Usability Evaluation and Interface Design – Proceedings of HCI International 2001 (Mahwah, NJ: Lawrence Erlbaum)

Psychometric Properties of the Mean Opinion Scale

James R. Lewis

IBM Voice Systems
1555 Palm Beach Lakes Blvd.
West Palm Beach, Florida
jimlewis@us.ibm.com

Abstract

The Mean Opinion Scale (MOS) is a seven-item questionnaire used to evaluate speech quality. Analysis of existing data revealed (1) two MOS factors (Intelligibility and Naturalness, plus a single independent Rate item), (2) good reliability for Overall MOS and for subscales based on the Intelligibility and Naturalness factors, (3) appropriate sensitivity of MOS factors, (4) validity of MOS factors related to paired comparisons, and (5) validity of MOS Intelligibility related to intelligibility scores. In conclusion, the current MOS has acceptable psychometric properties, but adding items to the Naturalness scale and increasing the number of scale steps from five to seven should improve its reliability.

1. Introduction

The Mean Opinion Scale (MOS) is the method for evaluating text-to-speech (TTS) quality recommended by the International Telecommunications Union (ITU). The MOS is a Likert-style questionnaire, typically with seven 5-point scale items addressing the following TTS characteristics: (1) Global Impression, (2) Listening Effort, (3) Comprehension Problems, (4) Speech Sound Articulation, (5) Pronunciation, (6) Speaking Rate, and (7) Voice Pleasantness.

It might seem that articulation tests that assess intelligibility (such as rhyme tests) would be more suitable for evaluating artificial speech than a subjective tool such as the MOS. Most modern text-to-speech systems, although more demanding on the listener than natural speech (Paris, Thomas, Gilson, & Kincaid, 2000), are quite intelligible (Johnston, 1996). "Once a speech signal has breached the 'intelligibility threshold', articulation tests lose their ability to discriminate. … it is precisely because people's opinions are so sensitive, not just to the signal being heard, but also to norms and expectations, that opinion tests form the basis of all modern speech quality assessment methods." (Johnston, 1996, pp. 102, 103)

Developers of products that use artificial speech output need reliable and valid tools for evaluating the quality of TTS systems. When the tool is a questionnaire that collects subjective ratings (like the MOS), it is important to understand its psychometric properties. The goal of psychometrics is to establish the quality of psychological measures (Nunnally, 1978). Some of the metrics of psychometric quality are reliability (consistent measurement), validity (measurement of the intended attribute), and sensitivity (responds to specific experimental manipulations).

1.1. Brief Review of Psychometric Practice

Reliability. The most common measurement of a scale's reliability is coefficient alpha (Nunnally, 1978). Coefficient alpha can range from 0 (completely unreliable) to 1 (perfectly reliable). For purposes of research or evaluation in which the final score will be the average of ratings from more than one questionnaire, the minimally acceptable reliability is .70 (Landauer, 1988).

Validity. Researchers commonly use the correlation coefficient to assess criterion-related validity (the relationship between the measure of interest and a different concurrent or predictive measure). The magnitude of the correlation does not need to be large to provide evidence of validity, but the correlation should be significant.

Sensitivity. A measurement is sensitive if it responds to experimental manipulation. For a measurement to result in statistically significant differences in an experiment, it must be both reliable and valid.

Number of scale steps. All other things being equal, a greater number of scale steps will enhance scale reliability, but with rapidly diminishing returns. As the number of scale steps increases from two to twenty, there is an initially rapid increase in reliability that tends to level off at about seven steps (Nunnally, 1978). After eleven

steps there is very little gain in reliability from increasing the number of steps. Lewis (1993) found that mean differences between experimental groups measured with questionnaire items having seven steps correlated more strongly with the observed significance level of statistical tests than did similar measurements using items that had five scale steps.

Factor analysis. Factor analysis is a statistical procedure that examines the correlations among variables to discover groups of related variables (Nunnally, 1978). Because summated (Likert) scales are more reliable than single item scores and it is easier to interpret and present a smaller number of scores, it is common to conduct a factor analysis to determine if there is a statistical basis for the formation of measurement scales based on factors. Generally, a factor analysis requires five participants per item to ensure stable factor estimates (Nunnally, 1978). There are a number of methods for estimating the number of factors in a set of scores, including discontinuity and parallel analysis (Coovert & McNelis, 1988).

1.2. Previous Research in MOS Psychometrics

Reliability. A literature review turned up no previous work reporting MOS reliability in any form.

Validity. Salza et al. (1996) measured the overall quality of three Italian TTS synthesis systems with a common prosodic control but different diphones and synthesizers using both paired comparisons and the MOS. Their results showed good agreement between the two measurement methods, providing evidence for the validity of the MOS. Johnston (1996) had listeners judge the quality of natural speech degraded with time frequency warping. He found a significant relationship in the expected direction for judgements using the MOS Listening Effort item (greater degradation led to poorer ratings).

Sensitivity. Johnston (1996) found that the MOS Listening Effort item showed statistically significant differences among the ratings of three TTS systems, and that this item was more sensitive than a more general item asking listeners to rate the overall quality of the system. He also found that using sentences as stimuli yielded results that were just as sensitive as those using longer paragraphs.

Yabuoka et al. (2000) investigated the relationship between five distortion scales (differential spectrum, phase, waveform, cepstrum distance, and amplitude) and MOS ratings. They were able to calculate statistically significant regression formulas for predicting MOS ratings from manipulations of the distortion scales. Unfortunately, they did not report the exact type of MOS that they used in the experiment.

Factor structure. The factor structure of the MOS is currently in question. Kraft and Portele (1995), using an eight-item version of the MOS (with an additional 'Naturalness' item), reported two factors – one interpreted as intelligibility (segmental attributes) and one as naturalness (suprasegmental, or prosodic attributes). The Speaking Rate (Speed) item did not fall in either of the two factors. More recently, Sonntag et al. (1999), using the same version of the MOS (but with 6-point rather than 5-point scales), reported only a single factor.

1.3. Goals of the Current Research

The goals of the current research were to (1) evaluate the factor structure of the 7-item 5-point-scale version of the MOS (the version reported by Salza et al., 1996, adapted for use in our lab), (2) estimate the reliability of the overall MOS score and any revealed factors, (3) investigate the sensitivity of the MOS scores, and (4) extend the work on validity of the MOS.

2. Method

2.1. Factor Analysis and Reliability Evaluation

Over the last two years we have conducted a number of experiments in which participants have completed the MOS. In some of these experiments we have also collected paired-comparison data and, in the most recent (Wang & Lewis, 2001), we also collected intelligibility scores. Participants in these experiments have included in approximately equal numbers, males and females, persons older and younger than 40 years old, and IBM and non-IBM employees. Drawing from six of these experiments I assembled a database of 73 independent completions of the version of the MOS that we have been using (taken from Salza et al, 1996). (Note: Using the guideline that the number of completed questionnaires required for factor analysis is five times the number of items in the questionnaire (Nunnally, 1978), the minimum required number of MOS questionnaires is 35, well below the 73 questionnaires in the database.) This database was the source for a factor analysis, reliability assessment (both of the overall MOS and the factors identified in the factor analysis) and sensitivity investigation using analysis of variance on the independent variable of System.

2.2. Validity Evaluations

Relationship to paired comparisons. Data from a classified IBM report provided an opportunity to replicate the finding of Salza et al. (1996) that MOS ratings correlate significantly with paired comparisons. In the experiment described in the report, listeners provided paired comparisons after listening to samples from each of two TTS voices, then provided MOS ratings for each voice after hearing them a second time.

Relationship to intelligibility scores. Data from Wang and Lewis (2001) provided an opportunity to investigate the correlation between MOS ratings and intelligibility scores. In that experiment, listeners heard a variety of types of short phrases produced by four TTS voices, with the task to write down what the voice was saying. After finishing that intelligibility task, listeners heard the samples for each voice a second time and provided MOS ratings after reviewing each voice.

3. Results

3.1. Factor Analysis

After conducting a factor analysis of the MOS database, a parallel analysis (Coovert & McNelis, 1988) on the resulting eigenvalues indicated a three-factor solution that accounted for about 71% of the variance. Table 1 shows the results of the three-factor varimax-rotated solution, with bolded text to highlight the factor on which each item had the highest load. Note that the third factor only contains a single item. In normal use of the term, a factor has more than one contributing item, so in this report the conclusion is that the MOS has two factors with one item (Speaking Rate) not associated with either factor. Labeling factors is always a subjective exercise, but the factors do appear to be consistent with the factors reported by Kraft and Portele (1995), with items 2-5 (Listening Effort, Comprehension Problems, Speech Sound Articulation, Pronunciation) forming an Intelligibility factor and items 1 and 7 (Global Impression, Voice Pleasantness) forming a Naturalness factor.

Table 1. Three-Factor Varimax-Rotated Solution

	FAC1	FAC2	FAC3
MOS1	0.327	**0.900**	0.194
MOS2	**0.629**	0.370	0.427
MOS3	**0.693**	0.104	0.358
MOS4	**0.672**	0.433	0.294
MOS5	**0.746**	0.437	0.139
MOS6	0.322	0.204	**0.754**
MOS7	0.182	**0.665**	0.139

3.2. Reliability

Coefficient alpha for the overall MOS was 0.89, with 0.88 for the Intelligibility factor and 0.81 for the Naturalness factor. (It isn't possible to compute coefficient alpha for a single item.) Thus, the reliability of the MOS was acceptable. The reliability of the Naturalness subscale was somewhat lower than the Intelligibility subscale, probably due to it only having two items.

3.3. Sensitivity

Overall MOS rating. A between-subjects one-way analysis of variance on the overall MOS rating was statistically significant ($F(4, 68) = 7.6$, $p = .00004$). As expected, the recorded human voice (Wave) received the best rating, followed by the concatenative and formant-synthesized voices respectively.

Analysis by factor. Figure 1 shows the relationship among the TTS systems in the database and the MOS factors (including Speaking Rate). A mixed-factors analysis of variance indicated a significant main effect of System ($F(4, 68) = 9.6$, $p = .000003$), a significant main effect of MOS Factor ($F(2, 136) = 14.7$, $p = .000002$), and a significant System by Factor interaction ($F(8, 136) = 3.1$, $p = .003$).

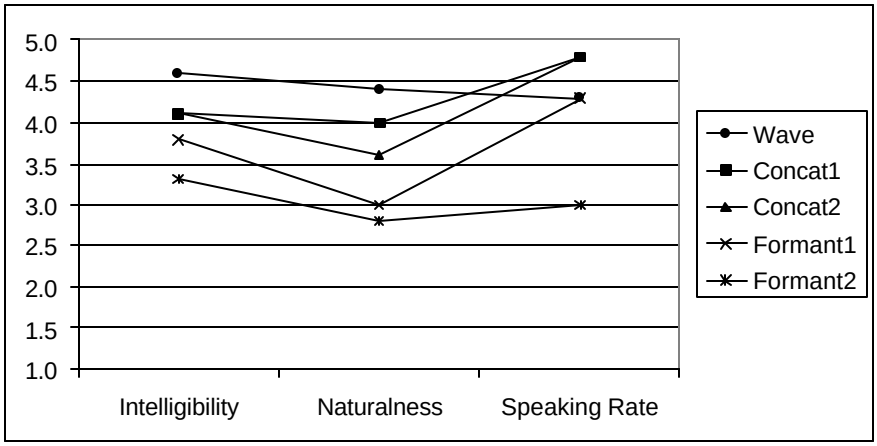
3.4. Validity

Correlation with paired comparisons. Correlations computed among the final preference votes (paired comparisons) of 16 listeners exposed to two distinctly different TTS systems and the mean difference scores for MOS ratings for both systems indicated that the validity coefficients for overall MOS, Naturalness and Intelligibility were

significant ($p < .10$, $r = .55$, .49, and .46 respectively). The correlation between paired comparisons and Speaking Rate was not significant ($r = .36$, $p = .172$).

Correlation with intelligibility scores from Wang and Lewis (2001). The only significant validity coefficient was that for Intelligibility ($r = -.43$, $p = .10$), which indicates evidence for both convergent and divergent validity. The evidence for convergent validity (having a significant relationship where expected) is the correlation between the MOS Intelligibility factor and the overall intelligibility score from Wang and Lewis. The evidence for divergent validity (failing to correlate significantly with scores hypothesized to tap into different constructs) is the non-significant correlations between the overall intelligibility score and the other MOS measurements (Overall MOS: $r = -.38$, $p = .15$; Naturalness: $r = -.19$, $p = .48$; Speaking Rate: $r = -.26$, $p = .33$).

Figure 1. Interaction of TTS System and MOS Factor



4. Discussion

The version of the MOS derived from Salza et al. (1996) seems to have reasonably good psychometric properties. The factor analysis of the current data resulted in a factor structure similar to that of Kraft and Portele (1995), specifically two factors (Intelligibility and Naturalness) and an unrelated item for Speaking Rate. The reliability of the overall MOS and its subscales is acceptable. Furthermore, the data replicated the validity result of Salza et al. by showing a significant correlation between paired comparison data and MOS data (Overall MOS, Naturalness, and Intelligibility). The data also indicated appropriate convergent and divergent validity for the intelligibility scores from Wang and Lewis (2001). Note that this result is similar to that reported by Johnston (1996), who found that the Listening Effort item (which is part of the Intelligibility factor) was more sensitive to degradation of speech intelligibility than the Global Effort item (which is part of the Naturalness factor).

Using principles from psychometrics (Nunnally, 1978), it should be possible to improve the reliability of the MOS. Rather than using 5-point scales with an anchor at each step, general reliability should improve slightly with a change to 7-point bipolar scales. Because the Naturalness factor had somewhat weaker reliability than the Intelligibility factor, it would be reasonable to add at least one more item to the MOS that is likely to tap into the construct of Naturalness.

The MOS Speaking Rate item failed to fall onto either the Intelligibility or Naturalness factor in both the current study and in Kraft and Portele (1995). This might have happened because Speaking Rate is truly independent of either of these constructs, or might have been an artifact due to the unique labeling of the scale points for this item. The other items have scales that have a clear ordinal pattern, such as "Excellent", "Good", "Fair", "Poor", and "Bad" for the Global Impression item. The labels for the Speaking Rate item are, in contrast, "Yes", "Yes, but slower than preferred", "Yes, but faster than preferred", "No, too slow", and "No, too fast", which do not have a clear top-to-bottom ordinal relationship. If the item assessing Speaking Rate had the same structure as the other items in the MOS, a future factor analysis could determine less ambiguously whether Speaking Rate is truly independent of Intelligibility and Naturalness.

5. A Proposed New Version of the MOS

The key proposals are to increase the number of scale steps per item from five to seven (using bipolar scales), to increase the number of items related to Naturalness, and to make the structure of the Speaking Rate item consistent with the other items. These modifications should improve the reliability of the MOS and, by extension, its other psychometric properties because reliability constrains the magnitude of validity coefficients (Nunnally, 1978) and limits a scale's sensitivity. The summary version of the revised MOS presented here shows the text of the item and its bipolar labels. (For the completely revised MOS and more details, see Lewis, 2001.)

1. *Global Impression:* Please rate the sound quality of the voice you heard. (Very Bad / Excellent)
2. *Listening Effort:* Please rate the degree of effort you had to make to understand the message. (Impossible Even with Much Effort / No Effort Required)
3. *Comprehension Problems:* Were single words hard to understand? (All Words Hard to Understand / All Words Easy to Understand)
4. *Speech Sound Articulation:* Were the speech sounds clearly distinguishable? (Not at All Clear / Very Clear)
5. *Pronunciation:* Did you notice any problems in the naturalness of sentence pronunciation? (Very Many Problems / Didn't Notice Any)
6. *Voice Pleasantness:* Was the voice you heard pleasant to listen to? (Very Unpleasant / Very Pleasant)
7. *Voice Naturalness:* Did the voice sound natural? (Very Unnatural / Very Natural)
8. *Ease of Listening:* Would it be easy to listen to this voice for long periods of time? (Very Difficult / Very Easy)
9. *Speaking Rate:* Was the speed of delivery of the message appropriate? (Poor Rate of Speech / Perfect Rate of Speech -- also include this additional sub-item: "If unsatisfactory, please circle one: Too Slow or Too Fast")

If the proposed changes work as expected, the revised MOS items 2-5 will continue to form an Intelligibility factor. Items 1 and 6-8 should form a Naturalness factor with substantially greater reliability (possibly in excess of .90) than the current Naturalness factor due to the additional items and the shift from five to seven scale steps. The change in the structure of item 9 (formerly item 6) should make it possible to determine whether Speaking Rate is truly independent of the other two factors without losing the ability to determine if a listener finds it too slow or fast.

6. References

Coovert, M. D., & McNelis, K. (1988). Determining the number of common factors in factor analysis: A review and program. *Educational and Psychological Measurement, 48*, 687-693.

Johnston, R. D. (1996). Beyond intelligibility: The performance of text-to-speech synthesisers. *BT Technology Journal, 14*, 100-111.

Kraft, V., & Portele, T. (1995). Quality evaluation of five German speech synthesis systems. *Acta Acustica, 3*, 351-365.

Landauer, T. K. (1988). Research methods in human-computer interaction. In M. Helander (Ed.), *Handbook of human-computer interaction*. New York: Elsevier.

Lewis, J. R. (1993). Multipoint scales: Mean and median differences and observed significance levels. *International Journal of Human-Computer Interaction, 5*, 383-392.

Lewis, J. R. (2001). *Psychometric properties of the Mean Opinion Scale* (Tech. Report in press -- will be available at http://sites.netscape.net/jrlewisinfl after publication).

Nunnally, J. C. (1978). *Psychometric theory*. New York: McGraw-Hill.

Paris, C. R., Thomas, M. H., Gilson, R. D., & Kincaid, J. P. (2000). Linguistic cues and memory for synthetic and natural speech. *Human Factors, 42*, 421-431.

Salza, P. L., Foti, E., Nebbia, L., & Oreglia, M. (1996). MOS and pair comparison combined methods for quality evaluation of text to speech systems. *Acta Acustica, 82*, 650-656.

Sonntag, G. P., Portele, T., Haas, F., & Kohler, J. (1999). Comparative evaluation of six German TTS systems. In *Eurospeech '99* (pp. 251-254). Budapest: Technical University of Budapest.

Wang, H., & Lewis, J. R. (2001). Intelligibility and acceptability of short phrases generated by text-to-speech (to appear in the conference proceedings for Human-Computer Interaction International '01).

Yabuoka, H., Nakayama, T., Kitabayashi, Y., & Asakawa, Y. (2000). Investigations of independence of distortion scales in objective evaluation of synthesized speech quality. *Electronics and Communications in Japan, Part 3, 83*, 14-22.