Available online at www.sciencedirect.com
SCIENCE DIRECT®
Computer Speech and Language 19 (2005) 129–146
COMPUTER SPEECH AND LANGUAGE
www.elsevier.com/locate/csl

On-line experimental methods to evaluate text-to-speech (TTS) synthesis: effects of voice gender and signal quality on intelligibility, naturalness and preference

Catherine Stevens [a,*], Nicole Lees [a], Julie Vonwiller [b], Denis Burnham [a]

[a] *MARCS Auditory Laboratories, School of Psychology, University of Western Sydney – Bankstown campus, Locked Bag 1797, Penrith South DC, NSW 1797, Australia*
[b] *Appen Pty Ltd., NSW, Australia*

Received 6 August 2003; received in revised form 15 March 2004; accepted 17 March 2004
Available online 19 June 2004

Abstract

Three experiments are reported that use new experimental methods for the evaluation of text-to-speech (TTS) synthesis from the user's perspective. Experiment 1, using sentence stimuli, and Experiment 2, using discrete "call centre" word stimuli, investigated the effect of voice gender and signal quality on the intelligibility of three concatenative TTS synthesis systems. Accuracy and search time were recorded as on-line, implicit indices of intelligibility during phoneme detection tasks. It was found that both voice gender and noise affect intelligibility. Results also indicate interactions of voice gender, signal quality, and TTS synthesis system on accuracy and search time. In Experiment 3 the method of paired comparisons was used to yield ranks of naturalness and preference. As hypothesized, preference and naturalness ranks were influenced by TTS system, signal quality and voice, in isolation and in combination. The pattern of results across the four dependent variables – accuracy, search time, naturalness, preference – was consistent. Natural speech surpassed synthetic speech, and TTS system C elicited relatively high scores across all measures. Intelligibility, judged naturalness and preference are modulated by several factors and there is a need to tailor systems to particular commercial applications and environmental conditions.

* Corresponding author. Tel.: +61-2-9772-6324; fax: +61-2-9772-6736.
E-mail address: kj.stevens@uws.edu.au (C. Stevens).

0885-2308/$ - see front matter © 2004 Elsevier Ltd. All rights reserved.
doi:10.1016/j.csl.2004.03.003

1. Introduction

Text-to-speech (TTS) synthesis is the automated process that produces spoken output from input text characters (Greene et al., 1986). The increasing number and uses of TTS in telecommunications, information services, and applications for the disabled give rise to the need for evaluation procedures that can reliably discriminate the effectiveness of the systems in various contexts (van Santen et al., 1998). The quality of TTS synthesis has increased over recent years (Gong and Lai, 2003), nonetheless, voice portals are far from perfect with human listeners detecting easily the odd intonation, clipped tones, and poorer clarity and prosody of synthesis relative to natural speech (Bailly, 2002; Koul and Allen, 1993; Pols et al., 1998).

The experimental paradigms of perceptual and cognitive psychology can be adapted to contribute to the development, evaluation, and refinement of TTS systems. For example, working with relatively early systems, Ralston et al. (1995) examined perceptual intelligibility, capacity demands of processing synthetic speech, and the effects of training. It is timely for these experimental methods to be applied to the new generation of TTS systems in an ecologically valid context (Cole et al., 1997; Roediger, 1997; Venkatagiri, 2003). Importantly, factors such as listeners' preference for different TTS systems and judgments about the naturalness of synthetic speech need also be examined methodically (Nass and Lee, 2001; Stern et al., 1999). Stern et al., for example, note that there is a substantial relationship between listeners' preferences for a particular synthetic speech system and the intelligibility of that system. Three experiments are reported that investigate the effect of the variables TTS system, voice gender, and signal quality on implicit measures of intelligibility (accuracy and speed), and judgments of naturalness and preference. The ultimate goal is to develop tasks and paradigms that can be used in a range of situations to evaluate an individual system or banks of TTS systems accurately, efficiently, and from the user's perspective.

1.1. Effects of voice gender and noise on intelligibility of synthetic speech

Hustad et al. (1998) compared the intelligibility of two synthesizers, DECtalk and Macintalk, using male, female and child voices. The descriptive results indicate that intelligibility scores are dependent on voice gender and interact with the system – in Macintalk a female voice appears to be more intelligible than a male voice and vice versa for DECtalk. The implication is that voice gender appears to affect intelligibility and should be considered in selecting a TTS system. However, the aim of Hustad, Kent and Beukelman was to compare TTS systems rather than explore the differences in intelligibility across gender. Given this emphasis, and a lack of statistical analysis of voice gender, it is not clear whether voice gender influences intelligibility significantly and/or consistently.

To examine the effect of noise on intelligibility, Koul and Allen (1993) added twelve-talker speech babble to the speech signal. Intelligibility increased as the signal-to-noise ratio increased suggesting that background noise interferes with intelligibility. The results also implied 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). Poorer signal quality is likely to increase demands on an operators' attentional capacity during performance of a task

(Kahneman, 1973; Pashler, 1998; Venkatagiri, 2003). The possibility that noise and voice gender factors interact and affect TTS intelligibility deserves further analysis especially in relation to other contemporary TTS systems.

1.2. Developing on-line methods to measure intelligibility from the user's perspective

Methods used to examine TTS synthesis 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., pad-pat). Delogu et al. (1998) have criticized these methods for not adequately representing continuous speech and for the limited number of response alternatives they afford. Other tests have employed a dictation task, for example the Spelling Alphabet Test (SpAT), and Phonetically Balanced Word Lists (PB). While beneficial in measuring intelligibility behaviourally and at the phoneme level, these tests are unnatural and remain limited in their representation of continuous speech.

The use of continuous or running speech is important both for ecological validity (i.e., being characteristic of real world settings) and because "durations of phonetic segments strongly depend on contextual factors such as the identities of surrounding segments, stress, accent, and phrase boundaries" (Bellegarda and Silverman, 1998). Kalikow et al. (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 et al., 1996) in a dictation task has been recommended for evaluating intelligibility of TTS synthesis in the context of continuous speech (Pols 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" and "Draw the house and the fact" (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 (e.g., voice gender, signal quality) and not to 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 and that minimizes the role of memory. One suitable method is the phoneme detection task that has been established in the field of speech perception (Cutler, 1976).

The phoneme detection task involves searching for, and indicating the occurrence of, a particular target phoneme embedded in an utterance (Cutler, 1976). 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 trials (that do not contain the target phoneme) can be used to obtain a precise measure of accuracy. This provides an on-line (low memory demand) measure of intelligibility that is implicit or hidden from the participants' awareness. Intelligibility was operationalized in Experiment 1 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.

2. Experiment 1

The aim of Experiment 1 was to investigate the intelligibility of male and female voice TTS systems in conditions of high and low signal quality focusing on the phoneme level in continuous speech. It was hypothesized that high quality signals are more intelligible than low quality (noise added) signals. The main effect of voice gender, and interactions between voice gender, signal quality and TTS system, were also investigated.

2.1. Method

2.1.1. Participants

Participants were 60 undergraduate Psychology students (53 females and 7 males) from the University of Western Sydney. Participants had a mean age of 22.58 years (SD = 7.17 years) with self-reported normal hearing and no experience in phonetics or any training with synthetic speech. All were native speakers of English.

2.1.2. Materials

Three SUS sentence frameworks (intransitive, imperative, interrogative; see Benoit et al., 1996) were used as a basis to construct a total of 72 novel sentences between six and eight words in length. Half of the sentences (experimental sentences) contained one of nine target phonemes (/p/, /f/, /i/, /m/, /dʒ/, /u/, /k/, /ɹ/, /ɑ/) in 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. The stimuli are listed in Appendix A.

Words containing targets were controlled according to written and spoken frequency using the Celex Lexical Database (Baayen et al., 1995). Sentences were synthesized from three anonymous concatenative TTS systems (A, B and C), creating a pool of 216 sentences. Voice accent was constant across systems A, B, and C. 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 quality condition (mean signal to noise ratio (SNR) = 8.79 dB, SD = 0.90). The 216 sentences were divided into three balanced versions of the experiment such that participants heard one of three experiment versions comprising 72 novel sentences with equal combinations of the variables system, gender and quality. The 216 sentences were presented across the sample with individual participants responding to a subset of 72 of the 216 sentence set. Natural recordings and synthetic versions of five sentences (in a male or female

voice) were used as practice items. However, to keep experiment length manageable natural speech was not used in the experimental trials.

2.1.3. Equipment

The experiment was programmed in DMDX v.2.9 experimental software and presented to participants on a Trident (PIII) PC through Sennheiser HD450II headphones.

2.1.4. Procedure

Participants were allocated randomly to one of three experiment versions and tested individually. They 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. Practice items and examples of errors were given to minimize errors based on orthography (e.g., false alarm by responding to the target phoneme "t" when words such as "thin" or "catch" were heard). Feedback was given on each trial to encourage vigilance. Each experiment version contained 10 practice items (five natural and five synthetic) followed by three blocks of 24 sentence trials. A 300 ms warning tone preceded each sentence by 1 s. Presentation of targets was randomised between trials. The experiment took 25 min.

2.2. Results

Mean accuracy scores, calculated as a discrimination index (DI), are shown in Table 1 as a function of TTS system, quality and gender. A DI of zero represents chance. Intelligibility was significantly higher for male voice ($M = 0.434$, $\mathrm{SD} = 0.068$) than female voice ($M = 0.361$, $\mathrm{SD} = 0.072$), $F(1, 59) = 9.30$, $p < 0.01$ and, as hypothesized, greater for high quality signals ($M = 0.453$, $\mathrm{SD} = 0.069$) than low quality signals ($M = 0.342$, $\mathrm{SD} = 0.075$), $F(1, 59) = 7.48$, $p < 0.01$. Natural speech items were compared directly with synthetic speech in the practice trials only. There was no evidence of a difference between the mean DI for natural ($M = 0.65$,

Table 1
Experiment 1: Mean accuracy scores (hit rate – false alarm rate) for male and female voice TTS synthesis in high and low quality signal conditions

	High quality	Low quality
System A		
Male	0.396	0.271
Female	0.390	0.175
System B		
Male	0.515	0.486
Female	0.429	0.384
System C		
Male	0.611	0.328
Female	0.378	0.407

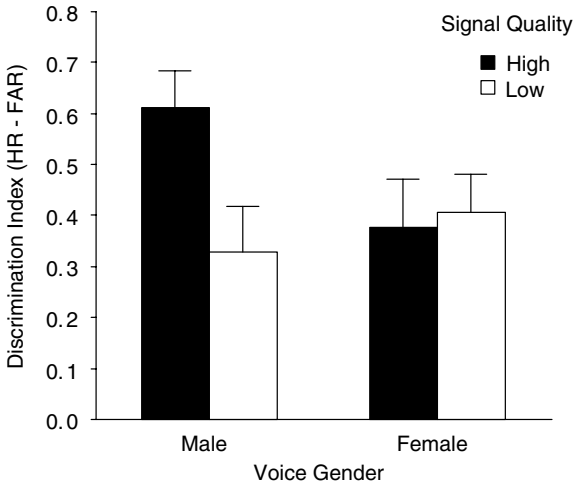

Fig. 1. Experiment 1: Accuracy scores for high and low quality male and female voice in TTS system C.

SD = 0.40) versus synthetic speech ($M = 0.60$, SD = 0.38), $F(1, 56) = 1.12$, $p > 0.05$ in the practice trials.

A significant three-way interaction indicates that the effects of voice gender and signal quality varies across the three TTS systems, $F(2, 59) = 6.83$, $p < 0.01$. As can be seen in Table 1 there was little difference between male and female voice 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 voice than female voice in both quality conditions. For TTS system C, the effect of voice gender and signal quality on accuracy scores was different. Fig. 1 shows that accuracy was significantly better for male than female voice in the high quality condition $t(58) = 3.06$, $p < 0.01$. 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.

The overall mean reaction time (recorded from the onset of the target phoneme to participant response) was 1323.65 ms. There were significant main effects of system and voice on reaction time (RT). Specifically, system C attracted the shortest RT ($M = 1297.51$ ms) followed by system B ($M = 1309.78$ ms) and then system A ($M = 1363.67$ ms), $F(2, 126) = 4.47$, $p < 0.05$. Reaction times were shorter in response to the male voice ($M = 1076.84$ ms) than the female voice ($M = 1178.31$ ms), $F(1, 126) = 27.10$, $p < 0.05$.

2.3. Discussion

The results of Experiment 1 indicate 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. As hypothesized, the voice gender of the TTS system and the quality of the signal do affect intelligibility of TTS synthesis. Generally, male voices are more intelligible than female

voices, 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 a standard or ideal setting.

The accuracy scores also suggest that voice gender and signal quality may interact within a particular TTS synthesis system. This was most apparent in the case of TTS system C where a male voice was more intelligible than a female voice in high quality signal conditions, but in a low quality signal condition the female voice was more intelligible than the male voice. Furthermore, in system C 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.

The use of SUS handicaps the TTS synthesis systems in that their production of a phrase such as "blue sky" is likely to have been checked and be more intelligible than the less probable "blue truth". Nonetheless, the method is an effective tool for fine-grained discrimination between systems and for examination of other variables that may interact. A future experiment could treat the degree of unpredictability of word clusters, phrases or sentences as an independent variable. Items from the domain in which the synthesis systems are to be used could also be adapted as stimuli.

Experiment 2 further investigated the variables, synthesis system, voice gender and signal quality, using ecologically valid, word categories as stimuli. Möbius (2003) states that a serious challenge for speech synthesis is the systematic treatment of events in speech and language that have low frequencies of occurrence. Accordingly, the real-world "call centre" stimuli used in Experiment 2 consist of discrete numerals, first and last names, international and local place names presented both as synthetic TTS and natural speech. Experiment 2 used an auditory search task based on previous studies involving natural speech and the identification of syllables in meaningful sentences (Davis, 1967) and vowel sounds (Charleston and Boyer, 1990). The auditory search task involved detecting a pre-established target phoneme embedded in a string of stimuli. A negative correlation has been demonstrated between search time and accuracy (Charleston and Boyer, 1990; Davis, 1967) – stimuli that are harder to detect should be associated with longer search times. Search time was used in Experiment 2 as a second implicit behavioural measure of intelligibility.

3. Experiment 2

The aim of Experiment 2 was to investigate the effect of TTS synthesis system, voice gender and signal quality on auditory search time and accuracy. The factorial design consisted of four levels of system (TTS systems A, B, C, and natural speech), two levels of voice gender (male, female), and two levels of signal quality (low, high) with repeated measures on the voice gender and signal quality factors. It was hypothesized that high quality signals are more intelligible than low quality signals and that voice gender contributes a main effect. Voice gender, signal quality and TTS system were also expected to interact.

3.1. Method

3.1.1. Participants

Sixty-eight undergraduate psychology students (64 females and 4 males) from the University of Western Sydney participated for course credit ($M = 19.61$ years; SD = 1.88). Participants had self-reported normal hearing, were native speakers of English and had no experience in phonetics or training with synthetic speech.

3.1.2. Materials

A stimulus pool of 70 words from call centre relevant categories (numerals, first names, surnames, international and local place names, see Table 2) were used to create 80 lists of 12 words, 16 lists for each of the five target phonemes (/t/, /s/, /n/, /v/, /i/). These 16 were subdivided into four lists according to the combination of voice gender and background noise. Within each of the four lists, the word containing the target phoneme was placed in the list early (items 1–3), middle (items 4–9), late (items 10–12), or not at all (no-target list). The sets of lists were recorded in each of three TTS systems as well as natural speech. The natural speech was recorded digitally in a sound attenuated booth by one male and one female speaker. Both spoke with a moderate Australian accent and were trained in voice, the former in linguistics and the latter in singing. All files were normalised for intensity using CoolEdit96. The addition of white noise for the low quality condition resulted in a mean SNR of 6.33 dB (SD = 1.24). Lists were presented in a new random order to each participant

3.1.3. Equipment

The stimuli were presented to participants through Sennheiser HD450II headphones, using DMDX v.2.9 software on a Trident (PIII) PC.

3.1.4. Procedure

Participants were assigned randomly to one of four experimental groups (TTS system A, B, C, or natural speech) and tested individually. They were instructed to listen for the sound of a

Table 2
Experiment 2: Target phonemes and stimulus items

/t/	/s/	/n/	/v/	/i/
Tanner	Seven	Natalie	Victoria	Egypt
Ten	Steve	Newport	Vienna	Egan
Two	Sweden	Nicholas	Ventura	Eden
Twelve	Saville	Neil	Victoria	Eastwood
Natalie	Castlecove	Hundred	Seven	Sweden
Victoria	Helsinki	Manly	Davis	Thirty-one
Curtis	Mascot	Vienna	Eleven	Steve
Ventura	Eastwood	Seventy	Saville	Neil
Egypt	Davis	Seven	Castlecove	Fifty
Eight	Curtis	Sweden	Steve	Manly
Mascot	Nicholas	Thirty-one	Hargrave	Helsinki
Newport	Paris	Egan	Olive	Seventy

phoneme target (presented visually at the start of each trial) and to press either the "YES" key or "NO" key after each word in the list, as quickly and accurately as possible, according to whether they heard the target phoneme. A practice trial of one list containing a mixture of voice gender and signal quality types was presented prior to the experiment to familiarize participants with the procedure. The 80 lists were divided into four groups of 20 lists to allow for breaks. The experiment took 30 min.

3.2. Results

Reaction time in a search task may be influenced by starting and stopping time (Latimer, 1972). This potential problem was circumvented by calculating search time across six items in the centre of late-target and no-target lists. That is, search time was calculated as the average time taken to search through items in positions 4–9 on late-target (after item 9), and no-target lists; dummy trials, where a target occurred in early or middle items of a list, were discarded from the analysis. Therefore, search time refers to the time taken to search through six items as a function of system, voice and quality, untainted by the presence of a target or the motor response required to react to a target. Accuracy was calculated as hit rate (HR) minus false alarm rate (FAR). A score of 1 reflects 100% accuracy and 0 is chance (equal HR and FAR).

Search times were analysed using a three-way analysis of variance. A significant main effect for voice gender was evident although, unlike Experiment 1, the means indicate that search time was shortest for female voice, $F(1,67) = 52.60$, $p < 0.01$. Accuracy, on the other hand, was significantly greater in response to male (0.73) than female (0.71) voice, $F(1,64) = 7.02$, $p < 0.01$. As hypothesized, a main effect of noise indicated that intelligibility was significantly greater in the high quality condition than in the low quality condition evident in both search time $F(1,67) = 23.21$, $p < 0.01$, and accuracy $F(1,64) = 50.03$, $p < 0.01$. Search time also reflected a main effect of system, $F(3,67) = 3.23$, $p < 0.05$. Search time was significantly faster in the natural speech condition than in the three TTS conditions, $F(1,67) = 8.32$, $p < 0.01$.

Of most interest are the interactions that occur between system and signal quality, and between signal quality and voice gender. There was a significant voice gender x system interaction evident in search time, $F(1,67) = 4.37$, $p < 0.01$ (Fig. 2), and mirrored in accuracy $F(1,64) = 3.56$, $p < 0.05$. In search time, there was also a significant voice gender x signal quality interaction $F(1,67) = 13.90$, $p < 0.01$. A post hoc analysis revealed that search time was facilitated when a female voice was presented as a high quality signal, $F(1,67) = 4.37$, $p < 0.01$. Fig. 3 illustrates that this combination of female voice and high signal quality was most effective whereas the other three conditions recorded approximately equal search times. System C was most intelligible under these conditions.

3.3. Discussion

An auditory search task has been used as a new method for on-line measurement of intelligibility using ecologically valid call centre items as stimuli. The results indicate that search time slows and accuracy decreases when targets are presented under noisy conditions. Search time was shortest for the female compared with the male voice in all three TTS synthesis systems and the natural speech condition. The male voice, in general, led to relatively good accuracy. Notably,

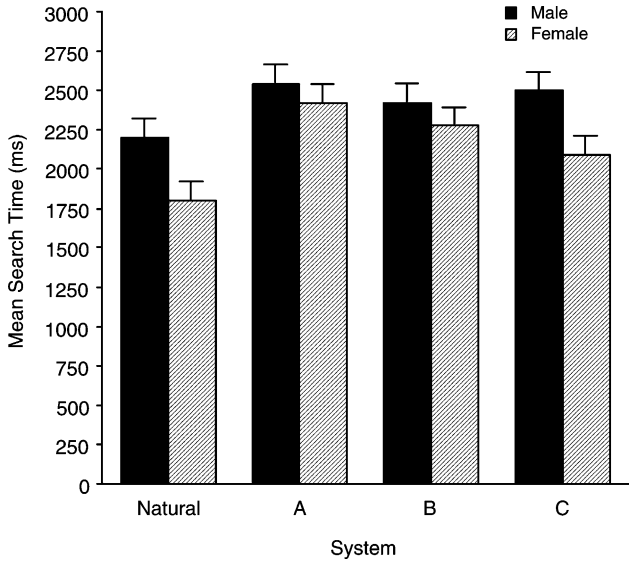

Fig. 2. Experiment 2: Search time as a function of speech system and voice gender.

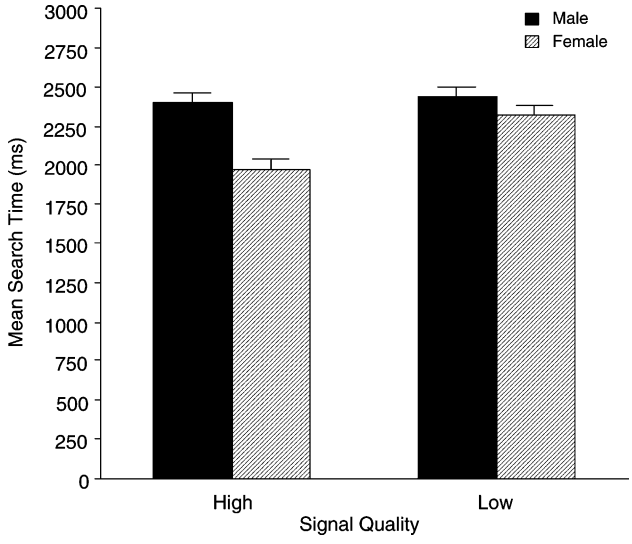

Fig. 3. Experiment 2: Mean search time interaction between voice gender and signal quality.

under noisy conditions, the female voice in TTS synthesis system C led to the shortest search times and greatest accuracy. With a predominance of female participants in the present sample, the results accord with those of Nass et al. (2003) wherein females were more sensitive to differences between synthetic and recorded speech.

Measures of naturalness and preference for sentence stimuli, generated by the three TTS synthesis systems and natural speech, were the focus of Experiment 3. TTS system (A, B, C,

natural speech), voice gender (male, female), and signal quality (low, high) were manipulated as independent variables. The method of paired comparisons (Cox, 1958; Hansen and Kollmeier, 1998, Wherry, 1938) was used in which participants make a series of relative judgments about the naturalness of, and their individual preference for, sentences generated by different TTS systems. The relative ranking of the TTS stimuli can then be deduced from the series of pairwise judgments. The aim of Experiment 3 was to investigate the possible differential effects and interactions of system, signal quality, and voice gender on independent rankings of naturalness and preference.

4. Experiment 3

4.1. Method

4.1.1. Participants

Eighty undergraduate Psychology students (66 females and 14 males) from the University of Western Sydney, Bankstown participated for 1% course credit. Participants had a mean age of 21.71 yrs (SD = 7.47 years) with self-reported normal hearing and no experience in phonetics or any training with synthetic speech. All were native speakers of English.

4.1.2. Materials

Two semantically unpredictable sentences "Drink the sky and the plant" and "Throw the day and the scream" were recorded from a male and female actor as well as synthesised from three anonymous TTS systems (A, B and C). For each TTS system, the sentences were synthesised using a female and male voice creating a total high quality pool of 16 sentence recordings (eight of each sentence). The low quality sentence items were created by mixing copies of the recordings with white noise using Cooledit96 (SNR $M = 8.63$, SD = 2.77). The 16 versions of each sentence were then paired according to all possible variations in only one of the system, gender or quality variables while the remaining variable levels were held constant. For example, the TTS A high quality male recording was paired with TTS B, C and Natural high quality male recordings, as a contribution towards examining the effect of system. Each pair was also reverse ordered resulting in a block of 80 randomly presented pairs for each sentence.

The two blocks of sentences were presented with a different set of instructions to avoid participant fatigue and to assist in separating the tasks. The instruction was either "Choose the sentence YOU PREFER out of the following pair" or "Choose the sentence YOU think is the most NATURAL sounding out of the following pair". The order of instructions and the instruction/sentence pairing were presented in all of four combinations across participants. Each version contained eight examples of high and low quality natural recordings (in a male or female voice) as practice items.

4.1.3. Equipment

The experiment was presented through Sennheiser HD450II headphones using DMDX version 2.9 experimental software on a Trident (PIII) PC.

4.1.4. Procedure

Participants were allocated randomly to one of four experiment versions and tested individually. Participants heard a pair of sentences (presented serially) and instructed to press one key marked either FIRST or SECOND according to which voice was rated most natural or most preferred. To further encourage vigilance and to avoid confusion between instructions, participants were encouraged to take a short break after each set of 20 items and were given a distracter task (memory for gestures) before starting with the alternative instructions. The experiment took 45 min.

4.2. Results

As all levels of the variables occurred an equal number of times in the paired comparisons, scores were calculated according to the proportion of times a variable was selected by each participant based on either preference or naturalness. A rating score (proportion) of 1 indicated that the factor was selected every time it appeared (i.e., most preferred or natural sounding), whereas a rating score of 0 indicated that the factor was never selected (i.e., least preferred or natural sounding).

The effects of system (TTS A, TTS B, TTS C, natural voice), voice gender (male, female) and signal quality (high, low) on naturalness and preference ranks were analysed using a $4 \times 2 \times 2$ within subjects analysis of variance (three-way ANOVA) for each dependent variable. The mean proportion ranks (out of a total of 1) for naturalness and preference obtained as a function of all three independent variables are shown in Table 3.

As shown in Fig. 4, there was a significant main effect of system type on both naturalness, $F(2.12, 80) = 269.43$, $p < 0.01$, and preference ranks, $F(2.05, 80) = 392.90$, $p < 0.01$. The natural

Table 3
Experiment 3: Mean proportions of naturalness and preference ranks

System x gender	Naturalness				Preference			
	High quality		Low quality		High quality		Low quality	
	M	SD	*M*	SD	*M*	SD	*M*	SD
Natural								
Male	0.81	0.11	0.63	0.13	0.74	0.15	0.62	0.10
Female	0.82	0.12	0.55	0.11	0.85	0.12	0.69	0.12
TTS A								
Male	0.58	0.13	0.48	0.16	0.54	0.12	0.42	0.12
Female	0.41	0.17	0.37	0.14	0.54	0.10	0.40	0.11
TTS B								
Male	0.31	0.15	0.28	0.23	0.43	0.11	0.32	0.17
Female	0.35	0.14	0.29	0.14	0.43	0.09	0.30	0.14
TTS C								
Male	0.52	0.21	0.43	0.13	0.55	0.14	0.38	0.09
Female	0.49	0.11	0.43	0.12	0.40	0.12	0.36	0.13

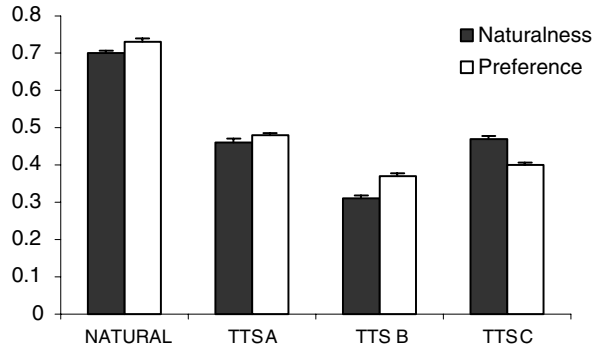

Fig. 4. Experiment 3: Mean naturalness and preference proportions for natural speech and three synthetic speech systems.

voice was ranked highest on both naturalness ($M = 0.70$, SD = 0.16) and preference ($M = 0.73$, SD = 0.15). TTS B was ranked lowest on both of these measures scoring a mean proportion of 0.31 (SD = 0.17) for naturalness and 0.37 (SD = 0.14) for preference. Post hoc analyses using Tukey's HSD revealed that there was no significant difference between TTS A and TTS C with respect to naturalness ranks. However, TTS A received significantly higher ranks on preference than TTS C.

There were also significant main effects for the two remaining independent variables, voice gender and quality. The male voice was ranked more highly than the female voice for naturalness $F(1, 80) = 28.33$, $p < 0.01$, although there was no significant difference between male and female voice with regard to preference. There was, however, an interaction of system with voice gender. The significant difference between male and female voices on the naturalness variable appears to be due to TTS A alone, with the male voice scoring a mean of 0.53 (MSE = 0.01) and the female voice scoring a mean of 0.39 (MSE = 0.01). There were no significant differences between male and female voice in any of the other systems.

As expected, high quality signals were consistently ranked more highly than low quality signals for both naturalness $F(1, 80) = 67.16$, $p < 0.01$ and preference $F(1, 80) = 209.51$, $p < 0.01$. However, the naturalness and preference ranks assigned to the low quality natural voice conditions often exceeded the ranks assigned to the high quality examples of synthetic voices.

4.3. Discussion

The paired comparison method used in Experiment 3 has enabled relative ranks of preference and naturalness to be deduced from a series of trials. Participants were not asked to explicitly rate the sentence stimuli (Viana et al., 2003) but rather the ranks were captured implicitly in a series of pairwise comparisons. As hypothesized, TTS synthesis system, signal quality and voice gender contributed individually and in combination to judgments of preference and naturalness. Natural speech surpassed all synthetic systems in relation to preference and naturalness. Importantly, system C ranked relatively highly when synthesized as either a male or female voice, and in the presence or absence of white noise. Each of the dependent measures recorded in Experiments 1–3

reflect a different aspect of perceptual performance – accuracy, speed, judged naturalness, preference. System C elicited high scores across all four measures.

5. General discussion

While there are many TTS synthesis 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 method for the evaluation of TTS systems that will provide an analysis of assured quality. This study applies the established experimental paradigms of cognitive psychology to investigate TTS intelligibility, naturalness and preference. The long-term goal is to develop a suite of programs that can be used in a variety of settings to evaluate an individual system or compare banks of TTS synthesis systems. Interactions between intelligibility and user preference may also be explored.

The results of Experiments 1 and 2 suggest that phoneme detection, using either sentence or call-centre stimuli, is a sensitive tool to partial out effects of voice gender, signal quality and TTS system. Specifically, the results indicate that the gender of the voice and the quality of the signal do affect intelligibility of TTS synthesis. In general, male voice is more intelligible than female voice, 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. This highlights the need to assess TTS systems relative to their intended application and environment, not just in a standard or ideal setting. Specifically, TTS systems that are to be used in noisy environments need to be evaluated in appropriately noisy, real-world contexts.

In addition to accuracy and search time as online, behavioural measures of intelligibility, preference and naturalness were deduced from relative ranks assigned to pairs of synthesized sentences (Experiment 3). This technique affords deduction of ratings without the use of explicit rating scales and is sensitive to the consistency of participant judgments. Just as TTS synthesis system, signal quality and voice gender affected accuracy and search time, so too were preference and naturalness ranks influenced by these variables, both in isolation and in combination. Importantly, the pattern of results across the three experiments indicates a consistency over the dependent measures. Specifically, system A and particularly system C yield good performance with respect to accuracy, search time, judged preference and naturalness.

One aim of the present study was to consider ecological factors in evaluation of TTS system intelligibility and naturalness. The results suggest that ecological validity is important not only in exploring signal and environmental variables that affect synthesis effectiveness but also in the methods used to measure system effectiveness. Other influential variables may now be investigated in naturalistic adaptations of the experimental materials used here – phoneme detection and auditory search tasks, the paired comparison method, semantically unpredictable sentences, and discrete call centre stimuli. For example, intelligibility of TTS synthesis measured in contexts that involve divided attention and high cognitive load are important real world issues (Delogu et al., 1998; Venkatagiri, 2003). Tasks that simulate the user situation should also be developed. Signal

quality and voice variables may be explored by manipulating different types of background noise and specific voice characteristics including prosody (Monaghan, 2003; Viana et al., 2003) and word rate. The four dependent measures discussed here – accuracy, search time, judged naturalness and preference – capture different aspects of human perceptual acuity and evaluative behaviour. The measures are recorded under conditions that minimise operator awareness and bias and provide readily interpretable indicators of system clarity and user preference.

Acknowledgements

This research was supported by the University of Western Sydney Research Partnerships Scheme and Appen Pty Ltd. The authors thank Caroline Jones, Bettina Keresztesi, Melinda Gallagher, Manjusha Vijayan, Rua Haszard Morris, Michael Tyler, Colin Schoknecht, Joseph Noyeaux and Marcus Mellick for assistance with stimuli, data collection or analysis, and Steve Greenberg and two anonymous reviewers for helpful comments. Conference papers reporting these findings have been presented at the Speech Science & Technology Conference, Melbourne 2002, the 8th Western Pacific Acoustics Conference (WESPAC8) Melbourne 2003, and Eurospeech 2003: 8th European Conference on Speech Communication & Technology, Geneva 2003. Further information may be obtained from Catherine Stevens, MARCS Auditory Laboratories, University of Western Sydney-Bankstown, Locked Bag 1797, Penrith South 1797, NSW, Australia. E-mail: kj.stevens@uws.edu.au; internet: http://marcs.uws.edu.au/.

Appendix A. Frameworks

1. Subject–verb–adverbial (intransitive structure)
   det + noun + verb (intr) + preposition + det + adj + noun
2. Q.word–verb–subject–direct object (interrogative structure)
   Quest.adv. + aux + det + noun + verb (trans) + det + adject + noun
3. Verb–direct object (imperative structure)
   verb (trans) + det + noun + conjunction + det + noun

Trial type	Framework		
	1	2	3
Targets: /p/, /f/, /i/			
Experimental	The plane smiled with the dry loan	How does the laugh sew the red cry	Stroke the home or the dream
Experimental	A drill looked in the iron drop	Where does the snake fix the long sound	See the smell and the blame
Control	A stare danced through the stolen wait	Where can the band spin the wide book	Bite the star and the truth

Appendix A (*continued*)

Trial type	Framework		
	1	2	3
Control	The mouth kicked with the late game	Why did the drive cook the light throat	Flood the stick or the king
Experimental	The door grew on the striped lung	How can the rifle ring the gold beat	Park the key and the swing
Experimental	A vest prayed with the short music	Why does the trail move the closed frog	Feel the night and the drink
Control	The bowl swings on the blank toy	Where can the hole lose the wet walk	Take the ghost or the talk
Control	A bell stayed in the thick boat	How did the tail climb the thin rage	Name the love and the chair
Targets: /m/, /ʤ/, /u/			
Experimental	A drill ran to the iron time	Where does the slope jiggle the free sleep	Glue the pain or the chance
Experimental	The mail crept on the dull cape	How does the judge staple the plain wind	Dent the sky and the fruit
Control	The thread stopped at the white price	When does the cave play the clean fire	Paint the shine and the fault
Control	A box ripped in the angry slide	Why did the tree find the dark cough	Ride the story and the claim
Experimental	A sneeze mixed with the slow blush	Why does the bluff hold the hot jail	Choose the blink and the heat
Experimental	The fight dropped from the slimy flash	How can the budgie win the round day	Dig the pool or the whale
Control	The tone crashed with the happy toast	How does the fur snap the quick shoe	Stop the square and the fluff
Control	A break spoke through the rough bath	How did the smoke feed the hard stench	Hit the stream and the journey
Targets: /k/, /ɹ/, /a/			
Experimental	The clamp fell in the fresh loop	Where can the crust chase the tall skin	Lift the jump and the farm
Experimental	A phone poured in the hungry clock	How can the rain hold the black meal	Fold heart and the fire
Control	A fish tapped with the easy golf	When does the flag tell the quiet stone	Open the step or the top
Control	The tape turned with the nasty leaf	Why did the spoon eat the hung wipe	Taste the loss and the soil
Experimental	The mood flew out the tricked song	When does the breath waste the heavy nose	Split the bath and the scene

Appendix A (*continued*)

Trial type	Framework		
	1	2	3
Experimental	A wheel cried on the loud dish	Why does the maze sweep the high robe	Blast the note and the grill
Control	The board hurt in the smooth log	When does the hoop cook the weak pen	Stir the ball and the disk
Control	A thrill blew through the pale month	How did the lake hold the sweet punch	Roll the goat and the salt

References

Baayen, R.H., Piepenbrock, R., Gulikers, L., 1995. The Celex lexical database (Release 2) [CD-ROM]. Linguistic Data Consortium, University of Pennsylvania [Distributor], Philadelphia, PA.

Bailly, G., 2002. Close shadowing natural versus synthetic speech. Int. J. Speech Technol. 6, 11–19.

Bellegarda, J.R., Silverman K.E.A., 1998. Improved duration modeling of English phonemes using a root sinusoidal transformation. In: Proc. of the Intl. Conf. on Speech & Language Processing (ICSLP98), Sydney, Australia, Paper 135.

Benoit, C., Grice, M., Hazan, V., 1996. The SUS test: A method for the assessment of text-to-speech synthesis intelligibility using Semantically Unpredictable Sentences. Speech Commun. 18, 381–392.

Charleston, D.E., Boyer, R.W., 1990. Auditory search using vowel sounds. Perc. Motor Skills 70, 1289–1290.

Cole, M., Hood, L., McDermott, R.P., 1997. Concepts of ecological validity: their differing implications for comparative cognitive research. In: Cole, M., Engestroem, Y. (Eds.), Mind, Culture and Activity: Seminal Papers from the Laboratory of Comparative Human Cognition. Cambridge University Press, New York, pp. 49–56.

Cox, D.R., 1958. Planning of Experiments. Wiley, New York.

Cutler, A., 1976. Phoneme-monitoring reaction time as a function of preceding intonation contour. Perc. Psychophys. 20, 55–60.

Davis, J., 1967. Auditory search for syllables embedded within meaningful sentences. JASA 41, 1277–1282.

Delogu, C., Conte, S., Sementina, C., 1998. Cognitive factors in the evaluation of synthetic speech. Speech Commun. 24, 153–168.

Gong, L., Lai, J., 2003. To mix or not to mix synthetic speech and human speech? Contrasting impact on judge-rated task performance versus self-rated performance and attitudinal responses. Int. J. Speech Technol. 6, 123–131.

Greene, B.G., Logan, J.S., Pisoni, D.B., 1986. Perception of synthetic speech produced automatically by rule: intelligibility of eight text-to-speech systems. Behav. Res. Methods, Instruments, Computers 18, 100–107.

Hansen, M., Kollmeier, B., 1998. Perception of band-specific speech quality distortions: detection and pairwise comparison. Acustica 86, 24.

Hustad, K.C., Kent, R.D., Beukelman, D., 1998. DECTalk and MacinTalk speech synthesisers: intelligibility differences for three listener groups. J. Speech, Lang. Hearing Res. 41, 744–752.

Kahneman, D., 1973. Attention and Effort. Prentice-Hall, Englewood Cliffs, NJ.

Kalikow, D.N., Stevens, K.N., Elliott, L.L., 1977. Development of a test of speech intelligibility in noise using sentence materials with controlled word predictability. JASA 61, 1337–1351.

Koul, R.K., Allen, G.D., 1993. Segmental intelligibility and speech interference thresholds of high quality synthetic speech in presence of noise. J. Speech Hearing Res. 36, 790–798.

Latimer, C.R., 1972. Search time as a function of context letter frequency. Perception 1, 57–71.

Möbius, B., 2003. Rare events and closed domains: two delicate concepts in speech synthesis. Int. J. Speech Technol. 6, 57–71.

Monaghan, A.I.C., 2003. A metrical model of prosody for multilingual TTS. Int. J. Speech Technol. 6, 73–81.

Nass, C., Lee, K., 2001. Does computer-synthesized speech manifest personality? Experimental tests of recognition, similarity-attraction, and consistency-attraction. J. Exp. Psychol.: App. 7, 171–181.

Nass, C., Robles, E., Heenan, C., Bienstock, H., Treinen, M., 2003. Speech-based disclosure systems: effects of modality, gender of prompt, and gender of user. Int. J. Speech Technol. 6, 113–121.

Pashler, H.E., 1998. The Psychology of Attention. MIT Press, Cambridge, MA.

Pols LCW, van Santen JPH, Abe, M., Kahn, D., Keller, E., 1998. The use of large text corpora for evaluating text-to-speech systems. In: Proc. of the First Int. Conf. on Language Resources and Evaluation, Granada, Spain.

Ralston, J.V., Pisoni, D.B., Mullennix, J.W., 1995. Perception and comprehension of speech. In: Syrdal, A.K., Bennet, R., Greenspan, S. (Eds.), Applied Speech Technology. CRC Press, Boca Raton, FL, pp. 233–288.

Roediger, H.L., 1997. The future of cognitive psychology? In: Solso, R.L. (Ed.), Mind and Brain Sciences in the 21st Century. MIT Press, Cambridge, MA, pp. 175–198.

Stern, S.E., Mullennix, J.W., Dyson, C.-L, Wilson, S.J., 1999. The persuasiveness of synthetic speech versus human speech. Human Factors 41, 588–595.

van Santen JPH, Pols LCW, Abe, M., Kahn, D., Keller, E., Vonwiller J., 1998. Report on the 3rd ESCA TTS workshop evaluation procedure. In: Proc. of the Third ESCA Workshop on Speech Synthesis, Jenolan Caves, Australia.

Venkatagiri, H.S., 2003. Segmental intelligibility of four currently used text-to-speech synthesis methods. J. Acoust. Soc. Am. 113, 2095–2104.

Viana, M.C., Oliveira, L.C., Mata, A.I., 2003. Prosodic phrasing: machine and human evaluation. Int. J. Speech Technol. 6, 83–94.

Voiers, W.D., 1983. Evaluating processed speech using the Diagnostic Rhyme Test. Speech Technol. (Jan/Feb), 30–39.

Wherry, R.J., 1938. Orders for the presentation of pairs in the method of paired comparison. J. Exp. Psychol. 23, 651–660.