**COMMITTEE T1 - TELECOMMUNICATIONS STANDARDS CONTRIBUTION**

| | | |
|---|---|---|
| DOCUMENT NUMBER: | T1A1.7/97-042 | |
| T1BBS FILES: | 7a170420.doc, 7a170420.pdf | |
| DATE: | October 22, 1997 | |
| STANDARDS PROJECT: | T1A1-17 | |
| TITLE: | Summary of objective audio quality measure performance data presented to T1A1 | |
| SOURCE: | Institute for Telecommunication Sciences<br>National Telecommunications and Information Administration<br>U.S. Department of Commerce | |
| AUTHORS: | D.J. Atkinson and Stephen Voran | |
| CONTACT: | D.J. Atkinson<br>NTIA/ITS.N3<br>325 Broadway<br>Boulder, CO 80303-3328 | <br>Tel +1 303 497 5281<br>Fax +1 303 497 5323<br>email dj@its.bldrdoc.gov |
| DISTRIBUTION: | T1A1.7, T1BBS | |

ABSTRACT: This contribution aggregates the available performance data on the MNB and P.861 objective speech quality measures. Specifically, results presented in contributions T1A1.7/97-032 and T1A1.7/97-034 are examined. Based on examination of the aggregated data presented in these two contributions, recommendations are made relating to the advancement of the work of project T1A1-17 and the development of a new ANS on objective audio quality measures.

## Introduction

The purpose of this contribution is to further the work outlined in T1A1.7 Standards Project 17, titled: "Objective measures for the assessment of audio quality." (see T1A1.7/97-020R1.) Specifically, section 2.2.1 identifies 4 proposed areas of study, the first of which involves the identification, evaluation and standardization of an objective algorithm for use in assessing the quality of telephone-band (300-3400 Hz) speech signals. This contribution seeks to enable the identification and evaluation tasks by gathering and summarizing available results on the behavior of the Perceptual Speech Quality Measure (PSQM) family of objective algorithms and a Measuring Normalizing Block (MNB) algorithm. It is expected that this information can significantly facilitate the completion of those tasks. The text of project T1A1-17 suggests that a draft standard should be available by Q4, 1997. All information in this contribution was originally provided to T1A1.7 in contributions T1A1.7/97-032 and T1A1.7/97-034.

## Correlation vs. RMS error

There has been significant amounts of discussion on how best to determine how closely the results of objective algorithms that assess speech quality agree with subjective test results. Specifically, in previous discussions during meetings of T1A1.7, the choice between correlation coefficient and RMS error has been discussed. In fact, both forms of comparison are available to T1A1.7 through contributions T1A1.7/97-032 and T1A1.7/97-034. Contribution T1A1.7/97-032 evaluates eight different objective algorithms and objective measure performance is computed by calculating coefficients of correlation between objective algorithm results and MOS results for the eight different subjective tests. Contribution T1A1.7/97-034 evaluates five objective algorithms, and objective measure performance is evaluated by applying a fitting procedure to map all objective results onto the MOS scale. Then RMS errors are calculated in the MOS domain for 12 different subjective tests.

While RMS error does bring different information than the coefficient of correlation, there is a relationship between the two statistics. Specifically, if the relationship between objective and subjective values is modeled as an independent additive noise source, then the correlation $\rho$ and the RMS error (RMSE) are related by

$$\text{RMSE} = \sigma \cdot \sqrt{1-\rho^{2}} \quad ,$$

where $\sigma$ is the standard deviation of the subjective scores. Thus, for a *single* subjective test, $\sigma$ is fixed, and one can expect a monotonic relationship between RMSE and $\rho$ . When this relationship holds, selection of an objective algorithm by lowest RMSE or highest $\rho$ will give the same result.

We have the opportunity to confirm this relationship. Test 8 in T1A1.7/97-032 is the same as Test 9 in T1A1.7/97-034. Thus we have both RMSE and $\rho$ values for the three objective algorithms that these two contributions address in common.

Table 1 shows that rank ordering the objective algorithms by decreasing RMSE or increasing $\rho$ gives the same results. Further, Table 1 allows us to estimate $\sigma$ as being in the 0.46 to 0.54 range.

|  | P.861 | MNB-1 | MNB-2 |
|---|---|---|---|
| RMSE | 0.28 | 0.18 | **0.15** |
| $\rho$ | 0.857 | 0.927 | **0.945** |
| $\frac{\text{RMSE}}{\sqrt{1-\rho^{2}}} = \hat{\sigma}$ | 0.54 | 0.48 | 0.46 |

Table 1. RMSE and $\rho$ compared for a single subjective test and 3 objective algorithms.

Based on this relationship, we believe that either statistic could be used to select the best objective algorithm for a fixed subjective test. Furthermore, having both correlation and RMS error available strengthens the ability to make a selection.

**Overview of Data**

In the data provided in T1A1.7/97-034, there are two tests that involve echo cancelers. It is proposed that this data be excluded from the discussion of objective measures for the following reason: comprehensive echo canceler performance evaluation requires both conversation-based and listening-only tests. Therefore, objective measurement of echo canceler performance would require analysis of 4 signals (two input and two output) and the relationship between the input/output pairs. All proposed objective algorithms however, consider only input and output of a single speech path. At this time, including (or proposing to include) echo canceler evaluation in the scope of a standard might lead to hasty conclusions based on insufficient data. Because T1A1.7/97-034 tests 3 and 4 both involve echo cancelers, we have excluded them from further discussions of the results presented in T1A1.7/97-032 and T1A1.7/97-034. We note however, that we would not expect their inclusion to lead to markedly different conclusions. The MNB2 algorithm outperformed all others on test 3, while a PSQM algorithm worked best on test 4. We propose that echo canceler evaluation should be listed as for “further study” in the scope of any standard developed based on the results presented in these two documents.

Combining the remaining results in T1A1.7/97-032 and T1A1.7/97-034 gives us 17 distinct subjective tests. Ten of these involve only error-free channels, and seven include errored channels. T1A1.7/97-034 compares five objective algorithms. These include ITU-T Rec. P.861, (denoted in T1A1.7/97-034 as PSQM20) and two variants, denoted by PSQM0 and PSQM40. These variants were included because they are related to ITU-T Rec. P.861 through the adjustment of a single variable. T1A1.7/97-034 also shows results for two MNB algorithms, MNB1 and MNB2. T1A1.7/97-032 compares ITU-T Rec. P.861, MNB1 and MNB2 and SNR, SNRseg, Perceptually-weighted SNRseg, Cepstral Distance, and Bark Spectral Distortion. Results in this contribution suggest that only P.861, its variants, and the MNB algorithms should be considered further by T1A1.7. Further, both contributions make it clear that MNB2 is more effective than MNB1 in almost every situation. Thus in our discussion, we include P.861, PSQM0, PSQM40 and MNB2.

**Discussion:**

Table 2 presents results from the 10 tests in both T1A1.7/97-032 and T1A1.7/97-034 that contained only clear-channel conditions. There were 297 channel conditions in these 10 experiments (some of them replicated between experiments, e.g., MNRU). Conditions ranged from 2.4 kbit/s to 64 kbit/s coders in both the waveform and non-waveform families, plus MNRU, homogeneous and mixed tandems, and IRS and unweighted filter characteristic conditions. For each test, the most effective algorithm is indicated by boldface type. The results are summarized at the bottom of the table where the percentage of time that an algorithm was the best choice for a test is indicated. Please note that the percentages add to more than 100% because they consider only those tests for which the algorithm was evaluated. For example, PSQM0 results are reported for 5 tests. Of those 5 chances, the algorithm performed best twice, yielding a 40%

rating.  On the other hand, results for the MNB2 algorithm are available for 10 tests.  It performed best in six of those, yielding a 60% rating, and providing the best overall applicability in these 10 experiments.

Table 3 shows results for 7 tests that included the above mentioned conditions plus errored-channel conditions including 1% to 5% frame error ratio with varying levels of burstiness, and bit errors resulting from C/I in the range of 8-20 dB.  The data in this table represents a total of 251 channel conditions.  As in Table 2, the most effective objective algorithm for each test is indicated by boldface type. Also as in Table 2, the results are summarized at the bottom of the table where the percentage of time that an algorithm was the best choice for a test is indicated.  As in the clear-channel conditions, the MNB2 algorithm performs better more consistently than the PSQM algorithms.  In this case, PSQM40 was best in 40% of its opportunities while MNB2 was best in 71% of its opportunities.  This leads to the conclusion that MNB2 is the best choice for overall applicability in errored-channel conditions.

Tables 4 and 5 summarize results across all 17 tests.  Table 4 examines the average correlation and average RMSE across all tests.  In addition, it examines which algorithm was most effective across all 17 tests.  Once again, the most effective objective algorithm is indicated by boldface type.  The average performance of MNB2 algorithm exceeds the averages of the other three measures.  Perhaps more significant, however is that each of the three variations of the PSQM was best twice, while the MNB2 was best 11 times, or 65% of the experiments in which it was evaluated.  This illustrates a clear trend that the MNB is a more generally applicable measure than P.861 and its variants.  Table 5 echoes this trend by summarizing the percentage of time that an algorithm was chosen as best.  The table separates the results into four-way comparisons (i.e., MNB2 and 3 variants of PSQM) using RMSE and two-way comparisons (i.e., MNB2 and P.861)using correlation.  The data in the table shows that the MNB is the best option in 60% of the cases for the four-way comparison and 75% of the cases for the two way comparison.

**Summary**

For each of the 17 tests, the performance differences between the four algorithms can range from small to large. Across the 17 tests though, there is a very clear trend:  MNB2 performs best in 11 of the 17 tests.  P.861 and the two variants are all tied for second place with each performing best for only two tests.

In this light, the work shown in T1A1.7/97-032 and T1A1.7/97-034 suggests that an ANSI Standard on Objective Audio Quality Measures could be applied to the following conditions:

Waveform and non-waveform coders at bitrates from 2.4 to 64 kb/s
Bursty and distributed frame error ratios up to 5%
Live channel conditions up to a C/I ratio of 8 dB

Also indicated by the data in the tables, the MNB2 algorithm might be recommended as the algorithm of preference for clear channels, errored channels, and overall measurement tasks.

Table 2. Ten tests over waveform and non-waveform coders (2.4 to 64 kbps) and MNRU. No channel errors. 297 conditions total.

| Test | P.861 | | PSQM0 | | PSQM40 | | MNB2 | |
|---|---|---|---|---|---|---|---|---|
| | ρ | RMSE | ρ | RMSE | ρ | RMSE | ρ | RMSE |
| 034-1† | | 0.21 | | 0.23 | | 0.20 | | **0.16** |
| 034-2 | | 0.27 | | 0.23 | | 0.34 | | **0.17** |
| 034-5 | | 0.33 | | 0.32 | | 0.33 | | **0.27** |
| 034-6 | | 0.14 | | **0.13** | | 0.17 | | 0.16 |
| 034-7 | | 0.15 | | **0.14** | | 0.15 | | 0.22 |
| 032-1‡ | 0.929 | | | | | | **0.952** | |
| 032-4 | 0.973 | | | | | | **0.980** | |
| 032-5 | **0.985** | | | | | | 0.981 | |
| 032-6 | **0.986** | | | | | | 0.981 | |
| 032-7 | 0.976 | | | | | | **0.983** | |
| Average ρ | 0.970 | | | | | | **0.975** | |
| Average RMSE | | 0.22 | | 0.21 | | 0.24 | | **0.20** |
| # Times Best | 2 | | 2 | | 0 | | **6** | |
| % Times Best | 2/10 = 20% | | 2/5 = 40% | | 0/5 = 0% | | **6/10 = 60%** | |

† The notation “034-n” indicates that the test is described in contribution T1A1.7/97-034 as test n.
‡ The notation “032-n” indicates that the test is described in contribution T1A1.7/97-032 as test n.

Table 3. Seven tests over waveform and non-waveform coders (2.4 to 64 kbps) and MNRU. Channel errors. 251 conditions total.

| Test | P.861 | | PSQM0 | | PSQM40 | | MNB2 | |
|---|---|---|---|---|---|---|---|---|
| | ρ | RMSE | ρ | RMSE | ρ | RMSE | ρ | RMSE |
| 034-8† | | 0.29 | | 0.28 | | 0.32 | | **0.19** |
| 034-10 | | 0.16 | | 0.17 | | **0.15** | | 0.20 |
| 034-11 | | 0.22 | | 0.29 | | **0.16** | | 0.22 |
| 034-12 | | 0.36 | | 0.36 | | 0.41 | | **0.28** |
| 034-9 = 032-8‡ | 0.857 | 0.28 | | 0.26 | | 0.33 | **0.945** | **0.15** |
| 032-2 | 0.941 | | | | | | **0.946** | |
| 032-3 | 0.795 | | | | | | **0.938** | |
| Average ρ | 0.864 | | | | | | **0.943** | |
| Average RMSE | | 0.26 | | 0.27 | | 0.27 | | **0.21** |
| # Times Best | 0 | | 0 | | 2 | | **5** | |
| % Times Best | 0/7 = 0% | | 0/5 = 0% | | 2/5 = 40% | | 5/7 **= 71%** | |

† The notation "034-n" indicates that the test is described in contribution T1A1.7/97-034 as test n.

‡ The notation "032-n" indicates that the test is described in contribution T1A1.7/97-032 as test n.

Table 4. Results of all 17 tests shown in Tables 2 and 3. 548 conditions total.

| | P.861 | | PSQM0 | | PSQM40 | | MNB2 | |
|---|---|---|---|---|---|---|---|---|
| Average ρ | 0.930 | | | | | | **0.963** | |
| Average RMSE | | 0.24 | | 0.24 | | 0.26 | | **0.20** |
| # Times Best | 2 | | 2 | | 2 | | **11** | |
| % Times Best | 2/17 = 12% | | 2/10 = 20% | | 2/10 = 20% | | 11/17 **=65%** | |

Table 5. Percentages of the 17 tests where each of the four algorithms performs best.

| | P.861 | PSQM0 | PSQM20 | MNB2 |
|---|---|---|---|---|
| 4-Way RMSE Comparisons | 0% | 20% | 20% | **60%** |
| 2-Way ρ Comparisons | 25% | | | **75%** |