

## EFFECT OF SPEAKER AND SAMPLING RATE ON MOS-X RATINGS OF CONCATENATIVE TTS VOICES

James R. Lewis  
International Business Machines Corp.  
Boca Raton, Florida

The MOS-X is a recently-developed questionnaire used to evaluate the quality of artificial speech. In this experiment, participants listened to audio files produced by concatenative text-to-speech voices for the purpose of assessing the effect of Speaker and Sampling Rate on MOS-X ratings. The concatenative voices were developed from recordings of three different human speakers (code named AF, AM, and B) and produced using two different sampling rates (8 kHz and 22 kHz). Six independent groups of raters participated, one group for each combination of speaker and sampling rate. Analyses of variance indicated a significant main effect of Voice, but no significant main effect of Sampling Rate and no significant Voice by Sampling Rate interaction. The results indicate that independent groups of raters are sensitive to speaker differences in concatenative text-to-speech (TTS) voices, but not to differences in these sampling rates.

### INTRODUCTION

For many years, the dominant method for producing artificial speech was formant synthesis (Klatt, 1980). A voice generated by formant synthesis is completely artificial, with the acoustics produced by phoneme models based on the analysis of formants (the spectral characteristics of the fundamental units of speech). An alternative approach is concatenative synthesis (Spiegel and Streeter, 1997). In concatenative synthesis, the text-to-speech (TTS) engine produces acoustics by concatenating (joining) pieces of speech recorded by a human speaker. The size of the concatenated units can vary from a single phoneme to an entire phrase. For producing speech from unrestricted text, the most common unit is the diphone (pair of phonemes), because this is smallest unit that can account for immediate effects of coarticulation (the changes in how a person produces a sound due to the articulatory effects of the immediately surrounding sounds).

Until recently, the biggest problem with artificial voices was intelligibility (Francis and Nusbaum, 1999). Most current high-quality formant TTS systems can produce intelligible speech, but the speech sounds synthetic and unnatural. Naturalness has consequently become a goal for the development of better TTS systems. There are many factors that can affect the perception of naturalness, from the fundamental acoustics of speech production to the prosodic contour of a phrase. With regard to fundamental speech

acoustics, concatenative systems have an advantage because they are derived from recordings of human speech.

The primary purpose of the current evaluation was to investigate listener sensitivity to the differences between concatenative text-to-speech voices with sampling rates of 8 and 22 kHz. Previous research investigating the psychometric properties of the MOS-X, a recently-developed questionnaire used to assess the quality of artificial speech (Polkosky and Lewis, 2003), 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 sampling rate.

Higher sampling rate leads to greater audio fidelity, but at a cost of increased requirements for storage and transmission capacity. For example, a sampling rate of 22 kHz is appropriate for use in desktop systems, but sampling rates greater than 8 kHz are not suitable for transmission over standard phone lines, which have a bandwidth of only 3.2 kHz. To accommodate a desired bandwidth, the sampling rate must be at least twice the desired bandwidth (Denes and Pinson, 1993).

The main psychological effect of a limited bandwidth is the loss of high-frequency information in the speech signal. For isolated words, this could cause a loss of intelligibility – especially for words that differ in high-frequency consonants such as ‘s’ and ‘f’. Additional context provided in longer

passages will usually allow listeners to recover from the loss of this high-frequency information. The upper limit of frequencies produced in human speech is around 7 kHz (Denes and Pinson, 1993).

The main concern in our lab, however, was the extent to which sampling rate might affect listener ratings of voice quality. At times the only speech samples available to us for comparison in competitive evaluations have had different sampling rates, and this situation could occur again in the future. If sampling rate 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 sampling rate. If sampling rate has little or no effect on listener ratings (at least, using our data collection methodology), then we could confidently conduct future comparative studies in which the speech samples differed in sampling rate.

## METHOD

### Participants

The participants in this experiment were IBM employees randomly selected from all IBM employees in the United States, invited by e-mail (400 invitations per voice) to visit a web site from which they could download the assigned audio file and complete the MOS-X questionnaire items (shown in Appendix A). The total number of respondents who completed the MOS-X for each voice were:

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

Responses were completely anonymous, as were the environments in which and equipment on which respondents listened to the audio samples.

### Stimuli

The stimulus for each group was an audio file of a synthetic voice speaking texts used in previous evaluations of synthetic voices (Polkosky, 2003).

There were two versions for each speaker, one with a sampling rate of 22 kHz and one with a sampling rate of 8 kHz, for a total of six audio files of artificial voices (see above). After listening to their assigned voice, participants completed a set of seven-point bipolar rating scales that included the 15 MOS-X items. Appendix B contains the test text (which takes about a minute to play after conversion to artificial speech).

### Procedure

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-X items for that voice.

Note that this procedure is a between-subjects design. In our initial studies comparing listener ratings of competitive artificial voices (starting in 2001), we used carefully counterbalanced within-subjects designs that included MOS ratings and forced-choice paired comparisons, with 16 participants per study. To reach a larger sample size listening to audio samples in more realistic environments, we switched to a between-subjects web-based approach. In our initial usage of the new method (starting in 2002), we had several opportunities to compare the ratings from our lab-based within-subjects studies with the ratings of the same voices using the web-based between-subjects design, and found that both methods resulted in very similar ratings. Due to the advantage in power and generalizability of results gained with the between-subjects design, we prefer the use of that method.

Use of the between-subject design with participants spread over the entire United States does lead to a loss of control over the participant's listening environment and equipment. Even inexpensive speakers, however, can reach outputs of 22 kHz (but vary somewhat in the lower frequency ranges – usually 30-50 Hz) (Cheap Computer Systems Guide, 2003). The frequency response of

inexpensive headphones is similar to that of inexpensive speakers. Thus, it is very likely that all or almost all respondents used equipment that was capable of playing 22 kHz samples.

The MOS-X is an expanded version of the standard MOS used for the subjective evaluation of the quality of artificial voices and speech degraded by electronic transmission. The items of the standard MOS align with two factors: intelligibility and naturalness (Lewis, 2001). Previous research indicated a significant correlation between the MOS intelligibility scale and more direct measures of intelligibility (Wang and Lewis, 2001). The MOS-X has items that align with these traditional factors as well as new factors for Prosody and Social Impression (see Appendix A). For the procedures used to develop the scales for these additional factors, see Polkosky and Lewis (2003).

**RESULTS**

**Ratings by Scales**

A mixed-model analysis of variance, with Speaker and Sampling Rate 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 Sampling Rate and Scale and the interaction among Sampling Rate, 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 (and Figures 1 and 2).

Table 1. Mean Ratings by MOS-X Scales

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

Figure 1. Speaker by Scale Interaction (significant)

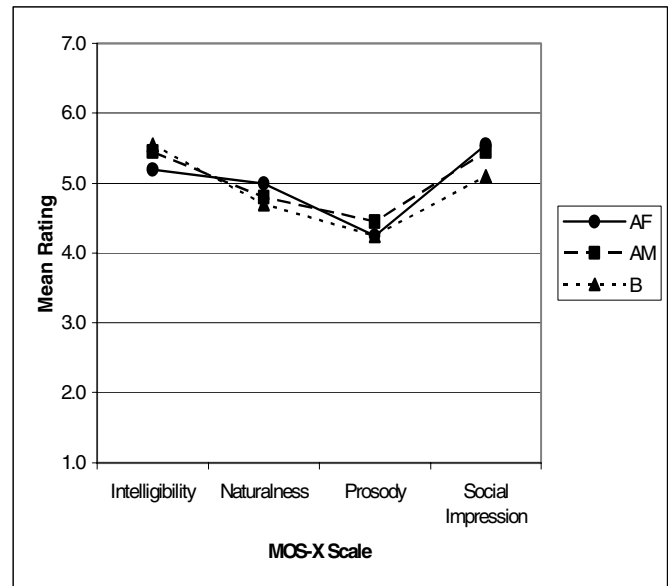
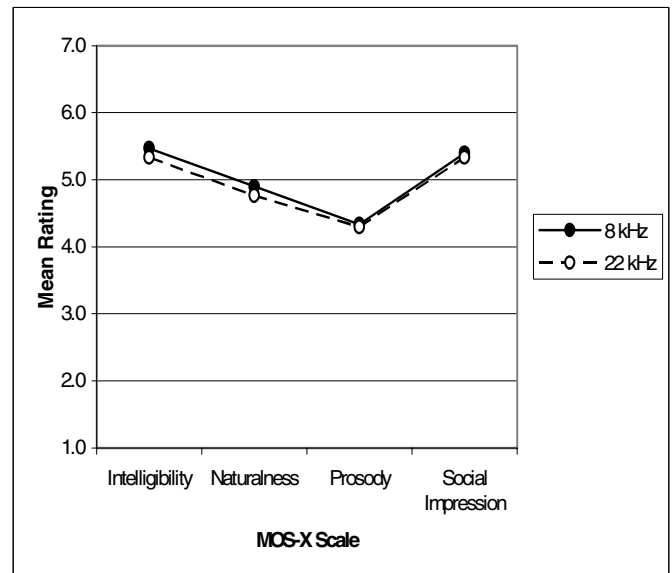


Figure 2. Sampling Rate by Scale Interaction (nonsignificant)



**Separate Comparisons of Sampling Rate by Speaker**

Table 2 shows the results of three analyses of variance (one for each speaker). 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 Sampling Rate. There were consistent main effects of Scale, showing that listeners did not routinely give the same value to each item when

rating a voice. The main effect of Sampling Rate and the interactions with Sampling Rate were consistently nonsignificant.

Table 2. Separate Analyses of Ratings as a Function of Voice and Sampling Rate

ANOVA Factor	AF	AM	B
Sampling Rate	$F(1,76)=1.1, p=.30$	$F(1,59)=0.29, p=.59$	$F(1,100)=0.84, p=.36$
Scale	$F(3,228)=38.4, p<.0001$	$F(3,177)=31.4, p<.0001$	$F(3,300)=41.2, p<.0001$
Sampling Rate by Scale	$F(3,228)=0.94, p=.42$	$F(3,177)=0.39, p=.76$	$F(3,300)=1.1, p=.37$

**DISCUSSION**

The consistently significant main effect of Scale and the consistently significant interactions between this effect 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 Sampling Rate and its interactions with Speaker and Scale provide evidence that this variable does not have a strong effect on listener ratings of speech quality made by independent groups (at least, not in the range of 8 kHz to 22 kHz).

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 sampling rate in the range of 8 to 22 kHz, then the sampling rate 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**

Cheap Computer Systems Guide. (2003). Cheap computer speakers ([http://www.cheap-computers-guide.net/cheap\\_computer\\_speakers.htm](http://www.cheap-computers-guide.net/cheap_computer_speakers.htm)).

Denes, P. B., and Pinson, E. N. (1993). *The speech chain*. New York, NY: W. H. Freeman.

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* (pp. 63-97). Boston, MA: Kluwer.

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

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

Polkosky, M. D., and 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.

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

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* (pp. 144-148). Mahwah, NJ: Lawrence Erlbaum.

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

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 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**      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**

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 B. THE TEST TEXTS**

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?