Assessment of the Intelligibility and Perceived Quality of Speech Produced by Text-to-Speech Engines

Michael D. Hall, Erica B. Stevens, Richard Eyraud, Denise Padden, and Patricia K. Kuhl

Department of Speech and Hearing Sciences, CHHD Box 357920, University of Washington, Seattle, WA 98195

Abstract: The intelligibility and quality of speech produced by three text-to-speech (TTS) engines was assessed for CVC words, meaningful sentences, semantically anomalous sentences, and paragraphs. Intelligibility was evaluated using transcription accuracy and ratings of articulation clarity. Additionally, the "naturalness" of speech was rated. Consistent differences in intelligibility and naturalness were found across engines. Implications for TTS synthesis are discussed.

As the intelligibility of speech generated by TTS engines continues to improve, it will become increasingly important to assess the quality of speech, including perceived naturalness, relative to human speech. Yet, evaluations of TTS synthesis generally have not included measures of perceptual quality (1). The current study sought to assess both the intelligibility and "naturalness" of speech generated by three current TTS engines: the Whisler engine in the Microsoft Speech SDK (Version 3.0B), AT&T's engine within the Watson Advanced Speech Applications Platform (Version 2), and Elan Informatique's ProVerbe Speech Engine SDK (Version 1.200).

METHOD

Twenty-four monolingual English-speaking adults with no known hearing deficits were presented 16-bit WAV files synthesized using the default male voice from three TTS engines. The Microsoft engine produced speech at a 22 kHz sampling rate with an average fundamental frequency (F0) of 113 Hz at a rate of 130 wpm. AT&T's engine produced speech with an average F0 of 111 Hz at 170 wpm, and Elan's engine produced speech with an average F0 of 106 Hz at 180 wpm; both engines produced tokens at an 11 kHz sampling rate.

Each engine produced materials from long-standing tests of intelligibility that encompassed a wide range of speech complexity from isolated words to meaningful sentences, semantically anomalous sentences, and paragraphs. Intelligibility of phonemes and words was assessed using 150 CVC syllables from the Modified Rhyme Test (MRT, 2). Listeners identified each CVC, and then rated the clarity of its articulation (from 1 to 7). Sixty items from the Harvard Sentences (3) were used to evaluate speech perception in meaningful context. Each sentence was transcribed and rated for naturalness (1-7). Semantically anomalous sentences (60 items) were taken from the Haskins Laboratories Syntactically Normal Sentence Test (4). Listeners transcribed 4 target words from each anomalous sentence. Finally, perception of discourse was evaluated using ratings of intelligibility (1-100) and naturalness (1-7) for three passages (5).

RESULTS AND DISCUSSION

Intelligibility in the MRT was moderate (see Figure 1, left panel). Accuracy for items produced by the Elan engine was poorer than for items from the other engines (p<.01). The Microsoft engine also produced more clearly articulated speech than the other engines (p<.01), and items were more clearly articulated by the AT&T engine than the Elan engine (p<.01; see Figure 1, right panel).

The intelligibility of phonemes within words was consistent with the overall pattern of MRT results. Medial vowels produced by the Microsoft and AT&T engines were perceived more accurately than from the Elan engine.
Word-final consonants produced by the Microsoft engine were perceived more accurately than phonemes produced by the other engines \((p<0.05)\). The effect of engine was marginal for initial consonants.

Words in meaningful sentences were quite intelligible (see Figure 2, left panel); near perfect performance was obtained for the Microsoft and AT&T engines. Naturalness ratings (see Figure 2, middle panel) ranged from poor \((2.29)\) to moderate \((4.07)\). Both intelligibility and naturalness were reduced for the Elan engine \((p<0.01)\). Anomalous sentences were less intelligible than meaningful sentences (see Figure 2, right panel). Accuracy was greater for items from the Microsoft engine than the other engines \((p<0.05)\).

Finally, the passages were rated as highly intelligible, although somewhat less so for the Elan engine \((p<0.05;\) Figure 3, left panel). However, listeners still judged the discourse to be unnatural (see Figure 3, right panel). Items from the Elan engine were rated as less natural than items from the other engines \((p<0.05)\).

CONCLUSIONS

The Microsoft engine generated the highest scores on all dependent measures. There are several potential explanations for this finding. First, the Microsoft engine may have a better algorithm for synthesizing units of speech, resulting in fewer phoneme confusions. Second, the Microsoft engine used a higher sampling rate. Third, the slower speaking rate of the Microsoft engine could have improved clarity. Finally, differences between engines in prosodic boundaries also may have influenced intelligibility and naturalness in sentences and paragraphs. While it is unknown which factor(s) contributed to the current results, these results do indicate that (a) TTS speech is relatively unnatural, and (b) ratings of clarity and naturalness paralleled intelligibility across engines. It thus appears that such ratings represent useful measures in assessing the quality of speech produced by TTS engines.

REFERENCES