



Expanding the MOS: Development and Psychometric Evaluation of the MOS-R and MOS-X

MELANIE D. POLKOSKY AND JAMES R. LEWIS

IBM Corporation, 8051 Congress Ave, Suite 2227, Boca Raton, FL 33487, USA

jimlewis@us.ibm.com

Abstract. The Mean Opinion Scale (MOS) is a questionnaire used to obtain listeners' subjective assessments of synthetic speech. This paper documents the motivation, method, and results of six experiments conducted from 1999 to 2002 that investigated the psychometric properties of the MOS and expanded the range of speech characteristics it evaluates. Our initial experiments documented the reliability, validity, sensitivity, and factor structure of the P.L. Salza et al. (*Acta Acustica*, Vol. 82, pp. 650–656, 1996) MOS and used psychometric principles to revise and improve the scale. This work resulted in the MOS-Revised (MOS-R). Four subsequent experiments expanded the MOS-R beyond its previous focus on Intelligibility and Naturalness, to include measurement of the Prosody and Social Impression of synthetic voices. As a result of this work, we created the MOS-Expanded (MOS-X), a rating scale shown to be reliable, valid, and sensitive for high-quality evaluation of synthetic speech in applied industrial settings.

Keywords: Mean Opinion Scale (MOS), subjective assessment of synthetic speech, psychometric evaluation

Introduction

Evaluation of intelligibility and acceptability is a vital component of effective synthetic speech development and use. Schmidt-Nielsen (1995) distinguishes between two major types of evaluations, intelligibility tests and acceptability tests. Various intelligibility tests, which evaluate how well listeners understand words or phrases spoken by a synthetic voice, provide highly correlated yet inconsistent results despite the various types of stimuli used (e.g., rhyming words, phrases, syllables, sentences) (Schmidt-Nielsen, 1995). In addition, listeners may understand a synthetic voice but still indicate that it has unacceptable quality, a subjective characteristic that intelligibility tests cannot assess. Acceptability tests provide a subjective measure of synthetic voice quality, typically using comparison rankings or rating scales. While listener variability has been identified as a common problem with subjective rating scales for acceptability measurement, they also provide an efficient and rapid method of synthetic speech evaluation (Schmidt-Nielsen, 1995).

Until relatively recently, the most pronounced problem with artificial voices was intelligibility (Francis

and Nusbaum, 1999). It might seem, therefore, that articulation tests that assess intelligibility (such as rhyme tests) would be more suitable for evaluating artificial speech than a subjective acceptability tool. However, most current text-to-speech systems, although more demanding on the listener than natural speech (Paris et al., 2000), are quite intelligible: "Once a speech signal has breached the 'intelligibility threshold', articulation tests lose their ability to discriminate. . . . it is precisely because people's opinions are so sensitive, not just to the signal being heard, but also to norms and expectations, that opinion tests form the basis of all modern speech quality assessment methods" (Johnston, 1996). In addition to the technological gains made in synthetic speech over the past decade, researchers have also recognized the need for increasingly sophisticated perceptual assessment methods (Pisoni, 1997).

The Mean Opinion Scale (MOS)

The Mean Opinion Scale (MOS) is a widely used method for evaluating the quality of telephone systems

and synthetic speech, recommended by the International Telecommunications Union (Schmidt-Nielsen, 1995; ITU, 1994; van Bezooijen and van Heuven, 1997). The MOS is a Likert-style questionnaire, typically with seven 5-point scale items addressing the following TTS characteristics: (1) Global Impression, (2) Listening Effort, (3) Comprehension Problems, (4) Speech Sound Articulation, (5) Pronunciation, (6) Speaking Rate, and (7) Voice Pleasantness. In the most typical use of the MOS, naïve listeners assign scores for each item after listening to speech stimuli, usually sentences (Schmidt-Nielsen, 1995). Documented use of the MOS in empirical studies of synthetic speech has been somewhat limited (Cartier et al., 1992, as cited by Salza et al., 1996; Goldstein, 1995; Goodman and Nash, 1982 as cited by Schmidt-Nielsen, 1995; Johnston, 1996; Kraft and Portele, 1995; Moller et al., 2001; Pascal and Combescure, 1988, as cited by Schmidt-Nielsen, 1995; Salza et al., 1996; Yabuoka et al., 2000), although it was shown to be a recognizable assessment method in a survey of worldwide speech laboratories (Pols and Jekosch, 1997).

The primary advantage of the MOS lies in its efficiency: it quickly provides feedback on a synthetic voice's intelligibility and naturalness as evaluated by listeners. Researchers can also compare ratings of several voices and determine the general aspects that differentiate the voices. Unlike other acceptability evaluation procedures, the MOS does not require time-consuming listener calibration and standardization procedures (as required by the Diagnostic Acceptability Measure), pre-specified speech stimuli or testing environments, or other rigid procedural requirements (Schmidt-Nielsen, 1995). Thus, this tool is flexible for a wide range of evaluative goals, synthetic speech applications, and user groups. When analyzing MOS scores, inferential statistics (such as ANOVA), compare the amount of within-group listener variability to the amount of variability among ratings of different synthetic voices. If listener variability is greater than or equal to variability among synthetic voices, the ANOVA will be nonsignificant; otherwise, it will be statistically significant. These procedures give the researcher specific methods of determining whether the recognized problem of listener variability adversely affects observed MOS ratings. Despite these potential advantages, however, little is known about the psychometric features of the MOS, a potential deterrent to its use in speech system evaluation.

Psychometrics and the MOS

Psychometrics is the use of statistical analysis to evaluate the quality of practitioner-created scales and other psychological measures (Nunnally, 1978). If a measurement tool, such as the MOS, has not been subjected to psychometric analysis, developers of speech products cannot be confident of a product evaluation that uses the tool. Some of the metrics of psychometric quality are:

1. Reliability—consistency of ratings performed on the same speech system at different times;
2. Validity—measurement of the specific aspects of a speech system that the developer intended; and
3. Sensitivity—ability of the tool to detect differences among speech systems.

These psychometric characteristics each are important to the overall quality of a scale, but they may not occur in the same scale. For example, a scale may result in the same ratings for the same system each time it is used (reliability), but it may not measure particular characteristics of interest to a developer (validity). It also may not discriminate among different speech systems (sensitivity). Therefore, it is important for developers to understand the reliability, validity, and sensitivity of subjective scales used for speech system evaluation.

Despite the recognized importance of psychometric review, previous research documenting the reliability, validity, and sensitivity of the MOS has been sparse and unsystematic. Cartier et al. (1992, as cited by Salza et al., 1996), in their evaluation of several synthetic systems, found reliable and reproducible results, but no additional or more specific comments on MOS reliability have appeared in the previous literature. Salza et al. (1996) measured the overall quality of three Italian TTS synthesis systems with a common prosodic control but different diphones and synthesizers using both paired comparisons and the MOS. Their results showed good agreement between the two measurement methods, providing some evidence for the validity of the MOS. Johnston (1996) also showed evidence of validity and sensitivity, but only for the MOS Listening Effort item. Yabuoka et al. (2000) investigated the relationship between five distortion scales (differential spectrum, phase, waveform, cepstrum distance, and amplitude) and MOS ratings. They calculated statistically significant regression formulas for predicting MOS ratings from manipulations of the distortion

scales, suggesting that the MOS was sensitive to differences created by the distortion scales. Unfortunately, Yabuoka et al. (2000) did not report the exact type of MOS that they used in the experiment. In general, this research does little to clarify the quality of the MOS for use as a measure in applied industrial settings.

Another statistical method often associated with scale development is factor analysis. Factor analysis is a statistical procedure that examines the correlations among variables to discover groups of related scale items, known as factors (Nunnally, 1978). This procedure allows individual scale items to be grouped categorically by identifying the abstract variables that a scale effectively measures. For a complex scale with a large number of individual items, the presence of factors permits greater ease in interpreting its results. Few researchers have investigated the factor structure of the MOS. Kraft and Portele (1995) found that an eight-item version of the MOS had two factors—one they called Intelligibility (segmental) and one called Naturalness (suprasegmental, or prosodic attributes). The Speaking Rate (Speed) item did not fall in either of the two factors, suggesting that this item was unrelated to the other seven items on the MOS. More recently, Sonntag et al. (1999), using the same version of the MOS (but with 6-point rather than 5-point scales), reported only a single factor. Therefore, the factor structure of the MOS continues to be unclear and limited, another potential drawback to its use in speech system evaluations.

Expanding the Scope of Coverage of the MOS

In addition to a paucity of research in the psychometrics of the MOS, the content of its items has received virtually no attention in previous research. Item content related to only two factors limits the instrument's ability to discriminate among voices with similar intelligibility and naturalness. Researchers in the late 1980s and early 1990s acknowledged that the intelligibility of synthetic speech can rival that of human speech (Greene et al., 1986; Murray and Arnot, 1993). As synthetic speech development has become increasingly sophisticated, it has become clear that intelligibility does not usually differentiate among modern synthetic voices. With the introduction of concatenative voices, naturalness also is becoming less of a discriminating factor. These trends suggest that the MOS, which measures only Intelligibility and Naturalness, may not effectively discriminate among current synthetic voices, nor future voices.

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 and Arnot, 1993), the relationship between vocal speech and impression formation or personality perception (for a review, see Brown et al., 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 and Bedrosian, 1992; Gorenflo and Gorenflo, 1997). All of these areas of research suggest additional measurement items that may capture 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 those related to prosody (Bradlow et al., 1996; Brown et al., 1973; Hosman, 1989; Koopmans-Van Beinum, 1992; Martin and Haroldson, 1992; Pelachaud et al., 1996; Slowiaczek and Nusbaum, 1985; Yaeger-Dror, 1996) and voicing characteristics (Bloom et al., 1999; Bradlow et al., 1996; Granstrom and Nord, 1992; Hieda and Kuchinomachi, 1997; Higashikawa and Minifie, 1999; Hillenbrand, 1988; Klatt and Klatt, 1990; Lavner et al., 2000; Page and Balloun, 1978; Robinson and McArthur, 1982; Slowiaczek and Nusbaum, 1985; Whalen and Hoequist, 1995). Other researchers have investigated the numerous social-emotional perceptions conveyed by speech (Berry, 1992; Ekman et al., 1991; Holtgraves and Lasky, 1999; Johnson et al., 1986; Massaro and Egan, 1996; Miyake and Zuckerman, 1993; Murray and Arnot, 1995; Murray et al., 1996; Paddock and Nowicki, 1986; Stern et al., 1999; Tartter and Braun, 1994; Whitmore and Fisher, 1996; Zuckerman et al., 1991). In general, the literature points to a number of perceptual characteristics that are not measured by acceptability tests or the current version of the MOS, but may further discriminate among synthetic voices.

Research Goals

The goals of our research program were to investigate and systematically improve the psychometric properties of the Mean Opinion Scale (MOS), and to expand the content of the MOS beyond its current content to include items that measure subtle vocal and social-emotional aspects of speech. Reliable and

valid measurement of these characteristics is important for understanding listeners' impressions of synthetic speech, developing increasingly sophisticated synthetic speech, and discriminating effectively among competitive artificial voices. Although the focus of our research was the MOS scale itself, the work proceeded during the emergence of commercial applications using synthetic speech, using voices created at IBM,

Nuance, SpeechWorks, AT&T, and other speech technology companies during the years 1998 to 2002. In many respects, the development of the MOS parallels the development of synthetic speech as it has become a viable and increasingly sophisticated technology.

Table 1 summarizes the focus of revisions and psychometric evaluation during each experiment, and provides an overview the organization of our experiments.

Table 1. Summary of MOS psychometric evaluations and revisions.

Study	Factor	Items	Item scale	Focus of psychometric evaluation			
				Reliability	Validity	Sensitivity	Factor structure
MOS (Salza et al., 1996) Experiment 1	Intelligibility	Global Impression	5-point scale	✓	✓	✓	✓
		Listening Effort					
		Comprehension Problems					
		Speech Sound Articulation					
		Pronunciation					
	Naturalness	Voice Pleasantness					
		Speaking Rate					
MOS-R Experiment 2	Intelligibility	Global Impression	7-point scale ^a	✓	✓	✓	
		Listening Effort					
		Comprehension Problems					
		Speech Sound Articulation					
		Pronunciation					
	Naturalness	Speaking Rate ^a					
		Global Impression ^a					
		Voice Pleasantness					
		Voice Naturalness ^a					
		Ease of Listening ^a					
MOS-R3 Experiment 3	Intelligibility	Listening Effort	7-point scale	✓	✓		✓
		Comprehension Problems					
		Speech Sound Articulation					
		Precision					
	Naturalness	Voice Pleasantness					
		Voice Naturalness					
		Humanlike Voice ^a					
		Voice Quality ^a					
	Social Impression ^a	Trust ^a					
		Confidence ^a					
	Fluency ^a	Emphasis ^a					
		Rhythm ^a					
		Intonation ^a					
	Voice ^a	Loudness ^a					
		Depression ^a					

(Continued on next page.)

Table 1. (Continued).

Study	Factor	Items	Item scale	Focus of psychometric evaluation			
				Reliability	Validity	Sensitivity	Factor structure
MOS-R4 Experiment 4	Intelligibility	Listening Effort	7-point scale	✓		✓	✓
		Comprehension Problems					
		Speech Sound Articulation					
		Precision					
	Naturalness	Voice Pleasantness					
		Voice Naturalness					
		Humanlike Voice					
		Voice Quality					
	Social Impression	Trust					
		Confidence					
		Enthusiasm ^a					
		Persuasiveness ^a					
	Negativity ^a	Depression ^a					
		Fear ^a					
MOS-X Experiments 5 and 6	Intelligibility	Listening Effort	7-point scale	✓		✓	✓
		Comprehension Problems					
		Speech Sound Articulation					
		Precision					
	Naturalness	Voice Pleasantness					
		Voice Naturalness					
		Humanlike Voice					
		Voice Quality					
	Prosody ^a	Emphasis ^a					
		Rhythm ^a					
		Intonation ^a					
	Social Impression	Trust					
		Confidence					
		Enthusiasm					
Persuasiveness							

^aIndicates a change in item or factor from the previous version of the MOS.

Our strategy for the research was to begin by evaluating the psychometric properties of the Salza et al. (1996) MOS, which appears in Appendix A. The next step was to develop a version of the MOS, the MOS-Revised (MOS-R), with improved psychometric properties for its traditional focus on intelligibility and naturalness (Experiments 1 and 2). The final step was to expand the content of the MOS-R to broaden its evaluative scope (Experiments 3 through 6) by generating new items and creating additional factors for the scale. The resulting scale is known as the MOS-Expanded (MOS-X) (see Appendix B). We provide an interpretive summary of

our statistical analysis of scale drafts as a rationale for our revisions in each experiment; please refer to the technical appendix for detailed statistics (Appendix C).

Experiment 1: The MOS Psychometric Evaluation

Initially, the research focused on determining the MOS's quality and identifying its factor structure. The specific goals of Experiment 1 were to evaluate the factor structure of the 7-item 5-point-scale version of the MOS (the version reported by Salza et al. (1996)

adapted for use in our lab—see Appendix A), estimate the reliability of the overall MOS score and any revealed factors, investigate the sensitivity of the MOS scores, and extend the work on the validity of the MOS.

Method

Factor Analysis and Reliability Evaluation. Over the period 1999 to 2001, we conducted a number of experiments in which participants completed the standard MOS. In some of these experiments, we also collected paired-comparison data and, in the most recent (Wang and Lewis, 2001), we collected intelligibility scores. Participants in these experiments have included in approximately equal numbers, males and females, persons older and younger than 40 years old, and IBM and non-IBM employees. The rated speech samples included concatenative and formant-synthesized voices and, in one case, a recorded human voice (non-professional speaker). Drawing from six of these experiments, we assembled a database of 73 independent completions of the Salza et al. (1996) MOS. This database was the source of data for a factor analysis, reliability assessment (both of the overall MOS and the factors identified in the factor analysis) and sensitivity investigation using analysis of variance.

Validity Evaluations. Re-analysis of our data from the previous studies in which listeners provided both MOS ratings and paired comparisons allowed us to replicate the finding of Salza et al. (1996) that MOS ratings correlate significantly with paired comparisons.

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

Results and Discussion

Factor Analysis. The factor analysis confirmed the previous results of Kraft and Portele (1995), and indicated that the MOS had two factors, Intelligibility and

Naturalness. One item (Speaking Rate) was not associated with either factor.

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

Reliability. Reliability (coefficient alpha) for the overall MOS was 0.89. Respective reliabilities for Intelligibility and Naturalness were 0.88 and 0.81. Using the minimum criterion of 0.70 (Landauer, 1988; Nunnally, 1978), all reliabilities were acceptable but could be improved.

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

Validity. Significant positive correlations between participants’ preference votes and MOS scores showed that the MOS was a valid measure of voices (overall MOS, $r(14) = .55$, $p = .03$; Naturalness factor, $r(14) = .46$, $p = .07$; Intelligibility factor, $r(14) = .49$, $p = .05$).

Additional data from Wang and Lewis (2001) indicated a marginally significant correlation between intelligibility scores from listener transcriptions and their MOS ratings of Intelligibility ($r(14) = .43$, $p = .10$), which is evidence for convergent validity. None of the other correlations between these scores and MOS

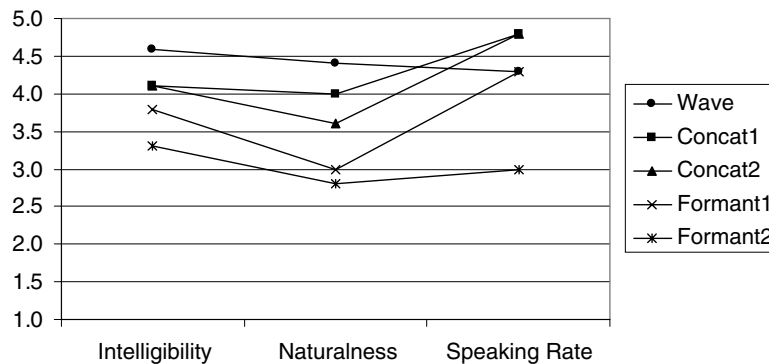


Figure 1. Voice by factor interaction (MOS).

ratings (Naturalness, Speaking Rate) were significant (all $p > .33$), which is evidence for divergent validity.

Sensitivity. Figure 1 illustrates the relationship among the five voices in our database and the MOS factors (plus Speaking Rate). Of these voices, the two formant voices had poorer ratings for Intelligibility and Naturalness, as compared to the two concatenative voices and a human (male, non-professional) voice. The Intelligibility, Naturalness, and Speaking Rate profile of each system was unique, suggesting that the MOS was sufficiently sensitive to discriminate among these voices.

Summary of Results. The version of the MOS derived from Salza et al. (1996) seemed to have reasonably good psychometric properties. The factor analysis of the current data resulted in a factor structure similar to that of Kraft and Portele (1995), specifically two factors (Intelligibility and Naturalness) and an unrelated item for Speaking Rate. The reliabilities of the overall MOS and its Intelligibility and Naturalness subscales were acceptable (greater than the minimal standard of 0.70). The evidence for sensitivity and validity was strong.

Revisions to the MOS. We made several changes to the MOS to improve its reliability. These changes included adding two items, Voice Naturalness and Ease of Listening, and increasing the number of scale steps from five to seven. In addition, we changed in the structure of the Speaking Rate item to make it more compatible with the other six items and allow us to determine if it is truly independent of the other two factors. We named this version of the questionnaire the MOS-Revised (MOS-R).

Experiment 2: The MOS-Revised (MOS-R) Psychometric Evaluation

The purpose of the Experiment 2 was to determine if the changes made in Experiment 1 worked as expected. Specifically, the expected consequences of the revision were:

- The reliability of the overall MOS would improve;
- The revised MOS items 2–5 would continue to form an Intelligibility factor; and
- Items 1 and 6–8 would form a Naturalness factor with substantially greater reliability (possibly in excess of .90) than the current Naturalness factor.

Additionally, the change in the structure of the Speaking Rate item would make it possible to determine whether that item is truly independent of the other two factors.

Method

The data came from three different studies of concatenative TTS voices, each with 16 participants. For all three studies, the participant groups had full factorial crossing of the following three independent variables (enhancing the generalizability of results):

- Gender (half were male, half were female),
- Age (half under 40 years of age, half over 40 years of age), and
- Employment (half were employees of a temporary help agency, half were IBM employees),

In all three studies, participants listened to four different TTS voices speaking four different text samples

and provided MOS-R ratings for each voice. We combined the data from the three independent experiments and performed a separate analysis for each set of ratings within participants across experiments for a total of four independent analyses. Each of these analyses included 48 sets of ratings.

Results and Discussion

Factor Analyses. Four separate (independent) factor analyses were conducted on the data, and three of these procedures indicated that the MOS-R continued to have two factors, Intelligibility and Naturalness. The analyses further indicated that two items, Global Impression and Speaking Rate, were somewhat problematic in that they did not consistently fall into either of the two main factors. However, three factor analyses showed that Global Impression aligned with both the Intelligibility and Naturalness factors and Speaking Rate aligned with Intelligibility.

Reliability. For the combined analysis, the reliability values (coefficient alpha) for the overall MOS-R scale, Intelligibility, and Naturalness were, respectively, .93, .89, and .95. These values show increases in reliability for the MOS-R scale and both of its factors as compared with the MOS.

Summary of Results. The intended changes did result in the expected improvement to the MOS scale and its two factors, Intelligibility and Naturalness. Combining the psychometric work done in Experiments 1 and 2, we documented the scale's factor structure, validity, and sensitivity, and also measured and improved the scale's reliability.

Revisions to the MOS-R. With continued development of synthetic speech and decreasing differences among the Intelligibility and Naturalness of synthetic voices, we decided to broaden the scope of characteristics evaluated by the MOS-R. The remaining experiments describe the complex, iterative tasks required to achieve this goal.

Experiment 3: The MOS-R Expansion and Psychometric Evaluation

The purpose of Experiment 3 was to add perceptual speech characteristics and social impression items not

previously measured by the MOS-R, creating a questionnaire named the MOS-R3. 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 (200 individuals in each of five groups) invited to participate in the study. Of this sample, 204 individuals completed the study questions (20% return rate).

Materials. 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
- C: concatenative male
- D: concatenative male
- E: formant male.

All voices had an 8 kHz sampling rate and 16-bit dynamic range. Voices A and B used the same underlying TTS technology and source voice. Voices C and D used different underlying TTS technologies (different from Voices A and B and different from each other).

The initial item set for the MOS-R3 included the nine items from the MOS-R and an additional item (Human-Like Voice) expected to align with Naturalness. The set also included eight new items based on clinical evaluation of human speech characteristics: voice (Loudness, Emphasis, Voice Quality, Pitch), fluency (Interruptions, Rhythm, Intonation), and articulation (Precision) (Shiple and McAfee, 1992). If the evaluation of human speech is similar to synthetic speech evaluation, we would predict that the three 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 new items related to the social impression created by human voices. These items were selected based on the review of previous literature and characteristics identified as relevant to application development (Topic Interest, Trust, Confidence, and

Depression). Thus, the initial item set for the MOS-R3 contained 22 items.

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 a link that caused the synthetic voice file to play on the participant's audio player application. They then completed the MOS-R3 items for that voice.

Results and Discussion

Due to an error with the data collection software, the data for Voice A was not collected and could not be used in the analysis.

Factor Analysis. The factor analysis revealed that the new scale included five factors (increased from two factors in the MOS-R). We named these factors Intelligibility, Naturalness, Fluency, Voice, and Social Impression. Somewhat problematic was the association of Voice Quality with Naturalness (instead of Voice) and Interruptions with Intelligibility (instead of Fluency). This result demonstrated that voice, fluency, and articulation may be problematic factor labels because of

their specificity. 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.

Reliability. After item deletions to create a final measure with 15 total items, four factors (Intelligibility, 0.91; Fluency, 0.88; Social Impression, 0.87; Naturalness, 0.89) and the Overall (0.93) score had coefficient alphas greater than 0.70, demonstrating reliabilities adequate for usability evaluation (Landauer, 1988). However, the Voice factor had inadequate reliability (0.46) based on this criterion.

Sensitivity. Again, the MOS-R revisions made in this experiment indicated that the scale continued to be sensitive to differences in synthetic voices, with Voices B and C rated most positively for all factors. Figure 2 shows that the new factors created unique profiles for four TTS voices, although the profiles of Voices D and E were similar.

Summary of Results. Although Experiment 3 was successful in broadening the scope of the MOS-R to include items related to human speech evaluation (Voice, Fluency) and social perception of speech (Social Impression), the resulting scale demonstrated several problems. Primarily, the Voice factor had weak

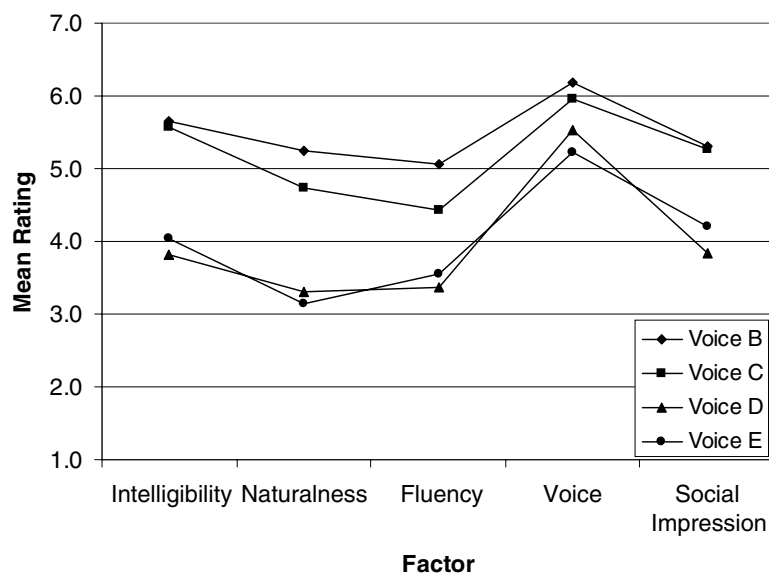


Figure 2. Voice by factor interaction (MOS-R3).

reliability and the Social Impression factor included too few items to be a valid measure of this complex variable.

Revisions to the MOS-R3. We targeted the addition of several new items that we expected to align with the Voice factor and improve its reliability. We also included items we expected would align with Social Impression.

Experiment 4: The MOS-R4 Expansion and Reliability Improvement

The purpose of Experiment 4 was to evaluate the effects of adding items to improve the reliability of the Voice factor and increase 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 Experiment 3), with 200 individuals in each of five groups invited to participate in the study. Of this sample, 138 individuals completed the study questions (14% return rate).

Materials. 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 Experiment 3. The technology used to produce Voice C2 was the same as that used to produce Voice C in Experiment 3, but the source voice for Voice C2 was different.

To create the initial item set for the MOS-R4, we retained the 15 items from the final version of the MOS-R3 (Listening Effort, Comprehension, Articulation, Pleasantness, Voice Naturalness, Humanlike Voice, Loudness, Emphasis, Voice Quality, Rhythm, Intonation, Precision, Trust, Confidence, and Depression). As in Experiment 3, 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 relied significantly less on the specific areas of human speech evaluation, making the items in this study qualitatively different than those explored in Experiment 3. Therefore, one possible outcome of Experiment 4 was that we would not retain the Voice and Fluency factors.

Procedure. The procedure was identical to that of Experiment 3.

Results and Discussion

Factor Analysis. As in Experiment 3, we again found that the scale included five factors, three of which were the familiar Intelligibility, Naturalness, and Social Impression. Interestingly, Factor 4 included a single item (Loudness) from the earlier Voice factor, and Factor 5 included two Negativity items (Depression, Fear). This alignment was unexpected, based on our goal of generating additional items that would align with the Voice factor.

Because only one item associated with Factor 4 (Voice), we omitted Loudness and performed a second factor analysis, forcing a four-factor solution (eliminating the Voice factor). The four-factor model appeared to be more consistent with the results of Experiment 3 and the theoretical association of items in the literature, and included more than one item per factor (but the Negativity factor only included two items). Therefore, this analysis indicated that the MOS-R4 contained the four factors Intelligibility, Naturalness, Negativity, and Social Impression.

Reliability. We again adjusted items to create an efficient, but reliable scale. Following the manipulation, three factors (Intelligibility, 0.84; Social Impression, 0.84; Naturalness, 0.85) and the Overall score (0.89) demonstrated adequate reliability above 0.70 (Landauer, 1988). The reliability of the Negativity factor (0.65) was below this criterion.

Sensitivity. A mixed model ANOVA indicated the extent to which the final version of the MOS-R4 discriminated among the five synthetic voices. Figure 3

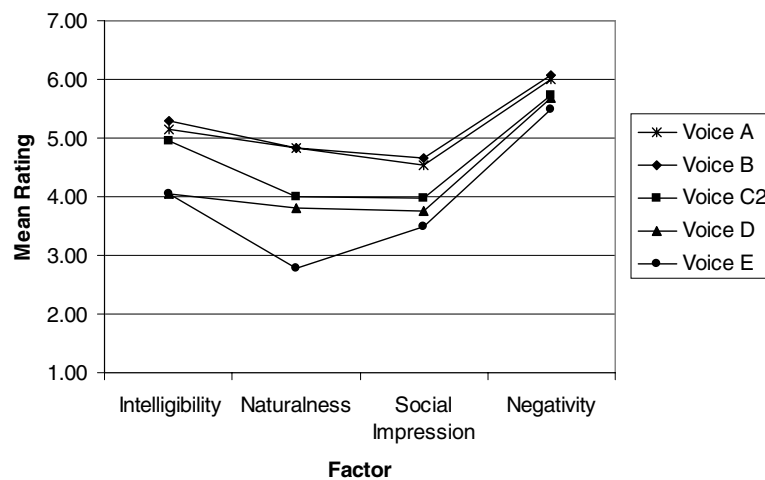


Figure 3. Voice by factor interaction (MOS-R4).

illustrates the similarity between Voices A and B (as expected because they used the same core TTS technology and the same source voice). 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 (a low-quality concatenative voice). Of the four factors, only Negativity seemed to be relatively insensitive to the differences among the voices.

Summary of Results. The outcomes of Experiments 3 and 4 were encouraging, but not completely satisfying. The addition of the new items in each experiment led to the emergence of new factors (Fluency, Voice, and Social Impression in Experiment 3; Negativity and Social Impression in Experiment 4). In Experiment 3, 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 Experiment 4, the final MOS-R4 had four items supporting the Social Impression factor, but the Negativity factor only had two items and low reliability and sensitivity.

The primary goal of Experiments 3 and 4 was to expand the coverage of the MOS to include new factors that have become 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-R3 included items that contribute to prosody (Emphasis, Rhythm, Intonation, Interruptions), yet these items did not clearly align in a single factor. In Experiment 4, the stronger loadings of Social Impression items (likely due to the larger effect sizes of social impression as compared with vocal speech perceptions) resulted our deletion of all items that could be related to prosody (Emphasis, Voice Quality, Rhythm, Intonation, Monotone Quality), as required to create an efficient, yet reliable scale. Recently, researchers have begun to acknowledge that prosodic qualities are vital for acceptable synthetic speech and to develop algorithms that approximate human prosody (Portele and Heuft, 1997; Sonntag and Portele, 1998). In addition to our goals to measure social aspects and prosody of synthetic speech, we required that each factor have acceptable reliability and the support of at least three items.

Revisions to the MOS-R4. We targeted an analysis of the combined data from Experiments 3 and 4 for our next revisions to the MOS. As a consequence of the iterative evaluation process of Experiments 3 and 4, the complete item sets for the studies had 14 items in common. The common items were four items associated with Intelligibility, four items associated with Naturalness, three items associated with Fluency, two items associated with Social Impression, and the Depression item (associated with the Voice factor in the MOS-R3 and the Negativity factor in the MOS-R4).

Because these items were common across both studies, the sample size for their psychometric evaluation

was the sum of the sample sizes for Experiments 3 and 4 (342 complete and independent sets of responses). The factor analyses of Experiments 3 and 4 strongly suggested that the Intelligibility, Naturalness, and Fluency factors would remain intact in an analysis of the combined data. It also seemed likely that the two items associated with Social Impression in the MOS-R3 and MOS-R4 would continue to align. The expected behavior of the Depression item was harder to predict. If it aligned with the Social Impression factor and the Social Impression factor's reliability exceeded 0.70, then this version of the MOS would meet the initial goals of our research program, producing an Expanded MOS (MOS-X).

Experiment 5: The Initial MOS-Expanded (MOS-X)

The goal of this experiment was to complete a psychometric analysis of the combined data from Experiments 3 and 4, with the purpose of uncovering a stable, logical, and theoretically-supported factor structure for the new scale.

Method

To perform this analysis, we created a new database from the results of Experiments 3 and 4. The 14 items included the items 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. The analysis indicated that the new scale contained four factors, which we named Intelligibility and Naturalness (the consistent factors throughout our research), Prosody, and Social Impression. 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. 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 Experiments 3 and 4. At the same time, the Depression item was a concern, since it was not as clearly a part of Social Impression as the other items in this factor.

Reliability. Coefficient alpha for each factor indicated acceptable reliability (Overall: .92, Intelligibility: .88, Naturalness: .87, Prosody: .85, Social Impression: .71), although Social Impression was a clear candidate for improvement.

Sensitivity. Statistical analysis confirmed that the new scale effectively discriminated among the six synthetic voices used in Experiments 3 and 4. Figure 4 shows 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 Impression of Voice E were identical to that of concatenative Voice D (a low-quality concatenative voice). All four factors seemed to be reasonably sensitive to the differences among the voices.

Summary of Results. The combined analysis of data from Experiments 3 and 4 produced a four-factor scale, but a single item (Depression) had a somewhat tenuous association with the Social Impression factor. The validity and sensitivity of the scale seemed adequate, but the reliability was relatively low for the Social Impression factor. To perform this combined analysis, we had to remove several items from the scale that did not appear in both of the previous experiments (Enthusiasm and Persuasiveness), and these items may have been better contributors to the Social Impression factor.

Revision to the Initial MOS-X. We proposed a final new study to see if adding Enthusiasm and/or Persuasiveness items would improve the reliability of Social Impression.

Experiment 6: Final MOS-X Psychometric Evaluation

We undertook a final study to determine if we could improve the reliability of Social Impression by adding Enthusiasm and Persuasiveness to the MOS-X scale.

Method

Participants. The sample included complete sets of ratings from 327 randomly selected IBM employees.

Materials. The study used a between-subjects design with ten levels of the independent variable of synthetic

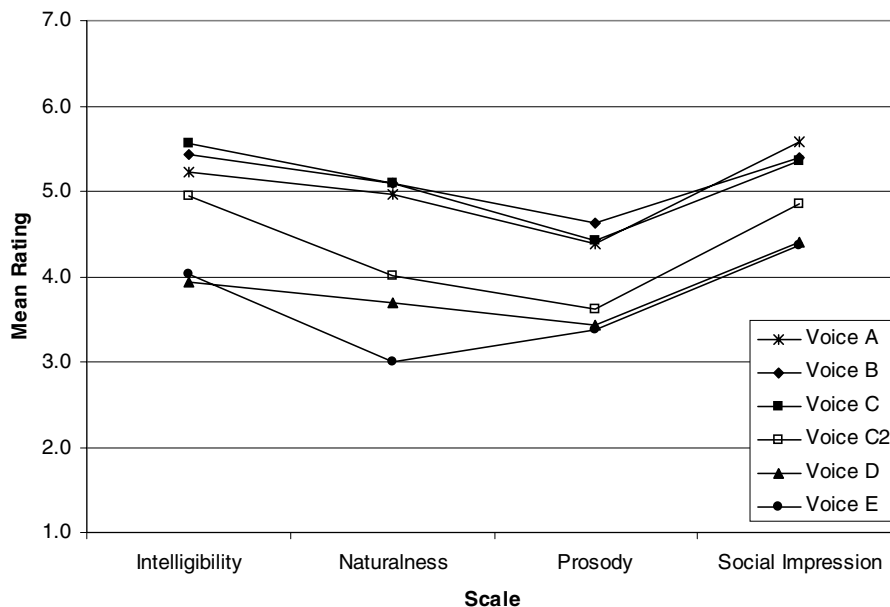


Figure 4. Voice by factor interaction (initial MOS-X).

voice. The voices and their key characteristics were:

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

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

Procedure. Participants received an email inviting them to participate in the study and directing them to a web page (one page for each participant group) with instructions, a link to a recording of one of the synthetic voices, and the rating scales. After accessing the web page, participants clicked the link that caused the

synthetic voice file to play on the participant's audio player application. They then completed the 16 items for that voice.

Results and Discussion

Factor Analysis. For the factor analysis, we forced a 4-factor solution due to the results of a previous study (Experiment 3), showing that Enthusiasm and Persuasiveness loaded on Social Impression. As predicted, the factor analysis showed that both Enthusiasm and Persuasiveness loaded on Social Impression. This final analysis confirmed that the MOS-X had four clear factors: Intelligibility, Naturalness, Prosody, and Social Impression.

Reliability. We removed Depression to create a 15-item scale and calculated reliability statistics. Intelligibility (0.88), Naturalness (0.86), Prosody (0.86), Social Impression (0.86), and the Overall score (0.93) had coefficient alphas much greater than 0.70, demonstrating reliabilities adequate for usability evaluation (Landauer, 1988).

Sensitivity. The statistical analysis suggested that the factor profiles for the voices were significantly different, as shown in Fig. 5. Generally, participants rated

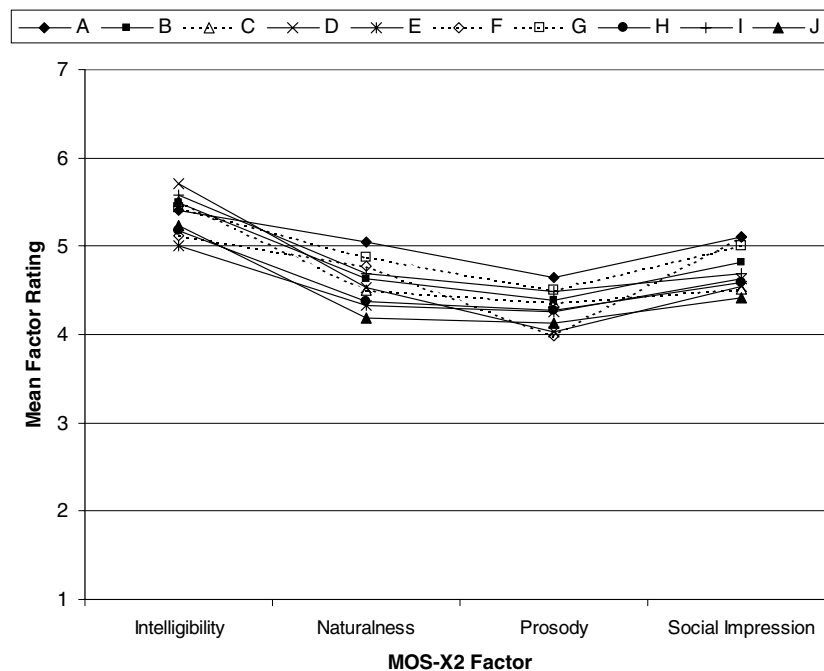


Figure 5. Voice by factor interaction (final MOS-X).

Intelligibility most positively and Prosody most poorly of the four factors.

Summary of Results. This final study showed that this version of the MOS-X achieved a theoretically-derived, psychometrically sound factor structure, and it is a reliable and sensitive instrument. With these results, we achieved the goal of revising and expanding the MOS for use in applied measurement of synthetic speech.

General Discussion

Psychometric evaluations of the original MOS provided some indication of its strength for evaluating synthetic speech, but this research had not been systematic nor repeated to ensure that the scale was reliable, valid, and sensitive for applied measurement. Although the MOS's psychometric properties were acceptable, Experiment 2 showed that it was possible to improve its psychometric properties, leading to the development of the Revised MOS (MOS-R). Additional studies expanded and validated the instrument, resulting in a more comprehensive measure of synthetic speech, the MOS-X.

The final version of the MOS-X contained two important advancements over the initial MOS-R. First, Experiments 3 and 4 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 Intelligibility and Naturalness factors from the MOS and MOS-R.

Experiments 3 and 4 also resolved several 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 and Portele, 1995; Lewis, 2001a,b), loaded on the Fluency factor in Experiment 3 (MOS-R3). We later excluded Speaking Rate from the MOS-R3 without significant loss of reliability. The Humanlike Voice item loaded strongly on the Naturalness factor, and we retained it through the MOS-R3 and MOS-R4 into the MOS-X. Finally, the Global Impression item

consistently loaded on more than one factor, although its strongest loading tended to be on the Intelligibility factor. We removed this item during the efficiency phase of Experiment 3 and found that reliability 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-R3 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 and Arnot, 1993). This associative pattern partially prompted the Voice factor name in Experiment 3. However, when we removed Pitch in Experiment 4, the Loudness item became a separate factor and Depression associated with the Social Impression factor.

The elusive Voice factor was also apparent in Experiment 4. In this version of the MOS, Fear and Depression aligned in a factor we named Negativity. Both of these items elicit listeners' perceptions of negative emotion, which distinguishes them from the items associated with the social-personality inferences elicited by other 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 and 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-R4, although covertly.

A second observation concerns the type of items that we removed from the MOS-R in Experiments 3 and 4. 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 and Arnott, 1993).

The final items included in the MOS-X resulted in a blend of the factors present in the MOS-R of Experiments 3 and 4. 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 because both types of ratings point to acoustic modifications that may be made in a voice, and indicate the social attributions listeners make about a voice they hear in applications that use synthetic speech. The final evaluation in Experiment 6 confirmed that our final version is an instrument with strong psychometric properties, making it well-suited for high-quality evaluations in industrial settings.

In general, this body of research has met our initial goal of strengthening the MOS for speech system evaluation. The MOS-R measures the same factors as the standard MOS, but with improved psychometric properties. The MOS-X broadened the range of variables this instrument measures reliably and is sensitive enough to detect key differences among a set of artificial voices. Munsterberg asserted in 1913 "the whole world of industry will have to learn the great lesson, that of the three great factors, material, machine, and man, the man is not the least, but the most important" (p. 593). In the world of speech technology, reliable, valid, and sensitive evaluation of listener perception is vital to the usability of applications that use synthetic speech.

Appendix A: The Standard MOS (Salza et al., 1996)

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

- Excellent
- Good
- Fair
- Poor
- Bad

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

- No effort required
- Slight effort required
- Effort required
- Major effort required
- Message not understood with any feasible effort

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

- None
- Few
- Some
- Many
- Every word

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

- Yes, very clearly
- Yes, clearly enough
- Fairly clear
- No, not very clear
- No, not at all

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

- No
- Yes, but not annoying
- Yes, slightly annoying
- Yes, annoying
- Yes, very annoying

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

- Yes
- Yes, but slower than preferred
- Yes, but faster than preferred
- No, too slow
- No, too fast

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

- Very pleasant
- Pleasant
- Fair
- Unpleasant
- Very unpleasant

Appendix B: The MOS-X (Final Version)

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

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

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

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

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

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

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

SLURRED OR									PRECISE
IMPRECISE	1	2	3	4	5	6	7		

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

VERY									VERY
UNPLEASANT	1	2	3	4	5	6	7	PLEASANT	

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

VERY									VERY
UNNATURAL	1	2	3	4	5	6	7	NATURAL	

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

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

8. <i>Voice Quality</i> : Did the voice sound harsh, raspy, or strained?									
SIGNIFICANTLY									NORMAL
HARSH/RASPY	1	2	3	4	5	6	7		QUALITY
9. <i>Emphasis</i> : Did emphasis of important words occur?									
INCORRECT									EXCELLENT USE
EMPHASIS	1	2	3	4	5	6	7		OF EMPHASIS
10. <i>Rhythm</i> : Did the rhythm of the speech sound natural?									
UNNATURAL OR									NATURAL
MECHANICAL	1	2	3	4	5	6	7		RHYTHM
11. <i>Intonation</i> : Did the intonation pattern of sentences sound smooth and natural?									
ABRUPT OR									SMOOTH OR
ABNORMAL	1	2	3	4	5	6	7		NORMAL
12. <i>Trust</i> : Did the voice appear to be trustworthy?									
NOT AT ALL									VERY
TRUSTWORTHY	1	2	3	4	5	6	7		TRUSTWORTHY
13. <i>Confidence</i> : Did the voice suggest a confident speaker?									
NOT AT ALL									VERY
CONFIDENT	1	2	3	4	5	6	7		CONFIDENT
14. <i>Enthusiasm</i> : Did the voice seem to be enthusiastic?									
NOT AT ALL									VERY
ENTHUSIASTIC	1	2	3	4	5	6	7		ENTHUSIASTIC
15. <i>Persuasiveness</i> : Was the voice persuasive?									
NOT AT ALL									VERY
PERSUASIVE	1	2	3	4	5	6	7		PERSUASIVE

MOS-X Scales

Overall: Average items 1–15

Intelligibility: Average items 1–4

Naturalness: Average items 5–8

Prosody: Average items 9–11

Social Impression: Average items 12–15

Appendix C: Statistical Appendix

This appendix provides additional detail on the statistics (factor loadings and ANOVA results) summarized in the Results and Discussion sections of each Experiment. Factor loadings greater than 0.50 (shown in bold) were considered aligned with the factor named in column headings.

Experiment 1

Table 2. MOS factor loading.

Item	Content	Intelligibility	Naturalness	Undefined
1	Global Impression	0.327	0.900	0.194
2	Listening Effort	0.629	0.370	0.427
3	Comprehension Problems	0.693	0.104	0.358
4	Speech Sound Articulation	0.672	0.433	0.294
5	Pronunciation	0.746	0.437	0.139
6	Speaking Rate	0.322	0.204	0.754
7	Voice Pleasantness	0.182	0.665	0.139

Sensitivity. A mixed-factors ANOVA indicated a significant main effect of System ($F(4, 68) = 9.6, p = .000003$), a significant main effect of MOS Factor ($F(2, 136) = 14.7, p = .000002$), and a significant System by Factor interaction ($F(8, 136) = 3.1, p = .003$).

Experiment 2

Table 3. MOS-R factor loading.

Item	Content	Intelligibility	Naturalness
1	Global Impression	0.56	0.51
2	Listening Effort	0.78	0.36
3	Comprehension Problems	0.85	0.23
4	Speech Sound Articulation	0.70	0.39
5	Pronunciation	0.57	0.48
6	Voice Pleasantness	0.32	0.88
7	Voice Naturalness	0.40	0.83
8	Ease of Listening	0.38	0.83
9	Speaking Rate	0.53	0.32

Experiment 3

Table 4. MOS-R3 factor loading.

Item	Content	Factor 1: Intelligibility	Factor 2: Fluency	Factor 3: Voice	Factor 4: Social impression	Factor 5: Naturalness
1	Global Impression ^a	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 ^a	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 ^a	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 ^a	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 ^a	0.516	0.306	-0.171	0.429	-0.047
15	Pitch ^a	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 ^a	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

^aItem removed to improve reliability and/or efficiency of scale.

Sensitivity. A mixed model ANOVA indicated the extent to which the MOS-R3 discriminated among the four synthetic voices. The ANOVA showed a main effect of synthetic voice ($F(3, 154) = 26.92, p < 0.0001$), factor ($F(4, 616) = 79.03, p < 0.0001$), and a significant interaction between these variables ($F(12, 616) = 4.56, p < 0.0001$, shown in Fig. 2).

Experiment 4

Table 5. MOS-R4 factor loading.

Item	Content	Factor 1: Social impression	Factor 2: Intelligibility	Factor 3: Negativity	Factor 4: 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 ^a	0.565	0.403	-0.221	0.173
9	Voice Quality	0.320	0.375	0.015	0.618
10	Rhythm ^a	0.406	0.406	0.082	0.536
11	Intonation ^a	0.325	0.478	0.019	0.505
12	Monotone Quality ^a	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 ^a	0.599	0.174	0.082	0.457
19	Persuasiveness	0.691	0.249	0.002	0.351
20	Impatience ^a	0.451	0.156	0.277	0.522
21	Fear	0.029	0.138	0.835	0.226

^aItem removed to improve reliability and/or efficiency of scale.

Sensitivity. The ANOVA showed a main effect of synthetic voice ($F(4, 124) = 9.18, p < 0.0001$), factor ($F(3, 372) = 101.92, p < 0.0001$), and a significant interaction between these variables ($F(12, 372) = 2.70, p = 0.002$).

Experiment 5

Table 6. Initial MOS-X factor loading.

Item	Content	Factor 1: Prosody	Factor 2: Intelligibility	Factor 3: Social impression	Factor 4: 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

(Continued on next page.)

Table 6. (Continued).

Item	Content	Factor 1: Prosody	Factor 2: Intelligibility	Factor 3: Social impression	Factor 4: Naturalness
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

Sensitivity. A mixed model ANOVA indicated the extent to which the MOS-X discriminated among the six different synthetic voices used in Experiments 3 and 4. The ANOVA showed a main effect of synthetic voice ($F(5, 275) = 27.5, p < 0.0000001$), factor ($F(3, 825) = 58.5, p < 0.0000001$), and a significant interaction between these variables ($F(15, 825) = 3.8, p = 0.000001$).

Experiment 6

Table 7. Final MOS-X factor loading.

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

^aItem removed to improve reliability and/or efficiency of scale.

Sensitivity. A mixed model ANOVA indicated the extent to which the MOS-X discriminated among the ten synthetic voices. The ANOVA showed a significant effect of factor ($F(3, 951) = 443.96, p < 0.0001$), and a significant interaction between factor and voice ($F(27, 951) = 2.37, p < 0.0001$). The main effect of synthetic voice was not significant ($F(9, 317) = 1.01, p = 0.44$), indicating a similar mean rating for the ten voices. The overall mean scores for the voices ranged from 4.50 (Voice J, least positive) to 5.10 (Voice A, most positive).

References

- 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*, 16:41–45.
- Bloom, K., Zajac, D., and Titus, J. (1999). The influence of nasality of voice on sex-stereotyped perceptions. *Journal of Nonverbal Behavior*, 23:271–281.
- Bradlow, A., Torretta, G., and Pisoni, D. (1996). Intelligibility of normal speech I: Global and fine-grained acoustic-phonetic talker characteristics. *Speech Communication*, 20:255–272.
- Brown, B., Strong, W., and Rencher, A. (1973). Perceptions of personality from speech: Effects of manipulations of acoustical parameters. *Journal of the Acoustical Society of America*, 54:29–35.
- Brown, B., Strong, W., and Rencher, A. (1975). Acoustic determinants of perceptions of personality from speech. *International Journal of the Sociology of Language*, 6:1–32.
- Cliff, N. (1987). *Analyzing Multivariate Data*. San Diego, CA: Harcourt Brace Jovanovich.
- Coovert, M.D. and 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., and Scherer, K. (1991). Face, voice, and body in detecting deceit. *Journal of Nonverbal Behavior*, 15:125–135.
- Francis, A.L. and Nusbaum, H.C. (1999). Evaluating the quality of synthetic speech. In D. Gardner-Bonneau (Ed.), *Human Factors and Voice Interactive Systems*. Boston, MA: Kluwer, pp. 63–97.
- Goldstein, M. (1995). Classification of methods used for assessment of text-to-speech systems according to the demands placed on the listener. *Speech Communication*, 16:225–244.
- Gorenflo, D. and Gorenflo, C. (1997). Effects of synthetic speech, gender, and perceived similarity on attitudes toward the augmented communicator. *AAC: Augmentative and Alternative Communication*, 13:87–91.
- Granstrom, B. and Nord, L. (1992). Neglected dimensions in speech synthesis. *Speech Communication*, 11:459–462.
- Greene, B., Logan, J., and 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. and 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. and Minifie, F. (1999). Acoustical-perceptual correlates of 'whisper pitch' in synthetically generated vowels. *Journal of Speech, Language, and Hearing Research*, 42:583–591.
- Hillenbrand, J. (1988). Perception of aperiodicities in synthetically generated voices. *Journal of the Acoustical Society of America*, 83:2361–2371.
- Hoag, L. and 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*, 35:1363–1366.
- Holtgraves, T. and Lasky, B. (1999). Linguistic power and persuasion. *Journal of Language and Social Psychology*, 18:196–205.
- Hosman, L. (1989). The evaluative consequences of hedges, hesitations, and intensifiers: Powerful and powerless speech styles. *Human Communication Research*, 15:383–406.
- 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 synthesizers. *BT Technology Journal*, 14:100–111.
- Johnson, W., Emde, R., Scherer, K., and Klinnert, M. (1986). Recognition of emotion from vocal cues. *Archives of General Psychiatry*, 43:280–283.
- Klatt, D. and Klatt, L. (1990). Analysis, synthesis, and perception of voice quality variations among female and male talkers. *Journal of the Acoustical Society of America*, 87:820–857.
- Koopmans-Van Beinum, F. (1992). The role of focus words in natural and in synthetic continuous speech: Acoustic aspects. *Speech Communication*, 11:439–452.
- Kraft, V. and 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., and Rosenhouse, J. (2000). The effects of acoustic modifications on the identification of familiar voices speaking isolated vowels. *Speech Communication*, 30:9–26.
- 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*. Mahwah, NJ: Lawrence Erlbaum, pp. 149–153.
- 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. and Haroldson, S. (1992). Stuttering and speech naturalness: Audio and audiovisual judgments. *Journal of Speech and Hearing Research*, 35:521–528.
- Massaro, D. and Egan, P. (1996). Perceiving affect from the voice and the face. *Psychonomic Bulletin & Review*, 3:215–221.
- Miyake, K. and 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*, 61:411–437.
- Moller, S., Jekosch, U., Mersdorf, J., and Kraft, V. (2001). Auditory assessment of synthesized speech in application scenarios: Two case studies. *Speech Communication*, 34:229–246.
- Munsterburg, H. (1913). Psychology and industrial efficiency. In L.T. Benjamin Jr. (Ed.), *A History of Psychology: Original Sources and Contemporary Research*, 2nd edn. Boston: McGrawHill, pp. 584–593.
- Murray, I. and 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*, 93:1097–1108.
- Murray, I. and 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., and Rohwer, E. (1996). Emotional stress in synthetic speech: Progress and future directions. *Speech Communication, 20*:85–91.
- Nunnally, J.C. (1978). *Psychometric Theory*. New York: McGraw-Hill.
- Paddock, J. and 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*:29–44.
- Page, R. and Balloun, J. (1978). The effect of voice volume on the perception of personality. *Journal of Social Psychology, 105*:65–72.
- Paris, C.R., Thomas, M.H., Gilson, R.D., and Kincaid, J.P. (2000). Linguistic cues and memory for synthetic and natural speech. *Human Factors, 42*:421–431.
- Pelachaud, C., Badler, N., and Steedman, M. (1996). Generating facial expressions for speech. *Cognitive Science, 20*:1–46.
- Pisoni, D. (1997). Perception of synthetic speech. In J. van Santen, R. Sproat, J. Olive, and J. Hirschberg (Eds.), *Progress in Speech Synthesis*. New York: Springer, pp. 541–560.
- Pols, L. and Jekosch, U. (1997). A structured way of looking at the performance of text-to-speech systems. In J. van Santen, R. Sproat, J. Olive, and J. Hirschberg (Eds.), *Progress in Speech Synthesis*. New York: Springer, pp. 519–528.
- Portele, T. and Heuft, B. (1997). Toward a prominence-based synthesis system. *Speech Communication, 21*:61–72.
- Salza, P.L., Foti, E., Nebbia, L., and Oreglia, M. (1996). MOS and pair comparison combined methods for quality evaluation of text to speech systems. *Acta Acustica, 82*:650–656.
- Schmidt-Nielsen, A. (1995). Intelligibility and acceptability testing for speech technology. In A. Syrdal, R. Bennett, and S. Greenspan (Eds.), *Applied Speech Technology*. Boca Raton: CRC Press.
- Shiple, K. and McAfee, J. (1992). *Assessment in Speech Language Pathology: A Resource Manual*. San Diego: Singular.
- Sonntag, G.P. and Portele, T. (1998). PURR—A method for prosody evaluation and investigation. *Computer Speech and Language, 12*:437–451.
- Sonntag, G.P., Portele, T., Haas, F., and Kohler, J. (1999). Comparative evaluation of six German TTS systems. *Eurospeech '99*. Budapest: Technical University of Budapest, pp. 251–254.
- Slowiaczek, L. and Nusbaum, H. (1985). Effects of speech rate and pitch contour on the perception of synthetic speech. *Human Factors, 27*:701–712.
- Stern, S., Mullennix, J., Dyson, C., and Wilson, S. (1999). The persuasiveness of synthetic speech versus human speech. *Human Factors, 41*:588–595.
- Tartter, V. and Braun, D. (1994). Hearing smiles and frowns in normal and whisper registers. *Journal of the Acoustical Society of America, 96*:2101–2107.
- van Bezooijen, R. and 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*. New York, NY: Mouton de Gruyter.
- Wang, H. and Lewis, J.R. (2001). Intelligibility and acceptability of short phrases generated by embedded text-to-speech engines. In *Proceedings of HCI International 2001: Usability Evaluation and Interface Design*. Mahwah, NJ: Lawrence Erlbaum, pp. 144–148.
- Whalen, D. and Hoequist, C. (1995). The effects of breath sounds on the perception of synthetic speech. *Journal of the Acoustical Society of America, 97*:3147–3153.
- Whitmore, J. and Fisher, S. (1996). Speech during sustained operations. *Speech Communication, 20*:55–70.
- Yabuoka, H., Nakayama, T., Kitabayashi, Y., and 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.
- van Riper, C. and Emerick, L. (1990). *Speech Correction*. Englewood Cliffs, NJ: Prentice Hall.
- Yaeger-Dror, M. (1996). Register as a variable in prosodic analysis: The case of the English negative. *Speech Communication, 19*:39–60.
- Zuckerman, M., Miyake, K., and Hodgins, H. (1991). Cross-channel effects of vocal and physical attractiveness and their implications for interpersonal perception. *Journal of Personality and Social Psychology, 60*:545–554.