T1A1.5/96-122
T1A1.7/96-036

COMMITTEE T1 - TELECOMMUNICATIONS STANDARDS CONTRIBUTION

DOCUMENT NUMBER:	T1A1.5/96-122 T1A1.7/96-036
T1BBS FILE:	6a151220.ps, 6a170360.ps
DATE:	October 29, 1996
STANDARDS PROJECT:	T1Q1.12, T1Y1.20
TITLE:	Informational Report on ITS Subjective Testing
SOURCE:	Institute for Telecommunication Sciences National Telecommunications and Information Administration U.S. Department of Commerce
AUTHOR:	Coleen Jones
CONTACT:	Coleen Jones 1-303-497-3764 (Phone) cjones@its.bldrdoc.gov
DISTRIBUTION:	T1A1.5, T1A1.7, T1BBS

ABSTRACT:

This contribution is provided for informational purposes. It contains a description of a subjective audio quality test conducted at the Institute for Telecommunication Sciences (ITS) that may be of interest to memebers of T1A1.7 and T1A1.5. A description of the subjective test is given along with results of the test. Also presented are some interesting sub-topics on hearing testing and Q-curves that resulted from this experiment.

Informational Report on Subjective Testing at ITS

1. Introduction

This informational contribution describes a subjective audio quality test conducted at the Institute for Telecommunication Sciences (ITS) that may be of interest to members of T1A1.7 and T1A1.5. This is the first subjective audio quality test conducted at ITS since the construction of our subjective testing facilities. Our facilities consist of two acoustically isolated listening chambers (for interactive testing) and associated audio equipment (PC with sound card, DAT machines, CD players, audio cassette recorders, mixers etc.). This subjective test was designed to estimate the calibration factors needed to map subjective scores from six different laboratories onto one scale, thus creating a single large database of audio files with relatively consistent subjective scores. This test is part of an on-going ITS effort aimed at developing objective measures of audio quality.

2. Purpose of Test

ITS has acquired several subjectively-scored audio data sets (audio files and subjective scores) from audio tests that were conducted at various laboratories.[1] All of these data sets are limited to 4 kHz bandwidth telephony speech samples. These data sets have been very useful in the development of our objective audio quality measures. However, to increase the usefulness of these data sets, we felt it necessary to calibrate these tests onto a single consistent rating scale. Thus, the purpose of this test was to conduct a subjective audio quality test at ITS that consisted of a subset from each of the six existing data sets. The results of this 'meta-test' allowed us to map the individual scores from each laboratory onto a single scale. This resulted in a single large data set that can be used for the development and testing of objective audio quality measures.

3. Test Design

The test pool consisted of six dissimilar audio data sets from six different laboratories (denoted as labs A-F). Five data sets were scored using the Absolute Category Rating (ACR) method, resulting in mean opinion scores (MOS). One was scored using the Diagnostic Acceptability Measure (DAM) method, resulting in composite DAM scores. This data set was included in spite of the fact that composite DAM and MOS are not purported to measure the same quantity. Subjectively rating every condition in each of the six data sets would have been prohibitive. Therefore, we chose a sample of conditions from each data set. The guidelines we used for selecting the sample conditions are listed below:

1. To get the most useful calibration factors, we chose to span the range of quality in our selections. Therefore, we chose approximately nine conditions per data set; three each of high, medium, and low quality. (High quality: $3.75 < MOS < 5.0$, medium quality: $2.25 < MOS < 3.75$, low quality $1.0 < MOS < 2.25$.)
2. Conditions that were common between at least two data sets were given priority. The option of IRS filtering was considered as part of the condition. For example, G.711, IRS filtered was considered a different condition than G.711, not IRS filtered.
3. If subjective data was available for the source, it was also included.

[1] We extend our thanks to those who have provided digitized speech samples and corresponding subjective scores to us. The work described would have been impossible without these vital contributions.

In most cases, once the common conditions were chosen (guideline 2), each quality range contained three conditions (guideline 1). The exception to the above guidelines was the DAM-scored data set for which we only chose three conditions, one high, one medium, and one low because it is a much smaller data set.

Of the 47 MOS-scored conditions chosen, the mean of the per condition mean opinion scores was 3.03. The standard deviation across the per condition mean opinion scores was 1.11.

The next element considered was the number of sentence pairs per condition and the number of listeners needed to obtain reasonable confidence intervals. Given the subjective score data we received from the other laboratories, we were able to set practical goals for the standard deviations and 95% confidence intervals in the meta-test. Assuming that our standard deviations would be similar to these laboratories, we calculated that we needed approximately 100 data points per condition. Due to the data available to us, we were limited to four sentence pairs per condition, thus necessitating 25 listeners.

4. Test Procedure

Approximately 80 listeners were randomly selected from the Department of Commerce Boulder Campus phone book. The random selection was forced to be half women and half men. Of these selected employees, appointments were made for 39 people. Twelve listeners were disqualified due to hearing (see below), and two failed our consistency checks. This gave us 25 qualified listeners (12 male, 13 female).

ITU-R Recommendation P.800 (formerly P.80) does not require that a listener's hearing be tested. However, it does refer to Annex B of ITU-R Recommendation P.78 as a possible method of hearing screening. Recommendation P.78 states that no listener "should exceed a hearing loss of a [sic] 15 dB at all frequencies up to and including 4 kHz and no more than 25 dB at 8 kHz." We felt this to be an extremely stringent requirement for our purposes (documentation accompanying our audiometer classifies a 15 dB hearing loss relative to the norm as "normal"). After some experimentation, we chose a threshold of 30 dBHL, which is classified as a "mild" hearing loss in the audiometer documentation. Each listener's pure-tone perception threshold was tested at 500, 1000, 2000, and 4000 Hz. If the threshold was above the norm by more than 30 dB (+30 dBHL) at any frequency in either ear, that listener was disqualified.

Each listener participated in two test sessions and scored 100 sentence pairs per session. A practice session of six sentence pairs was given at the beginning of each session. Each condition was represented by two sentence pairs in each session. Within this constraint, the order of presentation was randomized within each session for each listener. Each listener scored all 200 (50 conditions * 4 sentence pairs/condition) sentence pairs.

Listeners scored each sentence pair using an electronic form presented on a pen-based PC. The electronic form has several advantages over paper forms. The first is that the test can be self-paced. The controlling PC plays a sentence pair and when the listener scores the sentence pair, the controlling PC then plays the next sentence pair. If the listener does not score the audio file

within eight seconds, the electronic form times out, and a no-score is sent to the controlling PC. The electronic form is also useful because the controlling PC receives the scores and automatically records them, thus negating the need for data entry. Another advantage is that the forms stay synchronized with the audio. In other words, the listener cannot rate the current audio on the incorrect form.

Each listener's scores were checked for consistency. This subjective test contained four sentence pairs for each condition. For each listener, the spread in these four scores can be attributed to at least two sources. One source is the variation that the condition shows to differing speech sources. A second source is the lack of consistency in the listener's responses. While these two sources are not truly separable, the first source is the same for all listeners. Thus, the spread in the four scores for each condition might be viewed as a relative measure of consistency for each listener. Given that this proposed consistency measure is relative, no absolute thresholds can be set prior to the test, but outlier listeners might be detected after the test. Such listeners may be intentionally or unintentionally providing responses that do not actually reflect their opinions.

Thus, for each listener, the consistency per condition was calculated as the average absolute difference between unique pairs of the four sentence pairs (six combinations in all). The consistency per condition was then averaged over the 50 conditions to give an overall consistency measure for each listener. This overall consistency measure was then normalized by the standard deviation over all files a given listener scored (200 in all). This step attempts to compensate for each listener's personal "dynamic range". All but one of the listeners' consistency scores fell well within two standard deviations of the mean consistency value. A final listener had an unusually high consistency score that was 3.2 standard deviations above the mean. With all of the arguments for and against the removal of outliers duly noted, we removed that listener's scores from further consideration.

Using the electronic form, listeners were given eight seconds to score each file. If the listener did not score, the electronic form timed-out and returned a 'no-score' to the controlling PC. Using this information, each listener's scores were also checked to see how many 'no-scores' he or she had. Any listener with four or more 'no-scores' was disqualified. Using this criterion, one other listener was disqualified (seven 'no-scores').

5. Results

Table 1 below lists the summary statistics for this subjective test.

Table 1: ITS Subjective Test Summary Statistics

Grand Mean:	3.01
Average condition standard deviation:	0.67
Average condition 95% confidence interval half-width:	0.13

Our average condition standard deviation is very good relative to some of the other tests from which our test was derived. Our average confidence interval is in our targeted range. These are encouraging results for an initial subjective test in a new facility.

Figure 1 shows a plot of the confidence interval vs. condition MOS for the 50 conditions in this test. Note that at the end-points of the MOS scale, the confidence interval decreases, especially at the low end. In fact, there was one condition with an MOS of 1.00 and a calculated confidence interval of 0.00. Note also on the high end that our maximum MOS was 4.52 for a source file (confidence interval 0.12).

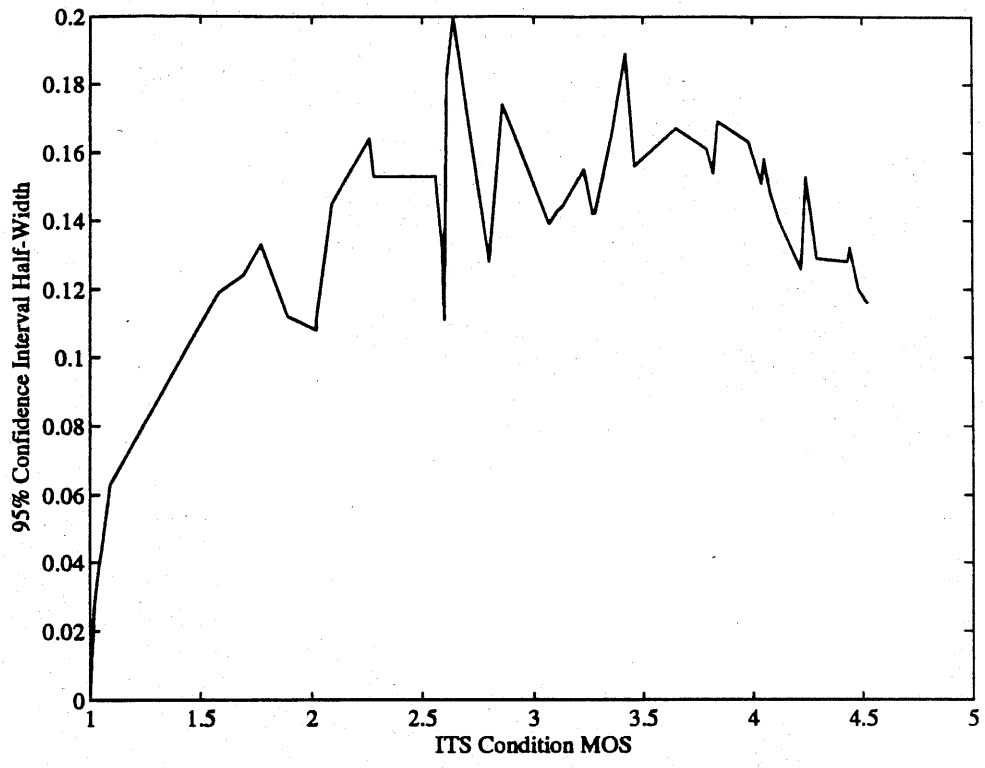

Figure 1: Per Condition Confidence Intervals

Figure 2 is a histogram of condition mean opinion scores. The large number of conditions falling into the bin from 4.0 to 4.5 are mostly source conditions, high bit rate codecs (at least 16 kbps), and higher MNRU values (30 dBQ and up).

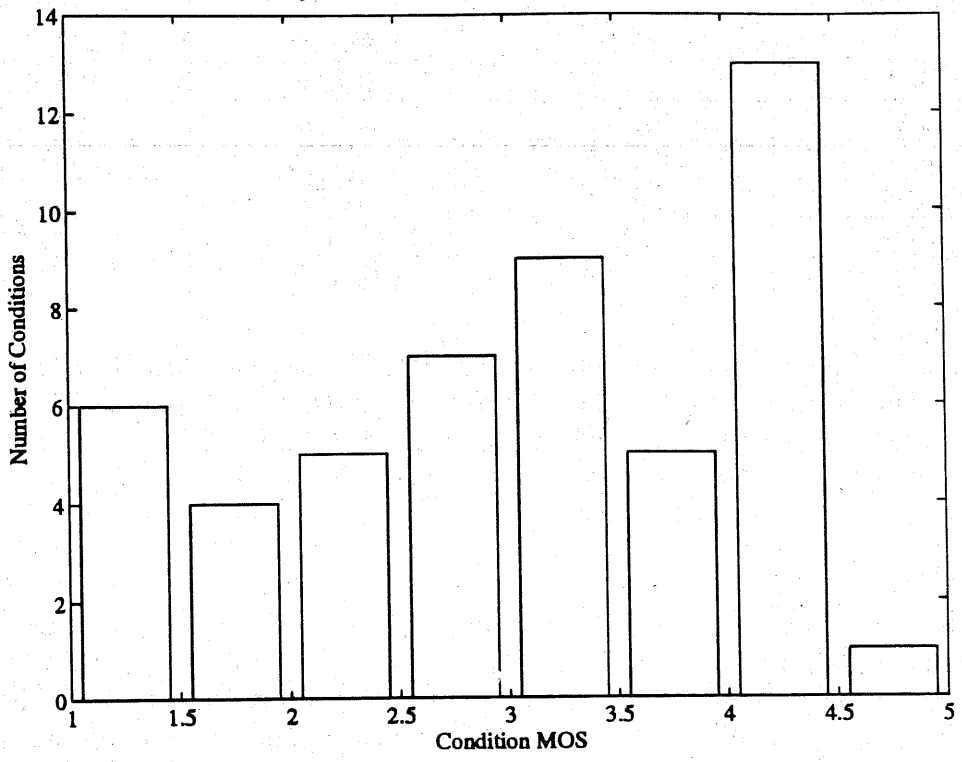

Figure 2: MOS Histogram

5.1 Main Test Results

We chose to use linear regression to map original scores to the scores gathered in the meta-test. Table 2 summarizes the regression between each laboratory and ITS. For example, if we denote ITS' MOS as I and Lab A's MOS as A, then $I \approx \hat{A} = 1.0317 * A - 0.1930$. The table also lists the correlation coefficients between the individual lab's condition mean opinion scores and ITS' condition mean opinion scores. Except for Lab C, approximately nine data points were involved in each of these least-squares fits. Note that our results were very similar to the results attained at the different laboratories. In general, the slopes are near unity, and the offsets are close to zero. Our results differ the most from Lab B's results. The slope and offset for Lab B indicate that ITS' listeners used a smaller range on the MOS scale, and that they were less critical than Lab B's listeners by almost one MOS unit. The correlation coefficients were also close to unity indicating that our results were similar to that of the other laboratories, resulting in good linear fits between each laboratory and ITS.

Table 2 Summary of Linear Least-Squares Fits (Individual laboratory to ITS)

Laboratory	Least-Squares Fit		Correlation
	Slope	Offset	Coefficient
Lab A	1.0317	-0.1930	0.9825
Lab B	0.8299	0.8668	0.9925
Lab C*	0.0467	-0.4473	0.9998
Lab D	0.9081	-0.0436	0.9701
Lab E	0.9824	-0.0052	0.9873
Lab F	1.0610	0.0059	0.9909

For reference, Figure 3 and Figure 4 are scatter plots of the meta-test results relation to original scores for data sets from Lab B (highest correlation coefficient excluding Lab C) and Lab D (lowest correlation coefficient). The boxes around the data points show the 95% confidence intervals on the original lab's MOS and on the meta-test MOS. The solid line is the least-squares linear fit between the two sets of scores.

5.2 Hearing

Figure 5 shows the correlation coefficient between each listener's condition mean opinion scores and the 25 qualified listeners' condition mean opinion scores. If the given listener was a qualified listener, his or her condition MOS was correlated with the condition MOS of the other 24 qualified listeners. Qualified listeners are plotted with an asterisk '*', and hearing disqualified listeners with a plus '+'. Both listeners 4 and 20 had hearing losses that exceeded our threshold across the bandwidth we tested (500, 1000, 2000, 4000 Hz). They performed noticeably worse than other listeners (as did qualified listener 15). Hearing disqualified listeners 19, 33, and 39 exceeded the threshold marginally or exceeded it in only one ear. Based on correlation coefficient, their hearing loss disqualification was apparently unnecessary as they performed on par with the qualified listeners. Hearing disqualified listeners 16, 28, and 31 had hearing losses only at 4000 Hz, and did quite well, relatively speaking. Interestingly enough, listeners 16, 28, and 31 had the three highest correlation coefficients of all listeners. Analysis of these listeners and hearing loss suggests that the threshold may be relaxed in future 4 kHz bandwidth telephony speech tests.

5.3 MNRU MOS Values

In earlier work, we had computed linear regressions between the available MNRU MOS values in the various data sets. The results of these regressions provided for a somewhat limited calibration procedure that sought to relate MOS values between the data sets. Once the meta-test was complete, we repeated some of those linear regressions, finding the linear least-squares fits between meta-test MNRU MOS values and the original MNRU MOS values.

* Lab C scores are composite DAM scores that can range from 0 to 100. The correlation coefficient is nearly unity because only three data points were used for this linear regression.

In Table 3, the results of these MNRU-based regressions are compared with the regressions of Section 5.1 in Table 2. Only the IRS-filtered MNRU conditions were used, limiting the results to the data sets from labs D, E, and F. In effect, the MNRU-based regressions find inter-lab relationships along a single dimension in speech quality space: signal-correlated white noise. Because they use a wider range of conditions, one would expect that the regressions of Section 5.1 would serve to characterize the inter-lab relationships more fully. The two regressions show similarity only in that they both identify slopes less than one for labs D and E and slopes greater than one for lab F. The offsets are small and do not show agreement between the two regressions.

Table 3: Linear Regression Results Using Meta-Test MOS Compared to MNRU MOS

	Regression Results Using MNRU MOS Values		Regression Results Using Meta-Test Conditions	
Laboratory	Slope	Offset	Slope	Offset
Lab D	0.8509	0.1229	0.9081	-0.0436
Lab E	0.9534	0.0115	0.9824	-0.0052
Lab F	1.0053	0.1927	1.0610	0.0059

Figure 6 is a plot of the original MNRU MOS values for these three laboratories and the meta-test. This plot should reinforce the importance of reference conditions within any subjective test.

6. Conclusions

ITS is enthusiastic about our new subjective testing facilities. We feel that this first test has given us useful information and has demonstrated that our methodology works well. We plan to proceed with additional subjective tests that assess speech and audio quality, as well as attributes of human auditory perception. As these tests are completed, we will inform T1A1.7 and T1A1.5 of our results.

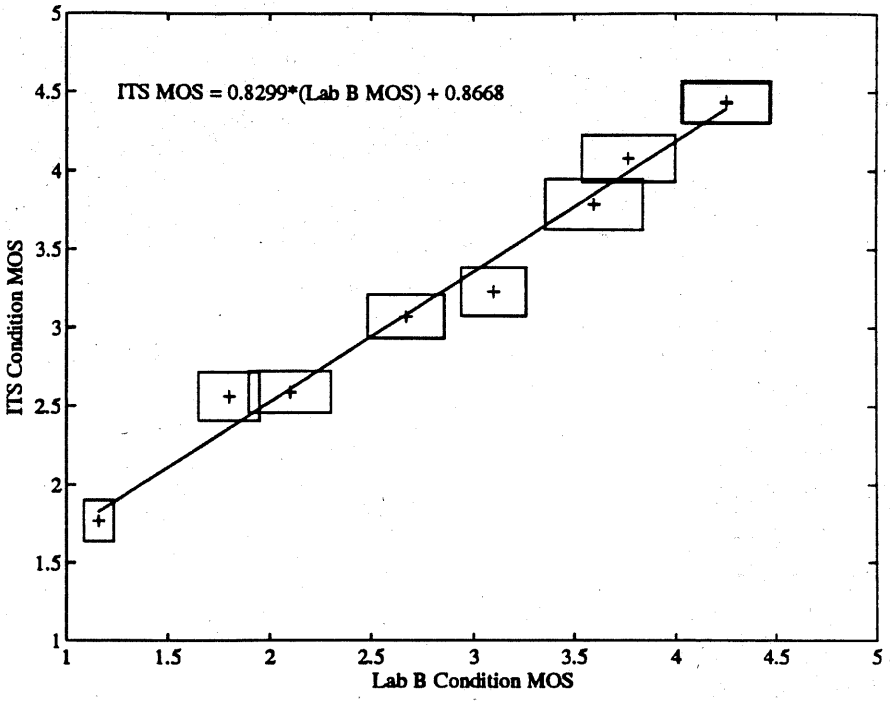

Figure 3: Lab B scatter plot

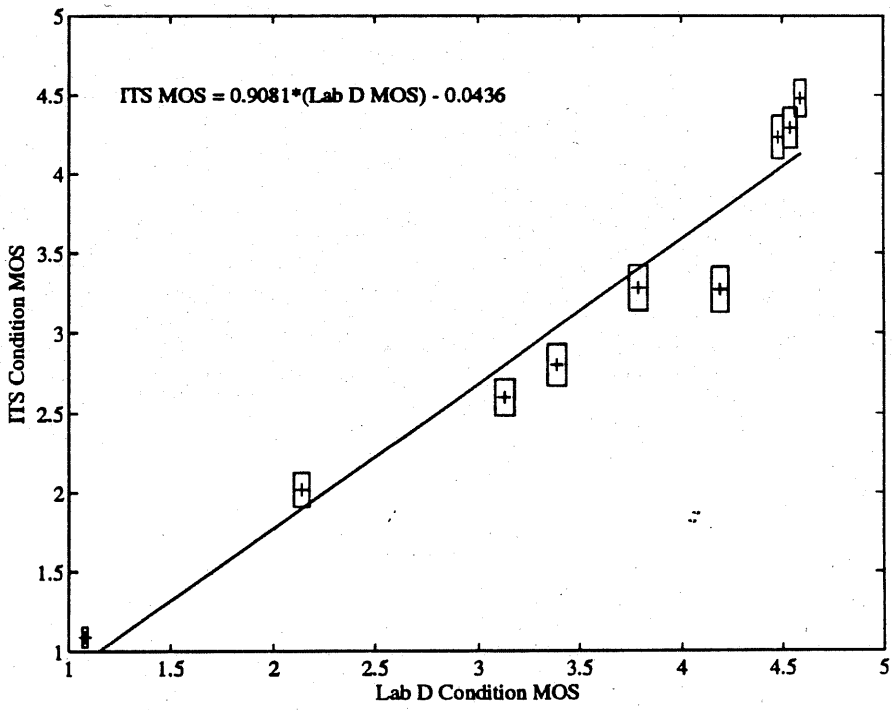

Figure 4: Lab D scatter plot

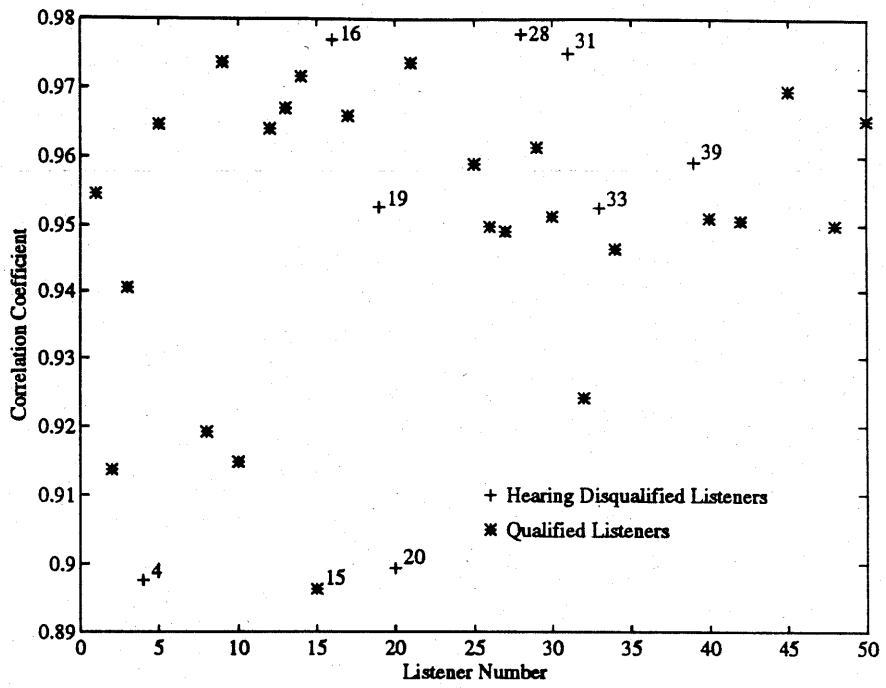

Figure 5: Hearing Scatter Plot

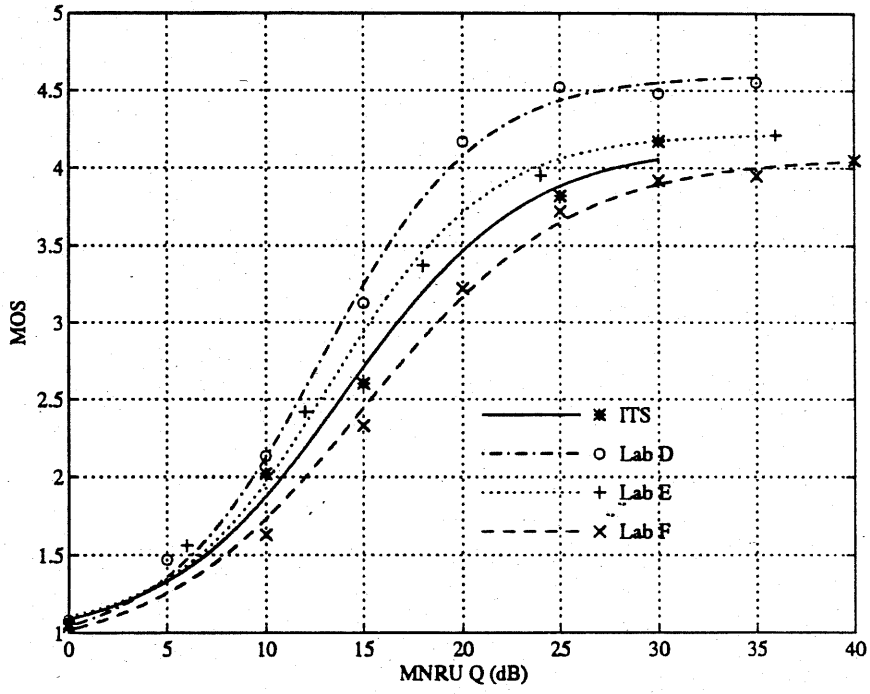

Figure 6: IRS-filtered MNRU MOS Values