Techniques for Comparing Objective and Subjective Speech Quality Tests

Stephen Voran, Institute for Telecommunication Sciences
325 Broadway, Boulder, Colorado 80303, USA, sv@bldrdoc.gov

Abstract

Objective (or instrumental) tests of speech quality have been proposed as ways to reduce the need for expensive and time-consuming subjective (or auditory) tests. Both types of tests attempt to quantify the range of opinions that listeners express in response to a group of speech transmission or storage devices, but objective test results often show measurable deviation from subjective test results. This deviation may be judged to be acceptable if the objective test offers significant savings of time and expense. This cost-performance judgement cannot be made without a meaningful statistical measure of the deviation that is likely to be associated with the objective test. This paper offers a number of techniques that compare both the central tendencies and the uncertainties of two tests. The resulting statistics have direct, intuitive interpretations.

1. Introduction

Subjective speech quality tests seek to quantify the range of opinions that listeners express when they hear speech transmission or storage devices that are under test. Some subjective testing procedures have been refined and standardized (see [1] for example). Although carefully conducted subjective tests are often rather involved, time-consuming, and expensive, they are generally considered to be the most relevant way to evaluate and compare new speech transmission and storage devices. Over the years, numerous objective tests of speech or audio quality have been proposed as ways to reduce the need for subjective testing. Some recent examples are given in [2-10]. When an automated implementation is available, these proposed objective tests often provide repeatable results, and require significantly less effort, time, and expense than the corresponding subjective tests. On the other hand, there is often considerable deviation between the objective and subjective test results. Subjective tests are generally considered to hold the "correct answer." Objective tests that generate results that closely approximate subjective test results are said to be more useful than objective tests that show larger deviations.

More generally, there is a cost-performance trade-off at work. Depending on the particular testing environment, one may choose to accept an approximate result in order to save time or expense. For some applications, a rather rough approximation may be accepted if the savings are great enough. If the approximation must be very accurate, then smaller savings might be acceptable. Of course, there will always be some testing situations where the uncertainty of the results must be minimized at all costs. Subjective tests will remain the choice for these situations.

If one wishes to make fully informed cost-performance decisions regarding subjective and objective testing options, one needs a comparison technique that indicates what deviations are likely to be associated with each of the objective testing options. The coefficient of correlation between subjective and objective test results is often used for this purpose. Higher values indicate smaller discrepancies, and variance partitioning interpretations exist. An analysis of variance can be performed on subjective and objective test results. This allows for the computation of more meaningful variance partitioning results. In particular, the uncertainties in both the subjective and objective results can be accounted for and compared. These comparison techniques may not be as meaningful or as intuitive as one might wish.

This paper offers three additional techniques for comparing objective and subjective speech quality tests at different levels of data aggregation. These techniques may provide more direct and intuitive results than those mentioned above. By comparing objective and subjective confidence intervals, these techniques compare both the central tendencies and the uncertainties of two tests. In 2.1, objective test results are compared with listeners' subjective votes. In 2.2, we consider the relationship between objective results and the estimated mean of listeners' votes. The objective-subjective comparison described in 2.3 examines pair-wise classifications made by subjective and objective tests. The

comparison tabulates the inconsistencies between the objective and subjective classifications, and reports them in the form of three different error rates. These comparisons should provide a meaningful basis for the cost-performance decisions faced by those who specify subjective and objective tests of speech quality.

2. Comparison Techniques

The discussion that follows uses example results from an actual subjective test and a corresponding objective test. The tests involve two low bit-rate speech coders operating over a family of simulated radio channels, and μ-law Pulse Code Modulation (PCM) operating over an error-free channel. The subjective tests are Absolute Category Rating tests using the Mean Opinion Score (MOS) scale (excellent, good, fair, poor, bad.) A total of 19 different speech coder-transmission channel combinations are tested. For simplicity, each of these 19 combinations will be referred to as a "device" in the examples that follow. Each device is tested with sentences from three female speakers and three male speakers. Since each of these North-American English speakers provides 5 sentences, a total of 30 sentences is used to test each of the 19 devices. Sixteen listeners rate each of the resulting 570 processed sentences.

The corresponding objective test provides a single result for each of the 570 sentences. The objective test is an early prototype from the perception-based objective audio quality assessment research that is being conducted at the Institute for Telecommunication Sciences. Digitized representations of each original sentence and the corresponding processed sentence are perceptually transformed and compared, resulting in an auditory distance value. The perceptual transformation includes auditory filtering, an equal-loudness transfer function, and a non-linear loudness perception function. A simple mathematical function is used to map each auditory distance value to an MOS result. Like the rest of this prototype objective test, the mapping is highly provisional; it has not yet been refined or optimized, and it is not yet based on an appropriate amount of data. It would be inappropriate to attempt to draw firm conclusions about the objective test. It is presented here only to demonstrate the objective-subjective comparisons that are proposed below.

2.1 Fraction of Outlier Objective Votes

The most basic question one might ask about a proposed objective test is: If we view this objective test as an additional listener in a group of test subjects, would a significant fraction of its votes tend to be outliers? If the answer is yes, then the objective measure is probably of very limited use. Roughly speaking, it brings us less information than a single, "well-behaved" (non-outlier) listener would. If the answer is no, then further, more demanding questions can be asked to ascertain how much information the objective test does carry.

Let 100-p be the percent of votes that are outliers. That is, a vote is an outlier if it falls below the $\frac{1}{2}\cdot(100-p)^{th}$ percentile, or above the $\frac{1}{2}\cdot(100+p)^{th}$ percentile. The answer to the outlier question clearly depends on the choice of p in this definition. As an example, the outlier test has been applied to subjective and objective results for each of the 570 test sentences described above. In Figure 1, the fraction of these 570 data points for which the objective test is an outlier is plotted as a function of p. The percentile values for each of the 570 subjective test results are estimated from the sample standard deviation of the 16 listener votes. A Gaussian distribution of votes is assumed:

$$\text{Outlier Fraction} = \frac{N(p)}{570}, \quad N(p): \text{ Number of sentences for which } |S_i - O_i| > \sigma_i \cdot G(p),$$

$$G(p) = x: \frac{1}{\sqrt{2\pi}} \int_{-x}^{x} e^{\frac{-t^2}{2}}\, dt = \frac{p}{100}, \tag{1}$$

S_i: Sample mean of listener votes for i^{th} sentence,

σ_i: Sample standard deviation of listener votes for i^{th} sentence,

O_i: Objective test result for i^{th} sentence.

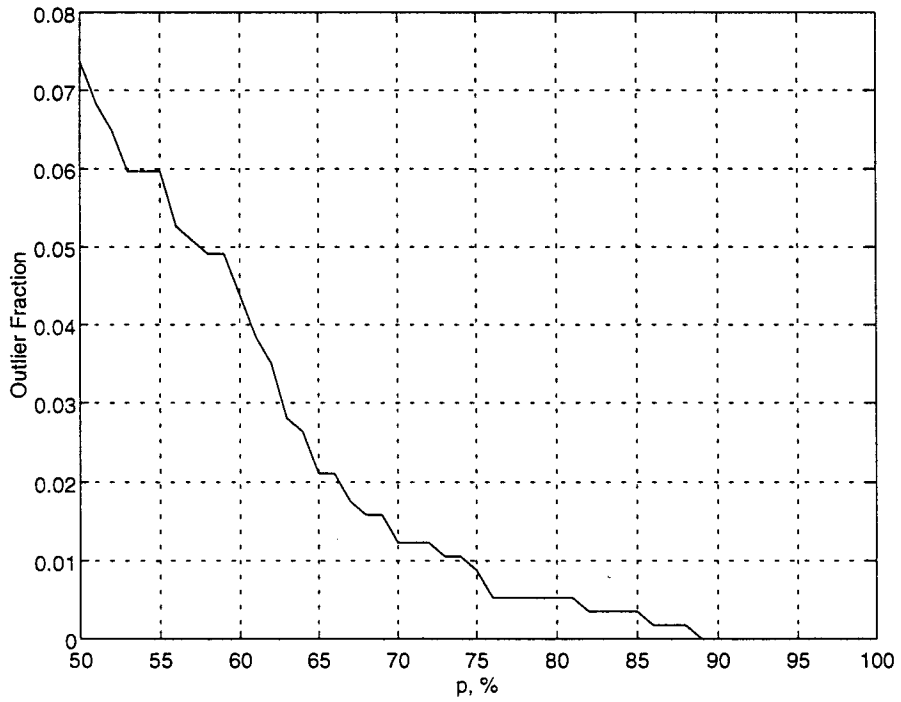

Figure 1: Fraction of outliers as a function of p.

In many disciplines, outliers are defined by p=99, 95, or 90%. For any of these definitions, Figure 1 shows that the objective test never generates an outlier vote. If one takes the extreme view that outliers are those data points outside the central 50% of the distribution (p=50%), then the objective test is an outlier for only 7% of the test sentences. We conclude that when compared to the group of test subjects, this objective test does not generate a significant number of outlier votes for this set of test sentences. In the more general case, to answer the outlier question with a firm yes or no, one would have to fix a value of p, and define what fraction of votes constitutes a "significant fraction." If the set of test conditions were broad enough, one might then be able to make a general statement about the objective test in answer to the question posed above.

2.2 Objective Results and Estimated Subjective Means

Once it is established that the objective test under consideration does not tend to vote as an outlier, one might wish to consider the level of agreement between the objective test results and the estimated means of the listeners' votes. The mean of the listeners' votes can be estimated by a sample mean, and that estimate has an uncertainty associated with it. One might ask: Does a significant fraction of the objective test results fall outside the p% confidence intervals for the estimated means of the subjective test votes? The answer provides an indication of the number of errors one might expect if the objective test results were used in place of the estimated subjective means. A yes or no answer requires that p be fixed, and that the term "significant fraction" be defined by a numerical threshold. As before, we offer a more general example, based on the test results for the 570 test sentences described above. Figure 2 shows the fraction of the 570 data points for which the objective test result falls outside the p% confidence interval on the estimated mean of the subjective votes. The confidence intervals (C.I.) for the sample means used in Figure 2 are based on the t-distribution with 15 degrees of freedom:

$$\text{Fraction Outside C.I.} = \frac{N(p)}{570}, \quad N(p)\text{: Number of sentences for which } |S_i - O_i| > \frac{\sigma_i}{\sqrt{15}} \cdot T(p),$$

$$T(p) = x : \int_{-x}^{x} \frac{\Gamma((15+1)/2)}{\sqrt{\pi\, 15}\ \Gamma(15/2)} \left(1 + \frac{\omega^2}{15}\right)^{\left(-\frac{15+1}{2}\right)} d\omega = \frac{p}{100}. \tag{2}$$

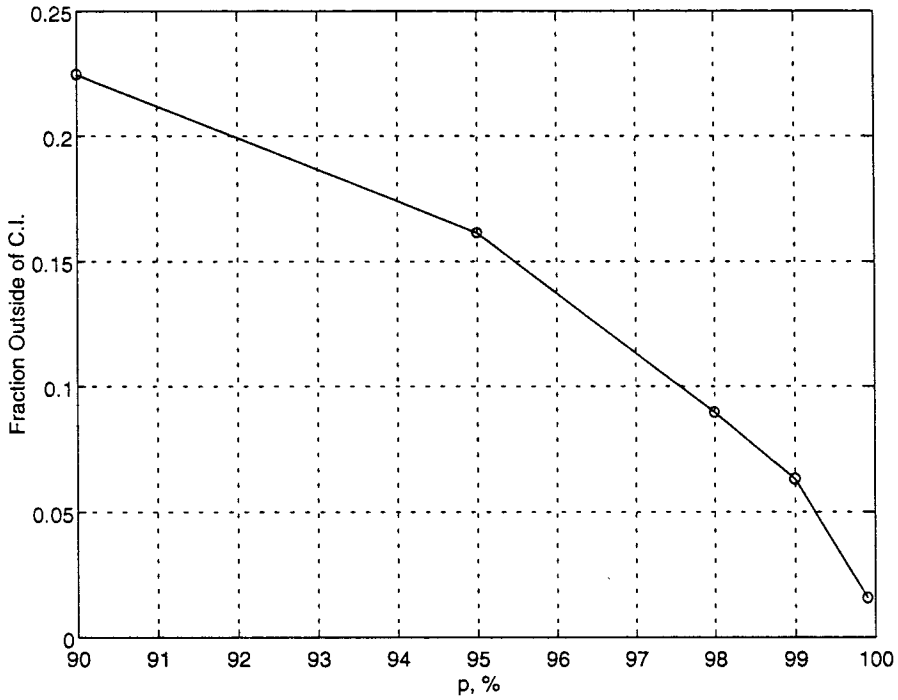

Figure 2: Fraction of objective test results that fall outside of the p% confidence interval for the subjective test estimated means. (Calculated for p=90, 95, 98, 99, and 99.9%.)

In the more general case, if a value of p and a threshold were specified, and if the set of test conditions were broad enough, then one might be able to make a general statement about the objective test in response to the question posed above.

2.3 Classification of Pairs via Subjective and Objective Tests

Subjective tests often seek to compare proposed new speech transmission or storage devices with those based on an existing accepted technology. Examples include code-excited linear predictor (CELP) based devices compared to adaptive differential PCM (ADPCM) based devices, and ADPCM compared to PCM. One popular question is: Does this proposed device sound significantly worse than the device that it might replace? Another question is: Given these two new candidate devices, does one sound significantly better than the other? Even when these pair-wise comparison questions are not explicitly asked, the use of reference conditions often allows one to use pair-wise comparisons to answer the questions that were asked.

Ideally, tests that compare two devices will cover all relevant operating conditions. Those conditions may include speakers, language, background noise, input levels, transmission channel conditions, and even different testing facilities or test procedures. When all the test results are appropriately averaged to create overall subjective results for each device, the results are only estimates of the true mean values, and there is some uncertainty in those estimates. When the confidence intervals about two means overlap, we say that the two devices are not statistically separable, or that they are "tied." When the confidence intervals for two devices do not overlap, we can classify the first device as "lower" or "higher" than the second device. Thus, in a test involving n devices, each of the $\frac{1}{2} \cdot n \cdot (n-1)$ distinct pairs of devices can be placed in one of three sets: L (lower), T (tied), or H (higher). Let s(x) be the estimated mean subjective score associated with device x, and let Δ_x be the width of the p% confidence interval about s(x). Then the classification rules are given by equation 3. The classification of certain pairs is used to answer questions like those listed above. The objective test results for each device also have a sample mean and a confidence interval about that sample mean. Using rules analogous to those given in equation 3, each pair of devices can be placed into one of the three sets (L,T,H) according to the results of the objective test.

$$
\begin{aligned}
s(a) - s(b) &< -\frac{\Delta_a + \Delta_b}{2} &&\rightarrow (a,b) \in L, \\
-\frac{\Delta_a + \Delta_b}{2} \le s(a) - s(b) &\le \frac{\Delta_a + \Delta_b}{2} &&\rightarrow (a,b) \in T, \\
\frac{\Delta_a + \Delta_b}{2} &< s(a) - s(b) &&\rightarrow (a,b) \in H.
\end{aligned}
\tag{3}
$$

If the objective and subjective classifications are not the same, then an error has occurred. In Table 1, three types of errors are defined.

	Obj: (a,b)∈L	Obj: (a,b)∈T	Obj: (a,b)∈H
Subj: (a,b)∈L	Correct Classification	False Tie	False Ranking
Subj: (a,b)∈T	False Differentiation	Correct Classification	False Differentiation
Subj: (a,b)∈H	False Ranking	False Tie	Correct Classification

Table 1: Nine possible outcomes for a pair of devices that are tested subjectively and objectively.

We suggest that the rates of occurrence of the three errors provide a highly relevant comparison between the objective and subjective test results. When pair-wise comparisons between devices are the motivation for a subjective test, the three error rates can be interpreted as the probability of making one of three different kinds of mistakes, if the objective test had been used in place of the subjective test. The three types of errors are probably not equally harmful. A false ranking would generally be considered the worst mistake one could make. A false ranking means that *a* is reported to be better than *b*, when in fact, the subjective test would have shown the opposite to be true. A false differentiation might be the next worst type of error. Here, one reports that *a* is better than *b* when the subjective test would have failed to differentiate them. (It might be argued that one cannot know for a fact that a "false differentiation" is truly an error. Perhaps a subjective test that used more listeners would have had narrower confidence intervals and hence would have resolved the two devices in the same way the objective test did.) The false tie might be considered the least dangerous of the three errors. In this event, one reports that two devices are equivalent, when the subjective test would have been able to separate them.

By explicitly comparing the classifications of device pairs, one is implicitly comparing both the means and uncertainties of the objective and subjective tests. This is very important since means alone will not allow one to make decisions like those described above. The three error rates are invariant to any fixed bias or positive linear scaling of objective test results. In fact, an objective test need not generate results on the subjective testing scale. As long as the objective test results include appropriate measures of central tendency and dispersion, one can use them to classify device pairs into the three sets. In effect, the three error rates measure the ability of an objective test to correctly answer basic questions about pairs of devices, rather than its ability to generate precise numerical results on a particular scale.

We offer a final example from the subjective and objective tests described above. For each of the 19 devices considered, the t-distribution can be used to estimate 95% confidence intervals for the subjective and objective sample means. Each of the 171 distinct pairs of devices can be classified both subjectively and objectively. When the differences between the two classifications are noted, we find that 36 false ties have occurred, resulting in a false tie rate of 0.21. One false differentiation was observed, and no false rankings occurred for this particular set of devices. The rather high false tie rate is consistent with our observation that for these devices, this particular objective test tends to have wider confidence intervals than the corresponding subjective test. This type of mismatch is probably one of the more desirable ones,

since it minimizes the occurrences of the potentially more harmful false rankings and false differentiations.

3. Summary

We have presented three techniques for comparing objective and subjective speech quality tests. The comparisons operate at different levels of data aggregation, and hence they answer very different questions. In general, the most appropriate choice for comparing two tests is strongly influenced by the questions that the tests were originally designed to answer. By comparing objective and subjective confidence intervals, one can simultaneously compare both the central tendencies and the uncertainties of two tests. In effect, one can then compare decisions rather than numbers. Since decisions are often the ultimate output of a subjective testing and data analysis procedure, we might say that the procedure of 2.3 has moved the objective-subjective comparison up to the highest level, and perhaps most relevant and intuitive level. We suggest that the comparisons described here can form a meaningful basis for the cost-performance decisions that are intrinsic to the selection of subjective and objective speech quality testing procedures.

References

[1] CCITT, Recommendation P.83, "Subjective Performance Assessment of Telephone-Band and Wideband Digital Codecs," International Telegraph and Telephone Consultative Committee, Geneva, Switzerland, 1992.

[2] J. Beerends and J. Stemerdink, "Modelling a Cognitive Aspect in the Measurement of the Quality of Music Codecs," *Audio Engineering Society 96th Convention*, Amsterdam, The Netherlands, Feb. 1994.

[3] U. Halka and U. Heute, "A New Approach to Objective Quality-Measures Based on Attribute-Matching," *Contribution to ITU-T, SG-12 Speech Quality Experts Group*, Geneva, Switzerland, Feb. 1992.

[4] J. Herre, E. Eberlein, H. Schott, and K. Brandenburg, "Advanced Audio Measurement System Using Psychoacoustic Properties, "*Audio Engineering Society 92th Convention*, Vienna, Austria, March 1992.

[5] B. Paillard, B. Mabilleau, and S. Morissette, "PERCEVAL: Perceptual Evaluation of the Quality of Audio Signals," *J. Audio Eng. Soc.*, vol 40, no. 1, Jan. 1992.

[6] S. Wang, A. Sekey, and A. Gersho, "An Objective Measure for Predicting Subjective Quality of Speech Coders," *IEEE J. on Sel. Areas in Comm.*, vol.10, no.5, June 1992.

[7] A. De, and P. Kabal, "Auditory Distortion Measure for Speech Discrimination Information Approach," *IEEE Globcom '92* Orlando, USA, Dec. 1992.

[8] M. Hollier, M. Hawksford, and D. Guard, "Characterization of Communications Systems Using a Speech-Like Test Stimulus," *93rd Convention of Audio Engineering Society*, San Francisco, USA, Oct. 1992.

[9] R. Kubichek, "Mel-Cepstral Distance Measure for Objective Speech Quality Assessment," *Proc. Pacific Rim Conference on Communications, Computers, and Signal Processing,* Victoria, Canada, May 1993.

[10] CCITT Recommendation P.11, Supplement No. 3, Annex G, "Objective Method of Estimating the Quality of Speech Degraded by Nonlinear Distortion," International Telegraph and Telephone Consultative Committee, Geneva, Switzerland, 1992.