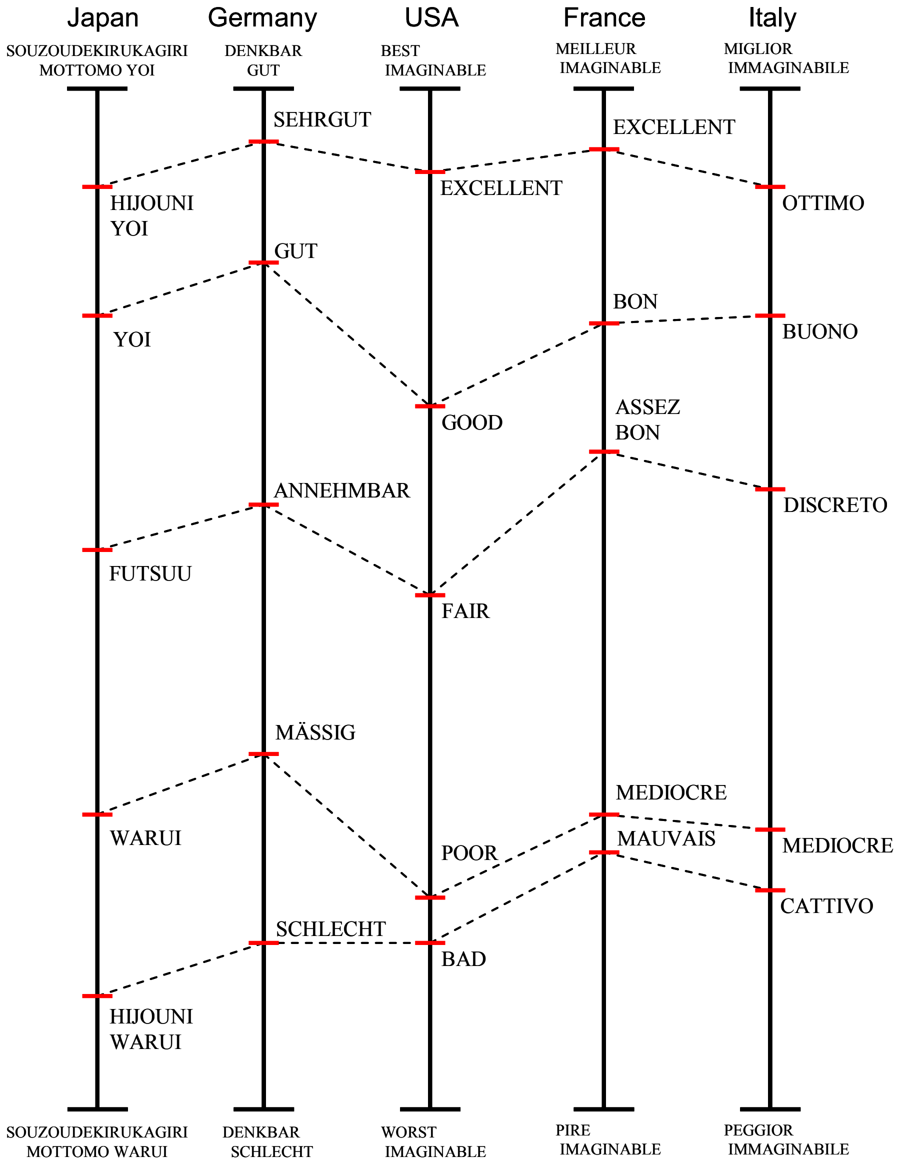


Figure 1 Graphic scales of quality-scale terms from ref. [1]-[3].

B: Difference in speech quality assessment values among countries

The differences between the Japanese and Western (French, Canadian English, and German) MOS values for speech signals with the same degradation conditions are shown in Fig. 2 [4].

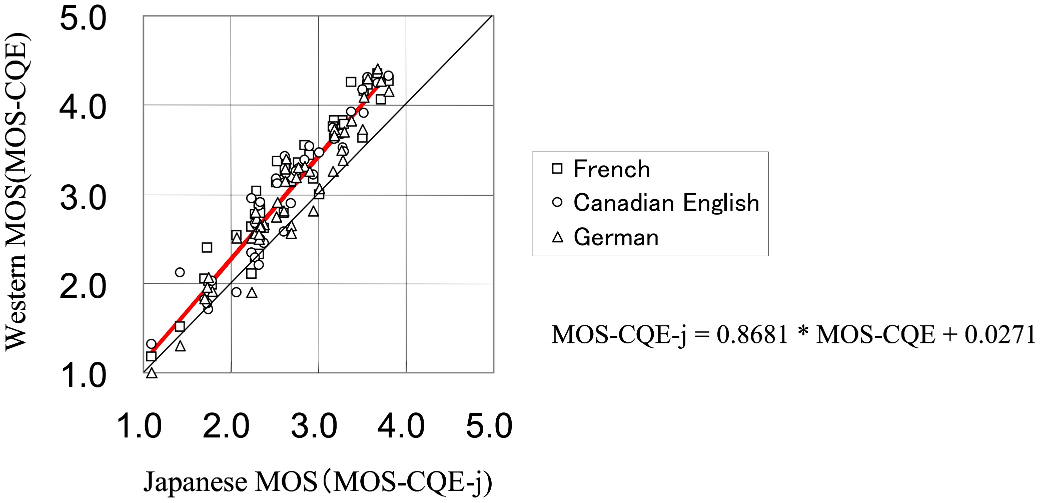
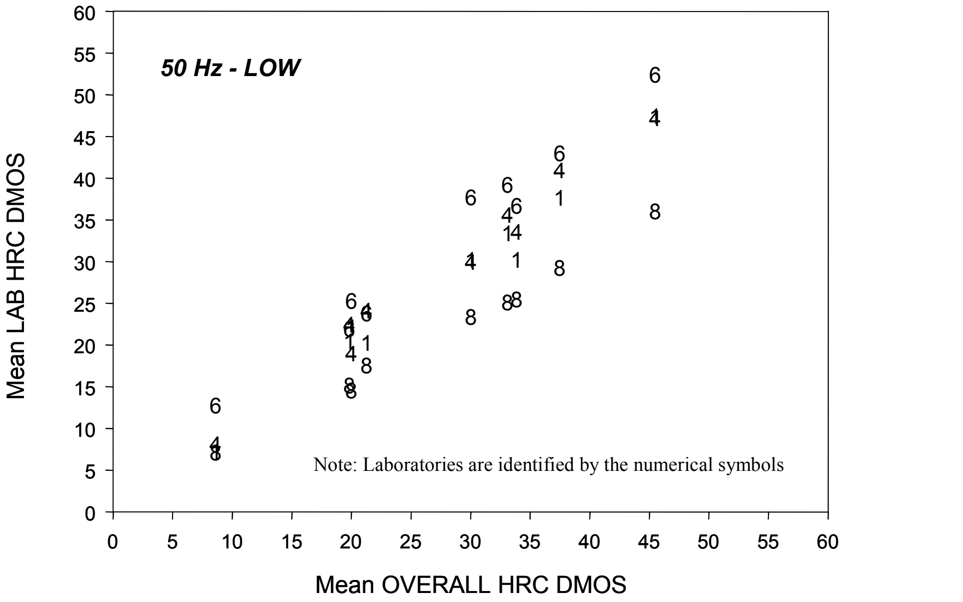


Figure 2 Relationship between Western and Japanese MOS from ref. [4].

C: Difference in video quality assessment values among laboratories

The assessment result in the VQEG Phase I test is shown in Fig. 3. Differences in quality assessment values among laboratories for the same HRCs were reported [5].



Note: Laboratories are identified by the numerical symbols

Figure 3 Difference of DMOS between laboratories in Phase I test from ref. [5].

Appendix 2   Types of scores for reference quality characteristics

It may be possible to use two types of scores for reference quality characteristics. Discussion and a confirmation test will be necessary. In particular, MOS is an ordinal scale, so its values do not necessarily represent exact differences between perceived magnitudes. Interval scale values (psychometrics values) should be constructed from the MOS or DMOS values by utilizing the law of categorical judgment [6].

Type 1:   Averaging DMOS

Reference quality characteristics will be calculated by averaging the DMOS of common PVSs for each PVS.

- Advantage: The calculation is simple.
- Disadvantage: A model of the proponent laboratory, which will assess the largest number of PVSs, may gain an advantage because the adjustment width is small.

Type 2:   Averaging psychometrics values

DMOS will be transformed to successive scale values (psychometrics values) for each experiment and will then be averaged for each PVS.

- Advantage: It is hard to depend on a specific experiment. This way we can expect fair transformation.
- Disadvantage: A performance comparison will be performed on successive scale values, not on the DMOS scale. The transformation to successive scale values is a complicated calculation [6].