IMPLICIT MEASUREMENT OF INTELLIGIBILITY OF MALE AND FEMALE VOICE TEXT-TO-SPEECH (TTS) SYNTHESIS IN NOISE USING A PHONEME DETECTION TASK

Nicole Lees ¹, Catherine Stevens¹, & Julie Vonwiller ²

¹MARCS Auditory Laboratories, University of Western Sydney
²APPEN Speech Technology

ABSTRACT: Given the increasing application of TTS synthesis in commercial and clinical settings, there is a need to develop methods of evaluation from the user's perspective. An experiment is reported that compares the effect of two factors, voice gender and signal quality, on the intelligibility of three TTS systems from the user’s point of view. It was hypothesised that male voiced TTS would be more intelligible than female voiced TTS and that low quality signals would reduce intelligibility. Accuracy scores were obtained from a phoneme detection task using Semantically Unpredictable Sentences (SUS). The results indicate that an interaction between voice gender and signal quality is dependent on the TTS system. The results suggest that intelligibility, from the user’s point of view, is modulated by several factors. Consequently there is a need to evaluate TTS systems in the context of commercial requirements and specific applications.

Text-to-Speech (TTS) synthesis is the automated process that produces spoken output from input text characters (Greene, Logan & Pisoni, 1986). The increasing number and uses of TTS in telecommunications, information services, and applications for the disabled give rise to a need for evaluation procedures that can reliably discriminate the effectiveness of the systems in various contexts (van Santen, Pols, Abe, Hahn, Keller, & Vonwiller, 1998). This need is brought into sharp focus by the cost of TTS systems and because many TTS systems are still less effective than systems using natural speech (Koul & Allen, 1993; Pols, van Santen, Abe, Kahn & Keller, 1998). The problem to be addressed is that conventional methods of evaluation do not consider ecological or contextual factors that may influence TTS effectiveness.

Evaluation by the end user involves several dimensions such as whether the TTS output can be understood (intelligibility), whether it sounds like natural speech (naturalness) or whether the user likes the TTS output or not (preference) (Koul & Allen, 1993; Pavlovic, Rossi, & Espesser, 1990), although more specific categories of evaluation have also been proposed (see Pols et al, 1998). These dimensions of evaluation can be applied to different levels of an utterance such as the phoneme (individual speech sound), syllable, word, sentence or paragraph level (Goldstein, 1995).

The typical methods used to examine intelligibility include the Diagnostic Rhyme Test (DRT) and its derivative the Modified Rhyme Test (MRT; Voiers, 1983). Both the DRT and the MRT employ a two-alternative forced-choice discrimination task using monosyllabic words that differ in initial or final phoneme in a word (e.g. pack-pat). Delogu, Conte, and Sementina (1998) have criticized these methods for not adequately representing continuous speech and for the limited number of response alternatives they afford. Other tests such as the Standard Segmental Test and the Minimal Pairs Intelligibility Test (van Santen, 1993) include more response alternatives. Alternatively, some tests have employed a dictation task, for example the Spelling Alphabet Test (SpAT), and Phonetically Balanced Word Lists (PB). While beneficial in measuring intelligibility at the phoneme level, these tests remain limited in their representation of continuous speech.

Recent studies examining intelligibility have found differences between two synthesisers (DECtalk and Macintalk) producing male, female and child voices (Hustad, Kent & Beukelman, 1998). The descriptive results that were provided indicated that intelligibility scores are dependent on voice gender and interact with the system - in Macintalk a female voice appears to be more intelligible than the male voice and vice versa for DECtalk. The implication here is that voice gender appears to affect intelligibility and should be considered in selecting a TTS system. Indeed Reeves and Nass (1996) point to a possible interaction between voice gender and message content. However, the aim of the Hustad et al
experiment was to compare TTS systems rather than exploring the differences in intelligibility across gender itself. Given this emphasis, and the lack of statistical tests for gender it is not yet clear whether voice gender significantly influences intelligibility. An earlier study by Koul and Allen (1993) acknowledged the value of comparing male and female voicing produced by DECTalk, finding that the male voice was more intelligible than the female voice. However, systematic investigation of voice gender in other TTS systems is yet to be made.

Koul and Allen (1993) also examined the effect of noise on intelligibility by adding twelve-talker speech babble to the speech signal. Intelligibility increased as the signal-to-noise ratio increased suggesting that background noise interferes with intelligibility. Interestingly, the results also suggested that the impact of added noise is more deleterious on the DECTalk male voice than the DECTalk female voice. That is, the female voice was slightly more intelligible than the male voice under noisy conditions (a reversal of order compared with the no noise condition). The idea that noise and voice gender factors interact to affect TTS intelligibility deserves further attention especially in relation to other TTS systems. Moreover, if ecological factors are to be considered then it is appropriate to develop a method that is more representative of the natural flow of speech in real-world environments such as the office. Therefore, an alternative to the MRT method used by Koul and Allen (1993) is necessary.

The use of continuous or running speech is particularly important because “durations of phonetic segments strongly depend on contextual factors such as the identities of surrounding segments, stress, accent, and phrase boundaries” (Bellagarda & Silverman, 1998, p.1). Kalikow, Stevens and Elliot (1977) developed a test of synthetic speech presented under varying noise conditions (Speech In Noise test: SPIN). The test items in this case were sentences rather than monosyllabic words. The predictability of the sentences was manipulated (low, medium and high predictability) to assess the impact of context (use of linguistic-situational cues) on intelligibility. Participants listened to sentences and were asked to write down the final word in each sentence. There was a significant effect of both predictability and noise, with low predictability and high noise reducing intelligibility. However, the effect of noise was more prominent in low predictability sentences and did not have a significant effect on highly predictable sentences. More recently, the use of Semantically Unpredictable Sentences (SUS; Benoit, Grice, & Hazan, 1996) in a dictation task has been recommended for evaluating intelligibility of TTS synthesis in the context of continuous speech (van Santen et al, 1998).

The SUS method preserves sentence syntax while disrupting the semantic cohesiveness of the sentence, for example “The table walked through the blue truth” (Benoit et al, 1996). The disruption to the semantic flow prevents the perceiver from using context cues to understand the utterance. Thus differences in listeners’ perception can be attributed to changes in variables being manipulated (such as voice gender, signal quality) and not variations in word or item probability. A dictation task involving SUS sentences is a difficult task and is therefore useful in discriminating between TTS systems that appear close in intelligibility on other tests. However, it has been suggested by Kalikow et al (1977) that “A test of a listener’s ability to understand everyday speech must…assess both the acoustic-phonetic and the linguistic-situational components of the process” (p.1). This can be achieved by combining SUS sentences with an alternative task that targets intelligibility at the phoneme level. An ideal method is the phoneme detection task that has been established in the field of speech perception (Cutler, 1976; Costa, Cutler, & Sebastian-Galles, 1998).

The phoneme detection task involves searching for, and indicating the occurrence of, a particular target phoneme embedded in an utterance (Costa, et al 1998). A wide selection of phoneme targets is possible, with flexibility in placing these targets in various positions in both the word and the sentence. This makes the phoneme detection task suited to SUS material. An additional benefit is that control or catch sentence trials (that do not contain the target phoneme) can be used to obtain a precise measure of accuracy. This provides an implicit measure of intelligibility, which is operationalized as the proportion of correct identifications (hit rate) minus the proportion of false positives (false alarm rate). A false positive refers to responding as if a target phoneme is present when it is in fact absent.

To date, the potential interaction between TTS system, voice gender and signal quality has not been explored systematically. The present study aims to assess intelligibility of male and female voiced TTS systems in conditions of high and low signal quality focusing on the phoneme level in continuous speech. It is hypothesized that male voicing will be more intelligible than female voicing and that high
quality signals will be more intelligible than low quality (noise added) signals. However, it is likely that there will be interactions between voice gender and signal quality in different TTS systems.

METHOD

Participants

Sixty undergraduate Psychology students (53 females and 7 males) from the University of Western Sydney, Bankstown participated for 1% course credit. Participants had a mean age of 22.58 yrs (SD = 7.17yrs) with self-reported normal hearing and no experience in phonetics or any training with synthetic speech. All were native speakers of English.

Materials

Three SUS sentence structures (intransitive, imperative, interrogative; see Benoit et al, 1996), were used as a basis to construct a total of 72 novel sentences between 6 and 8 words in length. Half of the sentences (experimental sentences) contained one of nine target phonemes (/p/, /f/, /m/, /dZ/, /k/, /c/, /i/, /a/, /u/) in either the beginning, middle or end of a word, that was in turn either at the beginning, middle or end of the sentence. The remaining 36 control sentences did not contain the target phoneme.

Words containing targets were controlled according to written and spoken frequency using the Cielex Lexical Database (Baayen, Piepenbrock & Gulikers, 1995). Sentences were synthesised from three anonymous TTS systems (A, B and C), creating a pool of 216 sentences. For each system, half of the sentences were spoken in a female voice, the other half in a male voice. Half of the sentences in each voice gender were then mixed with white noise in Cooledit96 to create a low signal quality condition (3.89 dB signal to noise ratio). Natural recordings and synthetic versions of six sentences (in a male or female voice) were used as practice items. The 216 sentences were divided into three balanced versions of the experiment for presentation through Sennheiser HD450II headphones using DMDX version 2.9 experimental software (Forster, 2002) on a Trident (PIII) PC.

Procedure

Participants were allocated randomly to one of three experiment versions and tested individually. Participants were instructed to listen for the sound of a letter target (provided visually on screen) and to press a key marked YES as soon as they heard the target phoneme, or to press NO at the end of the sentence if they did not detect the target. To encourage vigilance, correct/incorrect feedback was given on each trial. Each experimental version contained 20 practice items (10 natural and 10 synthetic) followed by three blocks of 24 sentence trials. A 300 msec warning tone preceded each sentence by 1 sec. Presentation of targets was randomised between trials. The experiment took 25 minutes.

RESULTS

Mean accuracy scores are shown in Table 1, as a function of TTS System, Quality and Gender. In accordance with the hypotheses intelligibility was significantly higher for male voicing (M = .434, SD = .068) than female voicing (M = .361, SD = .072), F(1, 59) = 9.30, p<.01, and was higher for high quality signals (M = .453, SD = .069) than low quality signals (M = .342, SD = .075), F(1, 59) = 7.48, p< .01. There was also a significant main effect for TTS system, F(2, 59) = 7.57, p< .01.

A significant three-way interaction indicates that the effects of voice gender and signal quality varies across the three TTS systems, F( 2, 59) = 6.83, p<.01. As can be seen in Table 1 there was little difference between male and female voicing for TTS system A in the high quality condition. However, in the low quality condition, accuracy appears to be better for the male voice than the female voice. In TTS system B, there was a tendency for accuracy to be better for male voices than female voices in both quality conditions.
Table 1. Mean accuracy scores (hit rate – false alarm rate) for male and female voiced TTS synthesis in high and low quality signal conditions.

<table>
<thead>
<tr>
<th></th>
<th>High Quality</th>
<th>Low Quality</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>TTS A</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Male</td>
<td>.396</td>
<td>.271</td>
</tr>
<tr>
<td>Female</td>
<td>.390</td>
<td>.175</td>
</tr>
<tr>
<td><strong>TTS B</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Male</td>
<td>.515</td>
<td>.486</td>
</tr>
<tr>
<td>Female</td>
<td>.429</td>
<td>.384</td>
</tr>
<tr>
<td><strong>TTS C</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Male</td>
<td>.611</td>
<td>.328</td>
</tr>
<tr>
<td>Female</td>
<td>.378</td>
<td>.407</td>
</tr>
</tbody>
</table>

For TTS system C, the effect of voice gender and signal quality on accuracy scores was somewhat different. Figure 1 shows that accuracy was significantly better for male than female voicing in the high quality condition, $t(58) = 3.06, p< .01$. Interestingly, low quality female voices appear to be more intelligible than low quality male voices and high quality female voices, although this effect was not significant.

![Figure 1: Accuracy scores for male and female voicing in high and low quality signals for TTS System C](image-url)
DISCUSSION

While there are many TTS systems in the market today, it is widely agreed that no product is mature and of outstanding quality. Until a user installs a specific TTS system in their environment, they do not know its suitability, capabilities, strengths and weaknesses. At present there is no systematic information available to indicate what user perception and preferences are, and what factors significantly affect perception and performance. What is needed is an independent mechanism for the evaluation of TTS systems that will provide an analysis of assured quality. This study is a first step in applying the established experimental paradigms of cognitive psychology to investigate TTS intelligibility and effectiveness. The long term goal is to develop a suite of programs that can be used in a variety of settings to evaluate and compare banks of TTS systems.

The first aspect that is clear from the present study is that there was considerable but systematic variation in intelligibility scores. The pattern of results suggests that the method – phoneme detection using semantically unpredictable sentences – is a sensitive tool to partial out effects of voice gender, signal quality and TTS system. We have shown that the gender of the TTS system voice and the quality of the signal do affect intelligibility of TTS synthesis. Generally, male voicing is more intelligible than female voicing, and high quality signals are more intelligible than low quality signals. More importantly, there is evidence that the effect of voice gender and signal quality is not consistent in all TTS systems. For example, voice gender affected intelligibility of TTS Systems B and C presented as high quality signals whereas there was no effect of voice gender in TTS System A presented as a high quality signal. Different TTS systems appear to have strengths under particular conditions. This finding highlights the need to assess TTS systems relative to their intended application and environment, not just in some standard or ideal setting.

Our results also suggest that voice gender and signal quality can interact within a particular TTS synthesis system. This is most apparent in the case of TTS System C where a male voice is more intelligible than a female voice in high quality signal conditions, but in a low quality signal condition the female voice is more intelligible than the male voice. Furthermore, the female voice as a low quality signal was more intelligible than a female voice as a high quality signal. The implication of this result is that TTS systems that are to be used in noisy environments need to be evaluated in appropriately noisy, real-world contexts.

One aim of the present study was to consider ecological factors in evaluation of TTS system effectiveness. These first results of their type suggest that ecological validity is important not only in exploring signal and environmental variables that affect TTS effectiveness but also in the methods used to measure system effectiveness. There are limitations to the present study that will be addressed in future studies. First, we have manipulated only two of many possible ecological and contextual variables. It is not uncommon to be performing other tasks, such as paperwork or even driving a car, while listening to messages. Intelligibility when divided attention is required and cognitive load varies will be investigated. Signal quality and voice variables will also be explored in more depth. For example, different types of background noise and specific voice characteristics will be manipulated. Second, the ecological validity of utterance content warrants consideration. A phoneme detection task in continuous speech is valuable because it includes both the acoustic-phonetic and linguistic components of speech perception while controlling the effect of context cues in the identification of speech sounds. However, SUS sentences are not natural in content. It is necessary to use utterance types that are characteristic of specific applications and environments. Finally, intelligibility is one important dimension of TTS synthesis effectiveness. Changes in preference and naturalness ratings as a function of signal quality and voice gender will now also be determined.

ACKNOWLEDGEMENTS

This research was supported by a grant from the University of Western Sydney Research Partnerships Scheme and Appen Pty Ltd. The authors thank Denis Burnham and Caroline Jones for assistance with stimuli and experimental design. Further information may be obtained from Catherine (Kate) Stevens, email: kj.stevens@uws.edu.au; www.uws.edu.au
REFERENCES


