Psychometric Properties of the Mean Opinion Scale

TR 29.3403
March 12, 2001

James R. Lewis

IBM Voice Systems

West Palm Beach, Florida

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 marginally acceptable reliability for 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. The current MOS has acceptable psychometric properties, but adding items to the naturalness scale and increasing the number of scale steps should improve its reliability.

ITIRC Keywords

Mean Opinion Scale
MOS
Artificial speech
Synthetic speech
Text-to-speech
Psychometric evaluation

Contents

Executive Summary

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 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.

The factor structure of the MOS is currently in question. Kraft and Portele (1995), using an eight-item version of the MOS, reported two factors – one interpreted as intelligibility and one as naturalness (with speaking rate remaining as a single item not strongly associated with either factor). More recently, Sonntag et al. (1999), using the same version (but with 6-point rather than 5-point scales), reported only a single factor. The MOS has had some recent validation by several independent laboratories (Johnston, 1996; Salza, Foti, Nebbia, & Oreglia, 1996; Yabuoka, Nakayama, Kitabayashi, and Asakawa, 2000). The goals of the current research were to evaluate the factor structure of the 7-item version of the MOS (using 5-point scales), to estimate the reliability of the overall MOS score and any revealed factors, and to extend the work on sensitivity and validity of the MOS.

Analysis of data from 73 participants gathered over the last two years from six IBM internal evaluations of TTS systems revealed (1) two MOS factors (Intelligibility with 4 items and Naturalness with 2 items plus Speaking Rate as an independent item not strongly associated with either factor), (2) good reliability for Overall MOS (α=.89) and for the Intelligibility and Naturalness factors (.88 and .81, respectively), (3) appropriate sensitivity of MOS factors to manipulation of TTS system, (4) validity of MOS factors related to paired comparisons (replicating Salza et al., 1996), and (5) validity of the MOS Intelligibility factor related to intelligibility scores (from Wang & Lewis, 2001). The data indicate that the current MOS has acceptable psychometric properties, but also suggest that adding items to the naturalness scale and increasing the number of scale steps should improve its reliability.

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).

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 statistically significant.

Sensitivity. A measurement is sensitive if it responds to experimental manipulation. For a measurement to be sensitive enough 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 only 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).

Previous Research in MOS Psychometrics

Reliability. A literature review turned up no previous work reporting the reliability of the MOS 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 some 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.

Goals of the Current Research

The goals of the current research were to:

- 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)
- estimate the reliability of the overall MOS score and any revealed factors
- investigate the sensitivity of the MOS scores.
- extend the work on validity of the MOS.

Method

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, shown in Appendix A). (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.) The database appears in Appendix B, and 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.

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 all the samples of a given voice for 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.

Results

Factor Analysis

Figure 1 shows the scree plot from a factor analysis of the MOS database. The results of a parallel analysis (Coovert & McNelis, 1988) indicated a three-factor solution accounting 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.

Figure 1. Scree Plot from Factor Analysis

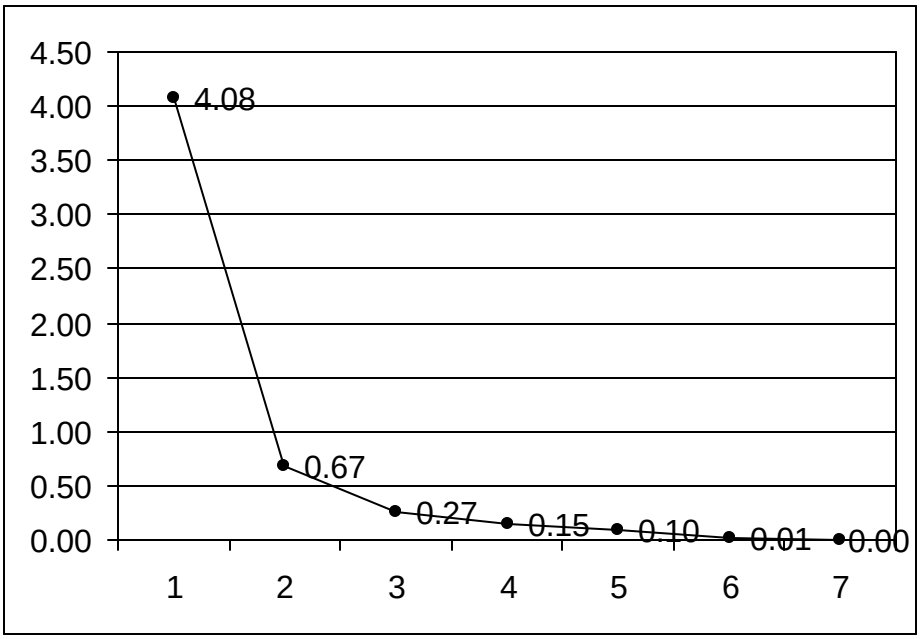


Table 1. Three-Factor Varimax-Rotated Solution

	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

Reliability

Table 2 shows coefficient alpha for the overall MOS and for each of the factors. (It isn't possible to compute coefficient alpha for a single item.)

The reliabilities of the overall MOS and subscales based on the Intelligibility and Naturalness subscales are acceptable (greater than .70).

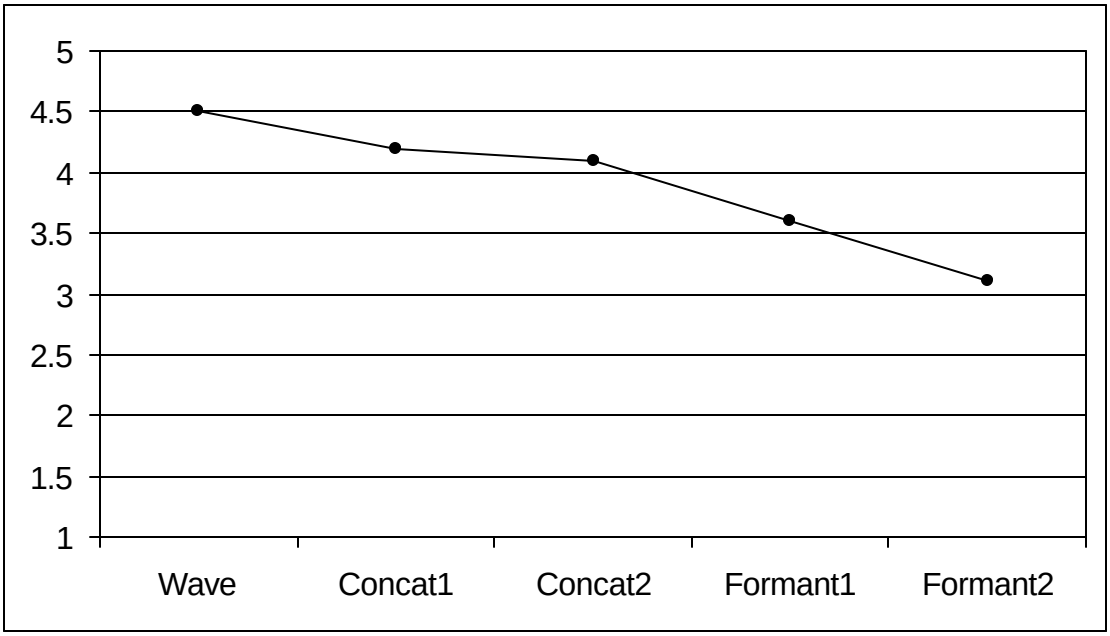
Table 2. MOS Reliability Estimates

Factor	Alpha
Overall	0.89
Intelligibility	0.88
Naturalness	0.81

Sensitivity

Overall MOS rating. Figure 2 shows the overall mean rating for each TTS system in the database. An associated between-subjects one-way analysis of variance 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.

Figure 2. Overall Mean Ratings for Database TTS Systems

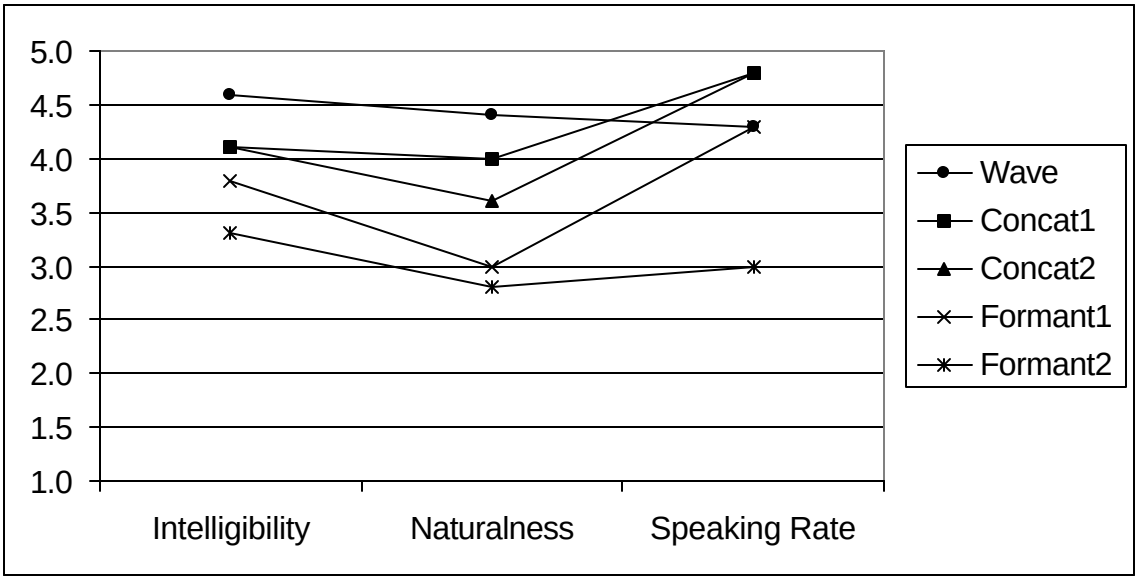


Analysis by factor. Table 3 and Figure 3 show the relationship among the TTS systems in the database and the MOS factors (including Speaking Rate). A mixed-factors analysis of variance indicated a significant main effect of System ($F(4, 68) = 9.6$, $p = .000003$), a significant main effect of MOS Factor ($F(2, 136) = 14.7$, $p = .000002$), and a significant System by Factor interaction ($F(8, 136) = 3.1$, $p = .003$).

Table 3. TTS System by MOS Factor Interaction

System	Intelligibility	Naturalness	Speaking Rate
Wave	4.6	4.4	4.3
Concat1	4.1	4.0	4.8
Concat2	4.1	3.6	4.8
Formant1	3.8	3.0	4.3
Formant2	3.3	2.8	3.0

Figure 3. Interaction of TTS System and MOS Factor

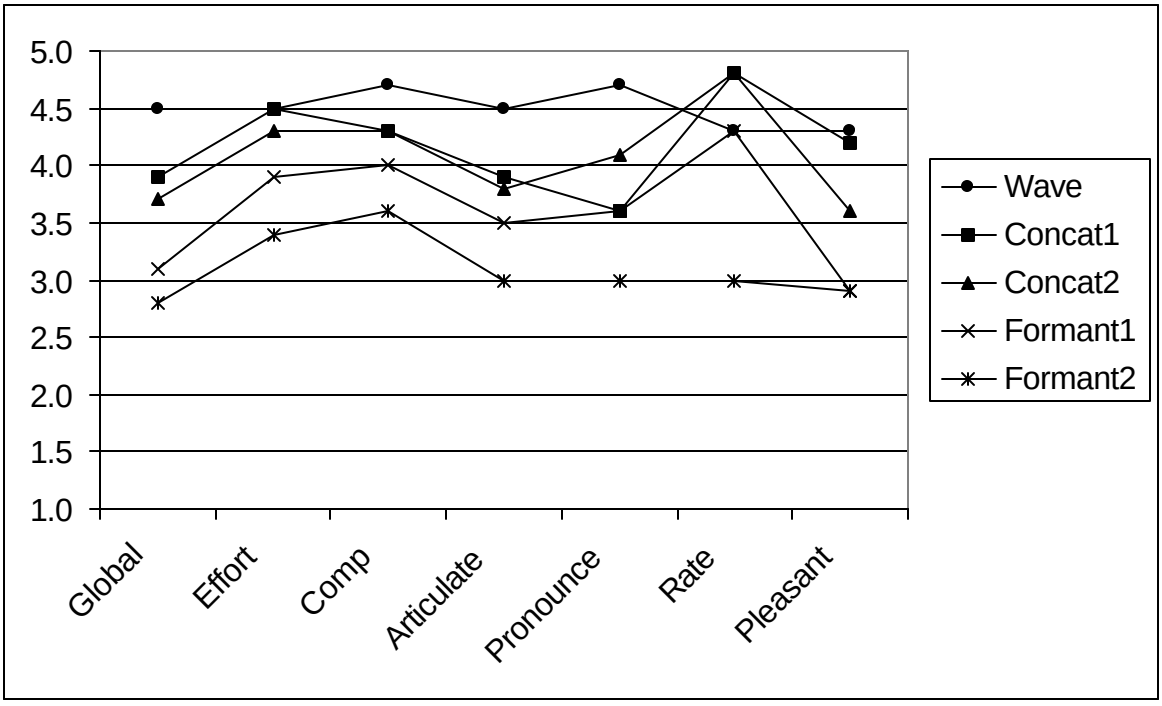


Analysis by item. Table 4 and Figure 4 show the interaction between TTS System and MOS Item. An analysis of variance on this level of the data a significant main effect of System ($F(4, 68) = 7.6$, $p = .00004$), a significant main effect of MOS Factor ($F(6, 408) = 4.9$, $p = .00007$), and a significant System by Factor interaction ($F(24, 408) = 1.8$, $p = .01$).

Table 4. TTS System by MOS Item Interaction

System	Global	Effort	Comp	Articulate	Pronounce	Rate	Pleasant
Wave	4.5	4.5	4.7	4.5	4.7	4.3	4.3
Concat1	3.9	4.5	4.3	3.9	3.6	4.8	4.2
Concat2	3.7	4.3	4.3	3.8	4.1	4.8	3.6
Formant1	3.1	3.9	4.0	3.5	3.6	4.3	2.9
Formant2	2.8	3.4	3.6	3.0	3.0	3.0	2.9

Figure 4. Interaction of TTS System and MOS Item



Validity

Correlation with paired comparisons. Table 5 shows the correlations 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, organized by overall MOS rating, the Intelligibility factor, the Naturalness factor, and the Speaking Rate item. The validity coefficients for overall MOS, Naturalness and Intelligibility were significant ($p < .10$). The correlation between paired comparisons and Speaking Rate was not significant ($p = .172$).

Table 5. Validity Coefficients for MOS Measurements and Paired Comparisons

	MOS	Naturalness	Intelligibility	Speaking Rate
Correlation	0.55	0.49	0.46	0.36
Probability	0.028	0.054	0.073	0.172

Correlation with intelligibility scores. Table 6 shows the correlations among the intelligibility scores from Wang and Lewis (2001). The only significant validity coefficient was that for Intelligibility (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.

Table 6. Validity Coefficients for MOS Measurements and Intelligibility

	MOS	**Naturalness**	**Intelligibility**	**Speaking Rate**
Correlation	-0.38	-0.19	-0.43	-0.26
Probability	0.15	0.48	0.10	0.33

Discussion

Summary

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 reliabilities of the overall MOS and its Intelligibility and Naturalness subscales are acceptable (greater than the minimal standard of .70). Consistent with this finding, the evidence for appropriate sensitivity was strong. 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).

Improving the Reliability of the MOS

Naturalness. 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, overall 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 two more items to the MOS that are likely to tap into the construct of Naturalness.

Speaking rate. 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
- 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
- No, too fast

which do not have a clear top-to-bottom ordinal relationship. If the item(s) 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 independent of Intelligibility and Naturalness, or whether it is actually associated with one of these two subscales of the MOS.

A Proposed New Version of the MOS

This section contains proposed modifications of the MOS to improve its reliability and, by extension, its other psychometric properties because reliability constrains the magnitude of validity coefficients (Nunnally, 1978) and also limits a scale's sensitivity.

1. *Global Impression:* Please rate the sound quality of the voice you heard.

VERY BAD	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**EXCELLENT**

2. *Listening Effort:* Please rate the degree of effort you had to make to understand the message.

IMPOSSIBLE EVEN WITH MUCH EFFORT	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**NO EFFORT REQUIRED**

3. *Comprehension Problems:* Were single words hard to understand?

ALL WORDS HARD TO TO UNDERSTAND UNDERSTAND	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**ALL WORDS EASY**

4. *Speech Sound Articulation:* Were the speech sounds clearly distinguishable?

NOT AT ALL CLEAR	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**VERY CLEAR**

5. *Pronunciation:* Did you notice any problems in the naturalness of sentence pronunciation?

VERY MANY PROBLEMS	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**DIDN'T NOTICE ANY**

6. *Voice Pleasantness:* Was the voice you heard pleasant to listen to?

VERY UNPLEASANT	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**VERY PLEASANT**

7. *Voice Naturalness:* Did the voice sound natural?

VERY UNNATURAL	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**VERY NATURAL**

8. *Ease of Listening:* Would it be easy to listen to this voice for long periods of time?

VERY DIFFICULT	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**VERY EASY**

9. *Speaking Rate:* Was the speed of delivery of the message appropriate?

POOR RATE OF SPEECH	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**PERFECT RATE OF SPEECH**

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 and, with the shift from five to seven scale steps, should achieve reliability in excess of .90. 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 addition of two 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 that item is truly independent of the other two factors without losing the ability to determine if a listener is satisfied with the speaking rate or finds it too slow or fast.

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.

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.

Appendix A. The MOS

The MOS uses 5-point scales. For this analysis, a higher number indicates a better rating. The seven MOS items (from Salza et al., 1996) are:

1. Global Impression: Your answer must indicate how you rate the sound quality of the voice you have heard.

[ ] Excellent

[ ] Good

[ ] Fair

[ ] Poor

[ ] Bad

2. Listening Effort: Your answer must indicate the degree of effort you had to make to understand the message.

[ ] No effort required

[ ] Slight effort required

[ ] Effort required

[ ] Major effort required

[ ] Message not understood with any feasible effort

3. Comprehension Problems: Your answer must indicate if you found single words hard to understand.

[ ] None

[ ] Few

[ ] Some

[ ] Many

[ ] Every word

4. Speech Sound Articulation: Your answer must indicate if the speech sounds are clearly distinguishable.

[ ] Yes, very clearly

[ ] Yes, clearly enough

[ ] Fairly clear

[ ] No, not very clear

[ ] No, not at all

5. Pronunciation: Your answer must indicate if you noticed any anomalies in the naturalness of sentence pronunciation.

[ ] No

[ ] Yes, but not annoying

[ ] Yes, slightly annoying

[ ] Yes, annoying

[ ] Yes, very annoying

6. Speaking Rate: Your answer must indicate if you found the speed of delivery of the message appropriate.

[ ] Yes

[ ] Yes, but slower than preferred

[ ] Yes, but faster than preferred

[ ] No, too slow

[ ] No, too fast

7. Voice Pleasantness: Your answer must indicate if you found the voice you have heard pleasant.

[ ] Very pleasant

[ ] Pleasant

[ ] Fair

[ ] Unpleasant

[ ] Very unpleasant

Appendix B. Database of 73 Independent MOS Questionnaires

PART	STUDY	SYSTEM	MOS1	MOS2	MOS3	MOS4	MOS5	MOS6	MOS7
1	TTS99	FORM1	3	5	5	4	5	5	2
2	TTS99	FORM1	4	5	5	4	5	5	3
3	TTS99	FORM1	2	4	5	3	3	5	3
4	TTS99	FORM1	1	3	3	3	1	4	1
5	TTS99	FORM1	2	4	4	3	3	5	2
6	TTS99	FORM1	4	5	5	4	5	4	4
7	TTS99	FORM1	3	4	4	4	4	5	2
8	TTS99	FORM1	5	5	5	5	5	5	4
9	TTS99	FORM1	3	4	4	3	3	4	3
10	TTS99	FORM1	3	4	4	4	3	5	3
11	TTS99	FORM1	4	5	5	4	4	5	3
12	TTS99	FORM1	3	5	5	5	5	5	3
13	TTS99	FORM1	4	5	5	5	5	4	4
14	TTS99	FORM1	4	5	5	4	5	5	4
15	TTS99	FORM1	2	4	5	3	5	5	2
16	SHORT	FORM2	4	4	5	4	5	5	4
17	SHORT	FORM2	2	4	5	3	3	3	2
18	SHORT	FORM2	2	2	5	3	2	1	2
19	SHORT	FORM2	3	3	2	2	3	1	3
20	SHORT	FORM2	3	4	4	4	4	3	3
21	SHORT	FORM2	2	2	3	2	2	1	2
22	SHORT	FORM2	3	5	1	3	3	3	3
23	SHORT	FORM2	1	3	2	2	2	1	2
24	SHORT	FORM2	1	3	3	2	1	5	1
25	SHORT	FORM2	3	3	3	3	3	5	3
26	SHORT	FORM2	4	5	5	5	5	4	4
27	SHORT	FORM2	3	3	5	3	2	3	4
28	SHORT	FORM2	3	4	4	4	4	3	3
29	SHORT	FORM2	2	2	3	2	2	2	3
30	SHORT	FORM2	5	4	3	3	3	3	2
31	SHORT	FORM2	4	4	4	3	4	5	5
32	CON1	CONCAT1	4	5	5	4	5	5	4
33	CON1	CONCAT1	3	4	5	3	5	3	3
34	CON1	CONCAT1	3	4	5	4	3	5	2
35	CON1	CONCAT1	4	4	3	3	3	5	3
36	CON1	CONCAT1	3	4	4	3	3	4	3

Note: The STUDY column contains codes for the six experiments used as sources for the database. The SYSTEM column contains codes for the selected voice that listeners rated with the MOS. FORM indicates a formant-synthesized voice, CONCAT indicates a concatenative voice, and WAVE indicates a recorded human voice.

PART	STUDY	SYSTEM	MOS1	MOS2	MOS3	MOS4	MOS5	MOS6	MOS7
37	CON1	CONCAT1	4	4	5	4	4	5	5
38	CON1	CONCAT1	3	3	3	4	3	4	4
39	CON1	CONCAT1	4	5	5	4	4	5	3
40	CON1	CONCAT1	4	5	5	4	5	5	4
41	CON1	CONCAT1	4	5	4	4	4	5	5
42	CON1	CONCAT1	3	4	4	4	5	5	3
43	CON1	CONCAT1	3	4	4	3	4	5	3
44	CON1	CONCAT1	4	4	4	4	4	5	4
45	CON1	CONCAT1	4	4	4	4	4	5	4
46	CON1	CONCAT1	4	5	5	4	4	5	3
47	CON1	CONCAT1	5	5	4	5	5	5	4
48	CON2	CONCAT2	4	4	4	4	3	5	4
49	CON2	CONCAT2	4	3	4	4	5	5	5
50	CON2	CONCAT2	4	4	4	4	3	5	5
51	CON2	CONCAT2	4	5	4	4	4	5	4
52	CON2	CONCAT2	5	5	4	5	5	5	4
53	CON2	CONCAT2	3	4	4	3	2	5	4
54	CON2	CONCAT2	4	5	5	5	4	4	4
55	CON2	CONCAT2	5	5	4	4	3	5	5
56	CON2	CONCAT2	3	5	4	2	3	5	5
57	CON2	CONCAT2	2	4	4	4	3	4	3
58	CON2	CONCAT2	4	5	4	4	4	5	4
59	CON2	CONCAT2	4	5	4	3	3	5	4
60	CON2	CONCAT2	3	3	4	4	2	4	2
61	CON2	CONCAT2	4	5	5	4	4	5	4
62	CON2	CONCAT2	4	5	5	5	5	5	5
63	CON2	CONCAT2	5	5	5	4	5	5	4
64	CONCUR	WAVE1	3	3	4	3	3	4	4
65	CONCUR	WAVE1	5	5	5	5	5	5	4
66	CONCUR	WAVE1	4	5	5	4	5	5	4
67	CONCUR	WAVE1	5	4	4	5	5	5	4
68	CONCUR	WAVE1	5	5	5	5	5	3	5
69	CONCUR	WAVE1	5	5	5	5	5	4	5
70	WEB1	FORM1	3	2	2	2	2	1	3
71	WEB1	FORM1	3	2	1	2	2	2	3
72	WEB1	FORM1	2	2	1	2	3	2	4
73	WEB1	FORM1	3	2	3	3	1	5	3

Note: The STUDY column contains codes for the six experiments used as sources for the database. The SYSTEM column contains codes for the selected voice that listeners rated with the MOS. FORM indicates a formant-synthesized voice, CONCAT indicates a concatenative voice, and WAVE indicates a recorded human voice.