Voice Interaction Design

The Journal of AVIXD Association for Voice Interaction Design

Vol. 2, Issue 2, March 2018 pp. 1-22

Investigating MOS-X Ratings of Synthetic and Human Voices

James R. Lewis
Senior HF Engineer
IBM Corp.
5901 Broken Sound
Parkway Suite 514C
Boca Raton, FL
USA
jimlewis@us.ibm.com

Abstract

The objectives of this research were to (1) evaluate and compare two versions of the expanded Mean Opinion Scale, one using the original 15 items (MOS-X) and the other a four-item version with one item for each of the four factors of the MOS-X (the MOS-X2), and (2) establish preliminary benchmarks for the interpretation of ratings collected using these questionnaires. Respondents (n = 865) provided ratings for 56 thirty-second recordings of speech samples – 53 from recordings of synthetic voices made from 2001 through 2017 and three from professional human voice talents.

Both questionnaires had acceptable psychometric quality (reliability and validity), but the factor structure of the MOS-X did not exactly match the expected structure. The MOS-X2 had a stronger statistical relationship to outcome metrics of Likelihood-to-Recommend (LTR) and Overall Quality than the MOS-X. The very old samples (those using technologies from 2001-2002) received consistently poor ratings. A few of the synthetic voice samples came close to the ratings given to the professional human voice talents.

Either questionnaire version is acceptable for use, but due to its stronger statistical relationship to the key outcome metrics of LTR and Overall Quality and its shorter length, it is more effective and efficient to use the MOS-X2. The mean MOS-X and MOS-X2 for the ratings of the professional human voices were both about 85 (after conversion to a 0-100 point scale), so a synthetic voice with mean ratings at or approaching 85 would be very good. Ratings over 70 are, relative to the set of voice samples in this study, above average.

Keywords

TTS, text-to-speech, synthetic voice, mean opinion scale, MOS-X benchmarks

Introduction

This report describes a study conducted to compare two standardized versions of questionnaires designed for the rating of the quality of artificial voices, both based on the Mean Opinion Scale (MOS). In addition to this comparison, a second objective was to use the ratings of professional human voice talents to aid in the interpretation of ratings collected using these questionnaires.

The MOS was originally published in the 1990s to provide an instrument for the subjective assessment of speech over noisy or otherwise degraded channels, but was soon adopted for the evaluation of synthetic speech (Francis & Nussbaum, 1999; ITU, 1994; Johnston, 1996; Lewis, 2011; Schmidt-Nielsen, 1995; van Bezooijen & van Heuven, 1997). The most common form of the original version was a questionnaire with seven 5-point items. Although not designed to be a multidimensional metric, factor analysis has typically indicated the two underlying constructs of Intelligibility and Naturalness (Kraft & Portele, 1995; Lewis, 2001a). Figure 1 shows the typical MOS items (from Salza, Foti, Nebbia, & Oreglia, 1996) and the factor with which each was associated in Lewis (2001a).

Item	Content	1	2	3	4	5	Factor
1	*Global Impression: Your answer must indicate how you rate the sound quality of the voice you have heard.*	Bad	Poor	Good	Fair	Excellent	N
2	*Listening Effort: Your answer must indicate the degree of effort you had to make to understand the message.*	Message not understood with any feasible effort	Major effort required	Effort required	Slight effort required	No effort required	I
3	*Comprehension Problems: Your answer must indicate if you found single words hard to understand.*	Every word	Many	Some	Few	None	I
4	*Speech Sound Articulation: Your answer must indicate if the speech sounds are clearly distinguishable.*	No, not at all	No, not very clear	Fairly clear	Yes, clearly enough	Yes, very clearly	I
5	*Pronunciation: Your answer must indicate if you noticed any anomalies in the naturalness of sentence pronunciation.*	Yes, very annoying	Yes, annoying	Yes, slightly annoying	Yes, but not annoying	No	I
6	*Speaking Rate: Your answer must indicate if you found the speed of delivery of the message appropriate.*	No, too fast	No, too slow	Yes, but faster than preferred	Yes, but slower than preferred	Yes	U
7	*Voice Pleasantness: Your answer must indicate if you found the voice you have heard pleasant.*	Very unpleasant	Unpleasant	Fair	Pleasant	Very pleasant	N

Figure 1. The original Mean Opinion Scale (MOS) -- notations in the Factor column represent Naturalness (N), Intelligibility (I), and Unrelated (U)

Standardized questionnaires should have acceptable reliability and validity (for a summary of psychometric development using classical test theory, see Lewis, 2016). Throughout this paper, reliability was assessed using coefficient alpha, a metric of internal consistency used to estimate measurement reliability (consistency) which can range from 0 to 1. The usual criterion for acceptable reliability is a coefficient alpha of at least 0.70 (Nunnally, 1978). There are several different ways to assess validity. In this paper, the focus will be on concurrent validity (the correlation between measurements taken at the same time) and construct validity (the pattern of item-factor loadings from factor analysis). For concurrent validity, the typical minimum criterion to confirm a relationship is a correlation with an absolute magnitude of 0.30, and for construct validity the success criterion is that the pattern of item-factor loadings either makes sense (for exploratory analysis) or matches an expected pattern (for confirmatory analysis).

Lewis (2001a) reported that coefficient alpha for the overall MOS was 0.89, with 0.88 for the Intelligibility factor and 0.81 for the Naturalness factor, all indicative of an acceptable level of reliability. There was also evidence of concurrent validity with paired comparison data and sensitivity to manipulation (significant differences between ratings for a recorded human voice and two types of text-to-speech voices).

Despite these acceptable psychometric properties, there were a number of weaknesses in the original version of the MOS. Reported validity coefficients were marginally significant, and the failure of the Speaking Rate item to align with any other items could have either been due to its different response item structure or actual independence from the other items. These reported weaknesses spurred additional research.

Polkosky and Lewis (2003) used psychometric principles to revise and improve the MOS with a series of studies that led to the MOS-Expanded (MOS-X), which includes measurement of the prosody and social impression of synthetic voices in addition to their intelligibility and naturalness. The MOS-X has a total of 15 items, with four for Intelligibility, four for Naturalness, three for Prosody, and four for Social Impression (see Figure 2), with 7-point bipolar item formats (7-point items are slightly more reliable than 5-point items -- Lewis, 1993; Nunnally, 1978).

The evaluation of the final version of the MOS-X (n = 327, between-subjects online assessment of 10 TTS voices) indicated that it had acceptable psychometric properties (Polkosky & Lewis, 2003). Its overall reliability was 0.93, and the coefficient alpha for each factor exceeded 0.85 (Intelligibility: 0.88, Naturalness: 0.86, Prosody: 0.86, and Social Impression: 0.86). Item alignment on factors indicated a high degree of construct validity. MOS-X ratings were sensitive to differences among the 10 TTS voices.

The MOS-X appeared to be useful for research purposes, but has two shortcomings for practical user experience (UX) work. One is the number of items that study participants need to rate to get an MOS-X score. The other is the level of detail in the item content, which might be difficult for some participants and would be time-consuming for all. Even if these are not actually show-stopping problems, it might be difficult to convince a client that they are not.

To address these issues, I developed a variation of the MOS-X with one item per MOS-X factor (the MOS-X2 – see Figure 3). To partially compensate for reducing the number of items, I used 11-point (0-10) scales for the items (for recent information on the optimal number of response options for these types of questions, see Lewis & Erdinç, 2017). Another advantage of using 0-10 point scales is that they are easily transformed to a 0-100 point scale (just multiply by 10).

The key objectives of this research were to (1) evaluate and compare the MOS-X and the MOS-X2 to examine their psychometric properties, and (2) establish preliminary benchmarks for the interpretation of ratings collected using these questionnaires. If the MOS-X2 were to have comparable psychometric quality as the MOS-X, then due to its shorter length and reduced complexity, it would be the better questionnaire to use for future evaluation of synthetic voices.

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 the 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

Figure 2. The Mean Opinion Scale-Expanded (MOS-X) -- the four factors are Intelligibility (Items 1-4), Naturalness (Items 5-8), Prosody (Items 9-11), and Social Impression (Items 12-15) -- factor scores are the means of their item scores; the overall score is the mean of the factor scores

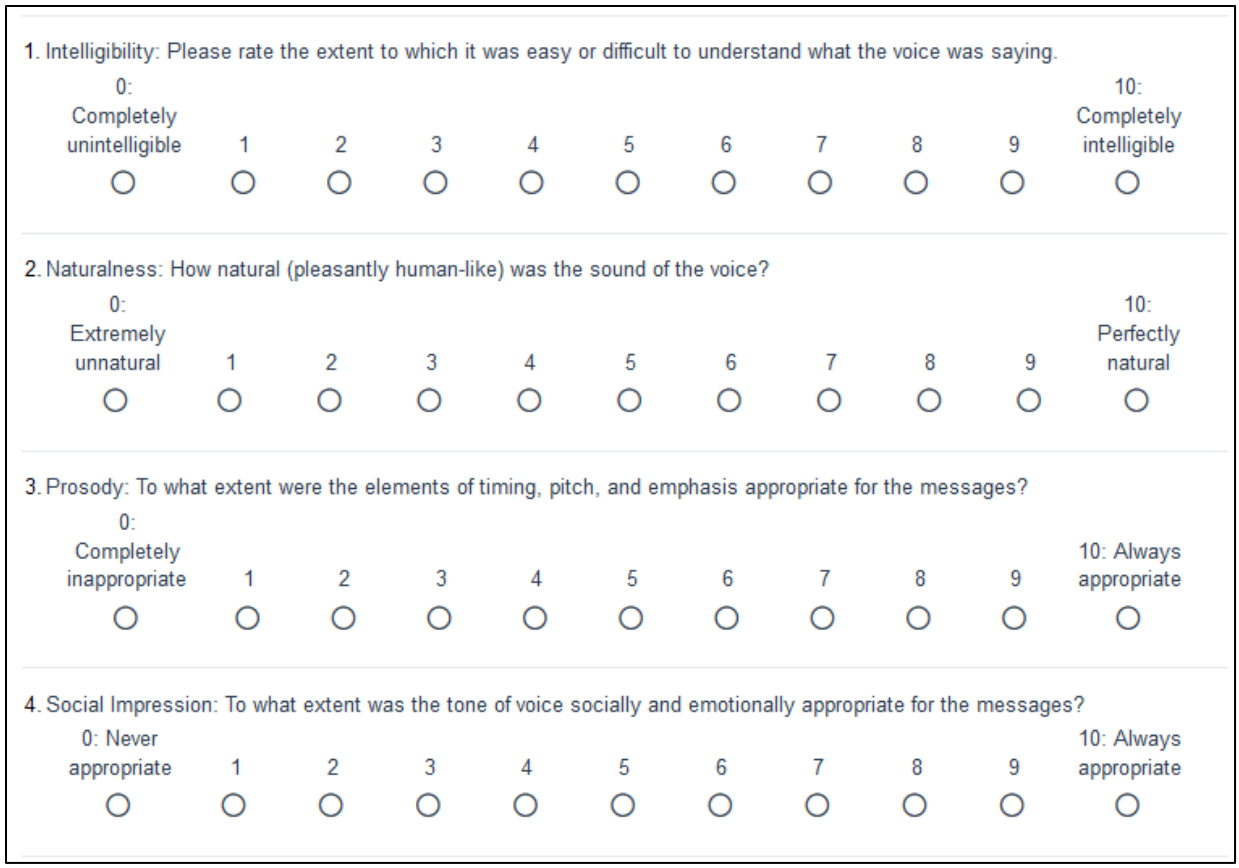

Figure 3. The Mean Opinion Scale-Expanded Version 2 (MOS-X2) -- the four items reflect the content of the four factors of the MOS-X (Intelligibility, Naturalness, Prosody, and Social Impression) – item scores are the given rating times 10, and the overall score is the mean of the item scores, with a potential range from 0-100

Method

The following steps were taken to achieve these objectives:

- Established an IBM User Experience (UX) panel to ensure the availability of a large sample of participants
- Collected a large number of speech recordings (mostly synthetic, but a few human)
- Randomly assigned recordings to members of the UX panel (one recording per member)
- Sent emails to UX Panel members with the randomly assigned recording and a link to a SurveyGizmo survey that included a few demographic questions, the MOS-X, the MOS-X2, and ratings of two outcome metrics (likelihood to recommend the voice for use in a service application and the overall quality of the voice).
- Put the data obtained into an SPSS file for statistical analysis (using Version 24).

The IBM User Experience Panel

I sent invitations to join the panel to 20,000 randomly selected IBM internal emails (United States only). About 10% of those invited agreed to join the panel (1940 members). Of these 1940 members, 865 (44.6%) responded to an email about this study of synthetic voices which would require them to review their assigned recording and complete the survey.

The Voice Samples

For the voice samples, I went through all the recordings of synthetic voices I had worked with from 2000 to 2017 (Lewis, 2001a-b, 2002, 2004; Polkosky & Lewis, 2002a-c). Previous research has shown that listeners make their quality judgements quickly (Polkosky & Lewis, 2002c; Wang & Lewis, 2001), so the samples were edited to have a length of about 30 seconds. In addition to the synthetic voices, I also put together three samples of professional human voice talents who had either recorded segments for an interactive voice response system or had provided recordings for the production of a synthetic voice. Table 1 lists the voice samples and the following key characteristics:

VoiceCode: This is the code given to survey participants to identify the voice they were rating (with 30 added to the base voice number to avoid having a number that stood out to participants, such as Sample01).

Year: This is the year the voice recording was made.

Gender: This is the gender (male or female) of the voice.

Type: This was the type of voice. Formant voices use an older technology of synthesis-by-rule (Klatt, 1980) rather than the currently standard method of combining bits of speech recorded by a professional voice talent (standard concatenative method – see Spiegel & Streeter, 1997). Expressive refers to a method developed at IBM to enhance the emotional expressiveness of otherwise standard concatenative voices by (1) having the voice talent provide additional recordings spoken with the target expression, (2) having a set of markup tags to direct the underlying text-to-speech (TTS) engine to produce the target expression by preferentially selecting from the appropriate recordings and performing additional acoustic modelling to achieve the desired expression (Pitrelli, Bakis, Eide, Fernandez, Hamza, & Picheny, 2006). The Expressive samples in this study had Good News content, and were either untagged (Expressive) or included Good News tagging (ExpressiveTagged). Human refers to the voices recorded by professional human voice talents.

TextCode: This documents the textual content of the recording. Over the years of conducting evaluations of synthetic voices, there has been variation in the test text, as shown below:

AirFrance_YouHave: "Air France flight zero nine five departs from Miami International at eight-fifty p.m. and arrives at Charles De Gaulle in Paris at eleven-ten a.m. the next day. There are five coach class tickets available for June sixth, but there are no aisle seats. You have requested a payment of one-hundred-eighty-seven dollars and fifty-six cents to BellSouth on March twelfth from your checking account. The resulting balance will be eight-hundred seventy-seven dollars and ninety-eight cents. Would you like information about a car loan?"

LetsReview: "Let's review to be sure I heard you correctly. Are you planning to pick up the car at the Dallas/Fort Worth airport, May 30, 2004, United Airlines flight four eight two and returning to the Dallas/Fort Worth airport, June 7, 2004, 10 AM, using your current profile? The approximate total charge for this reservation is three hundred fifty one dollars, with unlimited mileage. Would you like to hear a breakdown of the charges? Goodbye!"

ICanHelp: "I can help you make, change, or cancel reservations. What would you like to do? Is there anything else I can do for you today? Goodbye! Making new reservation … from which city or airport will you pick up the car? Do you know your arrival airline and flight number? Let's review to be sure I heard you correctly. Do you want to use your current profile for this reservation? Would you like to hear a breakdown of the charges? Would you like me to make this reservation?"

Hello_YouHaveRented: "Hello, thank you for calling the automated voice center. May I have your customer number, please? You have rented a four-door Chrysler sedan for fifty-nine dollars per day with drop off at Baltimore International Airport on May third. What's new and hot in health and beauty? Visit Top Sellers. You'll find the hottest makeup and skin care, customer favorites from hair care and shaving, and other great products. This attractive phone is designed for high-speed Internet access."

Hello_AirFrance: "Hello, thank you for calling the automated voice center. May I have your customer number, please? Air France flight zero nine five departs from Miami International at eight-fifty p.m. and arrives at Charles De Gaulle in Paris at eleven-ten a.m. the next day. There

are five coach class tickets available for June sixth, but there are no aisle seats. You have rented a four-door Chrysler sedan for fifty-nine dollars per day with drop off at Baltimore International Airport on May third."

ByPayingUpFront: "By paying up front you can avoid additional service fees and depending on your state this could save you up to fifty dollars a year. We're happy to offer you a free quote. Got it! Thank you very much! That's wonderful! Great! Great news! That will help us get you the best rate possible. There are many reasons to switch. For starters, most customers save money, and we are highly rated in customer satisfaction. There are no fees for you to get a quote with us. You can get your free quote in fifteen minutes or less."

ThatsAGreatChoice: "That's a great choice, and you will always have the option to change it at any time. We are so happy that you're joining our family. Can you believe that you'll get this coverage for only fifteen dollars a month? Wow! Our top priority is to give you the best service possible. Great! Just one more question and you'll be on your way to top quality coverage with us. We really appreciate your payment. Thank you!"

Table 1. Voice Samples

Voice	VoiceCode	Year	Gender	Type	TextCode
1	Sample31	2001	Male	Formant	AirFrance_YouHave
2	Sample32	2001	Female	Standard	AirFrance_YouHave
3	Sample33	2001	Male	Standard	AirFrance_YouHave
4	Sample34	2001	Female	Standard	AirFrance_YouHave
5	Sample35	2001	Male	Standard	AirFrance_YouHave
6	Sample36	2001	Female	Standard	AirFrance_YouHave
7	Sample37	2001	Male	Standard	AirFrance_YouHave
8	Sample38	2001	Female	Standard	AirFrance_YouHave
9	Sample39	2001	Male	Standard	AirFrance_YouHave
10	Sample40	2002	Female	Standard	AirFrance_YouHave
11	Sample41	2002	Male	Standard	AirFrance_YouHave
12	Sample42	2002	Male	Standard	AirFrance_YouHave
13	Sample43	2005	Female	Human	LetsReview
14	Sample44	2005	Female	Human	ICanHelp
15	Sample45	2005	Female	Standard	AirFrance_YouHave
16	Sample46	2005	Female	Standard	AirFrance_YouHave
17	Sample47	2005	Female	Standard	Hello_YouHaveRented
18	Sample48	2005	Female	Standard	Hello_YouHaveRented
19	Sample49	2006	Female	Standard	Hello_YouHaveRented
20	Sample50	2006	Female	Standard	Hello_YouHaveRented
21	Sample51	2006	Female	Standard	Hello_YouHaveRented
22	Sample52	2009	Female	Standard	Hello_AirFrance
23	Sample53	2009	Male	Standard	Hello_AirFrance
24	Sample54	2016	Female	ExpressiveTagged	ByPayingUpFront

Table 1. Voice Samples (cont.)

Voice	VoiceCode	Year	Gender	Type	TextCode
25	Sample55	2016	Female	ExpressiveTagged	ThatsAGreatChoice
26	Sample56	2016	Female	Expressive	ByPayingUpFront
27	Sample57	2016	Female	Expressive	ThatsAGreatChoice
28	Sample58	2016	Female	Concatenative	ByPayingUpFront
29	Sample59	2016	Female	Concatenative	ThatsAGreatChoice
30	Sample60	2017	Female	Human	ByPayingUpFront
31	Sample61	2017	Female	ExpressiveTagged	ByPayingUpFront
32	Sample62	2017	Male	Concatenative	ByPayingUpFront
33	Sample63	2017	Male	Concatenative	ByPayingUpFront
34	Sample64	2017	Male	Concatenative	ThatsAGreatChoice
35	Sample65	2017	Male	Concatenative	ByPayingUpFront
36	Sample66	2017	Female	Concatenative	ByPayingUpFront
37	Sample67	2017	Female	Concatenative	ByPayingUpFront
38	Sample68	2017	Female	Concatenative	ByPayingUpFront
39	Sample69	2017	Female	Concatenative	ByPayingUpFront
40	Sample70	2017	Female	Concatenative	ByPayingUpFront
41	Sample71	2017	Female	Concatenative	ByPayingUpFront
42	Sample72	2017	Female	ExpressiveTagged	ThatsAGreatChoice
43	Sample73	2017	Female	Concatenative	ByPayingUpFront
44	Sample74	2017	Male	Concatenative	ThatsAGreatChoice
45	Sample75	2017	Female	Concatenative	ThatsAGreatChoice
46	Sample76	2017	Male	Concatenative	ThatsAGreatChoice
47	Sample77	2017	Female	Concatenative	ThatsAGreatChoice
48	Sample78	2017	Female	Concatenative	ThatsAGreatChoice
49	Sample79	2017	Female	Concatenative	ThatsAGreatChoice
50	Sample80	2017	Female	Concatenative	ThatsAGreatChoice
51	Sample81	2017	Female	Concatenative	ThatsAGreatChoice
52	Sample82	2017	Female	Concatenative	ThatsAGreatChoice
53	Sample83	2017	Female	Expressive	AirFrance_YouHave
54	Sample84	2017	Female	Expressive	Hello_YouHaveRented
55	Sample85	2017	Female	ExpressiveTagged	AirFrance_YouHave
56	Sample86	2017	Female	ExpressiveTagged	Hello_YouHaveRented

Here are a few important notes about some of the samples:

Sample31: This was the only voice produced using a formant TTS engine.

Sample43: This was a human recording made by splicing together a large number of smaller recordings. The recordings were made with the intention of being spliced as smoothly as possible.

Sample44: This was the same human voice as Sample43, but is continuous without any inter-sentence splicing.

Sample60: This was a human recording made without any inter-sentence splicing. The audio in this recording was part of the audio used to create the Expressive TTS voice described below for Sample61.

Sample61: This was a voice converted from a standard to Expressive TTS in early 2017. This is the only sample in this study where the synthesized sentences exactly matched sentences used to build the voice and the text was enclosed in Good News tags (in fact, the sentences match those in the human recording Sample60). Due to this matching, this sample was expected to receive relatively high ratings.

Sample72 and 83-86: These samples used the same voice as Sample61, but the synthesized sentences were not in the set of sentences used to build the voice (some samples were tagged with Good News and others were not, as indicated in the table).

Results

The data analyses focused on construct validity, scale reliability, regression analyses of scales with the outcome metrics of likelihood-to-recommend (LTR) and overall quality, and finally, benchmarking and normative analyses.

Construct Validity

Table 2 shows the loadings from a factor analysis of the MOS-X items (unweighted least squares method with varimax rotation).

Table 2. Structure of MOS-X Four-Factor Solution

Item	Factor 1	Factor 2	Factor 3	Factor 4
1. Effort	0.215	**0.784**	0.108	0.117
2. Comprehension	0.142	**0.832**	0.098	0.130
3. Articulation	0.241	**0.869**	0.177	0.141
4. Precision	0.190	**0.771**	0.203	0.221
5. Pleasantness	**0.614**	0.309	*0.496*	0.183
6. Naturalness	*0.493*	0.247	**0.701**	0.314
7. Humanlike	*0.485*	0.223	**0.667**	0.260
8. Harsh	**0.432**	0.379	0.333	0.149
9. Emphasis	**0.541**	0.317	0.149	*0.486*
10. Rhythm	*0.430*	0.307	0.331	**0.643**
11. Intonation	*0.487*	0.308	0.361	**0.642**
12. Trust	**0.766**	0.163	0.335	0.175
13. Confidence	**0.773**	0.253	0.246	0.222
14. Enthusiasm	**0.746**	0.213	0.178	0.250
15. Persuasive	**0.828**	0.210	0.242	0.238

Note: Bold indicates the strongest loading for an item on a factor; italics indicates a secondary loading that exceeded .400 in strength

The observed factor structure was similar but not identical to that observed by Polkosky & Lewis (2003). The first four items (Intelligibility) clustered together alone on one factor. The last four items (Social Impression) also clustered together on a factor, but that factor also had strong

item loadings from some items formerly associated with Naturalness and Prosody (Items 5, 8 and 9).

Table 3 shows the factor loadings for the combined items of the MOS-X and the MOS-X2. The most interesting finding here is that the four items of the MOS-X2 were distributed among the four factors with the expected alignment pattern. This illustrates the structural relationship between the MOS-X and the MOS-X2, which was derived from the MOS-X.

Table 3. Structure of MOS-X and MOS-X2 Combined Four-Factor Solution

MOS-X Item	**Factor 1**	**Factor 2**	**Factor 3**	**Factor 4**
1. Effort	0.205	**0.814**	0.124	0.100
2. Comprehension	0.129	**0.837**	0.121	0.137
3. Articulation	0.231	**0.841**	0.197	0.173
4. Precision	0.182	**0.746**	0.216	0.244
5. Pleasantness	**0.613**	0.305	*0.491*	0.187
6. Naturalness	*0.487*	0.237	**0.731**	0.277
7. Humanlike	*0.476*	0.212	**0.702**	0.219
8. Harsh	**0.436**	0.383	0.317	0.144
9. Emphasis	**0.530**	0.305	0.169	*0.527*
10. Rhythm	*0.427*	0.295	0.368	**0.593**
11. Intonation	*0.485*	0.299	0.392	**0.600**
12. Trust	**0.774**	0.155	0.333	0.168
13. Confidence	**0.770**	0.250	0.249	0.238
14. Enthusiasm	**0.734**	0.221	0.207	0.237
15. Persuasive	**0.818**	0.210	0.271	0.226
MOS-X2 Item	**Factor 1**	**Factor 2**	**Factor 3**	**Factor 4**
1. Intelligibility	0.216	**0.802**	0.134	0.130
2. Naturalness	*0.421*	0.307	**0.648**	0.320
3. Prosody	*0.407*	0.336	0.348	**0.525**
4. Social Impression	**0.599**	0.218	0.337	0.349

Note: Bold indicates the strongest loading for an item on a factor; italics indicates a secondary loading that exceeded .400 in strength

Parallel analysis of the eigenvalues for both the MOS-X and the MOS-X2 indicated a two-factor solution, with Intelligibility as one factor and all other items on a second factor. In other words, it appears that the construct of Intelligibility stands on its own, but Naturalness, Prosody, and Social Impression tend to blend together. On the other hand, the alignment of items with factors in Table 3 suggests that there might still be practical value in working with a model made up of Intelligibility, Naturalness, Prosody, and Social Impression, as shown below in the sections on reliability and regression analysis.

Reliability

Coefficient alpha tended to be higher for the MOS-X and its subscales than for the MOS-X2, but all the scales had acceptable reliability (coefficient alpha greater than 0.70 – see Table 4). Because it is not possible to compute coefficient alpha for a single item, the only value of coefficient alpha for the MOS-X2 is for its overall (composite) score.

Table 4. Coefficient for the MOS-X (Overall and Subscales) and the MOS-X2

Scales	Alpha
Overall MOS-X	0.95
Intelligibility	0.92
Naturalness	0.90
Prosody	0.90
Social Impression	0.93
Overall MOS-X2	0.85

Concurrent Validity

As shown in Table 5, all correlations exceeded the minimum criterion of r = 0.30 as indicative of concurrent validity, and all were statistically significant ($p < 0.0001$).

Table 5. Concurrent Validities for the MOS-X and MOS-X2

Scales/Items	LTR	Overall Quality
Overall MOS-X	0.82	0.86
Intelligibility scale	0.58	0.66
Naturalness scale	0.79	0.82
Prosody scale	0.75	0.79
Social Impression scale	0.73	0.74
Overall MOS-X2	0.85	0.91
Intelligibility item	0.53	0.61
Naturalness item	0.82	0.83
Prosody item	0.74	0.80
Social Impression item	0.72	0.75

Regression Analyses

The purpose of the regression analyses was to determine which questionnaire produced responses that were more predictive of ratings of LTR and Overall Quality. For the MOS-X, the predictors were the subscales defined in Polkosky and Lewis (2003). For the MOS-X2, the predictors were its four items. Thus, for both questionnaires, there were four predictors: Intelligibility, Naturalness, Prosody, and Social Impression.

Table 6 shows the results of the various analyses. All models were statistically significant ($p < 0.0001$), as were the beta weights of all predictors ($p < 0.001$). The models accounted (after adjustment) for from 68.5% to 82.5% of the variation in the predicted metric. For both LTR and Overall Quality the MOS-X2 models had slightly higher coefficients of determination (R^2) than the MOS-X models.

Table 6. Beta Weights and Coefficient of Determination for Four Regression Models

Questionnaire	Predicting	Intelligibility (beta weight)	Naturalness (beta weight)	Prosody (beta weight)	Social Impression (beta weight)	R^2 (adjusted)
MOS-X	LTR	0.133	0.428	0.229	0.134	0.685
MOS-X2	LTR	0.111	0.479	0.193	0.209	0.742
MOS-X	Overall Quality	0.221	0.396	0.261	0.108	0.753
MOS-X2	Overall Quality	0.190	0.397	0.280	0.206	0.825

To compare the MOS-X and MOS-X2 models, 95% confidence intervals were computed around the adjusted estimates of R^2:

- MOS-X predicting LTR: 0.650-0.719
- MOS-X2 predicting LTR: 0.712-0.771
- MOS-X predicting Overall Quality: 0.724-0.781
- MOS-X2 predicting Overall Quality: 0.805-0.846

As a test of statistical significance, confidence intervals that did not overlap were indicative of a statistically significant difference ($p < 0.05$). There was clear separation in favor of the MOS-X2 for prediction of Overall Quality. The intervals slightly overlapped for prediction of LTR (but 90% confidence intervals were separate, indicating a potential difference in favor of the MOS-X2 with $p < 0.10$).

Benchmark Analyses

The data collected in this study are amenable to two types of benchmarking: (1) comparison with ratings of the three recordings of professional human voice talents and (2) comparison with ratings of the other synthetic voices.

Ratings of Human Voices

Three of the recordings were of professional human voice talents (Sample43, Sample44, and Sample60). Each sample had different content, with Sample43 including a large number of inter-sentence splices, unlike the other two samples. Careless splicing can degrade the quality of speech, but in this case the audio segments had been designed for smooth splicing. Table 7 shows the mean ratings obtained for the three human voice samples.

To make it easier to compare ratings from the two questionnaires, ratings from both were manipulated to range from 0 to 100. For the MOS-X2 (with item scales ranging from 0 to 10), this only required multiplication by 10. For the MOS-X (with item scales ranging from 1 to 7), the transformation was a little more complex ($t = (x\text{-}1)(100/6)$ where t is the transformed score and x is the original rating on a 7-point scale).

An ANOVA with a between-subjects variable of Sample (43, 44, and 60) and within-subjects variables of Questionnaire (MOS-X and MOS-X2) and Subscale (Intelligibility, Naturalness, Prosody, and Social Impression) revealed a significant three-way interaction ($F(3, 213) = 3.8$, $p = 0.003$). The main effect of Subscale was, as is typically the case, highly significant ($F(3, 213) = 41.3$, $p < 0.0001$), but neither of the other main effects (Questionnaire or Sample) was significant. This indicates that the overall means of the three samples did not differ significantly, nor did the overall means of the two questionnaires, but the underlying patterns of subscale scores were different. On the basis of the nonsignificant main effect of Sample, the data from the three human voice samples were combined for the purpose of establishing an upper boundary for judging the quality of a synthetic voice (see the grand means in Table 7).

Table 7. Mean Ratings of Human Voices

Voice	Questionnaire	Overall	Intelligibility	Naturalness	Prosody	Social Impression
Sample43	MOS-X	83.9	93.5	86.0	74.6	81.5
	MOS-X2	85.4	96.4	77.1	78.6	89.6
Sample44	MOS-X	82.9	92.2	82.3	78.2	79.0
	MOS-X2	86.6	96.8	78.0	82.4	89.2
Sample60	MOS-X	88.4	94.6	82.9	87.4	88.5
	MOS-X2	85.9	96.2	80.5	81.0	88.1
Grand Mean	MOS-X	84.8	93.4	83.9	79.5	82.7
	MOS-X2	85.8	96.5	78.4	80.5	88.1

A series of *t*-tests conducted on the grand means indicated no significant differences between the questionnaires for Overall ($t(73) = -1.03$, $p = 0.31$) and Prosody ($t(73) = -0.52$, $p = 0.61$), but there were significant differences for Intelligibility ($t(73) = -3.4$, $p = 0.001$), Naturalness ($t(73) = 4.2$, $p < 0.0001$), and Social Impression ($t(73) = -3.4$, $p = 0.001$). Table 8 shows the resulting benchmarks for the interpretation of MOS-X and MOS-X2 ratings of synthetic voices relative to professional voice talents.

Table 8. MOS-X and MOS-X2 Benchmarks Based on Ratings of Professional Voice Talents

Questionnaire	Overall	Intelligibility	Naturalness	Prosody	Social Impression
MOS-X	85.3	93.4	83.9	80.0	82.7
MOS-X2	85.3	96.5	78.4	80.0	88.1

Ratings of Synthetic Voices and Preliminary Curved Grading Scales

The remaining 53 samples were of various synthetic voices recorded from 2001 through 2017 (see Table 1 for a description of the voices and Appendix A for a table of the overall and subscale means for each voice). Although not a random sample of synthetic voices, they are somewhat representative of the advances in TTS quality made over that period of time. Table 9 shows the samples arranged in order of descending MOS-X and MOS-X2 ratings. The left two columns of the table show the percentile and an associated grade, similar to the scheme Sauro and Lewis (2016) used to develop a curved grading scale for the System Usability Scale. Specifically, the top and bottom 15 percentiles were assigned respectively to A and F, the center 40 percentiles were assigned to C, and the remaining two groups of 15 percentiles were assigned to B and D. The percentiles for A, B, and C were further divided into plus, neutral, and minus grades.

Table 9. Overall Ratings of Synthetic Voices (* = IBM Expressive Voice with Tagging)

Percentile	Grade	MOS-X	Sample	MOS-X2	Sample
100.0%	A+	81.7	61*	84.1	53
98.1%	A+	81.4	53	83.1	61*
96.2%	A+	74.3	85*	79.8	85*
94.3%	A	73.5	72*	75.9	72*
92.5%	A	73.2	82	73.4	82
90.6%	A	72.7	86*	73.3	55*
88.7%	A-	72.4	55*	72.1	71
86.8%	A-	70.6	52	71.8	86*
84.9%	A-	69.8	71	71.5	80
83.0%	B+	68.7	77	71.0	52
81.1%	B+	68.5	79	70.9	77
79.2%	B	66.8	80	69.8	68
77.4%	B	65.4	64	69.8	79
75.5%	B	65.3	68	69.7	64
73.6%	B-	64.3	58	67.8	50
71.7%	B-	63.6	66	67.7	75
69.8%	C+	61.9	76	67.6	76
67.9%	C+	61.9	78	67.5	49
66.0%	C+	60.9	65	67.5	58
64.2%	C+	60.7	49	67.5	66
62.3%	C+	60.5	50	66.9	78
60.4%	C+	60.2	73	66.3	65
58.5%	C+	60.0	69	66.3	84
56.6%	C	59.8	84	66.2	59
54.7%	C	59.4	81	66.2	62
52.8%	C	59.4	75	65.1	69
50.9%	C	59.2	48	65.0	57
49.1%	C	59.2	59	64.7	48
47.2%	C	58.7	56	64.2	56
45.3%	C	58.5	67	63.8	67
43.4%	C	58.4	41	63.8	81
41.5%	C	57.1	70	63.5	83
39.6%	C-	56.7	57	62.6	41
37.7%	C-	56.5	62	62.3	35
35.8%	C-	56.3	54*	62.0	46
34.0%	C-	55.6	51	61.9	73
32.1%	C-	54.7	74	61.3	54*

Table 9. Overall Ratings of Synthetic Voices (cont.)

Percentile	Grade	MOS-X	Sample	MOS-X2	Sample
30.2%	D	53.6	46	60.2	74
28.3%	D	52.8	63	59.8	39
26.4%	D	52.8	83	59.0	51
24.5%	D	51.4	35	58.8	63
22.6%	D	50.9	42	58.3	42
20.8%	D	46.4	39	57.0	70
18.9%	D	45.4	40	55.8	40
17.0%	D	44.9	47	54.0	47
15.1%	F	42.5	34	50.6	33
13.2%	F	41.6	33	49.8	32
11.3%	F	40.3	32	49.8	38
9.4%	F	37.9	38	49.5	34
7.5%	F	31.4	31	39.7	45
5.7%	F	31.1	45	38.7	31
3.8%	F	28.9	36	37.5	37
1.9%	F	28.4	37	37.1	36

Table 10 summarizes the data from Table 9 in the form of preliminary curved grading scales for overall MOS-X and MOS-X2 scores. Note that for any given grade range the synthetic voices MOS-X scores tended to be lower than the MOS-X2 scores.

Table 10. Preliminary Curved Grading Scales for Overall MOS-X and MOS-X2

MOS-X	MOS-X2	Grade
74.3-100	79.8-100	A+
72.7-74.2	73.3-79.7	A
69.8-72.6	71.5-73.2	A-
68.5-69.7	70.9-71.4	B+
65.3-68.4	69.7-70.8	B
63.6-65.2	67.7-69.6	B-
60.0-63.5	66.3-67.6	C+
57.1-59.9	63.5-66.2	C
54.7-57.0	61.3-63.4	C-
44.9-54.6	54.0-61.2	D
0-44.8	0-53.9	F

Table 11 shows typical (means over all synthetic voices in this study) and above average (means of the top 10 rated synthetic voices) values for the four subscales. Though not as granular as the curved grading scale presented in Table 10 for overall ratings, the values in Table 11 may be useful for roughly interpreting the subscales, especially when simultaneously

considered with the values give in Table 8 (benchmarks based on ratings of human voice talents). The values in Table 11 can be used to judge a synthetic voice as average by comparison with the mean values computed across all 53 synthetic voice samples or as above average by comparison with the mean values computed for the top ten synthetic voice samples. For example, when using the MOS-X, a Social Impression score of 55.5 would be average, and one of 72.8 would be above average (and referring back to Table 8, a score of 82.7 would match the sample of human voices).

Table 11. Means of All Synthetic Voices and the Top Ten Rated Voices

Mean of	Questionnaire	Intelligibility	Naturalness	Prosody	Social Impression	Overall
Top 10	MOS-X	86.7	68.1	67.8	72.8	73.8
	MOS-X2	93.7	63.4	70.7	74.6	75.6
All	MOS-X	77.4	54.4	53.5	55.5	60.2
	MOS-X2	84.8	51.2	60.9	64.8	65.4

Discussion

The MOS has a rich history in the assessment of voices transmitted over various channels and as adapted to the assessment of synthetic voices. The objectives of this research were to (1) evaluate and compare two versions of the expanded MOS, one using the original 15 items (MOS-X) and the other a four-item version with one item for each of the four factors of the MOS-X (the MOS-X2), and (2) establish preliminary benchmarks for the interpretation of ratings collected using these questionnaires. Respondents (n = 865) provided ratings for 56 thirty-second recordings of speech samples – 53 from recordings of synthetic voices made from 2001 through 2017 and three from professional human voice talents.

Regarding the first objective, the analyses of reliability and validity supported the use of both questionnaires. There were, however, some issues with the construct validity of the MOS-X (specifically, the item-factor alignment for Naturalness and Prosody did not match expectation). Regression analyses of the relationships between the MOS-X and the MOS-X2 favored the MOS-X2. This latter finding was surprising given that the MOS-X is based on ratings from 15 items as opposed to just four items in the MOS-X2. Taken together, this suggests that although researchers might find more value in the MOS-X due to its richer set of items, user experience practitioners conducting listening tests should use the MOS-X2 due to its stronger statistical relationship to outcome metrics like LTR and Overall Quality and its shorter length, which makes it faster and easier for listening test participants to complete.

The benchmark analyses provided two ways to interpret MOS-X and MOS-X2 ratings. Statistical analysis supported the combination of the three human voice samples for benchmarking. Despite clear differences in the pattern of means for Intelligibility, Naturalness, Prosody, and Social Impression between the two questionnaires, their overall means for the human voices were almost the same – close enough that statistical analysis supported combining them for the purpose of benchmarking. Thus, if a synthetic voice had an overall mean MOS-X or MOS-X2 at or approaching 85.3, that synthetic voice would have a human-like score. The entries in Table 8 give researchers and practitioners a sense of the ratings for Intelligibility, Naturalness, Prosody, and Social Impression that would be required to support a claim of being human-like. In this study, there were a few voices that received scores approaching those of the human voices, specifically, Sample53 and Sample61 received mean MOS-X and MOS-X2 scores greater than 80 (their MOS-X2 means were, respectively, 84.1 and 83.1).

Benchmark analyses based on the ratings of the synthetic voices used in this study are on somewhat shakier ground because they may not be statistically representative of the full population of synthetic voices, but due to the number of voices and samples in this study, they may be of use to practitioners to grade synthetic voices that do not have human-like ratings

(e.g., Table 10). For example, a score of 72 on the MOS-X or MOS-X2 would receive a grade of A- -- well above average though short of human-like. On the other hand, a voice with a rating of 54 would be graded a D – well below average.

Recommendations for Future Research

There are some limits to generalization in this study that should be addressed by future research. This line of research would benefit from increasing the number of speech samples (both synthetic and human), striving to include the best voices currently available. It would also improve generalizability to sample from populations outside of IBM.

Conclusion

The results of this research have substantially improved our understanding of the research and practical properties of the MOS-X and MOS-X2 questionnaires. They have also provided a basis for the interpretation of MOS-X and MOS-X2 ratings from poor to average to good to human-like.

Both questionnaires had acceptable psychometric quality (reliability and validity), but the factor structure of the MOS-X did not exactly match the expected structure. The MOS-X2 had a stronger statistical relationship to outcome metrics of Likelihood-to-Recommend (LTR) and Overall Quality than the MOS-X. The very old samples (those using technologies from 2001-2002) received consistently poor ratings. Samples created using the relative new Expressive technologies (with expressive tagging) received consistently good ratings. A few of the synthetic voice samples came close to the ratings given to the professional human voice talents.

Either questionnaire version is acceptable for use, but due to its stronger statistical relationship to the key outcome metrics of LTR and Overall Quality and its shorter length, it is more effective and efficient to use the MOS-X2. The mean MOS-X and MOS-X2 for the ratings of the professional human voices were both about 85 (after conversion to a 0-100 point scale), so a synthetic voice with mean ratings at or approaching 85 would be very good. Ratings over 70 are, relative to the set of voice samples in this study, above average.

Tips for Practitioners

- For listening tests, use the MOS-X2 rather than the MOS-X due to its shorter length, less demand on study participants, and its stronger statistical relationship to the outcome metrics of LTR and Overall Quality.
- Be cautious when using the benchmarks presented in this paper. The human benchmark of an overall score of 85.3 for both questionnaires should be fairly stable, but the curved grading scale presented in Table 10 is quite preliminary (and could shift radically given ratings of additional voice samples).
- When creating a TTS voice for a specific project, it can be advantageous to add to the base set of recordings as many segments as possible from the planned dialog specification to increase the likelihood of an exact match between the audio data used to build the voice and the audio that the voice will produce.

Acknowledgements

Many thanks to members of the IBM UX Panel who participated in this study.

References

Francis, A. L., & Nusbaum, H. C. (1999). Evaluating the quality of synthetic speech. In D. Gardner-Bonneau (ed.), *Human Factors and Voice Interactive Systems* (pp. 63-97). Boston, MA: Kluwer.

International Telecommunication Union (1994). *A Method for subjective performance assessment of the quality of speech voice output devices* (ITU-T Recommendation, p. 85). Geneva, Switzerland: ITU.

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.

Klatt, D. H. (1980). Software for a cascade/parallel formant synthesizer. *Journal of the Acoustical Society of America*, *67*, 971-995.

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

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

Lewis, J.R. (2001a). Psychometric properties of the Mean Opinion Scale. In *Proceedings of HCI International 2001: Usability Evaluation and Interface Design* (pp. 149-153). Mahwah, NJ: Lawrence Erlbaum.

Lewis, J. R. (2001b). *The Revised Mean Opinion Scale (MOS-R): Preliminary psychometric evaluation* (Tech Report 29.3414). West Palm Beach, FL: International Business Machines Corp.

Lewis, J. R. (2002). *Effect of voice and bandwidth on MOS-X ratings* (Tech. Report 29.3550). Boca Raton, FL: International Business Machines Corp.

Lewis, J. R. (2004). Effect of speaker and sampling rate on MOS-X ratings of concatenative TTS voices. In *Proceedings of the Human Factors and Ergonomics Society 48th Annual Meeting* (pp. 759-784). Santa Monica, CA: HFES.

Lewis, J. R. (2011). Practical *speech user interface design*. Boca Raton, FL: Taylor & Francis Group.

Lewis, J. R. (2016). Standardized questionnaires for voice interaction design. *Voice Interaction Design*, *1(1)*, 1-16.

Lewis, J. R., & Erdinç, O. (2017). User experience rating scales with 7, 11, or 101 points: Does it matter? *Journal of Usability Studies*, *12(2)*, 73-91.

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

Pitrelli, J. F., Bakis, R., Eide, E. M., Fernandez, R., Hamza, W., & Picheny, M. (2006). The IBM expressive text-to-speech synthesis system for American English. *IEEE Transactions on Audio, Speech, and Language Processing*, *14(4)*, 1099-1108.

Polkosky, M. D., & Lewis, J. R. (2002a). *Development and psychometric evaluation of an expanded Mean Opinion Scale (MOS-X)* (Tech. Report 29.3499). Boca Raton, FL: International Business Machines Corp.

Polkosky, M. D., & Lewis, J. R. (2002b). *Enhancement of the Mean Opinion Scale – Expanded (MOS-X)* (Tech. Report 29.3542). Boca Raton, FL: International Business Machines Corp.

Polkosky, M. D., & Lewis, J. R. (2002c). *Toward a multimedia methodology for evaluating user preferences for synthetic voices* (Tech. Report 29.3486). Boca Raton, FL: International Business Machines Corp.

Polkosky, M. D., & Lewis, J. R. (2003). Expanding the MOS: Development and psychometric evaluation of the MOS-R and MOS-X. *International Journal of Speech Technology*, *6*, 161-182.

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.

Sauro, J., & Lewis, J. R. (2016). *Quantifying the user experience: Practical statistics for user research* (2nd ed.). Cambridge, MA: Morgan Kaufmann.

Schmidt-Nielsen, A. (1995). Intelligibility and acceptability testing for speech technology. In A. Syrdal, R. Bennett, and S. Greenspan (Eds.), *Applied Speech Technology* (pp. 195-232). Boca Raton, FL: CRC Press.

Spiegel, M. F., & Streeter, L. (1997). Applying speech synthesis to user interfaces. In M. Helander, T. K. Landauer, & P. Prabhu (eds.), *Handbook of Human-Computer Interaction* (pp. 1061-1084). Amsterdam: Elsevier.

van Bezooijen, R. & van Heuven, V. (1997). Assessment of synthesis systems. In D. Gibbon, R. Moore, and R. Winski (Eds.), *Handbook of Standards and Resources for Spoken Language Systems* (pp. 481-563). New York, NY: Mouton de Gruyter.

Wang, H., & Lewis, J. R. (2001). Intelligibility and acceptability of short phrases generated by embedded text-to-speech engines. In *Proceedings of HCII 2001* (pp. 145-149). Mahwah, NJ: Lawrence Erlbaum.

About the Author

James R. (Jim) Lewis
Jim is human factors engineer at IBM, specializing in voice interaction design and usability assessment. He is a past president of the Association for Voice Interaction Design (AVIxD). His books include *Practical Speech User Interface Design* (2011) and (with Jeff Sauro in 2012/2016), *Quantifying the User Experience*.

Appendix A: Overall and Subscale Means for Synthetic Voices

Sample	Questionnaire	Intelligibility	Naturalness	Prosody	Social Impression	Overall
31	MOS-X	53.7	19.4	26.5	26.1	31.4
	MOS-X2	63.5	13.5	38.8	38.8	38.7
32	MOS-X	56.9	34.4	31.9	37.8	40.3
	MOS-X2	70.0	32.0	47.3	50.0	49.8
33	MOS-X	62.2	33.9	35.8	34.4	41.6
	MOS-X2	69.4	29.4	50.6	53.1	50.6
34	MOS-X	62.9	38.6	30.8	37.5	42.5
	MOS-X2	70.9	33.6	42.7	50.9	49.5
35	MOS-X	67.3	42.6	46.4	49.3	51.4
	MOS-X2	79.3	44.3	56.4	69.3	62.3
36	MOS-X	40.7	27.2	22.6	25.0	28.9
	MOS-X2	50.0	18.5	36.9	43.1	37.1
37	MOS-X	46.4	23.0	20.2	24.1	28.4
	MOS-X2	55.7	18.6	36.4	39.3	37.5
38	MOS-X	52.5	33.3	30.0	35.8	37.9
	MOS-X2	59.0	37.0	49.0	54.0	49.8
39	MOS-X	63.6	39.4	40.4	42.0	46.4
	MOS-X2	80.0	46.4	51.8	60.9	59.8
40	MOS-X	62.2	36.7	40.3	42.4	45.4
	MOS-X2	74.4	38.1	55.0	55.6	55.8
41	MOS-X	70.5	55.2	52.3	55.6	58.4
	MOS-X2	81.7	55.0	63.3	62.5	65.6
42	MOS-X	68.8	45.1	42.0	47.7	50.9
	MOS-X2	77.5	46.9	50.6	58.1	58.3
45	MOS-X	38.2	29.8	25.5	30.9	31.1
	MOS-X2	46.5	24.1	40.0	48.2	39.7
46	MOS-X	73.9	45.3	49.3	46.1	53.6
	MOS-X2	82.0	46.7	59.3	60.0	62.0
47	MOS-X	63.5	40.1	35.0	41.0	44.9
	MOS-X2	80.8	35.4	43.8	56.2	54.0
48	MOS-X	73.4	56.1	49.6	57.8	59.2
	MOS-X2	81.3	54.7	55.3	67.3	64.7
49	MOS-X	74.7	58.1	55.6	54.4	60.7
	MOS-X2	82.5	57.5	60.6	69.4	67.5
50	MOS-X	78.3	54.2	57.0	52.5	60.5
	MOS-X2	84.0	55.3	66.7	65.3	67.8

Sample	Questionnaire	Intelligibility	Naturalness	Prosody	Social Impression	Overall
51	MOS-X	68.6	52.2	48.1	53.3	55.6
	MOS-X2	76.0	49.3	52.0	58.7	59.0
52	MOS-X	81.9	63.9	64.8	72.0	70.6
	MOS-X2	87.3	55.3	66.0	75.3	71.0
53	MOS-X	95.2	77.7	69.4	83.3	81.4
	MOS-X2	97.9	77.1	78.6	82.9	84.1
54	MOS-X	74.1	50.2	46.9	53.9	56.3
	MOS-X2	83.9	46.1	53.3	61.7	61.3
55	MOS-X	86.8	67.1	66.7	69.1	72.4
	MOS-X2	94.4	58.9	67.2	72.8	73.3
56	MOS-X	78.1	53.4	53.5	50.0	58.7
	MOS-X2	86.9	51.9	51.9	66.3	64.2
57	MOS-X	86.9	45.3	45.1	49.4	56.7
	MOS-X2	93.8	53.8	55.4	56.9	65.0
58	MOS-X	88.7	55.1	60.8	52.8	64.3
	MOS-X2	88.9	54.4	61.1	65.6	67.5
59	MOS-X	84.7	53.6	49.6	48.9	59.2
	MOS-X2	91.3	52.7	62.0	58.7	66.2
61	MOS-X	93.8	78.4	73.6	81.3	81.7
	MOS-X2	97.5	73.8	77.5	83.8	83.1
62	MOS-X	82.8	46.1	48.0	49.1	56.5
	MOS-X2	92.7	47.3	60.7	64.0	66.2
63	MOS-X	80.1	42.4	46.2	42.6	52.8
	MOS-X2	86.9	40.8	53.1	54.6	58.8
64	MOS-X	87.2	54.9	59.4	59.9	65.4
	MOS-X2	92.5	53.8	65.6	66.9	69.7
65	MOS-X	82.5	58.6	52.5	50.0	60.9
	MOS-X2	88.9	50.5	62.1	63.7	66.3
66	MOS-X	72.9	57.6	57.6	66.1	63.6
	MOS-X2	83.1	55.0	62.5	69.4	67.5
67	MOS-X	72.2	49.2	57.0	55.7	58.5
	MOS-X2	84.0	50.7	58.0	62.7	63.8
68	MOS-X	78.6	63.3	58.5	60.6	65.3
	MOS-X2	87.3	58.7	65.3	68.0	69.8
69	MOS-X	80.4	55.4	48.0	56.4	60.0
	MOS-X2	81.8	52.9	60.0	65.9	65.1

Sample	Questionnaire	Intelligibility	Naturalness	Prosody	Social Impression	Overall
70	MOS-X	84.8	45.3	48.0	50.3	57.1
	MOS-X2	87.1	32.1	61.4	47.1	57.0
71	MOS-X	81.5	62.0	67.6	68.3	69.8
	MOS-X2	89.4	57.2	70.6	71.1	72.1
72	MOS-X	88.5	68.1	69.6	67.9	73.5
	MOS-X2	95.3	62.9	75.9	69.4	75.9
73	MOS-X	75.5	56.0	53.5	55.8	60.2
	MOS-X2	84.4	45.0	56.3	61.9	61.9
74	MOS-X	78.1	41.0	59.7	39.9	54.7
	MOS-X2	86.7	36.7	63.3	54.2	60.2
75	MOS-X	83.2	56.8	50.0	47.3	59.4
	MOS-X2	92.9	52.9	62.9	62.1	67.7
76	MOS-X	90.2	55.9	52.9	48.8	61.9
	MOS-X2	94.1	54.7	61.2	60.6	67.6
77	MOS-X	76.8	63.0	66.3	68.7	68.7
	MOS-X2	86.3	56.3	68.8	72.5	70.9
78	MOS-X	85.9	54.9	53.8	52.9	61.9
	MOS-X2	95.0	50.6	58.8	63.1	66.9
79	MOS-X	87.2	64.6	59.3	62.8	68.5
	MOS-X2	92.5	55.8	60.0	70.8	69.8
80	MOS-X	88.7	61.8	57.4	59.3	66.8
	MOS-X2	96.1	58.9	65.0	66.1	71.5
81	MOS-X	84.4	48.1	52.2	52.8	59.4
	MOS-X2	92.0	40.0	66.7	56.7	63.8
82	MOS-X	89.6	63.7	66.5	73.2	73.2
	MOS-X2	92.9	54.3	69.3	77.1	73.4
83	MOS-X	74.8	40.3	48.6	47.6	52.8
	MOS-X2	86.7	42.5	58.3	66.7	63.5
84	MOS-X	75.6	58.8	48.9	56.1	59.8
	MOS-X2	82.0	56.0	65.3	62.0	66.3
85	MOS-X	89.3	68.2	69.0	70.5	74.3
	MOS-X2	95.7	74.3	75.0	74.3	79.8
86	MOS-X	83.3	69.3	64.1	73.9	72.7
	MOS-X2	90.9	60.9	61.8	73.6	71.8