Intelligibility and Acceptability of Short Phrases Generated by Embedded Text-To-Speech Engines

Huifang Wang[1]
James R. Lewis [2]

[1]Cisco Systems, Inc.
255 W. Tasman Dr., SJC-J/2
San Jose, CA 95134
huifwang@cisco.com[1]

[2]IBM Voice Systems
1555 Palm Beach Lakes Blvd.
West Palm Beach, Florida
jimlewis@us.ibm.com

Abstract

We investigated the intelligibility and acceptability of three formant text-to-speech (TTS) engines suitable for use in devices with embedded speech recognition capability. Listeners transcribed and rated recordings of short phrases from four text domains (U.S. currency, dates, digits and proper names) produced by three commercially-available embedded TTS engines and a human speaker. The human voice received the best intelligibility and acceptability scores, and one of the TTS engines had superior intelligibility and acceptability relative to the two others. The results suggest that the ability to accurately produce names (the least constrained and least accurately transcribed text domain) was the system characteristic that best discriminated among the engines. The intelligibility and acceptability scores were generally consistent. Listeners transcribed shorter phrases more accurately than longer phrases, but acceptability ratings were independent of phrase length.

1. Introduction

Products with embedded speech capability typically have extreme resource constraints. For this reason, current text-to-speech engines for embedded products employ formant synthesis-by-rule to produce speech output. Because the competition in the embedded product space is fierce, we are always interested in understanding the strengths and weaknesses of our selected solution against the other solutions available in the marketplace.

The primary purpose of this experiment was to assess the intelligibility and acceptability of three formant text-to-speech (TTS) engines suitable for use in devices with embedded speech capability. Secondary purposes of the experiment were to:

- Benchmark the TTS engines against a recorded human voice
- Investigate any interaction among the engines and four different types of text commonly produced by embedded TTS engines (U.S. currency, dates, digits and proper names)
- Investigate the effects of phrase length on intelligibility and acceptability (measuring acceptability with the Mean Opinion Scale, or MOS)

We had an interest in the potential effects of phrase length on intelligibility because there are competing cognitive forces involved in the transcription of short phrases. One hypothesis is that transcription accuracy for shorter phrases should be better than for longer phrases because there is less material to remember. An alternative hypothesis is that longer phrases provide more acoustic and linguistic context, which should make it easier for listeners to compensate for any intelligibility problems in the production of artificial speech. It is likely that both hypotheses are true, but it wasn't clear which would exert the greater influence on transcription accuracy.

[1] At the time we designed this experiment, Huifang Wang was an employee of IBM Voice Systems in West Palm Beach, Florida.

With regard to acceptability rating with the MOS, Johnston (1996) had listeners judge the quality of natural speech degraded with time frequency warping. He found a significant relationship in the expected direction for judgements using the MOS Listening Effort item (greater degradation led to poorer ratings). He also found that using sentences as stimuli in an experiment of three TTS voices yielded results that were just as sensitive as those using longer paragraphs. The stimuli in the current experiment were shorter than those used by Johnston, ranging in length from 2 to 14 syllables. These stimuli provide an opportunity to extend the previous finding of the stability of MOS ratings to shorter speech samples and for the overall MOS (rather than a single item). Additional evidence that listener ratings of speech acceptability are reasonably independent of the length of the speech sample could have important implications for the design of more efficient experiments.

2. Method

2.1. Participants

Sixteen people participated in the experiment. Half of the participants were IBM employees and half were employees of a temporary employment agency, hired for the purpose of participating in this experiment. Each group had an equal number of males and females, with approximately equal age distribution across the groups (about half over and half under 40 years of age).

2.2. Materials

Each participant heard all four voices (human speech and three different commercially-available TTS voices, identified in this paper as TTS1, TTS2, and TTS3). The test texts were short phrases from four text domains: U.S. currency (e.g., "one dollar and twenty-three cents"), dates (e.g., "January first two thousand one"), numbers (e.g., "five five five one two three four") and proper names (e.g., "U.S. Forty Toll West"). The phrases in each domain had the same average number of syllables (6.375) to control for potential confounding due to phrase length differences. For purpose of controlling the distribution of phrase length across speech production conditions, we categorized the phrases by pairs as Short (1,2), Medium (3,4), Medium-Long (5,6) and Long (7,8).

2.3. Procedure

The experimental design used Latin squares to systematically counterbalance the presentation of test phrases among the various experimental conditions and to systematically counterbalance the order of presentation of the experimental conditions for both intelligibility and acceptability rating sessions for both groups of participants (IBM and agency employees). During an intelligibility session, participants listened to a test phrase from one of the text domains produced by one of the systems (or the human) and then transcribed it, with two trials per system (for a total of eight trials per session). Transcription accuracy scores were 1 for perfect transcription and 0 for incorrect transcription. After completing an intelligibility session, participants listened to both of the test phrases (from the assigned pair of phrases) produced by a given voice during the intelligibility session, and then rated the acceptability of that voice using the Mean Opinion Scale (MOS,), doing this for each voice. Participants continued in this fashion until they had completed the intelligibility and acceptability rating trials for all voices and for all text domains. (See the Appendix for the MOS items used in this experiment. See Lewis, 2001, for information about the psychometric properties of the MOS).

3. Results

The design of this experiment allowed two different arrangements of the data for the purpose of factorial analysis of variance: by Text Domain or by Phrase Length. Because our primary development interest was in the effects of Text Domain, that arrangement received the more detailed analyses.

3.1. Intelligibility

Arrangement by text domain. Analysis of variance on the intelligibility scores (averaged across the two trials) revealed significant main effects for both Voice ($F(3,36) = 3.98$, $p = .015$) and Text Domain ($F(3,36) = 29.3$, $p = .0000001$), and a marginally significant Voice by Text Domain interaction ($F(9,108) = 1.8$, $p = .08$). The between-subjects variables of Group and Gender were not significant ($p > .70$), and did not significantly interact with the other variables. A set of *t*-tests indicated that listening to both TTS2 and TTS3 resulted in significantly poorer transcription accuracy (77% and 70% respectively) than the 89% accuracy for the human voice (TTS2: $t(15) = 2.2$, $p =$

.04; TTS3: $t(15) = 4.6, p = .0003$), and TTS1 (84%) had somewhat better accuracy than TTS3 ($t(15) = 2.0, p = .06$). Transcription accuracies for all text domains were significantly different (all $p < .01$), with 95% accuracy for currency , 86% for times and dates, 78% for numbers and 63% for proper names. The nature of the Voice by Text Domain interaction on intelligibility was that all the voices showed similar patterns for the domains of currency, date and number, but were distinctly different for the domain of name, with accuracies ranging from 34% for TTS4 to 91% for the human voice (see Figure 1).

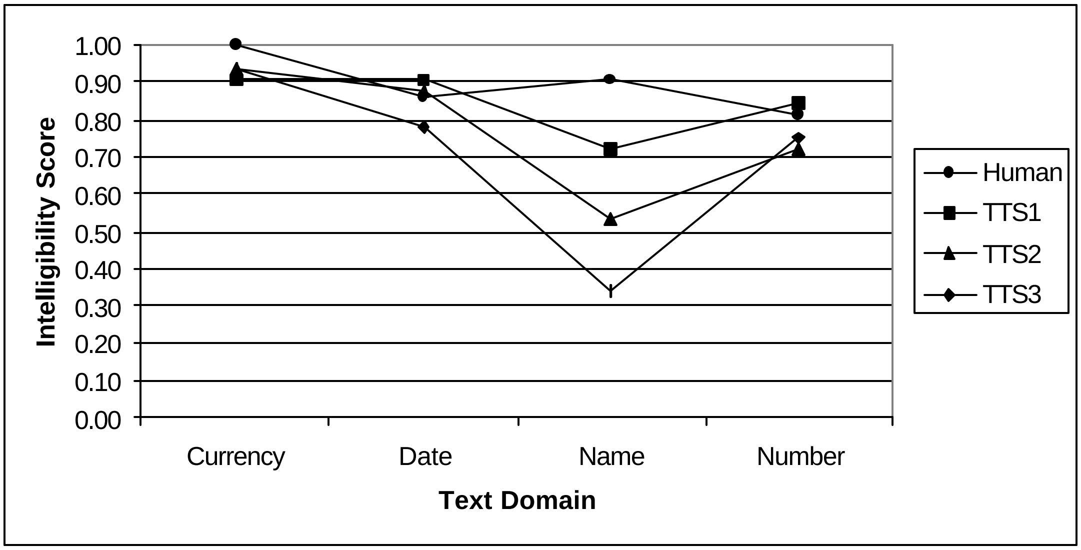

Figure 1. Voice by Text Domain Interaction (Intelligibility)

Arrangement by phrase length. An analysis of variance indicated a significant main effect of phrase length ($F(3,36) = 5.6, p = .003$), but no additional significant effects other than the previously described significant main effect of Voice. A set of *t*-tests indicated more accurate transcription for short (89%) than medium-long (79%) and long phrases (70%) ($t(15) = 2.67, p = .02$, and $t(15) = 3.83, p = .002$ respectively) and more accurate transcription for medium (83%) than long phrases ($t(15)=2.32, p=.04$).

3.2. Acceptability

Arrangement by text domain. Using the MOS factor structure described by Lewis (2001), analysis of variance detected significant main effects of Voice ($F(3,36) = 30.3, p = .0000000006$), Text Domain ($F(3,36) = 4.0, p = .015$), and Factor ($F(2,24) = 8.9, p = .001$). The main effects and interaction of Gender and Group were not significant (all $p > .50$). The Voice by Text Domain ($F(9,108) = 4.8, p < .00002$), Voice by Factor ($F(6,72) = 12.2, p = .000000002$), Text Domain by Factor ($F(6,72) = 2.8, p = .02$), and Voice by Text Domain by Factor ($F(18,216) = 10.0, p < .0000000001$) interactions were significant. There were also significant Factor by Gender by Group ($F(2,24) = 4.8, p = .02$)) and Voice by Factor by Gender by Group ($F(6,72) = 2.4, p = .04$) interactions. In these analyses MOS scores can range from 0 to 4, with lower numbers indicating better ratings. A set of *t*-tests on the main effect of Voice showed that all differences were statistically significant ($p < .10$ for TTS2 vs. TTS3, $p < .0004$ for all other differences, with an overall mean rating of 0.27 for the human voice, 1.09 for TTS1, 1.54 for TTS2, and 1.76 for TTS3). A similar set of tests on Text Domain showed poorer ratings for Date (1.29) and Name (1.35) than Number (1.06) or Currency (0.96) (all $p < .03$). A set of *t*-tests on the main effect of Factor indicated significantly poorer ratings for Naturalness (1.49) than Intelligibility (1.01) ($t(15) = 5.33, p = .0001$) and Speaking Rate (1.13) ($t(15) = 2.5, p = .02$). There is not sufficient space in this paper to describe all interactions in detail, but the key attributes of the significant two-way interactions were:

- The Voice by Text Domain interaction illustrated the clear separation between the human and TTS voices. For currency and number domains, TTS1 and TTS2 voices had almost equal acceptance. The greatest difference between TTS1 and TTS2 occurred for the production of the times and dates.
- Figure 2 shows the Voice by Factor interaction. Decomposing the overall MOS into its component factors (Intelligibility, Naturalness, and Speaking Rate) revealed a difference in profile between the artificial and human

voices. All of the voices occupied different vertical locations on the graph, consistent with the finding of a significant main effect of Voice. The human voice had about equal ratings for Intelligibility and Naturalness, with a somewhat poorer rating for Speaking Rate. All of the artificial voices received about equal ratings for Intelligibility and Speaking Rate, with somewhat poorer ratings for Naturalness.

- The most salient characteristic of the Text Domain by Factor interaction was the relatively poor rating for the name domain on the Intelligibility scale, which was consistent with the measured intelligibility of names relative to the other text domains.

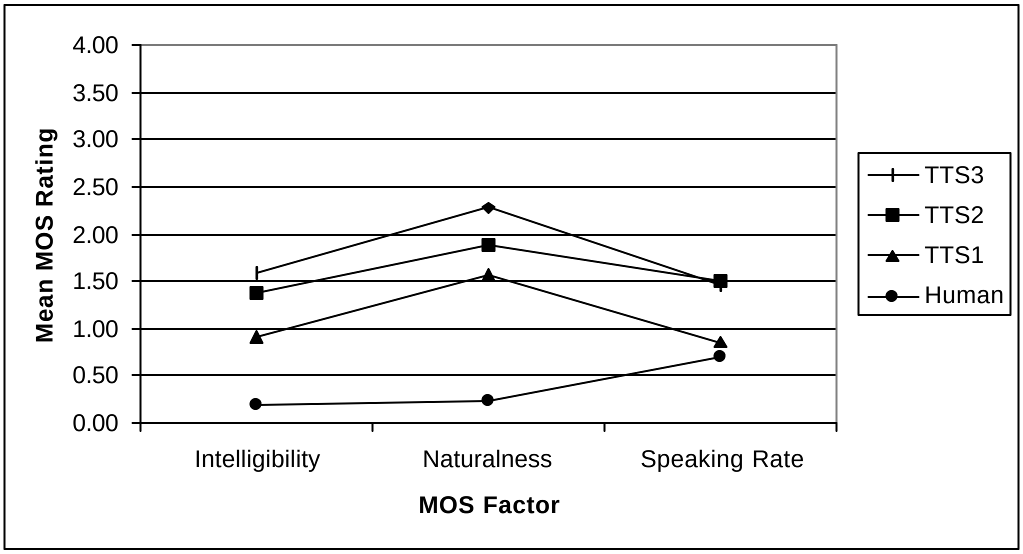

Figure 2. Voice by Factor Interaction (Acceptability)

Arrangement by phrase length. The most interesting finding from rearranging the data to investigate the effect of phrase length was that there was no significant effect of Phrase Length (*F*(3,36) = .08, *p* = .97, with mean ratings of 1.16 for Short, 1.18 for Medium, 1.15 for Medium-Long, and 1.18 for Long) and no Voice by Phrase Length interaction (*F*(9,108) = .54, *p* = .85), consistent with the findings of Johnston (1996).

4. Discussion

The key findings from the experiment were:

4.1. TTS1 was the Best Artificial Voice

Among the three TTS engines, TTS1 was the most intelligible and acceptable. The results of this experiment show no reason for us to consider using TTS2 or TTS3 in place of TTS1 (our current engine).

4.2. The Human Voice was Most Intelligible and Acceptable

TTS1 was the best artificial voice, but the recorded human voice had the best intelligibility scores and received the best acceptability ratings. Whenever possible (which could be rarely in embedded systems), developers should use a recorded human voice rather than an artificial voice.

4.3. The Least Intelligible Text Domain was that of Proper Names

The pattern of results by text domain for intelligibility scores was interesting in that the least constrained domain (names) had the poorest intelligibility overall and the greatest variability by voice. The significant Voice by Text Domain interaction suggests that the ability to accurately produce names was the system characteristic that best discriminated among the three TTS systems.

4.4. Memory Load Dominates Contextual Effects when Transcribing Short Phrases

The finding that longer phrases had poorer transcription accuracy than shorter phrases demonstrates that contextual effects are weaker than the effects of memory load when transcribing short phrases. The practical implication of this is that there is no reason to have systems produce longer phrases than necessary with the belief that this will enhance intelligibility. Consistent with the guidelines presented in Balentine and Morgan (1999), system messages should not be any longer than they need to be, whether produced by a human or artificial voice.

4.5. MOS Ratings are Stable over a Wide Variety of Phrase Lengths

The main effect of Phrase Length was not marginally nonsignificant -- it was dramatically nonsignificant (p = .97). The detection of other effects is evidence that the failure to detect effects of Phrase Length was not due to an insensitive experimental design. Listeners seem to arrive at their judgements of speech output quality quickly and produce consistent ratings even when the text sample is very short. The practical implication of this finding is that it should be possible to run very efficient studies comparing artificial voices when using the MOS as the instrument for assessing perception of voice quality.

4.6. MOS Ratings and Intelligibility Scores Tend to Lead to Similar Decisions

The results for intelligibility scores and acceptance ratings were highly similar, leading to identical conclusions regarding the relative quality of the voices. For example, the patterns of means for the significant main effects of Voice were virtually identical for both variables. Using data from the current experiment, Lewis (2001) found that intelligibility scores tend to correlate significantly with the MOS Intelligibility factor. This, along with the stability of the MOS over phrase lengths, supports the continued use of the MOS in experiments conducted for the purpose of evaluating the quality of TTS systems.

5. References

Balentine, B., & Morgan, D. P. (1999). *How to build a speech recognition application: A style guide for telephony dialogs*. San Ramon, CA: Enterprise Integration Group.

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

Lewis, J. R. (2001). *Psychometric properties of the MOS* (In the proceedings of this conference).

6. Appendix. MOS Used in the Experiment

The MOS text shown here is the same as that used in the experiment, but with the format modified to fit.

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