Enhancement of the Mean Opinion Scale - Expanded (MOS-X)

TR 29.3542
July 12, 2002

Melanie D. Polkosky
James R. Lewis

IBM Voice Systems
Boca Raton, Florida

Abstract

The Mean Opinion Scale-Expanded (MOS-X) is a questionnaire used to evaluate synthetic voices. Previous research demonstrated that the previous MOS-X had four factors (Intelligibility, Naturalness, Prosody, and Social Impression) but the Social Impression factor had relatively low reliability. In this study, we added two new items to the Social Impression factor (Enthusiasm and Persuasiveness) and improved its reliability. Deletion of the Depression item from this scale had no impact on the reliability of the scale, and created a concise 15-item measure. The resulting scale, the MOS-X, is a psychometrically sound instrument for measuring listeners' perceptions of synthetic speech in industrial settings.

ITIRC Keywords

Mean Opinion Scale
Mean Opinion Scale-Revised
Expanded Mean Opinion Scale
MOS
MOS-R
MOS-X
Artificial speech
Synthetic speech
Text-to-speech
Psychometric evaluation

Contents

Introduction

In previous work (Lewis, 2001a, 2001b; Polkosky & Lewis, 2001), we evaluated the psychometric properties of the Mean Opinion Scale (MOS), its revision (MOS-R), and its expanded version (MOS-X). These scales assess listeners' perceptions of the speech and social-perceptual characteristics of synthetic voices. The most recent expansion of the scale, the MOS-X (Polkosky & Lewis, 2001) included four scales measuring Intelligibility, Naturalness, Prosody, and Social Impression.

Based on a series of evaluations (Polkosky & Lewis, 2001), we recommended the addition of two scales (Enthusiasm and Persuasiveness) to improve the reliability of the Social Impression factor. Thus, the purpose of the current research was to further improve the psychometric properties of the MOS-X. Accurate and reliable measurement of the subtle perceptual characteristics is important for understanding listeners' impressions of synthetic speech, developing increasingly sophisticated synthetic speech, and discriminating effectively among IBM and competitors' artificial voices.

Method

Participants
The sample included complete sets of ratings from 327 participants, randomly selected from IBM Callup.

Design and Measures
The study used a between-subjects design with ten levels of the independent variable of synthetic voice. The voices and their key characteristics[1] were:

- A: concatenative female, 8 kHz
- B: concatenative male, 8 kHz
- C: concatenative male, 22 kHz
- D: concatenative male, 8 kHz
- E: concatenative male, 8 kHz
- F: concatenative female, 22 kHz
- G: concatenative male, 22 kHz
- H: concatenative male, 8 kHz
- I: concatenative male, 8 kHz
- J: concatenative male, 8 kHz.

The dependent measures were the ratings for the 14 MOS-X items (Listening Effort, Comprehension Problems, Articulation, Precision, Voice Pleasantness, Voice Naturalness, Humanlike Voice, Voice Quality, Emphasis, Rhythm, Intonation, Trust, Confidence, and Depression). In addition, we added the two proposed items, Enthusiasm and Persuasiveness, which we expected to align with the Social Impression factor.

Procedure
Participants received an email inviting them to participate in the study and directing them to a web page (one page for each participant group) with instructions, a link to a recording of one of the synthetic voices, and the rating scales. After accessing the web page, participants clicked the link that caused the synthetic voice file to play on the participant's audio player application. They then completed the 16 items for that voice.

Results and Discussion

The analysis consisted of a factor analysis to determine if the Enthusiasm and Persuasiveness scales clustered with the MOS-X Social Impression scale as expected. We also conducted additional statistical analysis to evaluate the reliability and sensitivity of the new MOS-X. We will refer to the final measure as the MOS-X for the remainder of this report.

Factor Analysis
For the factor analysis, we forced a 4-factor solution due to the results of a previous study (Polkosky & Lewis, 2001), showing that the two proposed scales loaded on Social Impression.

Table 1 shows the association of each item with each of the four factors; the highest loading (indicating strongest association) appears in bold. As shown, four items loaded on Factor 1 (items 1-4), three items loaded on factor 2 (items 9-11), five items loaded on Factor 3 (items 12-16), and four items loaded on Factor 4 (items 5-9). This pattern suggested that we retain the MOS-X factor structure and labels (Intelligibility, Prosody, Social Impression, and Naturalness). As predicted, both Enthusiasm and Persuasiveness loaded on Social Impression. Overall, this factor structure explained 65.56% of the variance in participant ratings.

[1] The purpose of this research was to evaluate the new MOS items – not to perform a competitive evaluation of voices. For this reason, we do not provide the details on the companies from which we obtained the voices.

Table 1. Factor Loadings for the MOS-X Four-Factor Solution

Item	Content	Factor1	Factor2	Factor3	Factor4
		Intelligibility	Prosody	Social Impression	Naturalness
1	Listening Effort	**0.730**	0.156	0.085	0.174
2	Comprehension	**0.808**	0.039	0.082	0.114
3	Articulation	**0.865**	0.095	0.103	0.151
4	Pronunciation	**0.716**	0.209	0.135	0.140
5	Pleasantness	0.218	0.181	0.477	**0.588**
6	Voice Naturalness	0.289	0.445	0.293	**0.670**
7	Humanlike Voice	0.240	0.376	0.294	**0.626**
8	Voice Quality	0.342	0.090	0.387	**0.466**
9	Emphasis	0.151	**0.662**	0.338	0.111
10	Rhythm	0.155	**0.744**	0.293	0.269
11	Intonation	0.189	**0.723**	0.343	0.256
12	Trust	0.229	0.338	**0.622**	0.316
13	Confidence	0.239	0.285	**0.691**	0.265
14	Depression	0.096	0.107	**0.631**	0.192
15	Enthusiasm	-0.002	0.311	**0.700**	0.150
16	Persuasiveness	0.060	0.379	**0.743**	0.154

Reliability

Table 2 shows reliability of each factor and the overall scale. Intelligibility (0.88), Naturalness (0.86), Prosody (0.86), Social Impression (0.86), and the Overall score (0.93) had coefficient alphas greater than 0.70, demonstrating reliabilities adequate for usability evaluation (Landauer, 1988).

To create a more efficient measure, we removed the Depression item from the Social Impression factor and recalculated coefficient alpha. This procedure allowed us to develop a measure with fewer items while maintaining consistent reliability. Social Impression retained its reliability (0.86), and the resulting 15-item scale continued to have high reliability (0.93).

As a result of this analysis, we retained 15 items in the final version of the MOS-X (see Appendix A): Listening Effort, Comprehension, Articulation, Precision, Pleasantness, Voice Naturalness, Humanlike Voice, Voice Quality, Emphasis, Rhythm, Intonation, Trust, Confidence, Enthusiasm, and Persuasiveness. The resulting factor structure and item alignment appears in Table 2.

Table 2. Four Factor MOS-X

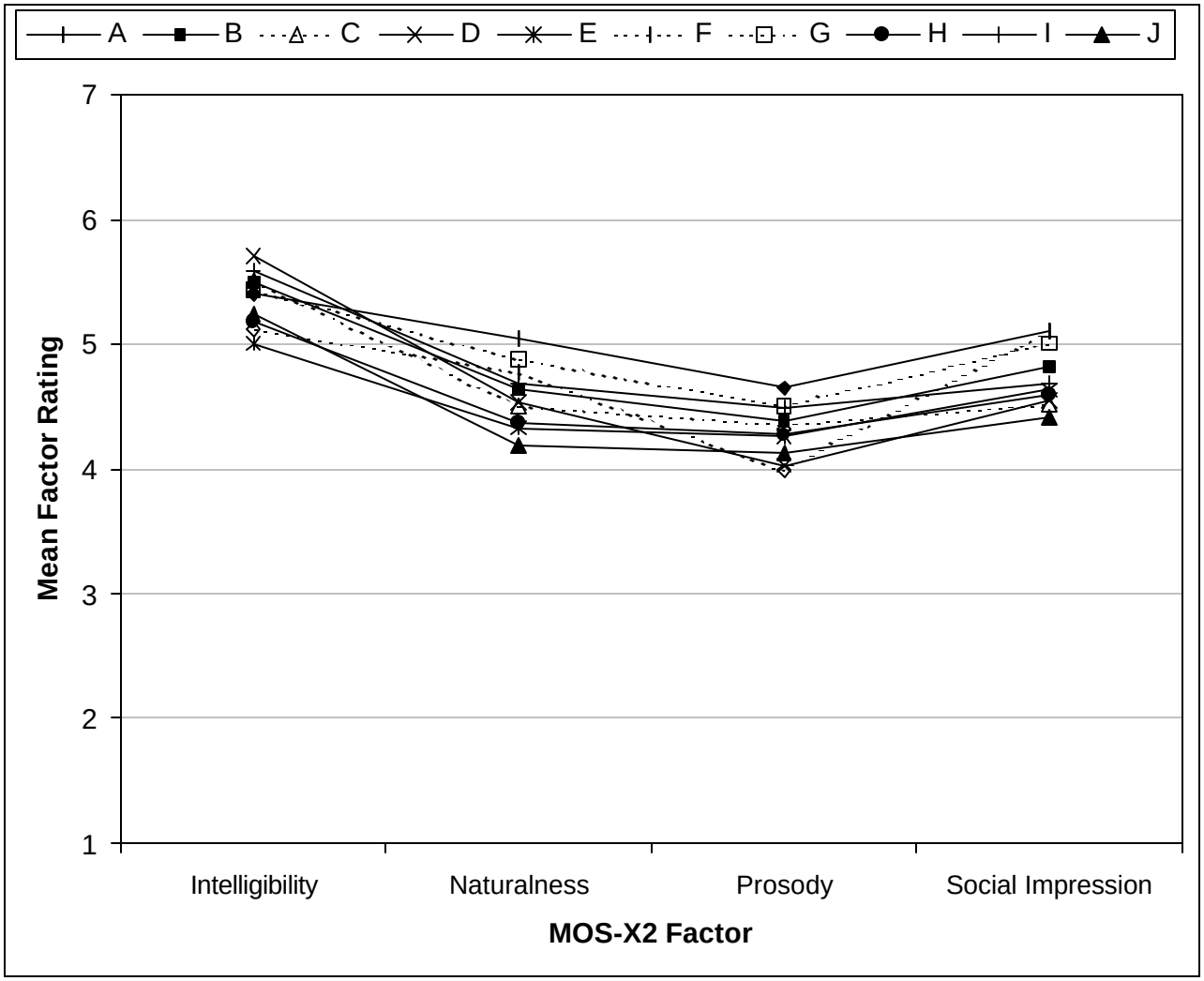
Intelligibility	Prosody	Social Impression	Naturalness
Listening Effort	Emphasis	Trust	Pleasantness
Comprehension Problems	Rhythm	Confidence	Naturalness
Articulation	Intonation	Enthusiasm	Humanlike Voice
Precision		Persuasiveness	Voice Quality

Sensitivity

A mixed model ANOVA indicated the extent to which the MOS-X discriminated among the ten synthetic voices. The ANOVA showed a significant effect of factor ($F(3,951) = 443.96$, $MSe = 3683.46$, $p < 0.0001$), and a significant interaction between factor and voice ($F(27,951) = 2.37$, $p < 0.0001$). The presence of these effects suggested that the factor profiles for the voices were significantly different. Generally, participants rated Intelligibility most positively and Prosody most poorly of the four factors. The main effect of synthetic voice was not significant ($F(9,317) = 1.01$, $MSe =$

175.10, p = 0.44), indicating a similar mean rating for the ten voices. Figure 1 illustrates these findings. The overall mean scores for the voices ranged from 4.50 (Voice J, least positive) to 5.10 (Voice A, most positive).

Figure 1. Voice by Factor Interaction (MOS-X)



Discussion

The current study continued previous psychometric work on the MOS scale and resolved a potential weakness of the most recent version. Polkosky and Lewis (2001) identified the relatively low reliability of the MOS-X Social Impression factor as a potential weakness, and suggested replacing the Depression item with Enthusiasm and Persuasiveness items. The present evaluation demonstrated that this change strengthened the reliability of the Social Impression factor, and simultaneously created a concise, psychometrically sound evaluation tool for synthetic speech. This line of research is important for reliable measurement of IBM's and competitors' current and future synthetic voices, and for appropriate diagnosis of perceptual limitations of artificial speech.

.

References

Lewis, J. R. (2001a). *Psychometric properties of the mean opinion scale* (Tech. Report 29.3403). Raleigh, NC: International Business Machines Corp.

Lewis, J. R. (2001b). *The revised mean opinion scale (MOS-R): Preliminary psychometric evaluation* (Tech. Report 29.3414). Raleigh, NC: International Business Machines Corp.

Polkosky, M. & Lewis, J. R. (2001). *Expansion and psychometric evaluation of the mean opinion scale – revised.* (Tech. Report 29.3414). Raleigh, NC: International Business Machines Corp.

Appendix A. Final Items and Item Arrangement for the MOS-X

1. *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**

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

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

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

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

4. *Precision*: Was the articulation of speech sounds precise?

SLURRED OR IMPRECISE	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**PRECISE**

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

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

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

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

7. *Humanlike Voice*: To what extent did this voice sound like a human?

NOTHING LIKE A HUMAN	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**JUST LIKE A HUMAN**

8. *Voice Quality*: Did the voice sound harsh, raspy, or strained?

SIGNIFICANTLY HARSH/RASPY	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**NORMAL QUALITY**

9. *Emphasis*: Did emphasis of important words occur?

INCORRECT EMPHASIS	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**EXCELLENT USE OF EMPHASIS**

10. *Rhythm*: Did the rhythm of the speech sound natural?

UNNATURAL OR MECHANICAL	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**NATURAL RHYTHM**

11. *Intonation*: Did the intonation pattern of sentences sound smooth and natural?

ABRUPT OR ABNORMAL	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**SMOOTH OR NORMAL**

12. *Trust*: Did the voice appear to be trustworthy?

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

13. *Confidence*: Did the voice suggest a confident speaker?

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

14. *Enthusiasm*: Did the voice seem to be enthusiastic?

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

15. *Persuasiveness*: Was the voice persuasive?

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

<u>MOS-X Scales</u>
Overall: Average items 1-15
Intelligibility: Average items 1-4
Naturalness: Average items 5-8
Prosody: Average items 9-11
Social Impression: Average items 12-15