Do Text-to-Speech Synthesisers Pronounce Correctly? A Preliminary Study

D.G. Evans, E.A. Draffan, A. James, and P. Blenkhorn

Evans, Draffan and Blenkhorn at School of Informatics
University of Manchester, United Kingdom
{david.g.evans, e.draffan, paul.blenkhorn}@manchester.ac.uk
James at IanSyst Ltd., Cambridge, UK
abi@dyslexic.com

Abstract. This paper evaluates 4 commercial text-to-speech synthesisers used by dyslexic people to listen to and proof read text. Two evaluators listened to 704 common English words and determined whether the words were correctly pronounced or not. Where the evaluators agree on incorrect pronunciation, the proportion of correct pronunciations for the four synthesisers is in the range 98.9% to 99.6% of the 704 words. The evaluators also listened to the same synthesisers speaking phrases in which there were 44 pairs of homographs and determined whether each instance of the homograph was correctly spoken or not. The level of correctness for the four synthesisers ranged from 76.3% to 91.3%.

1 Introduction

Text-to-speech synthesisers are used in a number of assistive technology systems. Although, the best-known application of text-to-speech synthesisers to assistive technologists is probably screen readers for blind people, perhaps the most wide-spread applications are those used to support people with other reading difficulties. This group includes people with learning disabilities, specific learning difficulties (such as dyslexia) and people who are learning a language. There are a number of systems that use speech synthesisers to speak text to a user, often visually highlighting the text on the screen as it is spoken. Examples include[1] Claro Software's ClaroRead Plus [1], Freedom Scientific's Wynn [2], Kurzweil Educational System's Kurzweil 3000 [3] and Texthelp's Read and Write Gold [4].

In this paper, we present our work in the context of a person with dyslexia using such text-to-speech systems to access text. The reason for setting this context, rather than a broader context of disabled people using speech synthesisers, is that the work reported here is the initial part of a larger study to investigate how dyslexic people use speech synthesisers and to determine whether there are issues arising in doing so.

[1] This list is not intended to be exhaustive, but merely illustrative of the list of products available.

K. Miesenberger et al. (Eds.): ICCHP 2006, LNCS 4061, pp. 855 – 862, 2006.

2 Using Text-to-Speech Synthesis in Assistive Technology

Over the past 20 years speech synthesisers have become ever more widely available for personal computers. Over this time, the quality of the speech has improved greatly. Today's speech synthesisers have a very high degree of naturalness (i.e. they sound much more like a human speaker than their predecessors). However, naturalness is not the only characteristic of a text-to-speech synthesiser; understandability (i.e. the degree to which the text is correctly perceived by the listener) is also of great importance. Naturalness and understandability are, to some extent, orthogonal. A text-to-speech synthesiser may sound exactly like a human speaker, but that does not mean that the text being spoken can be readily understood; conversely, a text-to-speech synthesiser that sounds robotic may have a very high degree of understandability but would never be mistaken for a human speaker. Some blind users of screen readers contend that the more robotic voices are more useful than the more natural voices because they are more consistent and understandable. This observation may be due, in part at least, to familiarity; the listener becoming conditioned to a particular voice. It may also be due to the fact that earlier, more robotic voices may have a more consistent (and less natural) prosody and pronunciation pattern than later systems.

A user with dyslexia will typically use text-to-speech in three situations – for further information see [5]. Firstly, he/she may be trying to understand a body of text, so, for example, he/she will use the system to read a complete document. This is useful to some dyslexic people, whose ability to process text in auditory form, or combined visual and auditory form, is better than their visual processing of text. Secondly, the user may wish to listen to an isolated word. Some users may read text visually and only call upon the text-to-speech system when a problematic word is reached, which requires clarification by listening to it. Isolated words are also spoken when a user is spell checking. In this case the user needs to be informed of the misspelled word and the options for its replacement. Thirdly, text-to-speech systems are used for proof reading. The user checks his/her writing for the correctness of sentence construction, grammar and punctuation. The prosodic variations of modern text-to-speech synthesisers assist in this task.

In order for a user to make full use of the text-to-speech system, the information that is presented must be clear and free from error. In the case of reading a body of text, incorrect pronunciation or other problems (for example, unusual prosodic effects) may distract the user from his/her major task of text comprehension. In the case of isolated word reading, the pronunciation must be correct. This is especially important in spell checking where the user should, if possible[2], be presented with distinctions between the misspelled word and the suggestions for replacement. The important question is, therefore, do text-to-speech synthesisers speak correctly?

[2] It may not be possible, for example a misspelling of the word 'fail' may be 'fale', by analogy with the pairings 'hail'/'hale' and 'mail'/male' there may be no distinction between the word and the misspelling. This is also true of homophones that may appear as corrections to the spelling. For example, the misspelling 'wether' will yield both 'whether' and 'weather' as suggested corrections. Note that 'wether' is actually a valid English word, meaning a 'castrated ram', but many spell checkers (including Microsoft Word's) mark it as an error.

3 The Study

Our initial work in this area, and the subject of this paper, was a study to determine the degree to which speech synthesisers pronounced words in an appropriate manner. Two separate trials were undertaken.

a) A list of 704 common English words was listened to in isolation to determine those words that had errors in pronunciation.
b) A set of 44 homographs[3] in the context of complete meaningful sentences.

Four text-to-speech synthesisers were used for this evaluation; they were chosen because they were commonly used by dyslexic people with text-to-speech systems. The synthesisers used were (the names used to refer to the synthesisers in the subsequent text are given in bold):

- AT&T Natural Voices *Audrey*, UK English
- Realspeak *Jane*, UK English
- Microsoft *Mary*, US English
- Plaintalk *Victoria*, US English

For both tests, two evaluators were used, one female and the other male. Both were native English speakers with experience in listening to speech synthesis systems; neither was dyslexic.

3.1 Isolated Word Test

The words were recorded using each of the speech synthesisers into audio files. The evaluators firstly listened to each audio file from start to finish with reference to a paper copy of the list of words. When the evaluator perceived that a word was unusually or incorrectly pronounced, he/she marked his/her list. All the words that were unmarked after this initial run were classified as being correctly pronounced. The evaluator then listened again to all the words that he/she had marked a problematic in the initial run. He/she then made a judgement to classify the word into one of three classes:

- Correct, the word is treated in the same way as those not marked after the initial run
- Incorrect, the word is incorrectly pronounced
- Partially correct, the pronunciation of the word is acceptable, but some characteristic of the pronunciation of the word causes the listener to note that something is unusual. This category is provided as such pronunciations may mislead a user with dyslexia when spell checking or distract the listener when reading a block of text.

An initial summary of the results is shown in Table 1. Table 2 shows the classification of words marked as by both of the evaluators for each of the four text-to-speech synthesisers.

[3] Words with same spelling, but different meaning and pronunciation (for example ‘moped’ which may be interpreted as the past tense of the verb ‘mope’ (to sulk) or the noun meaning a motor powered, two-wheeled vehicle with pedals).

Table 1. Summary of Isolated word classification by evaluator and synthesiser

	Female Evaluator			Male Evaluator		
	Correct	Partially Correct	Incorrect	Correct	Partially Correct	Incorrect
Audrey	98.2%	1.7%	0.1%	97.6%	1.1%	1.3%
Jane	99.1%	0.4%	0.4%	98.6%	0.6%	0.9%
Mary	99.1%	0.4%	0.4%	98.6%	0.6%	0.9%
Victoria	97.7%	2.0%	0.3%	96.4%	1.0%	2.6%

The overall level of correctness differs between evaluators, but the overall ranking of text-to-speech synthesisers is consistent Jane and Mary (joint first), Audrey then Victoria.

Table 2. Words marked Partially Correct (p) and Incorrect (i) by each evaluator (F = female, M= male)

Audrey			Jane			Mary			Victoria		
Word	F	M	Word	F	M	Word	F	M	Word	F	M
advertisement	p	p	altruistic	p		advertisement	p	p	advertisement		x
altruistic	p		apologise		p	apologise	x	x	automatics		x
bureaucracy		p	database	x	x	at		p	better	p	
cashier	p	x	dismissal		x	brochure		p	body		x
courier		p	enthusiasm	p		deliberate	p		both	p	x
discrepancy	p		expertise	x	x	despatch	p		brochure		p
exaggerate	p		fiancé	x	x	expertise	p		call	x	x
experienced	p	x	Florida	p		favourite	x	x	caught	p	x
expertise	x		fluctuate		x	from		p	cause		x
general	p	x	satisfactorily		p	hand		p	certificate	p	
glamorous	p	x	subtitles		p	jeopardise	x		certificates	p	
into		x	temperature		p	manufacturer		p	chose		x
itinerary	p		valuable		x	quantity		p	deliberate	p	
large	p	p	young		x	quarter		p	during	p	
quantity		p				recognise	p	p	enthusiasm	p	
secondary		p				revenue		p	example		p
success		x				subtle		p	expertise	p	
term		p				than		x	extension		p
thus	p	x				toward		p	far	p	
town	p	x				with		p	Florida	p	
woman		x				year		x	fluctuate		x
year		p				young		x	for		x
young		x							from		x
									little	p	x
									lose		x
									sceptical	p	p
									she		x
									simple		x
									subtle	x	x
									toward		p
									water		p
									whole		x
									woman		x
									young		x

It is clear from Table 2 that agreement between evaluators on Partially Correct or Incorrect words is relatively rare. The evaluation results can be combined such that errors are counted only when both evaluators indicate an error; where a word is classified as Partially Correct by one evaluator and Incorrect by the other, it is classified as partially correct. The results of this analysis are shown in Table 3.

Table 3. Classification of words for each synthesiser, errors marked only when evaluators agree

	Correct	Partially Correct	Incorrect
Audrey	98.9%	1.1%	0.0%
Jane	99.6%	0.0%	0.4%
Mary	99.4%	0.3%	0.3%
Victoria	99.1%	0.6%	0.3%

As can be seen from this table, the degree of correct pronunciation of common, isolated words by a range of text-to-speech synthesisers is very high.

3.2 Homograph Test

Homographs are words that are written in the same way, but which have different meanings and often different pronunciations – where pronunciation differs they are heteronyms. The degree to which the pronunciation varies is dependent on the heteronym. For example, the difference between the word 'moped' a verb (past tense) and as a noun results in different phonemes. In others, it is simply that syllable stress moves, for example the contrast between the noun and verb forms of the word 'project' (the stress is on the first syllable for the noun and the second for the verb).

Forty-four homographs were selected by choosing a fairly common and representative sample from the set provided in [6]. The selection of the words was in many ways arbitrary; however, it is argued that this is not important since the aim of the work is to gain some measure as to the degree to which homographs are correctly pronounced rather than to produce results for all.

A sentence or pair of sentences was then constructed that contrasted the different pronunciations of the homograph (for example 'He *moped*; his *moped* had been stolen'. The sentences were recorded into a single audio file for each of the synthesisers. The procedure for the isolated word matching was followed, with the evaluator marking words that he/she felt were in error, revisiting those considered to be in error and classifying as Correct, Partially Correct or Incorrect. Table 4 shows an initial summary of the results

Again, there is some variation between the evaluators, but the overall ranking of the synthesisers is consistent. There is again variation between the evaluators. Table 5, shows where both evaluators agree.

The combined results of both evaluators are shown in Table 6. In doing so, the results of the final row of Table 5 have been removed. The intention was that 'supply' was interpreted in the sense of being supple; however, the sentence can also be interpreted a 'supply' in the sense of a source and is thus removed.

Table 4. Initial results for the evaluation of homographs

	Female Evaluator			Male Evaluator		
	Correct	Partially Correct	Incorrect	Correct	Partially Correct	Incorrect
Audrey	87.2%	4.3%	8.5%	81.9%	2.1%	16.0%
Jane	87.2%	5.3%	7.4%	84.0%	2.1%	13.8%
Mary	71.3%	6.4%	22.3%	69.1%	4.3%	26.6%
Victoria	76.6%	10.6%	12.8%	77.7%	2.1%	20.2%

Table 5. Homograph classification where both evaluators agree. Cl. = classification with p = partially correct and x = incorrect. A = Audrey, J = Jane, M= Mary, V=Victoria. ✓ indicates error for that text-to-speech synthesiser.

Word	Context	Cl.	A	J	M	V
overall	He wore a red **overall**	p				✓
entrance	I like to **entrance** people	x			✓	
object	I want to **object**	x			✓	
present	Please **present** me …	x			✓	
record	I would like to **record** this session	x			✓	
refuse	I would like to **refuse** to…	p			✓	
second	We should **second** her to our department	x			✓	✓
		p	✓	✓		
subject	I know that I should not **subject** you to …	x			✓	
approximate	I would like you to **approximate** to it	x			✓	
	The amount is only **approximate**	x				✓
moderate	We need to **moderate** our output …	p			✓	
separate	I think we ought to **separate**	x			✓	
abuse	Don't give that **abuse**	p	✓			
close	I thought the door was going to **close**	x				✓
	That was **close**	x			✓	
diffuse	Particles will not **diffuse** in this atmosphere	p			✓	
house	We are not prepared to **house** him	x	✓			
learned	My **learned** father …	x	✓		✓	✓
bow	I need to **bow** out	x		✓		
	I'll take the red **bow**	x			✓	
invalid	It is **invalid** to call someone an invalid these days	p				✓
	It is invalid to call someone an **invalid** these days	x	✓		✓	
Lead	I need to **lead** you	x				✓
	The compass will be effected by the red **lead**	x		✓	✓	
Live	I like **live** music	x	✓	✓	✓	
moped	His **moped** had been stolen	x				✓
pasty	You are looking rather **pasty**	x		✓	✓	
	It must have been that **pasty** you ate	x	✓			✓
routed	The army was **routed** at the battle …	x		✓	✓	✓
	I then **routed** them via	x	✓			
wound	I **wound** some paper around it	x			✓	
august	I will ask the **august** man to speak	p		✓	✓	
polish	He is so good with boot **polish**	x				✓
wind	This is a **wind** up	x	✓			✓
supply	They move **supply** under pressure	x	✓	✓	✓	✓

It may not be clear from the results in tables 4 to 6 just how good modern speech synthesisers are at resolving homographs when words are given in valid context. To illustrate this, it worth identifying some tests that all synthesisers produced correct results with. These include:

- The wind will *buffet* us on the way to the *buffet* car.
- Don't give me that *abuse*, I do not *abuse* you.
- It was a *moderate* success. We need to *moderate* out output in future.
- I will *read* to you now. Just as I *read* to you yesterday.

The last of these is particularly impressive as the synthesiser determines the tense that is required from other cues in the sentence.

Table 6. Classification of homograph errors for synthesisers, where evaluators agree

	Correct	Partially Correct	Incorrect
Audrey	90.2%	2.2%	7.6%
Jane	91.3%	2.2%	6.5%
Mary	76.2%	4.2%	19.6%
Victoria	85.8%	2.2%	12.0%

4 Discussion

This limited evaluation of speech synthesisers does show that the levels of correctness for single, common words and for homophones in context is very high. This result was somewhat contrary to our preconception that rather more errors would be fond. However, this does not mean that users with dyslexia can successfully use text-to-speech systems to read text and address spelling errors. It simply means that the text is spoken to a very high standard; what is important, however, is how a user perceives and uses this information. The work reported here provides a starting point for further investigation with the knowledge that the text is almost always rendered correctly.

One issue that should be noted is that text-to-speech synthesisers are good at resolving homophones when supplied with context. However, when spell checking the text-to-speech synthesiser will be supplied with a single word (one of the correctly spelt options) with no context. It is impossible to determine how such words should be spoken without their contexts; consider, for example, the word 'read'. The text-to-speech synthesiser must default to one way of saying the word. This may be confusing to the user, who can see the context of the word on the screen and be able to hear the misspelled word in context. Further work is required to examine whether this is a significant issue and, if so, to determine ways in which users can best be supported.

References

1. Claro Software, http://www.clarosoftware.com, Accessed 19 Jan 2006
2. Freedom Scientific, http://www.freedomscientific.com/LSG/products/wynn.asp, Accessed 19 Jan 2006

3. Kurzweil Educational Systems, http://www.kurzweiledu.com/products_k3000win.asp, Accessed 19 Jan 2006
4. Texthelp, http://www.texthelp.com/rwg.asp?q1=products&q2=rwg, Accessed 19 Jan 2006
5. Fiddler, R.: An evaluation of the use of specialist support services by dyslexic students at a higher education institution. Skill Journal, March 2001
6. Higgins, J.: Homographs. http://www.marlodge.supanet.com/wordlist.homograph.html, Accessed 27 Aug 2004