Effect of Voice and Bandwidth on MOS-X Ratings

TR 29.3550
August 20, 2002

James R. Lewis

IBM Voice Systems
Boca Raton, Florida

Abstract

I used a subset of data from an earlier study of concatenative voices to evaluate the effect of Speaker (AF, AM, and B) and Bandwidth (8 kHz vs. 22 kHz) on MOS ratings. Six independent groups of raters participated, one group for each combination of speaker and bandwidth. Analyses of variance indicated a significant main effect of Voice, but no significant main effect of Bandwidth and no significant Voice by Bandwidth interaction. The results indicate that independent groups of raters are sensitive to differences in speakers, but not to differences in bandwidth.

ITIRC Keywords

Mean Opinion Scale-Extended
MOS-X
Artificial speech
Competitive evaluation
Synthetic speech
Speech bandwidth
Concatenative text-to-speech (TTS)

Contents

INTRODUCTION **1**

METHOD **3**

PARTICIPANTS **3**
STIMULI **3**
PROCEDURE **3**

RESULTS **5**

OVERALL RATING **5**

RATINGS BY SCALES **5**
RATINGS BY ITEMS **8**
SEPARATE COMPARISONS OF BANDWIDTH BY SPEAKER **11**

DISCUSSION **13**

REFERENCES **15**

APPENDIX A. THE MOS-X **17**

APPENDIX B. THE TEST TEXT **19**

Introduction

The primary purpose of this evaluation was to investigate listener sensitivity to the differences between concatenative voices with bandwidths of 22 and 8 kHz. Previous research investigating the psychometric properties of the MOS-X (Polkosky & Lewis, 2002a, 2002b), a questionnaire used to assess the quality of artificial speech, has demonstrated that listeners are sensitive to differences in concatenative voices due to differences in the human speaker used as a source for developing the voice. To date, however, there has been no comparable evaluation of listener sensitivity to differences in bandwidth.

Higher bandwidth leads to greater audio fidelity, but at a cost of increased requirements for storage and transmission capacity. For example, a bandwidth of 22 kHz is appropriate for use in desktop systems, but bandwidths greater then 8 kHz are not suitable for transmission over standard phone lines. The main concern in our lab, however, was the extent to which bandwidth might affect listener ratings of voice quality. There have been times in the past when the only speech samples available for comparison in competitive evaluations have had different bandwidths, and this situation is likely to occur again in the future. If bandwidth significantly affects independent listener ratings of speech quality, then we would know that we should not conduct studies in the future that compare samples that differ in bandwidth. If bandwidth has little or no effect on listener ratings, then we could confidently conduct future comparative studies in which the speech samples differed in bandwidth. Fortunately, results from previous evaluations in our lab provide the data necessary to investigate this effect. Specifically, we have data for three different speakers produced at both 22 and 8 kHz (for a total of six artificial voices).

Method

Participants

The total number of respondents who completed the MOS-X for each voice were:

- AF (22 kHz): 44
- AM (22 kHz): 41
- B (22 kHz): 36
- AF (8 kHz): 34
- AM (8 kHz): 20
- B (8 kHz): 66

Stimuli

The stimuli for each group was an audio file of a synthetic voice speaking texts used in previous evaluations of synthetic voices at IBM Voice Systems (Lewis & Polkosky, 2001; Polkosky & Lewis, 2001). The speakers were:

- AF – a female speaker
- AM – a male speaker
- B – a second male speaker

There were two versions for each speaker, one with a bandwidth of 22 kHz and one with a bandwidth of 8 kHz, for a total of six artificial voices. After listening to their assigned voice, participants completed a set of seven-point bipolar rating scales that included the 14 MOS-X[1] items (shown in Appendix A). Appendix B contains the test text (which takes about a minute to play after conversion to artificial speech).

Procedure

In previous studies, participants received an email inviting them to participate in the study and directing them to a web page containing the instructions, a link to one of the synthetic voices (one web page for each participant group), and the rating scales. After accessing the web page, participants clicked on a link that caused the synthetic voice file to play on the participant's audio player application (on a desktop computer system). They then completed the MOS items for that voice.

[1] This was a preliminary version of the MOS-X. The final version of the MOS-X is available in Polkosky and Lewis (2002b).

Results

Overall Rating

A one-way analysis of variance on all six voices using the overall MOS rating indicated no main effect of Speaker ($F(2, 235) = 0.41$, $p = .66$), no main effect of Bandwidth ($F(1, 235) = 0.43$, $p = .52$), and no Speaker by Bandwidth interaction ($F(2, 235) = 0.68$, $p = .51$).

Ratings by Scales

A mixed-model analysis of variance, with Speaker and Bandwidth as between-subjects independent variables and Scale as a within-subjects independent variable, indicated a significant main effect of Scale ($F(3, 705) = 93.8$, $p < .0001$) and a significant interaction between Speaker and Scale ($F(6, 705) = 4.2$, $p < .0001$). The interaction between Bandwidth and Scale and the interaction among Bandwidth, Scale, and Speaker were both nonsignificant ($F(3, 705) = 0.1$, $p = .96$, and $F(6, 705) = 1.2$, $p = .33$, respectively). The results for each voice appear in Table 1. Figures 1 and 2 show the Speaker by Scale and Bandwidth by Scale interactions, respectively.

Table 1. Mean Ratings by MOS-X Scales

Voice	Intelligibility	Naturalness	Prosody	Social Impression
AF 22 kHz	5.1	4.9	4.0	5.5
AF 8 kHz	5.3	5.1	4.5	5.6
AM 22 kHz	5.4	4.9	4.6	5.5
AM 8 kHz	5.5	4.7	4.3	5.4
B 22 kHz	5.5	4.5	4.3	5.0
B 8 kHz	5.6	4.9	4.2	5.2

Figure 1. Speaker by Scale Interaction (significant)

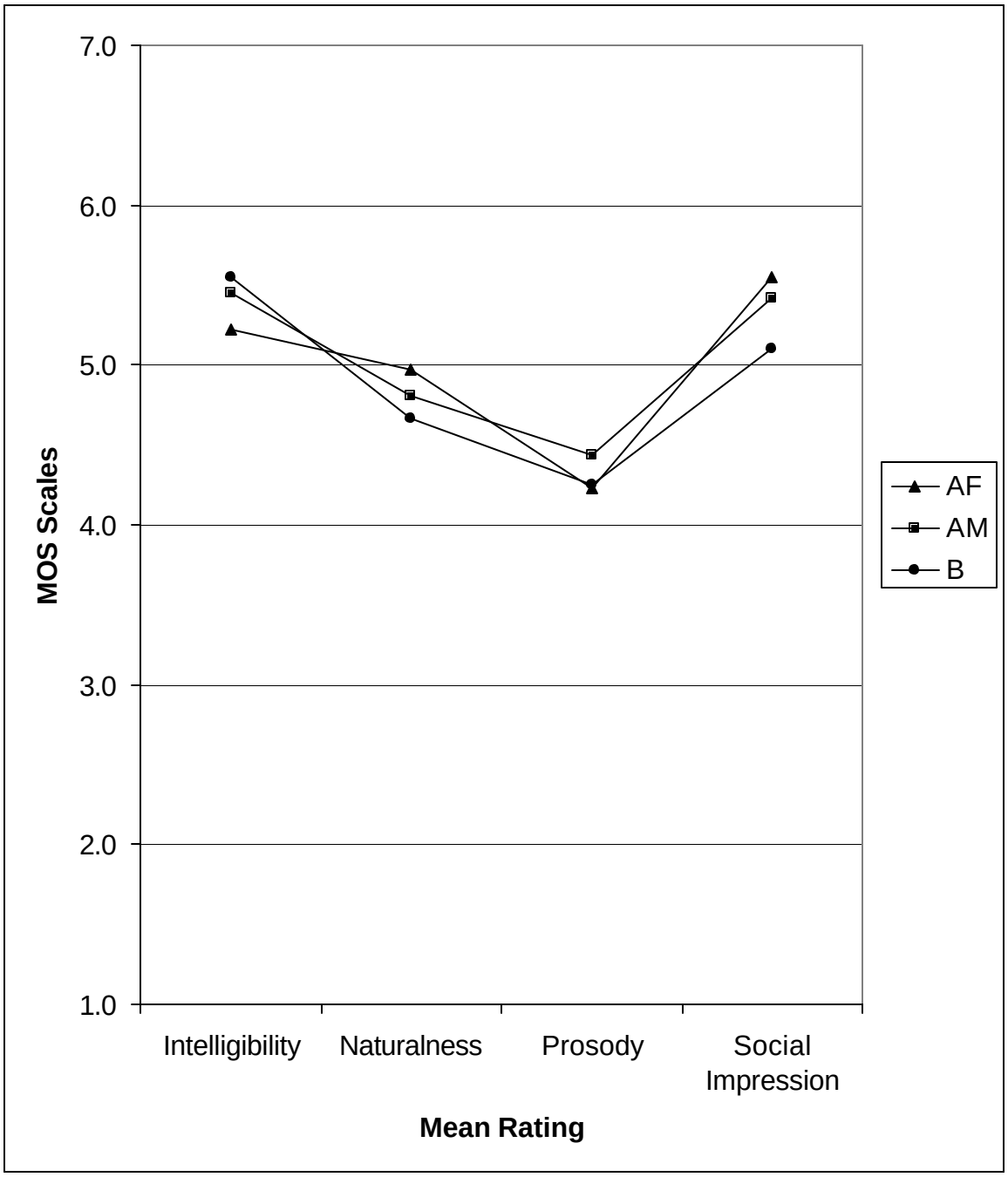
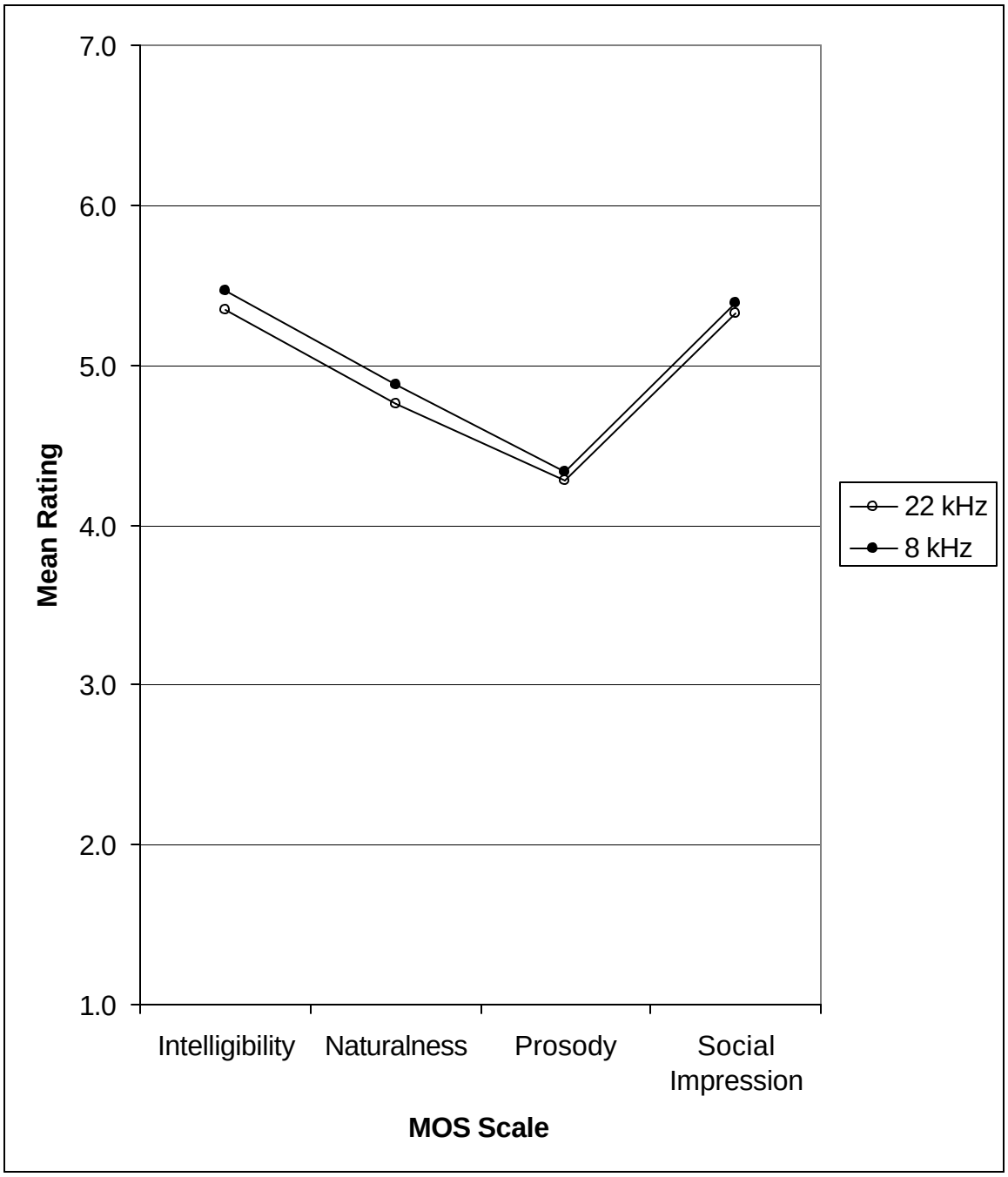


Figure 2. Bandwidth by Scale Interaction (nonsignificant)



Ratings by Items

The results for ratings at the item level were almost identical to those at the scale level. A mixed-model analysis of variance, with Speaker and Bandwidth as between-subjects independent variables and Scale as a within-subjects independent variable, indicated a significant main effect of Item ($F(13, 2886) = 57.4$, $p < .0001$) and a significant interaction between Speaker and Item ($F(26, 2886) = 3.6$, $p < .0001$). The interaction between Bandwidth and Item and the interaction among Bandwidth, Item, and Speaker were both nonsignificant ($F(13, 2886) = 0.58$, $p = .82$, and $F(26, 2886) = 1.2$, $p = .27$, respectively). The results for each voice appear in Table 2. . Figures 3 and 4 show the Speaker by Item and Bandwidth by Item interactions, respectively.

Table 2. Mean Ratings by MOS-X Items

Voice	1Effort	2Comp	3Artic	4Prec	5Pleas	6Nat	7Hum	8Qual	9Emp	10Rhy	11Inton	12Trust	13Conf	14Dep
AF 22	5.2	5.1	5.3	4.9	5.6	4.1	4.4	5.4	4.2	3.8	4.0	5.2	5.4	5.9
AF 8	5.4	5.3	5.6	5.0	5.7	4.3	4.5	5.5	4.6	4.3	4.5	5.5	5.4	5.9
AM 22	5.8	5.5	5.4	5.0	5.1	4.4	4.6	5.5	4.9	4.3	4.4	5.4	5.3	5.7
AM 8	5.6	5.6	5.7	5.0	4.9	4.1	4.5	5.5	4.4	4.3	4.4	4.9	4.9	6.3
B 22	5.5	5.6	5.4	5.4	4.9	3.8	4.3	4.9	4.2	4.3	4.3	4.8	5.0	5.1
B 8	5.6	5.6	5.6	5.6	4.9	4.2	4.6	5.7	4.4	4.1	4.2	5.1	5.3	5.3

Figure 5. Speaker by Item Interaction (significant)

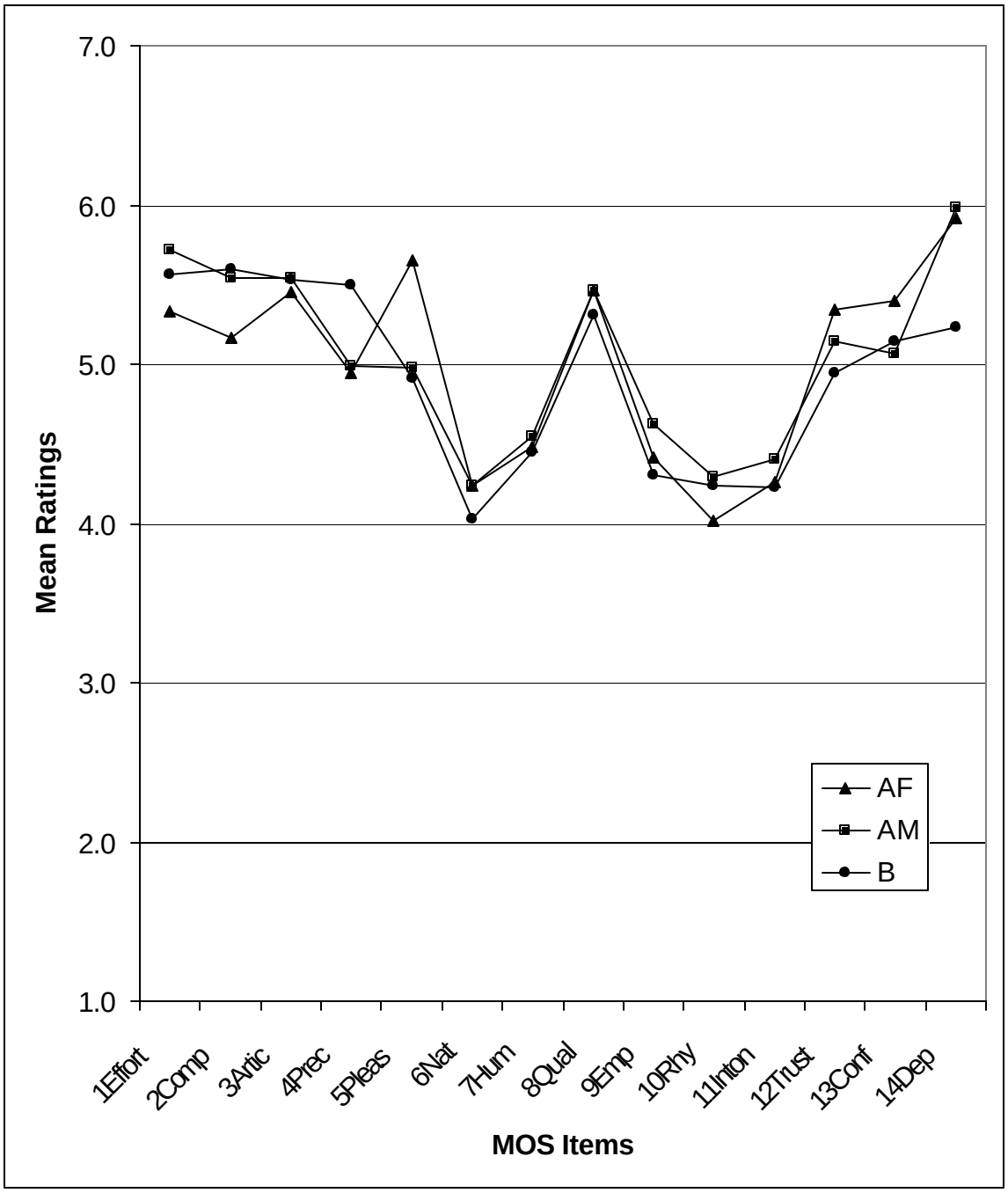
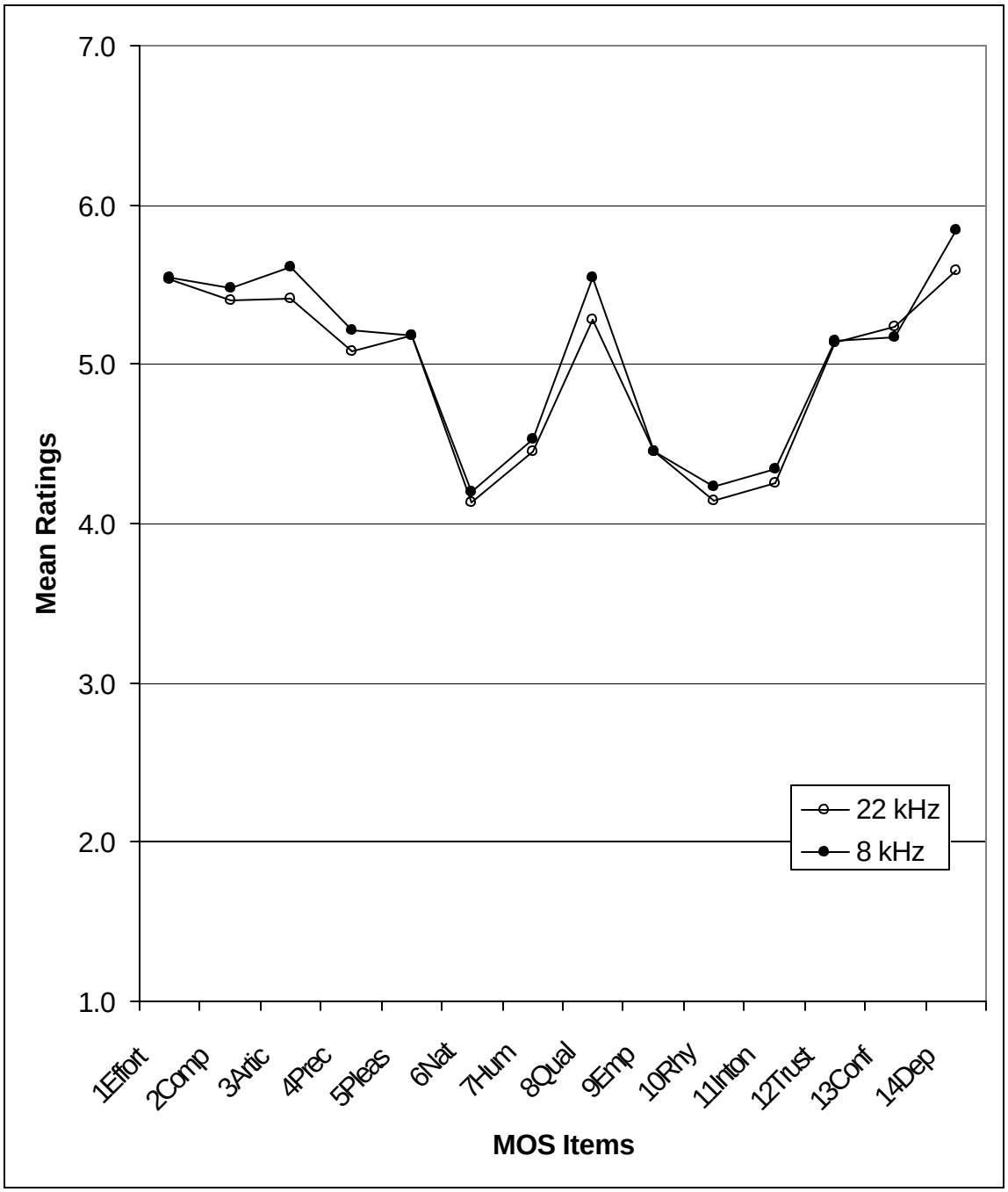


Figure 6. Bandwidth by Item Interaction (nonsignificant)



Separate Comparisons of Bandwidth by Speaker

Table 3 shows the results of six analyses of variance (three speakers by analyses at the scale and item level). The purpose of these analyses was to determine if any voice, analyzed independently from the other voices, showed evidence of a main effect or interaction with Bandwidth. There were consistent main effects of Scale and Item, showing that listeners did not routinely give the same value to each item when rating a voice. The main effect of Bandwidth and the interactions with Bandwidth were consistently nonsignificant.

Table 3. Separate Analyses of Ratings as a Function of Voice and Bandwidth

ANOVA Factor	**AF**	**AM**	**B**
Analysis by Scale			
Bandwidth	$F(1, 76)=1.1$, $p=.30$	$F(1, 59)=0.29$, $p=.59$	$F(1, 100)=0.84$, $p=.36$
Scale	$F(3, 228)=38.4$, $p<.0001$	$F(3, 177=31.4$, $p<.0001$	$F(3, 300)=41.2$, $p<.0001$
Bandwidth by Scale	$F(3, 228)=0.94$, $p=.42$	$F(3, 177)=0.39$, $p=.76$	$F(3, 300)=1.1$, $p=.37$
Analysis by Item			
Bandwidth	$F(1, 70)=1.2$, $p=.28$	$F(1, 55)=0.16$, $p=.70$	$F(1, 97)=0.85$, $p=.36$
Item	$F(13, 910)=23.5$, $p<.0001$	$F(13, 715)=20.4$, $p<.0001$	$F(13, 1261)=24.5$, $p<.0001$
Bandwidth by Item	$F(13, 910)=0.6$, $p=.82$	$F(13, 715)=1.3$, $p=.20$	$F(13, 1261)=1.0$, $p=.42$

Discussion

The consistently significant main effects of Scales and Items and the consistently significant interactions between these effects and Speaker provide evidence that the MOS-X is sensitive to Speaker differences, even when completely independent groups of raters have made the ratings.

In contrast, the consistently nonsignificant main effect of Bandwidth and its interactions with Speaker, Scale, and Item provide evidence that this variable does not affect listener ratings of speech quality made by independent groups.

It is possible that listeners exposed to both high-fidelity (22 kHz) and low-fidelity (8 kHz) versions of concatenative voices derived from the same speaker might be able to detect the difference. This type of exposure, however, does not typically happen during normal use of speech products.

These results indicate that if, in the future, we need to compare speech samples that differ in bandwidth, then the bandwidth differences are not likely to have any significant effect on the ratings as long as the ratings are provided by independent groups of listeners.

References

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

Appendix A. The MOS-X

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

<u>MOS-X Scales</u>
Overall: Average items 1-14
Intelligibility: Average items 1-3 and 11
Naturalness: Average items 4-6 and 8
Prosody: Average items 7 and 9-10
Social Impression: Average items 12-14

Appendix B. The Test Text

The moon had set by the time Peter and Cynthia returned from the lake. Through the darkness, two men approached the house. What can they be doing? They are up to no good! Tune in tomorrow for the exciting conclusion.

Trading on NASDAQ was lively today, February second, two thousand. Twenty-five million shares were traded, valued at over one-hundred-thirty-five million dollars. On Monday five-three two thousand, at nine-thirty Barbara Walters and Gerald Ford will ring the opening bell.

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?