Development and Psychometric Evaluation of an Expanded Mean Opinion Scale (MOS-X)

TR 29.3499
April 25, 2002

Melanie D. Polkosky
James R. Lewis

IBM Voice Systems
Boca Raton, Florida

Abstract

The Mean Opinion Scale-Revised (MOS-R) is a questionnaire used to evaluate synthetic voices. Previous research demonstrated that the MOS-R has adequate reliability and two factors: Intelligibility and Naturalness. In two studies, we expanded the content of the MOS-R to measure subtle vocal and social-emotional aspects of speech. Results indicated that the first revision had five factors (Intelligibility, Naturalness, Social Impression, Voice, and Fluency). The second revision had four factors (Intelligibility, Naturalness, Social Impression, and Negativity). A final analysis produced the Expanded MOS (MOS-X), which retained the traditional factors of Intelligibility and Naturalness and contained new Prosody and Social Impression factors.

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

The Mean Opinion Scale-Revised (MOS-R) is a ten-item questionnaire for the subjective evaluation of synthetic voices, developed at IBM[1] and adapted from the existing Mean Opinion Scale (MOS) scale (Kraft & Portele, 1995; Salza, Foti, Nebbia, & Oreglia, 1996). Several researchers have evaluated the MOS (Kraft & Portele, 1995; Lewis, 2001a) and improved its psychometric properties for use as a measurement tool in industrial settings (Lewis, 2001b). Although the measure has evolved from a seven-item scale of five-point bipolar ratings to a ten-item scale of seven-point bipolar ratings (see Table 1), the factor structure has remained relatively stable. The MOS and MOS-R measured two factors: Intelligibility and Naturalness. The scales also included a problematic Speaking Rate item that, until the most recent revision, did not consistently load on either Intelligibility or Naturalness (Lewis, 2001b). The MOS and MOS-R have been a primary method for measuring listener impressions of synthetic voices developed at IBM Voice Systems and elsewhere (Johnston, 1996; Kraft & Portele, 1995; Salza, Foti, Nebbia, & Oreglia, 1996; Yabuoka, Nakayama, Kitabayashi, & Asakawa, 2000).

Table 1. Summary of MOS and MOS-R Versions

Evaluation	**Items**	**Scale**	**Factors**
Mean Opinion Scale (Salza, Foti, Nebbia, & Oreglia, 1996)	1. Global Impression 2. Listening Effort 3. Comprehension Problems 4. Speech Sound Articulation 5. Pronunciation 6. Speaking Rate 7. Voice Pleasantness	5-point ordinal scales (except Speaking Rate)	Intelligibility (items 2-5) Naturalness (items 1, 7)
Mean Opinion Scale (Kraft & Portele, 1995)	1. Global Impression 2. Listening Effort 3. Comprehension Problems 4. Speech Sound Articulation 5. Pronunciation 6. Speaking Rate 7. Voice Pleasantness 8. Naturalness	5-point ordinal scales (except Speaking Rate)	Intelligibility Naturalness
Mean Opinion Scale (Sonntag, Portele, Haas, & Kohler, 1999)	1. Global Impression 2. Listening Effort 3. Comprehension Problems 4. Speech Sound Articulation 5. Pronunciation 6. Speaking Rate 7. Voice Pleasantness 8. Naturalness	6-point ordinal scales	Single factor
Mean Opinion Scale-Revised (Lewis, 2001b)	1. Global Impression 2. Listening Effort 3. Comprehension Problems 4. Speech Sound Articulation 5. Pronunciation 6. Speaking Rate 7. Voice Pleasantness 8. Naturalness 9. Ease of Listening 10. Humanlike Voice[2]	7-point ordinal scales	Intelligibility (items 1-6) Naturalness (items 7-9)

Previous evaluation and adaptation of the MOS has centered on improving its psychometric properties, especially its internal reliability and sensitivity. However, the content of MOS-R items has received comparatively little attention in

[1] IBM is a registered trademark of the International Business Machines Corp.
[2] Lewis (2001b) proposed addition of the Humanlike Voice item with the expectation that it would associate with the Naturalness factor.

previous research. The focus on only two factors may substantially limit the instrument's ability to discriminate among voices with similar intelligibility and naturalness. Indeed, researchers in the late 1980s and early 1990s acknowledged that the intelligibility of synthetic speech can rival that of human speech (Greene, Logan, & Pisoni, 1986; Murray & Arnot, 1993). As synthetic speech development has become increasingly sophisticated, it is reasonable to assume that intelligibility does not usually differentiate among current synthetic voices. With the introduction of concatenative voices, naturalness also is becoming less of a discriminating factor.

More recently, researchers have investigated the synthesis of more subtle and specific perceptual characteristics than intelligibility and naturalness. A significant psychological literature exists on the social-emotional aspects of human speech (for a review, see Murray & Arnot, 1993), the relationship between vocal speech and impression formation or personality perception (for a review, see Brown, Strong, & Rencher, 1975), and the social impact of speech disabilities, especially for individuals who use augmentative and alternative communication systems (synthetic voice prostheses) as a means of communication (Hoag & Bedrosian, 1992; Gorenflo & Gorenflo, 1997). All of these areas of research can inform measurement of listeners' vocal and social-emotional perceptions about synthetic speech. Numerous studies over the past three decades have investigated vocal speech characteristics that promote social-emotional perceptions, including:

- **Intonation, emphasis, or register** (Brown, Strong, & Rencher, 1973; Koopmans-Van Beinum, 1992; Pelachaud, Badler, & Steedman, 1996; Yaeger-Dror, 1996);
- **Fundamental frequency or pitch** (Bradlow, Torretta, & Pisoni, 1996; Hieda & Kuchinomachi, 1997; Higashikawa & Minifie, 1999; Slowiaczek & Nusbaum, 1985);
- **Speaking rate** (Bradlow, Torretta, & Pisoni, 1996; Brown, Strong, & Rencher, 1973; Slowiaczek & Nusbaum, 1985);
- **Timing** (Bradlow, Torretta, & Pisoni, 1996);
- **Intensity or loudness** (Granstrom & Nord, 1992; Page & Balloun, 1978; Robinson & McArthur, 1982);
- **Voice quality** (Hillenbrand, 1988; Klatt & Klatt, 1990; Lavner, Gath, & Rosenhouse, 2000; Whalen & Hoequist, 1995);
- **Nasality** (Bloom, Zajac, & Titus, 1999); and
- **Disfluency or hesitation** (Hosman, 1989; Martin & Haroldson, 1992).

Other researchers have investigated the social-emotional perceptions conveyed by speech, including:

- **Sadness** (Johnson, Emde, Scherer, & Klinnert, 1986; Murray & Arnot, 1995; Paddock & Nowicki, 1986);
- **Anger** (Johnson, Emde, Scherer, & Klinnert, 1986; Massaro & Egan, 1996; Murray & Arnot, 1995);
- **Fear** (Murray & Arnot, 1995);
- **Happiness** (Massaro & Egan, 1996; Murray & Arnot, 1995; Tartter & Braun, 1994);
- **Disgust** (Murray & Arnot, 1995);
- **Grief** (Murray & Arnot, 1995);
- **Stress** (Murray, Arnott, & Rohwer, 1996);
- **Fatigue** (Whitmore & Fisher, 1996);
- **Persuasiveness** (Holtgraves & Lasky, 1999; Stern, Mullennix, Dyson, & Wilson, 1999);
- **Attractiveness** (Berry, 1992; Miyake & Zuckerman, 1993; Zuckerman, Miyake, & Hodgins, 1991);
- **Truthfulness** (Ekman, O'Sullivan, Friesen, & Scherer, 1991); and
- **Gender** (Aronovitch, 1976; Newcombe & Arnkoff, 1979; Robinson & McArthur, 1982; Siegler & Siegler, 1976; Whiteside, 1999).

The primary purpose of the current research was to expand the content of the MOS-R to include items that measure subtle vocal and social-emotional aspects of speech. Accurate and reliable measurement of these perceptual characteristics is important to understanding listeners' impressions of synthetic speech, developing increasingly sophisticated synthetic speech, and discriminating effectively among IBM and competitors' artificial voices.

Study 1: MOS-R2a

The purpose of the first study was to add perceptual speech characteristics and social impression items not previously measured by the MOS-R. We expected that the new items would add new factors to the measure, which we hoped would improve its sensitivity and more clearly discriminate among user perceptions of synthetic voices. We limited the new items to primarily speech-based items consistent with the evaluative purpose of the previous MOS-R revisions.

Method

Participants
The sample consisted of 1000 randomly selected IBM employees, with 200 individuals in each of five groups. Of this sample, 204 individuals completed the study questions (20% return rate).

Design and Measures
The study used a between-subjects design with five levels of the independent variable of synthetic voice. The voices and their key characteristics[3] were:

- A: concatenative female
- B: concatenative female
- C: concatenative male
- D: concatenative male
- E: formant male

All voices had an 8 kHz sampling rate and 16-bit dynamic range. Voices A-D were concatenative; Voice E was formant. Voices A and B used the same underlying TTS technology. Voices C and D used different underlying TTS technologies (different from Voices A and B and different from each other).

The dependent measures were the ratings for the 22 MOS-R2a items shown in Appendix A. The items included 10 scales from the earlier version of the MOS (Global Impression, Listening Effort, Comprehension Problems, Articulation, Pronunciation, Voice Pleasantness, Voice Naturalness, Ease of Listening, Speaking Rate) and an item expected to align with Naturalness (Lewis, 2001b). We generated eight items based on clinical evaluation of human speech characteristics: voice (Loudness, Emphasis, Voice Quality, Pitch), fluency (Interruptions, Rhythm, Intonation), and articulation (Precision) (Shipley & McAfee, 1992). If human speech evaluation is similar to synthetic speech evaluation, we would predict that the fluency items would cluster with Speaking Rate to create a Fluency factor. Similarly, the new Precision item should align with the previous Intelligibility factor. Finally, we also generated four items related to the social impression created by human voices. These items were selected based on the review of previous literature and needs identified for application development (Topic Interest, Trust, Confidence, and Depression).

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 MOS-R2b items for that voice.

Results and Discussion

Due to a data collection error on its web page, we excluded the data for Voice A from the analysis and only analyzed the responses from the four remaining groups (a total of 160 participants).

[3] 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.

Factor Analysis

A discontinuity analysis (Cliff, 1987; Coovert & McNelis, 1988) indicated that the 22 items of the revised MOS-R measured five factors (accounting for 64.8% of the variance in the data).

Table 2 shows the association of each item with each of the five factors; the highest loading (indicating strongest association) appears in bold. As shown, seven items loaded on Factor 1 (items 1-5, 14, 18), five items loaded on factor 2 (items 10, 12, 16-17, 19), three items loaded on Factor 3 (items 11, 15, 22), two items loaded on Factor 4 (items 20-21), and five items loaded on Factor 5 (items 6-9, 13).

Table 2. Factor Loadings for the MOS-R2a Five-Factor Solution

Item	Content	Factor1	Factor2	Factor3	Factor4	Factor5
		Intelligibility	Fluency	Voice	Social Impression	Naturalness
1	Global Impression	**0.612**	0.237	0.253	0.193	0.448
2	Listening Effort	**0.712**	0.216	0.308	0.155	0.213
3	Comprehension	**0.742**	0.248	0.261	0.108	0.256
4	Articulation	**0.763**	0.203	0.209	0.158	0.340
5	Pronunciation	**0.487**	0.294	0.160	0.300	0.308
6	Pleasantness	0.243	0.217	0.315	0.219	**0.750**
7	Voice Naturalness	0.349	0.417	0.101	0.228	**0.605**
8	Ease of Listening	0.410	0.411	0.250	0.257	**0.511**
9	Humanlike Voice	0.398	0.342	0.073	0.214	**0.644**
10	Speaking Rate	0.306	**0.514**	0.264	-0.128	0.079
11	Loudness	0.171	0.105	**0.477**	0.086	0.069
12	Emphasis	0.181	**0.754**	0.197	0.202	0.182
13	Voice Quality	0.365	0.170	0.288	0.205	**0.524**
14	Interruptions	**0.516**	0.306	-0.171	0.429	-0.047
15	Pitch	0.274	0.248	**0.398**	0.044	0.319
16	Rhythm	0.267	**0.722**	-0.007	0.240	0.370
17	Intonation	0.282	**0.653**	-0.071	0.364	0.338
18	Precision	**0.612**	0.167	0.212	0.149	0.386
19	Topic Interest	0.068	**0.439**	0.285	0.410	0.266
20	Trust	0.173	0.202	0.246	**0.760**	0.352
21	Confidence	0.365	0.157	0.345	**0.662**	0.261
22	Depression	-0.104	-0.008	**-0.605**	-0.131	-0.133

Factor 1 included items previously associated with the MOS-R factor known as Intelligibility (Global Impression, Listening Effort, Comprehension, Articulation, Pronunciation), so we retained this label. The new item Precision associated with Intelligibility, as predicted. Similarly, Factor 5 included items consistent with the MOS-R factor called Naturalness (Pleasantness, Naturalness, Ease of Listening), adding the Humanlike Voice item (as predicted by Lewis, 2001b) and one additional item (Voice Quality). Therefore, we retained the Naturalness label for this factor. The remaining factors largely loaded according to the predicted factors of Fluency (Factor 2), Voice or phonation and its emotional correlates (Factor 3), and Social Impression (Factor 4). Of interest was the association of Voice Quality with Naturalness (instead of Voice) and Interruptions with Intelligibility (instead of Fluency). This result demonstrates that voice, fluency, and articulation may be problematic factor labels because of their specificity[4]. By contrast, Intelligibility and Naturalness are both broad and more abstract labels, since impairment in voice, fluency, and/or articulation diminishes both the intelligibility and naturalness of human speech.

[4] The previous MOS included the factor label Intelligibility. A similar labeling issue would occur if this factor had been previously labeled with the more specific and precise term Articulation. Although the items previously associated with this factor clearly relate to articulation (excluding other human speech characteristics), the more general label was provided.

Reliability
Table 3 shows reliability of each factor and the overall scale. Four factors (Intelligibility, Fluency, Social Impression, Naturalness) and the Overall score had coefficient alphas greater than 0.70, demonstrating reliabilities adequate for usability evaluation (Landauer, 1988). However, the Voice factor had inadequate reliability based on this criterion.

To create a more efficient measure, we removed items from each factor with the lowest loadings (or items that approximately equally loaded on more than one factor) and recalculated coefficient alpha. This procedure allowed us to develop a measure with fewer items while maintaining consistent reliability. Although the final instrument had six fewer items, the reliability of the factors and scale as a whole remained high (even improving for the Fluency factor), with the exception of the Voice factor.

Table 3. Original and Adjusted Reliability for the Five MOS-R2a Factors

Factor	Original Items	Original Coefficient Alpha	Retained Items	Adjusted Coefficient Alpha
Intelligibility	1-5, 14, 18	0.91	2-4, 18	0.91
Fluency	10, 12, 16-17, 19	0.85	12, 16-17	0.88
Voice	11, 15, 22	0.58	11, 22	0.46
Social Impression	20-21	0.87	20-21	0.87
Naturalness	6-9, 13	0.91	6-7, 9, 13	0.89
Overall	All 21 items	0.95	All 15 items	0.93

As a result of this analysis, we retained 15 items in the final version of the MOS-R2a (see Appendix B): Listening Effort, Comprehension, Articulation, Pleasantness, Voice Naturalness, Humanlike Voice, Loudness, Emphasis, Voice Quality, Rhythm, Intonation, Precision, Trust, Confidence, and Depression. The resulting factor structure and item alignment appears in Table 4. As shown, the removal of Speaking Rate from the Fluency factor and the association of Pleasantness, Humanlike Voice, and Voice Quality with Naturalness (instead of the subordinate factor Voice) weaken the descriptive quality of the Voice and Fluency labels.

Table 4. Five Factor MOS-R2a

Intelligibility	Fluency	Voice	Social Impression	Naturalness
Listening Effort	Emphasis	Loudness	Trust	Pleasantness
Comprehension Problems	Rhythm	Depression	Confidence	Naturalness
Articulation	Intonation			Humanlike Voice
Precision				Voice Quality

Inter-Factor Correlations
As is typical in the development of these types of instruments (Nunnally, 1978), the resulting factor (scale) scores for each factor had significant correlation with every other factor (n = 160, all $p < .004$, see Table 5). The magnitudes of the correlations were all significantly less than 1.0 ($p < .01$), avoiding the potential problem of multicollinearity in subsequent analyses.

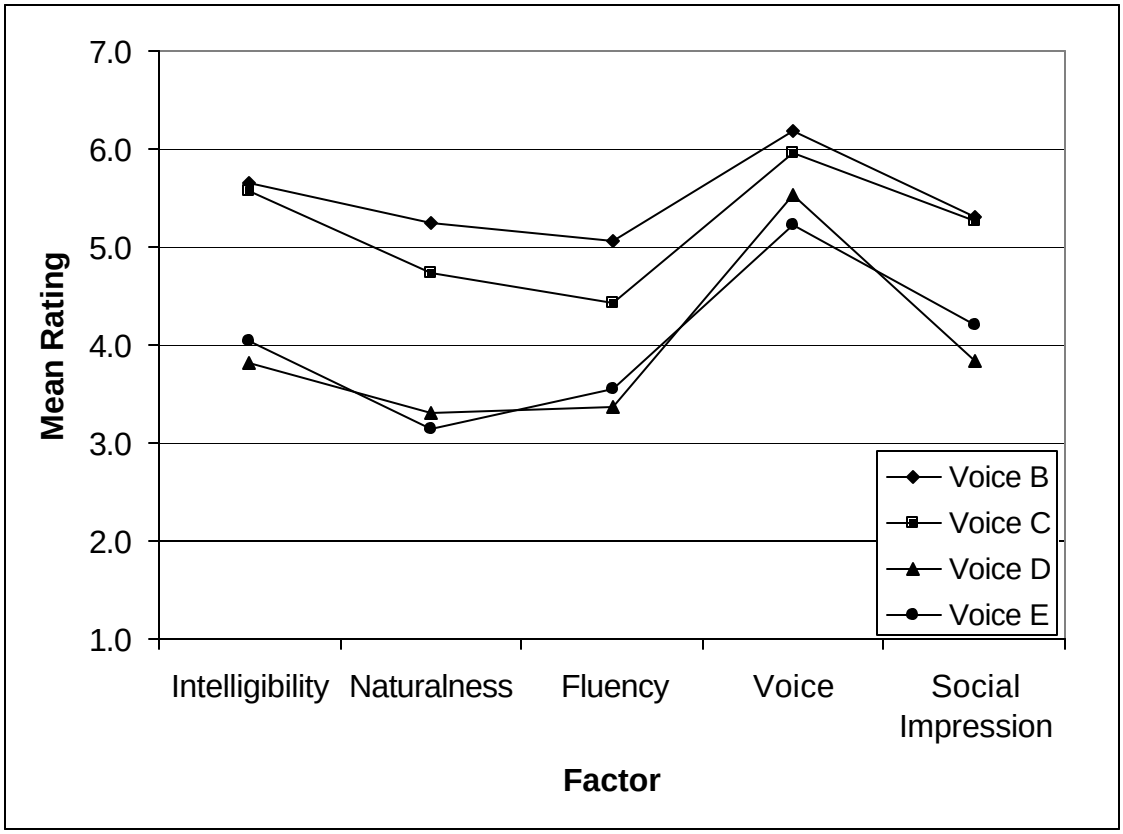
Table 5. Inter-Factor Correlations for the MOS-R2a

	Intelligibility			
Naturalness	0.71	**Naturalness**		
Fluency	0.59	0.83	**Fluency**	
Voice	0.40	0.36	0.23	**Voice**
Social Impression	0.59	0.66	0.57	0.43

Sensitivity
A mixed model ANOVA indicated the extent to which the MOS-R2a discriminated among the four synthetic voices. The ANOVA showed a main effect of synthetic voice ($F(3,154) = 26.92$, $MSe = 4.25$, $p < 0.0001$), factor ($F(4,616) =$

79.03, *MSe* = 0.85, $p < 0.0001$), and a significant interaction between these variables ($F(12,616) = 4.56$, $p < 0.0001$). The significant interaction appears in Figure 1, illustrating the superior ratings of Voices B and C to Voices D and E.

Figure 1. Voice by Factor Interaction (MOS-R2a)



Study 2: MOS-R2b

The purpose of the second study was to add items for the purpose of improving the reliability of the Voice factor and increasing the number of items associated with the Social Impression factor.

Method

Participants
The sample consisted of 1000 randomly selected IBM employees (none of whom were in the sample for Study 1), with 200 individuals in each of five groups. Of this sample, 138 individuals completed the study questions (14% return rate).

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

- A: concatenative female
- B: concatenative female
- C2: concatenative male
- D: concatenative male
- E: formant male

With the exception of Voice C2, the voices were the same as those used in Study 1. The technology used to produce Voice C2 was the same as that used to produce Voice C in the first study, but the source voice for Voice C2 was different.

The dependent measures were the ratings for the 22 MOS-R2b items shown in Appendix C. We retained 15 items from the final version of the MOS-R2a (Listening Effort, Comprehension, Articulation, Pleasantness, Voice Naturalness, Humanlike Voice, Loudness, Emphasis, Voice Quality, Rhythm, Intonation, Precision, Trust, Confidence, and Depression). As in our previous study, we generated additional items related to voice and its correlates in human speech (Monotone Quality, Attractiveness, Enthusiasm) and four additional social impression items (Persuasiveness, Enthusiasm, Impatience, and Fear). If the previous factor structure remained, we would expect the new items to align to the Voice and Social Impression factors, increasing their reliability. However, the new items rely significantly less on the specific areas of human speech evaluation, making the items in this study qualitatively different than the original Study 1 items. Therefore, we suspected that the Voice and Fluency factors would not be retained.

Procedure
The procedure was identical to that of Study 1.

Results and Discussion

Factor Analysis
As in Study 1, a discontinuity analysis indicated a five-factor solution, explaining 66% of the variance in the data. Table 6 shows the factor loadings (in bold) for each item in the MOS-R2b. Seven items loaded on Factor 1 (items 8, 12, 14-16, 18-19), five items loaded on factor 2 (items 1-3, 11, 13), one item loaded on Factor 3 (item 7), two items loaded on Factor 4 (items 17, 21), and six items loaded on Factor 5 (items 4-6, 9-10, 20).

Table 6. Factor Loadings for the MOS-R2b Five-Factor Solution

Item	Content	Factor1	Factor2	Factor3	Factor4	Factor5
		Social Impression	Intelligibility	Voice	Negativity	Naturalness
1	Listening Effort	0.202	**0.729**	0.344	0.083	0.262
2	Comprehension	-0.016	**0.802**	0.135	0.042	0.284
3	Articulation	0.045	**0.826**	0.023	0.083	0.208
4	Pleasantness	0.417	0.180	0.203	0.208	**0.576**
5	Voice Naturalness	0.216	0.278	0.019	-0.032	**0.852**
6	Humanlike Voice	0.184	0.282	0.025	-0.020	**0.775**
7	Loudness	0.150	0.107	**0.845**	0.035	0.046
8	Emphasis	**0.627**	0.420	-0.045	-0.149	0.087
9	Voice Quality	0.215	0.147	-0.005	0.313	**0.688**
10	Rhythm	0.406	0.406	-0.286	0.082	**0.536**
11	Intonation	0.420	**0.512**	-0.344	0.109	0.407
12	Monotone Quality	**0.620**	-0.004	-0.070	0.087	0.442
13	Precision	0.323	**0.638**	-0.103	0.366	0.044
14	Trust	**0.620**	0.252	0.335	0.249	0.217
15	Enthusiasm	**0.799**	-0.088	-0.001	0.164	0.232
16	Confidence	**0.651**	0.103	0.294	0.283	0.146
17	Depression	0.448	0.116	0.108	**0.718**	-0.023
18	Attractiveness	**0.585**	0.159	0.140	0.091	0.465
19	Persuasiveness	**0.703**	0.243	0.050	0.047	0.322
20	Impatience	0.416	0.140	0.247	0.248	**0.549**
21	Fear	-0.007	0.130	-0.005	**0.842**	0.240

Factor 1 included items related to Social Impression (Emphasis, Monotone Quality, Trust, Enthusiasm, Confidence, Attractiveness, Persuasiveness). Factor 2 included Intelligibility items (Listening Effort, Comprehension, Articulation, Intonation, Precision) and Factor 5 was similar to the previous Naturalness factor (Pleasantness, Naturalness, Humanlike Voice, Voice Quality, Rhythm, Impatience), so we again retained these labels. Interestingly, Factor 3 included only Loudness from the earlier Voice factor, and Factor 4 included two items of Negativity (Depression, Fear).

Because only one item associated with Factor 3 (Voice), we omitted Loudness and performed a second factor analysis, forcing a four-factor solution. The loadings appear in Table 7 and were similar to the association of items in the five-factor solution, except that the Voice factor no longer occurred. The four-factor model appeared to be more consistent with Study 1 results and the theoretical association of items in the literature, and included more than one item per factor (but note that the Negativity factor only included two items).

Table 7. Factor Loadings for Four-Factor Solution

Item	Content	Factor1	Factor2	Factor3	Factor4
		Social Impression	Intelligibility	Negativity	Naturalness
1	Listening Effort	0.241	**0.750**	0.113	0.236
2	Comprehension	-0.025	**0.797**	0.056	0.295
3	Articulation	0.039	**0.831**	0.070	0.211
4	Pleasantness	0.444	0.194	0.230	**0.553**
5	Voice Naturalness	0.192	0.268	-0.002	**0.849**
6	Humanlike Voice	0.165	0.275	0.008	**0.769**
8	Emphasis	**0.565**	0.403	-0.221	0.173
9	Voice Quality	0.320	0.375	0.015	**0.618**
10	Rhythm	0.406	0.406	0.082	**0.536**
11	Intonation	0.325	**0.478**	0.019	0.505
12	Monotone Quality	**0.579**	-0.017	0.043	0.493
13	Precision	0.321	**0.650**	0.295	0.063
14	Trust	**0.686**	0.292	0.250	0.178
15	Enthusiasm	**0.793**	-0.079	0.103	0.263
16	Confidence	**0.713**	0.141	0.270	0.114
17	Depression	0.500	0.142	**0.673**	-0.028
18	Attractiveness	**0.599**	0.174	0.082	0.457
19	Persuasiveness	**0.691**	0.249	0.002	0.351
20	Impatience	0.451	0.156	0.277	**0.522**
21	Fear	0.029	0.138	**0.835**	0.226

As compared with the results of Study 1, both the Intelligibility and Naturalness factors retained the core items from the earlier versions of the MOS-R. Two additional items, Intonation (Intelligibility) and Impatience (Naturalness), also loaded on these two factors. The Depression item moved from the Voice factor (Study 1) to a new factor in this data, pairing with Fear. We tentatively labeled this factor as Negativity, suggesting the negative valence associated with both these items and their contrast to the other social impression items (positive valence). Also of interest (yet somewhat expected during item generation) were the relatively large number of items that loaded on Social Impression and the apparent loss of the Fluency and Voice factors.

Reliability

Table 8 shows the reliability of each factor and the overall scale. Three factors (Intelligibility, Social Impression, Naturalness) and the Overall score demonstrated adequate reliability above 0.70 (Landauer, 1988). The reliability of the Negativity factor was just below this criterion.

To create a more efficient measure, we again removed items from factors with more than four items and recalculated coefficient alpha. The resulting instrument had eight fewer items, yet maintained a reliability of 0.89.

Table 8. Original and Adjusted Reliability for Four Factors

Factor	Original Items	Original Coefficient Alpha	Retained Items	Adjusted Coefficient Alpha
Intelligibility	1-3, 11, 13	0.84	1-3, 13	0.84
Negativity	17, 21	0.65	17, 21	0.65
Social Impression	8, 12,14-16, 18-19	0.85	14-16, 19	0.84
Naturalness	4-6, 9-10, 20	0.86	4-6, 9	0.85
Overall	All 22 items	0.90	All 14 items	0.89

As a result of this analysis, the resulting MOS-R2b included 14 items (see Appendix D): Listening Effort, Comprehension, Articulation, Pleasantness, Voice Naturalness, Humanlike Voice, Voice Quality, Precision, Trust, Enthusiasm, Confidence, Depression, Persuasiveness, and Fear. The resulting factor structure and item alignment appears in Table 9. This result illustrates the qualitative difference in our items in Study 1 and 2, in that the MOS-R2b items are less clearly related to human vocal characteristics (with the exception of Articulation, Precision, and Voice Quality) and more related to the social interpretations conveyed by human speech.

Table 9. Four Factor MOS-R2b

Intelligibility	**Negativity**	**Social Impression**	**Naturalness**
Listening Effort	Depression	Trust	Pleasantness
Comprehension Problems	Fear	Enthusiasm	Naturalness
Articulation		Confidence	Humanlike Voice
Precision		Persuasiveness	Voice Quality

Inter-Factor Correlations

As shown in Table 10, the resulting factor (scale) scores for each factor had significant correlation with every other factor (n = 138, all $p < .00003$). The magnitudes of the correlations were all significantly less than 1.0 ($p < .01$), avoiding the potential problem of multicollinearity in subsequent analyses.

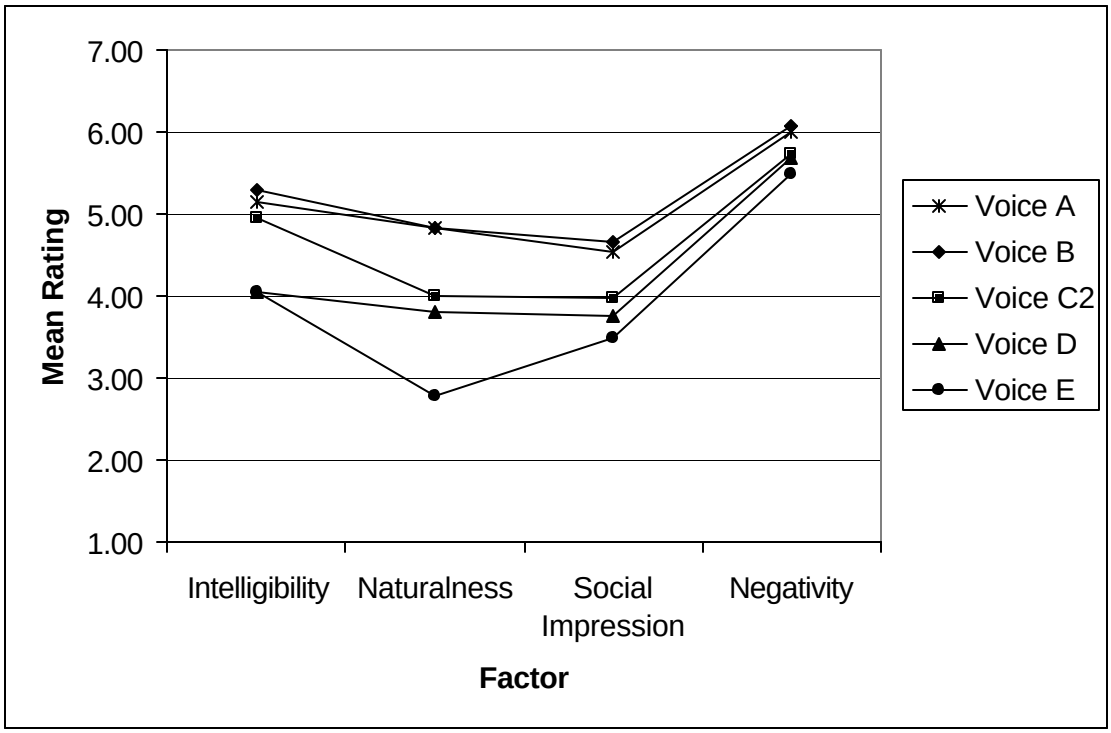
Table 10. Inter-Factor Correlations for the MOS-R2b

	Intelligibility		
Naturalness	0.57	**Naturalness**	
Negativity	0.35	0.36	**Negativity**
Social Impression	0.47	0.61	0.47

Sensitivity

A mixed model ANOVA indicated the extent to which the final version of the MOS-R2b discriminated among the five synthetic voices. The ANOVA showed a main effect of synthetic voice ($F(4,124) = 9.18$, $MSe = 2.93$, $p < 0.0001$), factor ($F(3,372) = 101.92$, $MSe = 0.75$, $p < 0.0001$), and a significant interaction between these variables ($F(12,372) = 2.70$, $p = 0.002$). Figure 2 shows the interaction (higher mean ratings are more positive), illustrating the similarity between Voices A and B (as expected because they used the same core TTS technology). Again consistent with expectation, the formant voice (Voice E) was the most poorly rated voice in terms of its perceived Social Impression and Naturalness. The perceived intelligibility of Voice E was identical with that of concatenative Voice D. Of the four factors, only Negativity seemed to be somewhat insensitive to the differences among the voices.

Figure 2. Voice by Factor Interaction (MOS-R2b)



Combining Studies 1 and 2: The Expanded MOS (MOS-X)

The outcomes of Studies 1 and 2 were encouraging, but not completely satisfying. The addition of the new items in each study led to the emergence of new factors (Fluency, Voice, and Social Impression in Study 1; Negativity and Social Impression in Study 2). In Study 1, the Voice factor did not have an acceptable level of reliability. The Social Impression factor was reliable, but had the support of only two items. In Study 2, the final MOS-R2b had four items supporting the Social Impression factor, but the Negativity factor only had two items, a relatively low reliability, and relatively low sensitivity.

The primary goal of this research was to expand the coverage of the MOS to include new factors that are becoming important in the evaluation of synthetic speech. To accomplish that goal, we felt that it was necessary to include a Social Impression factor and to include a factor related to the prosodic features of speech. Van Riper and Emerick (1990) define prosody as the "linguistic stress patterns [of speech] as reflected in pause, inflection, juncture" and the "melody or cadence of speech" (p. 491). Our initial MOS-R2a included items that contribute to prosody (Emphasis, Rhythm, Intonation, Interruptions), yet these items did not clearly align in a single factor but in two more precise and specific categories (Voice and Fluency). In Study 2, the stronger loadings of Social Impression items (likely due to the larger effect sizes of social impression as compared with vocal speech perceptions) resulted in removal of all items that could be related to prosody (Emphasis, Voice Quality, Rhythm, Intonation, Monotone Quality). Recently, researchers have begun to acknowledge that prosodic qualities are vital for acceptable synthetic speech and develop algorithms to approximate human prosody (Portele & Heuft, 1997; Sonntag & Portele, 1998). In addition to these content goals, each factor had to produce a scale with acceptable reliability and to preferably have the support of at least three items.

As a consequence of the iterative evaluation process of Studies 1 and 2, the complete item sets for the studies had 14 items in common (see Appendix E). The common items were the four items associated with Intelligibility in both the MOS-R2a and MOS-R2b, the four items associated with Naturalness in both the MOS-R2a and MOS-R2b, the three items associated with Fluency in the MOS-R2a, the two items associated with Social Impression in the MOS-R2a, and the Depression item (associated with the Voice factor in the MOS-R2a and the Negativity factor in the MOS-R2b).

Because these items were common across both studies, the sample size for their psychometric evaluation was the sum of the sample sizes for Studies 1 and 2 (342 complete and independent sets of responses). The factor analyses of Studies 1 and 2 strongly suggested that the Intelligibility, Naturalness, and Fluency factors would remain intact in an analysis of this combined data. It also seemed likely that the two items associated with Social Impression in the MOS-R2a and MOS-R2b would continue to align. The expected behavior of the Depression item was harder to predict. If, in the context of this subset of the overall data, it aligned with the Social Impression factor and the Social Impression factor's reliability exceeded .70, then this version of the MOS would meet the goals of our research, producing an Expanded MOS (MOS-X).

Method

To perform this analysis, we created a new database from the results of Studies 1 and 2. The 14 items included in the database addressed Listening Effort, Comprehension Problems, Articulation, Voice Pleasantness, Voice Naturalness, Humanlike Voice, Emphasis, Voice Quality, Rhythm, Intonation, Precision, Trust, Confidence, and Depression.

Results and Discussion

Factor Analysis and Reliability

A discontinuity analysis suggested either a three- or four-factor solution for these items. Because the three-factor solution mixed together the vocal speech and social impression items in a somewhat haphazard manner and we had prior expectation of a four-factor solution, we pursued the four-factor solution in subsequent analyses.

The four-factor solution accounted for 64% of the variance in the data. As shown in Table 11, the items aligned in a relatively clear pattern: Factor 1 included Emphasis, Rhythm, Intonation and Factor 3 included Trust, Confidence, and Depression. Factors 2 and 4 demonstrated a predictable clustering of items based on the relative stability of two factors throughout all modifications of the MOS. The Depression item aligned more strongly with the Social Impression factor

than with any other factor, but with a somewhat lower loading than the other two items. Coefficient alpha for each factor indicated acceptable reliability (Overall: .92, Intelligibility: .88, Naturalness: .87, Fluency/Prosody: .85, Social Impression: .71).

Table 11. Factor Loadings for the MOS-X Four-Factor Solution

Item	Content	Factor1	Factor2	Factor3	Factor4
		Prosody	Intelligibility	Social Impression	Naturalness
1	Listening Effort	0.18	**0.70**	0.28	0.23
2	Comprehension	0.24	**0.78**	0.11	0.23
3	Articulation	0.19	**0.82**	0.17	0.25
4	Pleasantness	0.21	0.27	0.40	**0.61**
5	Voice Naturalness	0.36	0.29	0.13	**0.79**
6	Humanlike Voice	0.30	0.34	0.20	**0.67**
7	Emphasis	**0.57**	0.23	0.28	0.17
8	Voice Quality	0.25	0.28	0.33	**0.50**
9	Rhythm	**0.73**	0.24	0.19	0.38
10	Intonation	**0.76**	0.25	0.23	0.30
11	Precision	0.23	**0.54**	0.31	0.25
12	Trust	0.20	0.19	**0.78**	0.29
13	Confidence	0.17	0.25	**0.68**	0.27
14	Depression	0.11	0.07	**0.40**	0.03

In the MOS-X, Factor 1 clearly included items related to prosody, indicating an obvious label. The elimination of the more specific and subordinate factors Voice and Fluency further pointed to the Prosody label, as well as a desire to keep the relative breadth of labels consistent across the four factors. The resulting factor structure and item alignment appears in Table 12. This result corresponds more successfully to our initial goal of improving the measurement of both perceptual speech and social impressions than the MOS revisions of Study 1 or 2.

Table 12. Four Factor MOS-X

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

Inter-Factor Correlations

As shown in Table 13, the resulting factor (scale) scores for each factor had significant correlation with every other factor (n = 281, all $p < .00001$). The magnitudes of the correlations were all significantly less than 1.0 ($p < .01$), avoiding the potential problem of multicollinearity in subsequent analyses.

Table 13. Inter-Factor Correlations for the MOS-X

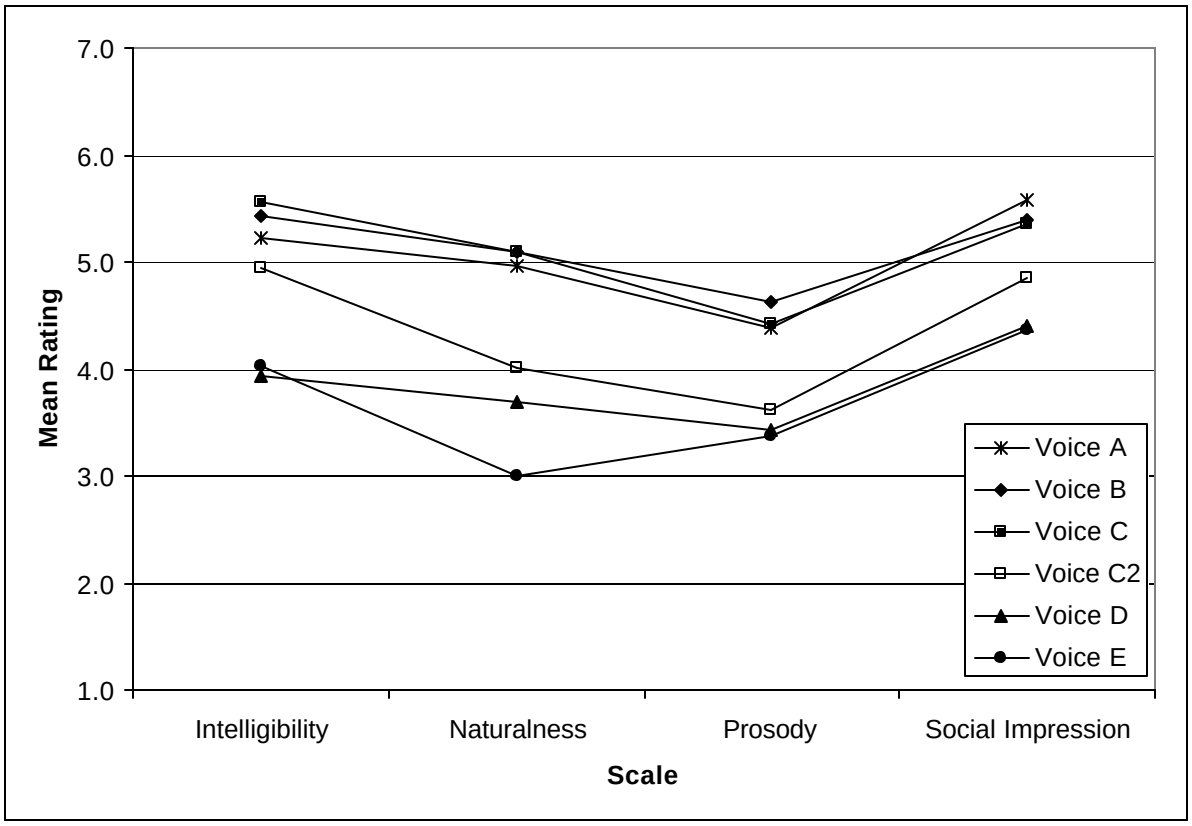
	Intelligibility		
Naturalness	0.66	**Naturalness**	
Prosody	0.57	0.68	**Prosody**
Social Impression	0.50	0.56	0.50

Sensitivity

A mixed model ANOVA indicated the extent to which the MOS-X discriminated among the six different synthetic voices used in Studies 1 and 2. The ANOVA showed a main effect of synthetic voice ($F(5,275) = 27.5$, $MSe = 3.4$, $p < 0.0000001$), factor ($F(3,825) = 58.5$, $MSe = 0.74$, $p < 0.0000001$), and a significant interaction between these variables ($F(15,825) = 3.8$, $p = 0.000001$). Figure 3 shows the interaction (higher mean ratings are more positive), illustrating

the similarity between Voices A and B (as expected because they used the same core TTS technology). Again consistent with expectation, the formant voice (Voice E) was the most poorly rated voice for perceived naturalness. The perceived intelligibility, prosody, and social implication of Voice E were identical to that of concatenative Voice D (a particularly poor concatenative voice). All four factors seemed to be reasonably sensitive to the differences among the voices.

Figure 3. Voice by Factor Interaction (MOS-X)



General Discussion

The current expansion and evaluation of the MOS-R revealed two important advancements over the previous MOS-R (Lewis, 2001b). First, both studies investigated a number of subtle vocal and social-emotional characteristics that past literature has validated as having an impact on listener perception of speech. Thus, using the literature and the results of these studies as a guide, we expanded the content of the current MOS to measure both prosodic and social impressions of listeners, producing the MOS-X. Developers of artificial voices can use these two new MOS factors to help guide the continued development of synthetic speech. In addition to expanding the scope of the MOS, the MOS-X retained the desirable psychometric properties of the MOS-R's Intelligibility and Naturalness factors.

These two studies also resolved several historical problems observed in the MOS and MOS-R. First, the Speaking Rate item, which did not clearly associate with either Intelligibility or Naturalness in earlier evaluations (Kraft & Portele, 1995; Lewis, 2001a, 2001b), loaded on the Fluency factor in Study 1 (MOS-R2a). We excluded Speaking Rate from the MOS-R2a without significant loss of reliability. As predicted by Lewis (2001b), the Humanlike Voice item loaded strongly on the Naturalness factor and we retained it through the MOS-R2a and MOS-R2b into the MOS-X. Finally, as noted by Lewis (2001b), the Global Impression item loaded on more than one factor, although its strongest loading was again on the Intelligibility factor. We removed this item during the efficiency phase of Study 1 and found that coefficient alpha improved, suggesting that the Global Impression item was at least partially responsible for the lower reliability of its associated factor in the previous evaluations.

We also generated several new and interesting problems. Most notably, the MOS-R2a Loudness item associated with Pitch and Depression. Loudness and pitch (and their acoustic correlates intensity and fundamental frequency, respectively) are typically measured in a clinical evaluation of human speech, particularly voice or phonation, and are indicative of a number of pathologies, including clinical depression (Baken, 1978; Murray & Arnot, 1993). This associative pattern partially prompted the Voice factor name in Study 1. However, when we removed Pitch, the Loudness item became a separate factor and Depression associated with the Social Impression factor (MOS-R2b).

The elusive Voice factor was also apparent in the MOS-R2b. In this version of the MOS-R2, Fear and Depression aligned in a factor we named Negativity. Both of these items elicit perceptions of emotion with negative valence, which distinguishes them from the items associated with the social-personality inferences elicited by Social Impression items. Fear and depression are signaled by voice characteristics: a rapid speaking rate, elevated pitch, wide pitch range, and irregular voicing conveys fear but a slow speaking rate, lowered pitch, reduced loudness, and downward inflections convey sadness or depression (Murray & Arnott, 1993). Thus, although we apparently eliminated the Voice factor, the inferences about a speaker's emotional state are derived from voice information. Thus, voice items remained in the MOS-R2b, although covertly.

A second observation concerns the type of items that we removed from the MOS-R2a and MOS-R2b. Most of the omitted items were perceptual ratings specific to the speech pattern itself and typical of evaluative judgments made of human speech disorders (Baken, 1978). Of the items ultimately removed from the revised scales, eight items were perceptual judgments made by speech-language pathologists in clinical evaluations (Speaking Rate, Loudness, Emphasis, Interruptions, Pitch, Rhythm, Intonation, Monotone Quality). All items remaining on the measure (except Voice Quality) appear to be more abstract interpretative qualities derived from speech. In many respects, this pattern of item exclusion is logical because naïve listeners do not have a clinical vocabulary or perceptual training to directly evaluate speech characteristics. The layperson is perhaps better suited to make inferences about a speaker's emotional state or social characteristics (even if the speaker is an abstraction), as shown by the vast literature on these topics (Murray & Arnott, 1993).

The items included in the MOS-X resulted in a blend of the factors present in the MOS-R2a and MOS-R2b. The Prosody factor targeted vocal speech perceptions and the Social Impression factor targeted social-emotional interpretations. The MOS-X became the most satisfying revision of the MOS-R because both types of ratings can help guide the continued development of synthetic speech. The only potential weakness of the MOS-X is the relatively low (though acceptable) reliability of its Social Impression factor, possibly due to the relatively low loading of Depression on that factor. Future work with the MOS-X should investigate the potential value of replacing Depression with the other items found to align with Social Implication in Study 2 (Enthusiasm and Persuasiveness).

In summary, the data from these analyses provide empirical evidence that the MOS-X has achieved the psychometric goals of (1) expanding MOS item coverage beyond the traditional factors of Intelligibility and Naturalness to include Prosody and Social Implication, (2) achieving adequate reliability for the measurement scales derived from the MOS-X factors, and (3) being sensitive enough to detect key differences among a set of artificial voices.

References

Aronovitch, C. (1976). The voice of personality: Stereotyped judgments and their relation to voice quality and sex of speaker. *Journal of Social Psychology, 99(2),* 207-220.

Baken, R. (1978). *Clinical measurement of speech and voice*. Boston: Allyn & Bacon.

Berry, D. (1992). Vocal types and stereotypes: Joint effects of vocal attractiveness and vocal maturity on person perception. *Journal of Nonverbal Behavior, 16(1*), 41-45.

Bloom, K., Zajac, D., & Titus, J. (1999). The influence of nasality of voice on sex-stereotyped perceptions. *Journal of Nonverbal Behavior, 23(4),* 271-281.

Bradlow, A., Torretta, G., & Pisoni, D. (1996). Intelligibility of normal speech I: Global and fine-grained acoustic-phonetic talker characteristics. *Speech Communication, 20(3-4),* 255-272.

Brown, B., Strong, W., & Rencher, A. (1973). Perceptions of personality from speech: Effects of manipulations of acoustical parameters. *Journal of the Acoustical Society of America, 54(1),* 29-35.

Brown, B., Strong, W., & Rencher, A. (1975). Acoustic determinants of perceptions of personality from speech. *International Journal of the Sociology of Language, 6(1),* 1-32.

Cliff, N. (1987). *Analyzing multivariate data*. San Diego, CA: Harcourt Brace Jovanovich.

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.

Ekman, P., O'Sullivan, M., Friesen, W., & Scherer, K. (1991). Face, voice, and body in detecting deceit. *Journal of Nonverbal Behavior, 15(2),* 125-135.

Gorenflo, D., & Gorenflo, C. (1997). Effects of synthetic speech, gender, and perceived similarity on attitudes toward the augmented communicator. *AAC: Augmentative and Alternative Communication, 13(2),* 87-91.

Granstrom, B., & Nord, L. (1992). Neglected dimensions in speech synthesis. *Speech Communication, 11(4-5),* 459-462.

Greene, B., Logan, J., & Pisoni, D. (1986). Perception of synthetic speech produced automatically by rule: Intelligibility of eight text-to-speech systems. *Behavior Research Methods, Instruments, and Computers, 18*, 100-107.

Hieda, I., & Kuchinomachi, Y. (1997). Preliminary study of relations between physical characteristics and psychological impressions of natural voices. *Perceptual and Motor Skills, 85*, 1483-1491.

Higashikawa, M., & Minifie, F. (1999). Acoustical-perceptual correlates of 'whisper pitch' in synthetically generated vowels. *Journal of Speech, Language, and Hearing Research, 42(3),* 583-591.

Hillenbrand, J. (1988). Perception of aperiodicities in synthetically generated voices. *Journal of the Acoustical Society of America, 83(6)*, 2361-2371.

Hoag, L., & Bedrosian, J. (1992). Effects of speech output type, message length, and reauditorization on perceptions of the communicative competence of an adult AAC user. *Journal of Speech and Hearing Research, 35(6),* 1363-1366.

Holtgraves, T., & Lasky, B. (1999). Linguistic power and persuasion. *Journal of Language and Social Psychology, 18(2)*, 196-205.

Hosman, L. (1989). The evaluative consequences of hedges, hesitations, and intensifiers: Powerful and powerless speech styles. *Human Communication Research, 15(3),* 383-406.

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

Johnson, W., Emde, R., Scherer, K., & Klinnert, M. (1986). Recognition of emotion from vocal cues. *Archives of General Psychiatry, 43(3),* 280-283.

Klatt, D., & Klatt, L. (1990). Analysis, synthesis, and perception of voice quality variations among female and male talkers. *Journal of the Acoustical Society of America, 87(2),* 820-857.

Koopmans-Van Beinum, F. (1992). The role of focus words in natural and in synthetic continuous speech: Acoustic aspects. *Speech Communication, 11(4-5),* 439-452.

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.

Lavner, Y., Gath, I., & Rosenhouse, J. (2000). The effects of acoustic modificiations on the identification of familiar voices speaking isolated vowels. *Speech Communication, 30(1),* 9-26.

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.

Martin, R., & Haroldson, S. (1992). Stuttering and speech naturalness: Audio and audiovisual judgments. *Journal of Speech and Hearing Research, 35(3),* 521-528.

Massaro, D., & Egan, P. (1996). Perceiving affect from the voice and the face. *Psychonomic Bulletin & Review, 3(2),* 215-221.

Miyake, K., & Zuckerman, M. (1993). Beyond personality: Effects of physical and vocal attractiveness on false consensus, social comparison, affiliation, and assumed and perceived personality. *Journal of Personality, 61(3),* 411-437.

Murray, I., & Arnott, J. (1993). Toward the simulation of emotion in synthetic speech: A review of the literature on human vocal emotion. *Journal of the Acoustical Society of America, 93(2),* 1097-1108.

Murray, I., & Arnott, J. (1995). Implementation and testing of a system for producing emotion-by-rule in synthetic speech. *Speech Communication, 16,* 369-390.

Murray, I., Arnott, J., & Rohwer, E. (1996). Emotional stress in synthetic speech: Progress and future directions. *Speech Communication, 20(1-2),* 85-91.

Newcombe, N., & Arnkoff, D. (1979). Effects of speech style and sex of speaker on person perception. *Journal of Personality and Social Psychology, 37(8),* 1293-1303.

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

Paddock, J., & Nowicki, S. (1986). Paralanguage and the interpersonal impact of dysphoria: It's not what you say but how you say it. *Social Behavior and Personality, 14(1),* 29-44.

Page, R., & Balloun, J. (1978). The effect of voice volume on the perception of personality. *Journal of Social Psychology, 105(1),* 65-72.

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.

Pelachaud, C., Badler, N., & Steedman, M. (1996). Generating facial expressions for speech. *Cognitive Science, 20(1)*, 1-46.

Portele, T., & Heuft, B. (1997). Toward a prominence-based synthesis system. *Speech Communication, 21(1-2),* 61-72.

Robinson, J., & McArthur, L. (1982). Impact of salient vocal qualities on causal attribution for a speaker's behavior. *Journal of Personality and Social Psychology, 43(2),* 236-247.

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.

Shipley, K. & McAfee, J. (1992). *Assessment in speech language pathology: A resource manual.* San Diego: Singular.

Siegler, D., & Siegler, R. (1976). Stereotypes of males' and females' speech. *Psychological Reports, 39(1),* 167-170.

Sonntag, G.P., & Portele, T. (1998). PURR – A method for prosody evaluation and investigation. *Computer Speech and Language, 12(4)*, 437-451.

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.

Slowiaczek, L., & Nusbaum, H. (1985). Effects of speech rate and pitch contour on the perception of synthetic speech. *Human Factors, 27(6),* 701-712.

Stern, S., Mullennix, J., Dyson, C., & Wilson, S. (1999). The persuasiveness of synthetic speech versus human speech. *Human Factors, 41(4),* 588-595.

Tartter, V., & Braun, D. (1994). Hearing smiles and frowns in normal and whisper registers. *Journal of the Acoustical Society of America, 96(4),* 2101-2107.

Van Riper, C. & Emerick, L. (1990). *Speech correction*, 4th ed. Englewood Cliffs, NJ: Prentice Hall.

Whalen, D., & Hoequist, C. (1995). The effects of breath sounds on the perception of synthetic speech. *Journal of the Acoustical Society of America, 97(5),* 3147-3153.

Whiteside, S.P. (1999). A comment on women's speech and its synthesis. *Perceptual and Motor Skills, 88,* 110-112.

Whitmore, J., & Fisher, S. (1996). Speech during sustained operations. *Speech Communication, 20*, 55-70.

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.

Yaeger-Dror, M. (1996). Register as a variable in prosodic analysis: The case of the English negative. *Speech Communication, 19(1),* 39-60.

Zuckerman, M., Miyake, K., & Hodgins, H. (1991). Cross-channel effects of vocal and physical attractiveness and their implications for interpersonal perception. *Journal of Personality and Social Psychology, 60(4),* 545-554.

Appendix A. Items for MOS-R2a Evaluation

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

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. *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**

10. *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 (RATING LESS THAN 6), PLEASE CIRCLE ONE: TOO SLOW or TOO FAST

11. *Loudness*: Was the voice appropriately loud?

INAPPROPRIATE LOUDNESS	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**APPROPRIATE LOUDNESS**

IF UNSATISFACTORY (RATING LESS THAN 6), PLEASE CIRCLE ONE: TOO LOUD or TOO SOFT

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

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

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

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

14. *Interruptions*: Did you notice interruptions in the speech, causing it to sound jerky or hesitant?

MANY INTERRUPTIONS	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**NO INTERRUPTIONS**

15. *Pitch*: Was the pitch of the voice appropriate?

INAPPROPRIATE PITCH	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**APPROPRIATE PITCH**

IF UNSATISFACTORY (RATING LESS THAN 6), PLEASE CIRCLE ONE: TOO HIGH or TOO LOW

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

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

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

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

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

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

19. *Topic Interest*: Did the voice show interest in the topic of conversation?

VERY UNINTERESTED	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**VERY INTERESTED**

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

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

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

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

22. *Depression*: Did the voice suggest a depressed speaker?

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

Appendix B. Final Items for the MOS-R2a

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

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

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

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

6. *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**

7. *Loudness*: Was the voice appropriately loud?

INAPPROPRIATE LOUDNESS	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**APPROPRIATE LOUDNESS**

IF UNSATISFACTORY (RATING LESS THAN 6), PLEASE CIRCLE ONE: TOO LOUD or TOO SOFT

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

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

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

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

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

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

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

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

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

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

15. *Depression*: Did the voice suggest a depressed speaker?

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

<u>MOS-R2a Scales</u>
Overall: Average items 1-15
Intelligibility: Average items 1-3 and 12
Naturalness: Average items 4-6 and 9
Fluency: Average items 8 and 10-11
Voice: Average items 7 and 15 (but transform 15: Score(15) = 7 – Rating(15) + 1)
Social Impression: Average items 13-14

Appendix C. Items for MOS-R2b Evaluation

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

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

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

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

6. *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**

7. *Loudness*: Was the voice appropriately loud?

INAPPROPRIATE LOUDNESS	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**APPROPRIATE LOUDNESS**

IF UNSATISFACTORY (RATING LESS THAN 6), PLEASE CIRCLE ONE: TOO LOUD or TOO SOFT

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

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

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

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

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. *Monotone Quality*: To what extent did the voice sound monotonous?

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

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

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

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

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

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

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

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

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

17. *Depression*: Did the voice suggest a depressed speaker?

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

18. *Attractiveness*: Did the voice suggest an attractive speaker?

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

19. *Persuasiveness*: Was the voice persuasive?

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

20. *Impatience*: Did the voice make you feel impatient?

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

21. *Fear*: Did the voice sound fearful?

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

Appendix D. Final Items for the MOS-R2b

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

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

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

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

6. *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**

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

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

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

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

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

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

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

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

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

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

12. *Depression*: Did the voice suggest a depressed speaker?

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

13. *Persuasiveness*: Was the voice persuasive?

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

14. *Fear*: Did the voice sound fearful?

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

<u>MOS-R2b Scales</u>
Overall: Average items 1-14
Intelligibility: Average items 1-3 and 8
Naturalness: Average items 4-7
Social Impression: Average items 9-11 and 13
Negativity: Average items 12 and 14

Appendix E. Final Items 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. *Voice Pleasantness*: Was the voice you heard pleasant to listen to?

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

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

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

6. *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**

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

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

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

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

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

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

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

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

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

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

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. *Depression*: Did the voice suggest a depressed speaker?

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

Appendix F. 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. *Depression*: Did the voice suggest a depressed speaker?

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

MOS-X Scales
Overall: Average items 1-14
Intelligibility: Average items 1-4
Naturalness: Average items 5-8
Prosody: Average items 9-11
Social Impression: Average items 12-14