Text-To-Speech Intelligibility across Speech Rates

Ann K. Syrdal\textsuperscript{1}, H. Timothy Bunnell\textsuperscript{2}, Susan R. Hertz\textsuperscript{3}, Taniya Mishra\textsuperscript{4}, Murray Spiegel\textsuperscript{5}, Corine Bickley\textsuperscript{6}, Deborah Rekart\textsuperscript{7}, Matthew J. Makashay\textsuperscript{7}

\textsuperscript{1}AT&T Labs – Research, Florham Park, NJ, USA
\textsuperscript{2}Nemours Biomedical Research, Alfred I. duPont Hospital for Children, Wilmington, DE, USA
\textsuperscript{3}NovaSpeech LLC and Dept. of Linguistics, Cornell University, Ithaca, NY, USA
\textsuperscript{4}Speech Applications Research, Applied Communication Sciences, Piscataway, NJ, USA
\textsuperscript{5}Dept. of Hearing, Speech and Language Sciences, Gallaudet University, Washington, DC, USA
\textsuperscript{6}AT&T Services, Dallas, TX, USA
\textsuperscript{7}Audiology & Speech Center, Walter Reed National Military Medical Center, Bethesda, MD, USA
syrdal@research.att.com, bunnell@asel.udel.edu, shertz@novaspeech.com, taniya@research.att.com, mspiegel@appcomsci.com, corine.bickley@gallaudet.edu, dr5519@att.com, matthew.j.makashay.civ@health.mil

Abstract

A web-based listening test measured intelligibility across speech rate of 8 TTS systems and of a linearly time-compressed human speech reference voice. The synthesis systems included 2 independent representatives of each of the following 4 synthesis methods: formant, diphone concatenation, unit selection concatenation, and HMM. For each TTS system, a female and a male American English voice were tested. Semantically unpredictable sentences were presented at 6 speech rates from 200 to 450 words per minute. In an open response format, listeners typed what they heard. Listener transcriptions were automatically scored at the word level, and a normalized edit distance per speech rate was calculated for each of 355 listeners. There were significant differences among the TTS systems. The 2 unit selection TTS systems were the most intelligible across speech rates; one was equivalent to human speech. Listeners’ native language, TTS familiarity, and audio equipment were also significant factors.

Index Terms: speech synthesis, text-to-speech, intelligibility, speech rate

1. Introduction

This paper systematically compares the intelligibility across six speech rates of eight different text-to-speech (TTS) systems. Each system is represented in the test by a male and a female American English TTS voice. Two independent systems representing each of the following four synthesis methods were included:

1. Formant synthesis \cite{1} (synthesis-by-rule) generates speech by predicting acoustic parameters that control a speech production model;

2. Diphone concatenation \cite{2} concatenates brief units from a small inventory of recorded speech and then applies signal processing to modify pitch and duration;

3. Unit selection concatenation \cite{3} concatenates units of recorded speech from a very large annotated corpus of continuous speech; units are selected by an algorithm to optimize auditory quality;

4. Hidden-Markov-model (HMM) synthesis \cite{4} uses a data-driven statistical machine learning approach with trained HMMs to learn a parametric model of speech for synthesis.

2. Experiment

TTS intelligibility was measured across six speech rates, ranging from the system’s (or human speaker’s) default rate, which was 200 wpm (words per minute) or less, to 250, 300, 350, 400, and 450 wpm. The listening test followed procedures recommended for evaluating TTS intelligibility by the authors, who constitute the ASA Standards Text-to-Speech Technology working group (S3-WG91). Our goals were to conduct a trial of the recommended procedures prior to submitting a formal standards proposal, to provide benchmark data on TTS intelligibility using the proposed methodology, and to document the effect of speech rate on TTS intelligibility across a wide range of synthesis systems and voices.

2.1. Task

The experimental procedure was an open response immediate recall test. The test was web-based and interactive, with trial-by-trial listener initiation of auditory stimulus presentation with no constraints on response interval. Each stimulus was limited to a single presentation. Participants were first presented a screen containing experimental instructions, followed by a form for entering demographic information. Next, six familiarization trials were presented, one per screen, each representing one of the six speaking rates to be tested. During the familiarization trials, listeners could adjust the equipment to their preferred level. Finally, 60 test trials were conducted; one unique test sentence was presented per trial. The first 10 test sentences were at the synthesis system’s or human speaker’s default rate (200 wpm or less). In the second set of 10 sentences, the rate was 250 wpm; in each successive set the rate increased by 50 wpm such that the sixth and final set was at 450 wpm. Stimulus order at each rate was randomized between listeners. A test session typically lasted 20 minutes.

The listener’s task in each trial was to listen to the sentence and type what was heard into a text box. If a word was unfamiliar or unclear, listeners were instructed to write what it sounded like as closely as possible, even if the result was not a known word. If they couldn’t recognize anything at all about a word, they were instructed to enter xxx in place of the word.
2.2. Text materials

Text input was in the form of 6- to 8-word semantically unpredictable sentences (SUSs) constrained to a maximum of 10 syllables. In SUSs, words of the appropriate syntactic categories defined by a grammatical sentence frame are randomly selected to yield sentences that are syntactically grammatical but likely to be semantically anomalous. Sentence frames included a variety of syntactic structures, such as active and passive sentences, assertions, imperatives, and interrogatives, and contained singular and plural nouns and pronouns and both transitive and intransitive verbs.

SUSs were generated with the SUSgen program [5]. The average number of words per sentence was 6.6 for each of the six speaking rates tested. The vocabulary set used by SUSgen contained familiar monosyllabic and polysyllabic words and a wide variety of phonemes and consonant clusters. Test sentences contained both stressed and unstressed syllables, and content and function words. An example of an SUS is: Hang the walls and the mud fields.

2.3. Speech synthesis systems

Eight TTS systems representing four synthesis methods were tested, along with a female human speech reference. Two TTS systems were tested for each method. A female and a male American English voice were tested for each of seven TTS systems. For one diphone concatenation system (C2), only the female voice at only four rates (200–350 wpm) was tested due to synthesis system constraints. With one known exception, all TTS systems were based on different speakers. The human speech reference (MJ) was a professional female speaker recorded reading the SUSs at a rate of 200 wpm; none of the TTS systems tested were based on the speaker. The human recordings were then linearly sped up with WSOLA (Waveform Similarity-based OverLap-Add), an algorithm [6] that modifies the time scale of the signal without changing the pitch. Linear time-compression of speech has been shown to have intelligibility advantages over some alternative methods that attempt to preserve certain aspects of natural speech temporal patterns [7].

Synthetic speech stimuli across all rates tested were generated by collaborating industry or academic laboratories; all audio files used 16-bit samples. The means used by a system to control its speech rate is proprietary knowledge not revealed to the authors. Speech rate accuracy was independently verified before testing as follows: For each target speech rate, the total duration of speech in the submitted audio files was divided by the total number of words of text in the constituent SUSs to yield an average word-per-minute measure.

A listener heard only one voice from one source during a test session. The synthesized (or sped-up human) voice tested in any given test session was determined automatically on a round-robin basis.

It was agreed to report TTS system results anonymously; only the synthesis technique is identified. Table 1 lists the TTS systems, their sampling rates, and the number of listeners and the breakdown of how many listeners heard a female voice (F.n) and how many heard a male voice (M.n).

2.4. Listener participants

Volunteer adult participants were solicited by email or web invitations. 374 listeners completed the test; 19 were eliminated because of a mean normalized word error score over 90 at 200 wpm or self-reported impaired hearing status. Data from the remaining 355 listeners are reported.

98 listeners were under 25 years of age, 206 were between 25 and 50, and 51 were over 50. Most participants were quite unfamiliar with synthetic speech: 46 reported they “never” listened to it, and 88 said that they “infrequently” did; 146 participants were “not sure”. 47 listened to TTS “occasionally”, 18 listened “frequently”, and 10 very “frequently”. 307 listeners identified themselves as native speakers of English, 48 as non-native. 218 listeners used headphones or ear buds (as recommended); 137 used speakers.

### Table 1: Stimulus information and listener counts.

<table>
<thead>
<tr>
<th>Stim</th>
<th>Method</th>
<th>Samp. Rate</th>
<th>F.n</th>
<th>M.n</th>
</tr>
</thead>
<tbody>
<tr>
<td>F1</td>
<td>Formant</td>
<td>11.025 kHz</td>
<td>23</td>
<td>20</td>
</tr>
<tr>
<td>F2</td>
<td>Formant</td>
<td>11.025 kHz</td>
<td>31</td>
<td>22</td>
</tr>
<tr>
<td>C1</td>
<td>Diphone Concat.</td>
<td>16 kHz</td>
<td>17</td>
<td>21</td>
</tr>
<tr>
<td>C2</td>
<td>Diphone Concat.</td>
<td>12 kHz</td>
<td>20</td>
<td>19</td>
</tr>
<tr>
<td>U1</td>
<td>Unit Sel. Concat.</td>
<td>16 kHz</td>
<td>17</td>
<td>28</td>
</tr>
<tr>
<td>U2</td>
<td>Unit Sel. Concat.</td>
<td>16 kHz</td>
<td>26</td>
<td>19</td>
</tr>
<tr>
<td>H1</td>
<td>HMM</td>
<td>16 kHz</td>
<td>24</td>
<td>25</td>
</tr>
<tr>
<td>H2</td>
<td>HMM</td>
<td>16 kHz</td>
<td>17</td>
<td>23</td>
</tr>
<tr>
<td>MJ</td>
<td>Human speech</td>
<td>16 kHz</td>
<td>22</td>
<td></td>
</tr>
</tbody>
</table>

2.5. Scoring

Each SUS transcription was automatically processed; orthographic accuracy was scored at the word level using the program WordScore [5], which calculates an edit distance by counting the minimum number of word insertions, deletions, and substitutions required to convert the listener’s transcriptions into the stimulus SUS text. Content word, function word, and word boundary errors were all taken into account. Homophones and some misspellings were predefined as admissible errors. For the set of 10 SUSs presented at a given speech rate, a normalized edit distance (NED) score was calculated by dividing the sum of the 10 SUS edit distances by the number of words in the set. In other words, the NED score is roughly the proportion of words misheard.

3. Results

A mixed effect ANOVA was run with one within-subject factor: Speech Rate (6 levels); and five between-subject factors: Speech Source (8 levels: 7 TTS, 1 human), Voice Gender (2), Audio Equipment (2), Listener Age (3), and Native Language (2). Note that TTS system C2 was not included in this analysis because it could only synthesize the lowest four rates.

There is a significant main effect for Speech Rate (F(5,1110) = 170.245, p < .0001); errors increase with speech rate. Post hoc Bonferroni multiple comparisons were run at the .05 significance level. Errors increase significantly for every 50 wpm rise in speech rate. There is a significant Speech Rate by Speech Source interaction (F(35,1110) = 4.64, p < .0001), illustrated in Figure 1, which plots error as a function of speech rate for the eight TTS systems and sped-up human voice (MJ). TTS system C2 is included in the plot. Although MJ and most TTS systems show a pattern of gradually increasing NED across speech rates, system H1 performs well at 200–250 wpm, but degrades rapidly at higher rates. In contrast, F2 shows a relatively small increase in NED.

There are three significant between-subjects main effects: (1) Audio Equipment (F(1,222) = 7.652, p < .006), with error scores 6.2% lower for headphones than speakers; (2) Native
Language ($F(1,222) = 7.721$, $p < .006$), with native English speaking listeners making 9.1% fewer errors than non-native; (3) Speech Source ($F(7,222) = 8.180$, $p < .0001$). The mean NED scores (averaged over 6 rates) are listed in Table 2. TTS system U2 and sped-up human speech MJ had significantly lower error scores than the other speech sources; U1 was statistically equivalent to M3, F2, and H1. Scores for F2, H1, H2, F1, and C1 were equivalent. There were no significant main effects of Voice Gender or Listener Age. There was one significant between-subject interaction: Audio Equipment by Speech Source ($F(7,222) = 2.766$, $p < .009$). Only the sped-up human voice, MJ, had a lower error score for speakers than for headphones; the reverse was true for synthetic voices.

**Figure 1: Word Error Rate across Speech Rate.**

Another mixed Speech Rate by Speech Source ANOVA was run that included the between-subject variable of TTS Familiarity (6 levels). There was a significant main effect of TTS Familiarity ($F(5,291) = 5.784$, $p < .0001$). The error rate declined as familiarity with TTS increased, but the only significant difference was between the two largest groups; those who listened ‘infrequently’ scored better than those ‘not sure’.

<table>
<thead>
<tr>
<th>Sp. Source</th>
<th>Mean NED</th>
<th>Std. Error</th>
</tr>
</thead>
<tbody>
<tr>
<td>C1</td>
<td>.591</td>
<td>.027</td>
</tr>
<tr>
<td>F1</td>
<td>.536</td>
<td>.023</td>
</tr>
<tr>
<td>H2</td>
<td>.524</td>
<td>.025</td>
</tr>
<tr>
<td>H1</td>
<td>.498</td>
<td>.022</td>
</tr>
<tr>
<td>F2</td>
<td>.480</td>
<td>.022</td>
</tr>
<tr>
<td>U1</td>
<td>.430</td>
<td>.022</td>
</tr>
<tr>
<td>MJ</td>
<td>.347</td>
<td>.035</td>
</tr>
<tr>
<td>U2</td>
<td>.332</td>
<td>.023</td>
</tr>
</tbody>
</table>

**Table 2: Mean normalized edit distance.**

4. Discussion

It is not surprising that the intelligibility of all systems and also of the sped-up human speech declined as speech rate increased. However, it is interesting that the intelligibility of systems did not degrade at the same rate.

The lack of a significant effect of TTS voice gender on intelligibility is noteworthy. It is interesting that the factor of listener age did not even approach significance ($F(2,222) = 1.000$, $p < .370$), nor was there an age by speech rate interaction. Although mean error scores increased with age, overall differences were rather small. The lack of an effect may be due to the greater familiarity with TTS of the older listeners. Of listeners over 50 years of age, 20% reported that they listened “frequently” or “very frequently” to TTS, as compared to 7% for the 25–50 year age group, and 3% for under-25 year age group.

The observation that both unit selection systems performed as well or better than the formant synthesis systems at the fastest speaking rates allows us to reject a commonly held claim that formant synthesis systems – as a class – are more intelligible than concatenative systems at fast speaking rates. Indeed, there is little evidence in these data that any one synthesis technology per se is inherently more intelligible than another over the conditions we have examined. While the large inventory unit selection system (U2) was clearly the best system overall, the smaller inventory system (U1) and the better formant synthesis system F2 were statistically indistinguishable in intelligibility at most rates. One of the HMM systems was as intelligible as the better unit selection system at the slowest speaking rate; but also posted the poorest intelligibility at the highest speaking rate. Thus, the results appear to be revealing differences among specific TTS systems rather than differences that are inherent to specific TTS technologies. Crucially, however, these results clearly show the power and sensitivity of the proposed standard for detecting differences among TTS systems and differences due to specific speech manipulations such as the one for speaking rate in the present experiment.

Results from the current experiment can be roughly compared with those from 36 listeners with early-onset blindness tested in a different experiment with some of the same synthesized SUS materials and TTS systems [9]. Formant and unit selection TTS voices were included in the blind listener study, with rates from 300 wpm to 550 wpm, and the cosine similarity metric was used to measure intelligibility. Despite these differences, some general comparisons can be made. The blind listeners, many of whom often use TTS at high speech rates, achieved higher intelligibility scores than the 355 listeners of the present study: even at rates of 500 wpm, transcription accuracy of blind listeners for all TTS systems was at or above 50%.

5. Summary and conclusions

TTS intelligibility was measured across six speech rates from 200 to 450 wpm. Transcription error rate (NED) was compared among eight synthesis systems and a human reference voice. The human voice was recorded speaking at 200 wpm, and was linearly time-compressed to achieve faster speech rates. Four synthesis methods (formant, small inventory diphone concatenative, unit selection concatenative, inventory and HMM) were each represented by two TTS systems. Each TTS system synthesized stimuli across rates with a female and a male American English voice (except for C2, for which only a female voice was tested); a different set of 10 Semantically Unpredictable Sentences (SUSs) was tested at each rate.

Results from the listening test are reported from 355 volunteer adult listeners. There were significant effects of speech rate, synthesis system, listener audio equipment, listener native language, and listener familiarity with TTS. There was a significant interaction between speech rate and synthesis system. Word error rate rose as speech rate increased, but the performance level and pattern across rates differed among TTS systems.
In general, the results demonstrate that the testing methodology discriminates nicely among synthesis systems, with no ceiling effects. One of the unit selection concatenative TTS systems had the best NED, while one of the small inventory diphone concatenative systems had the worst. The best system had statistically equivalent scores to the sped-up human speech reference. Not surprisingly, the test also demonstrated that word error rate rose as speech rate increased, though not always in a uniform fashion across systems.

6. Acknowledgements

We thank those who generously synthesized audio samples for this evaluation, the volunteer listeners, and the individuals who helped to spread the word about this test to potential participants.

The views expressed in this article are those of the authors and do not necessarily reflect the official policy or position of the Departments of the Navy, Army, or Air Force, the Department of Defense, or the US Government.

7. References