I.J. Information Engineering and Electronic Business, **2014, 4, 27-31**
Published Online August 2014 in MECS (http://www.mecs-press.org/)
DOI: 10.5815/ijieeb.2014.04.04

A Comparative Study of Arabic Text-to-Speech Synthesis Systems

Najwa K. Bakhsh
Computer Science Department, Faculty of Computing and Information Technology, King Abdulaziz University, Jeddah
nkbakhsh@kau.edu.sa

Saleh Alshomrani, Imtiaz Khan
Faculty of Computing and Information Technology, King Abdulaziz University, Jeddah
{ sshomrani , ihkhan }@kau.edu.sa

Abstract— Text-to-speech synthesis is the process of converting written text to speech. The lack of research on the growth of and the need for the Arabic language is notable. Therefore, this paper reports an empirical study that systematically compares two screen readers, namely, NonVisual Desktop Access (NVDA) and IBSAR. We measured the quality of these two systems in terms of standard pronunciation and intelligibility tests with visually impaired or blind people. The results revealed that NVDA outperformed IBSAR on the pronunciation tests. However, both systems gave competitive performance on the intelligibility tests.

Index Terms—Arabic Text-to-Speech; Speech Synthesis; Visually Impaired People; Pronunciation Test; Intelligibility Test; DRT

I. INTRODUCTION

Speech technology is an important subfield of natural language processing (NLP) that involves various techniques, such as speech synthesis, speech recognition and dialog systems. Text-to-speech (TTS) synthesis is the process of converting written text to speech [1-2]. Speech synthesis techniques aim to translate a chain of phonetic symbols to speech, to transform a given linguistic symbol and to generate speech automatically with information about intonation and stress to obtain the exact prosody. The quality of a speech synthesis system can be measured against different criteria, including pronunciation, comprehensibility and intelligibility. Research in this area has so far been mainly confined to English and other European languages. For the Arabic language, such tools are still in its infancy.

Arabic, which has rich morphology and syntax, is the fourth most widely spoken language in the world [3]. Recently, research on Arabic NLP has gained much attention, including the development of automatic TTS systems (for more details, see [4-5]).

We are in the process of developing a WebAnywhere system for the Arabic language that can help visually impaired people to access the Web. WebAnywhere is a Web-based, self-voicing browser that enables blind Web users to access the Web from almost any computer that can produce sound [6]. The performance of a WebAnywhere system depends, apart from other things, on the quality of the screen reader, which uses a TTS (or speech synthesiser) system.

In this article, we report two investigative studies to evaluate the performance of NonVisual Desktop Access (NVDA)[1] and IBSAR[2] (details to follow), two speech synthesiser systems used by blind people. The first study focused on the quality of the system in terms of pronunciation, and the second study focused on the intelligibility of the system. In both studies, our participants were blind or visually impaired people.

The rest of this article is organised as follows. In Section II, we present the background of speech synthesis and speech synthesis evaluation. Section III gives an overview of the two screen readers, namely, NVDA and IBSAR, which we compared in this study. The empirical study is described in Section IV. In Section V, we discuss the results. Section VI concludes.

II. BACKGROUND

A. Speech Synthesis Framework

TTS synthesis is an important subfield of NLP that aims to produce speech from some given textual input. To achieve this main goal, a TTS is broadly divided into two parts: 1) NLP module (also called front end) and 2) digital signal processing (DSP) module (also called back end). A schematic diagram of the system is presented in Figure 1.

This modular division is driven by necessity. The NLP module is mainly concerned with surface-level issues (e.g., morphology) related to linguistic analysis. The output of this module is a form of linguistic representation that includes, for example, information on the phonemes to be produced. The DSP module, the actual speech synthetic, takes this information and converts it to a speech waveform.

Two approaches are commonly used for generating synthetic speech waveforms: rule-based synthesis and concatenative synthesis [7]. The rule-based synthesis exploits the knowledge of speech experts to develop the synthesis system. In the rule-based synthesis, formant

[1] www.nvaccess.org.
[2] http://www.sakhr.com.

synthesis is the most commonly used method. In this approach [8-9], rule-based systems are space efficient because they eliminate the need to store speech segments. Conversely, concatenative synthesis is a data-driven approach in which parametrised short speech segments of natural speech are connected to form a representation of the synthetic speech [10]. It uses actual short segments of recorded speech that were cut from recordings and stored in a voice database as waveforms, for example.

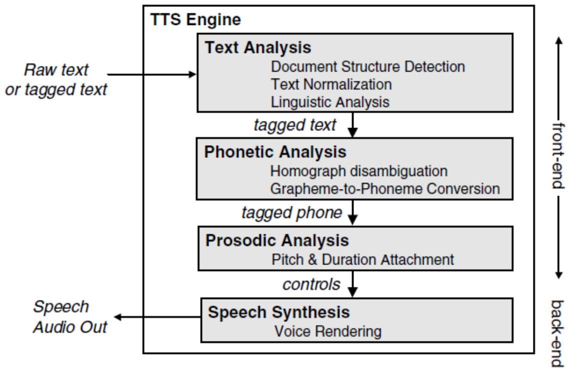

Fig. 1. Text-to-Speech Synthesiser System

B. Speech Synthesis Evaluation

A large body of work has been published on evaluating TTS synthesis systems [11-15] (for details, see [15]). Here, we review studies that evaluated speech synthesis systems for the Arabic language [5, 16-17].

In [5], the authors evaluated the performance of different Arabic TTS synthesisers using a pronunciation test. They reported that, overall, the performance of the competing synthesisers was reasonably well in both vowels and non-vowels. As they focused only on pronunciation, the quality of these synthesisers when other factors such as comprehensibility are considered remains unclear to us.

In [16], the authors evaluated the performance of an Arabic TTS using three criteria: intelligibility, naturalness and voice quality. The participants rated the performance of the system on a five-point scale. Overall, the system performed well on the subject criteria.

In [17], the authors evaluated a diphone speech synthesis system for the Arabic language using two tests: the diagnostic rhyme test (DRT), which measures the intelligibility of the synthesised speech, and the categorical estimation (CE) test, which measures the overall quality of the synthesised speech. The system performed reasonably well on the DRT test, but its performance was below par on the CE test.

These studies mainly focused on sighted people. In the current study, we compared the performance of two Arabic screen readers (NVDA and IBSAR) using blind or visually impaired people. Moreover, we used a balanced evaluation dataset and focused on pronunciation and intelligibility.

III. The Evaluated Systems

We evaluated the performance of two speech synthesis screen reader systems, namely, NVDA and IBSAR. NVDA is a Microsoft Windows-based screen reader especially developed for blind or visually impaired people. It enables Braille interfacing with the speech synthetic system, which allows visually impaired people to successfully browse the Web, for example. NVDA supports more than 35 natural languages, including Arabic. One of the major features of this system is that it is free and an open source.

IBSAR is another screen reader that reads documents aloud in one of its many high-quality voices. This screen reader is capable of reading aloud standard Windows dialog box messages, email systems and the Web. It supports both English and Arabic and can write audio data directly to Windows WAV files. Like NVDA, IBSAR also permits Braille interfacing that enables blind or visually impaired people to browse the Web comfortably.

The performance of these two systems has not been tested systematically. Therefore, how good these systems would be when they are integrated into a WebAnywhere system remains unclear. This uncertainty is one of the main motivations of our empirical study.

IV. Empirical Study

We compared the performance between NVDA and IBSAR in two experiments: 1) pronunciation test and 2) intelligibility test. Both experiments were conducted in King Abdulaziz University's experimental laboratory dedicated to people with cognitive disabilities.

A. Experiment 1: Pronunciation Test

A TTS system can be seen as starting with the words in the text, converting each word one-by-one into speech, and concatenating the result together. To this endeavour, however, this is imperative that each word should be pronounced correctly. So, the first legitimate test which comes to mind, to judge the quality of a TTS, is a pronunciation test.

We conducted two different pronunciation tests. In the isolated word test, each word is pronounced without any surrounding context. In the homograph test, *homographs* are presented in a one-sentence context. Homographs are words written in the same way but have different meanings and usually different pronunciations. Homographs pose a significant challenge in communication because of their similarity, making them the ideal candidate for a pronunciation test.

1. Materials and Design

A list of 30 words was carefully selected for the isolated word pronunciation test. This selection was made from a database of Arabic phonemes, which contain approximately 660 words, available at the computer research institute of the King Abdulaziz City of Science and Technology. The selected dataset was balanced in the sense that it contains 10 words that are easy to pronounce, 10 words that are slightly difficult to pronounce and 10 words that are difficult to pronounce. A truncated list (with each Arabic word transliterated in English) is shown in Table 1.

Similarly, a list of 10 homographs was pseudo-randomly selected. These homographs were embedded in a one-sentence context, with each sentence having one homograph (Table 1).

Table 1. Word lists for experiment 1

Isolated word list	Homographs
بَخْسِ (bakhs)	عَصَرَ (Asara) عُصِرَ (Usira)
أَغْنَى(Agna)	
أَنْذَرَ(Anthra)	سَلّمَ(Salma) سُلَّم(Solima)
بَدْرُ(Badr)	
بَذْخُ(Bathkh)	جَمَعَ(Jamaa) جَمّعْ(Jamma)
فوْز(Fawz)	
حَذْوَه(Hadwah)	رَبّعْ(Rabba) رُبّعْ(Robaa)
حَضِيض(Hadied)	
شَظْي(Shazi)	قَطَرْ(Qatar) قُطُر(Qutor)
كَافَأْت(Kafat)	
مَأْخُوذ(Makhoz)	كَتَبَ(Kataba) كُتُب(Kutiba)
جَذْب(Jathb)	

2. Participants and Procedure

Twelve visually impaired or blind undergraduate students took part in the experiment. The participants were native Arabic speakers. The experiment (i.e., isolated word test and homograph test), which lasted for approximately 45 min, was carried out in King Abdulaziz University's specialised experimental laboratory for disabled people.

Before running the experiment, the participants were briefed about the format and purpose of the experiment. The instructions were as follows:

In the first test, you will listen to some words in isolation, one word at a time. The same words that you have just heard will also be presented to you in Braille. Your task is to mark the word as follows:

- Correct: If you think that the pronounced word is the same as the one presented in Braille, mark it as correct.
- Incorrect: If you are convinced that the pronounced word is not the one that is presented in Braille, mark it as incorrect.
- Partially incorrect: If you are uncertain whether or not the pronounced word is the one presented in Braille, mark it as incorrect.

Similarly, mutatis mutandis for the homograph test. The homographs were embedded in a one-sentence context.

3. Results and Analysis

The responses were recorded as *correct*, *partially correct* or *incorrect*, for each test against NVDA and IBSAR. The percentage responses are shown in Table 2. The results showed that, for the isolated word pronunciation test, NVDA (above 63% correct responses) outperformed IBSAR (43% correct responses). A pair-wise t-test further revealed that these differences were highly significant ($p < 0.01$).

Similarly, for the homograph test, NVDA (above 83% correct responses) outperformed IBSAR (above 66% correct responses). Again, a pair-wise t-test revealed that these differences were highly significant ($p < 0.01$).

Table 2. Pronunciation test results (response in %)

Isolated words test			
Program	Correct	Partially Correct	Incorrect
IBSAR	43%	17%	40%
NVDA	63.33%	16.66%	20%
Homographs test			
Program	Correct	Incorrect	
IBSAR	66.66 %	33.33%	
NVDA	83.33%	16.66%	

B. Experiment 2: Intelligibility Test

Unlike the pronunciation test, the *intelligibility* test requires conducting elaborate listening tests. The intelligibility of the speech means whether or not the output of the synthesiser could be understood by a human listener. Intelligibility tests, based on testing speech coders, present word or sentence lists and enables subjects to transcribe the words they hear [18] on TTS-related testing. Generally, the intelligibility of a TTS system has many facets other than those covered by standard word-list driven intelligibility tests. At a sentence level, for example, a wrong prosody can destroy intelligibility.

Different intelligibility tests have been reported in the literature on speech synthesis. The most commonly used test is the DRT, which is one of the ANSI standards for measuring the intelligibility of speech in communication systems. In this test, rhyming word pairs are presented that differ only in their initial consonants (e.g., جِنان (Jannan)-حَنان (Hannan)). The sound of one of the rhyming words is played, and the task is to identify which rhyming word is pronounced.

A variant of DRT is the diagnostic medial consonant test (DMRT). In DMRT, word pairs such as stopper-stocker, which differ only in the intervocalic consonant, are examined.

In this experiment, we used both DRT and DMRT.

1. Materials and Design

In DRT, we manually constructed a list of 30 rhyming word pairs in Arabic that differ in their initial consonant. Again, the word pairs were carefully constructed to ensure a balanced representation of the difficult and easy rhyming words. A truncated list of the rhyming word pairs is shown in Table 3.

Similarly, in DMRT, a list of 20 rhyming word pairs was constructed that differ in the intervocalic consonant. A truncated list of these word pairs is shown in Table 3.

2. Participants and Procedure

Ten visually impaired or blind undergraduate students took part in this experiment. The participants were native Arabic speakers. The experiment lasted for approximately 30 min.

The participants were briefed about the format of the experiment. They were instructed to listen to a word and read two words in Braille. Their task would be to mark the word that they just heard.

Table 3. Word lists for experiment 2

DRT word pairs	DMRT word pairs
جِنان(Jenan) حَنان(Hanan)	كِتَاب(Ketab) كِتَان(Kenan)
هِبال(Hebal) حِبال(Hebal)	وِصَاف(Wesaf) وِصال(Wesal)
سراب(Sarab) صَراب(Serab)	كَسّاب(Kssab) كَسّار(Ksaar)

مستشار(Mostshar)
منشار(Menshar)
	رَزّانْ(Rzan)
	رَزّاقْ(Rezaq)

قَاسَمَ(Kasem) كَاسَمَ(Kasem)	مَلّعَب(Malab) مَلّعَقة(Malakah))

3. Results and data analysis

The responses were recorded as *correct* or *incorrect* for each test against NVDA and IBSAR. The percentage responses are shown in Table 4. The results showed that for both tests, both NVDA and IBSAR gave competitive performance.

Table 4. Intelligibility test results (response in %)

DRT		
Program	Correct	Incorrect
IBSAR	80%	20%
NVDA	76%	24%
DMRT		
Program	Correct	Incorrect
IBSAR	78%	22%
NVDA	86%	14%

V. Conclusion and Future Work

We empirically compared the quality of NVDA and IBSAR using a reasonable sample of visually impaired or blind people. We measured the quality of these two systems in terms of two pronunciation tests (isolated word test and homographs embedded in a one-sentence context) and intelligibility tests (DRT and DMRT). The experiments revealed that the overall quality of NVDA was better than that of IBSAR.

The study focused only on two aspects: pronunciation and intelligibility. In the future, we intend to extend this work in two ways. First, the quality of NVDA and IBSAR will be evaluated against other criteria, such as comprehensibility and naturalness, in a larger sample size. We aim to include both sighted and visually impaired or blind people in future studies.

Second, the overall best screen reader will be integrated with an Arabic WebAnywhere system to help visually impaired or blind people access the Web.

Acknowledgments

We would like to thank Mostafa Alshafi, Mansour Al-Ghamdi and Layla Bawazer for their valuable support in designing the experimental material and conducting the experiment in the specialised disability laboratory at King Abdulaziz University.

References

[1] D. H. Klatt Review of text-to-speech conversion for English. Journal of the Acoustical Society of America. Vol. 82(3), 1987.

[2] J. Allen, M. S. Hunnicutt and D Klatt. From Text to Speech. Cambridge University Press, Cambridge, 1987.

[3] F. A. Nwesri, S. M. M. Tahaghoghi, and F. Scholer. Stemming Arabic conjunctions and prepositions. In Mariano Consens and Gonzalo Navarro, editors, String Processing and Information Retrieval, 12th International Conference. Buenos Aires, Argentina, pp. 206-217, 2005.

[4] A. Youssef and O. Emam. An Arabic TTS System Based on the IBM Trainable Speech Synthesizer. In: Le traitement automatique de l'arabe, JEP–TALN 2004, Fès. 2004.

[5] Al-Wabil, H. Al-Khalifa and W. Al-Saleh. Arabic-Text-To-Speech Synthesis: A Preliminary Evaluation. In Proceedings of World Conference on Educational Multimedia, Hypermedia and Telecommunications. Vancouver, Canada, 2007, pp. 4423-4430.

[6] P. J. Bigham, M. P. Craig and E. L. Richard. Engineering a Self-Voicing, Web-Browsing Web Application Supporting Accessibility Anywhere. In Proceedings of the International Conference on Web Engineering. New York, USA, 2008.

[7] Dutoit. An Introduction to Text-to-Speech Synthesis. London: Kluwer Academic Publishers, 1997.

[8] J. Allen, M.S. Hunnicutt, and D. Klatt, From Text to Speech, The MITalk System, Cambridge: Cambridge University Press, 1987.

[9] J. N. Holmes. Formant Synthesizers: Cascade or Parallelm. Speech Communication, Vol 2, pp 251-273, 1983.

[10] M. M. Sondhi and D.J. Sinder. Articulatory modeling: a role in concatenative text-to-speech synthesis. Text to Speech Synthesis: New Paradigms and Advances, A. Alwan and S. Narayanan, Eds., Englewood Cliffs, Prentice Hall, 2003.

[11] L. C. W. Pols, J. P. H. Santen, M. Abe, D. Kahn and E. Keller . The use of large text corpora for evaluation text-to-speech systems. In Proceedings of the First International Conference on Language Resources and Evaluation, Granada, Spain, 1998.

[12] Y. V. Alvarez and M. Huckvale. The reliability of the ITU-T P.85 standard for the evaluation of text-to-speech systems. In Proceedings of ICSLP2002. Denver, Colorado, pp. 329-332, 2002.

[13] D. G. Evans, E. A. Draffan, A. James, and P. Blenkhorn. Do Text-to-Speech Synthesisers Pronounce Correctly? A Preliminary Study. In proceedings of Computers Helping People with Special Needs. Springer, lecture series pp. 855-862, 2006.

[14] Stevens, N. Lees, J. Vonwiller and D. Burnham. On-line experimental methods to evaluate text-to-speech (TTS) synthesis: effects of voice gender and signal quality on intelligibility, naturalness and preference. Computer Speech and Language. Vol. 19, pp. 129-46, 2005.

[15] Y. Chang. Evaluation of TTS Systems in Intelligibility and Comprehension. In proceedings of the 23rd Conference on Computational Linguistics and Speech Processing. pp 64-78, 2008.

[16] M. Zeki, O. O. Khalifa and A. W. Naji. Development of an Arabic text-to-speech system. International Conference on Computer Communication Engineering. pp. 1-5, 2010.

[17] M. Z. Rashad, H. M. El-Bakry, I. R. Isma'il. Diphone Speech Synthesis System for Arabic Using MARY TTS. International Journal of Computer Science & Information Technology. Vol 2(4), 2010.

[18] M. F. Spiegel, M.J. Altom, and M.J. Macchi. Comprehensive assessment of the telephone intelligibility of synthesized and natural speech. Speech Communication. Vol 9, pp. 279-291, 1990.

Authors' Profiles

Najwa K. Bakhsh was born in Jeddah, Kingdom of Saudi Arabia (KSA). She is currently a computer science post-graduate student at King Abdulaziz University, Jeddah, KSA. She obtained her bachelor's degree in computer science from King Abdulaziz University, Jeddah, KSA, in 2003. She is currently working as a programmer at King Abdulaziz University. Her research interests include Web mining, Web accessibility, cloud technology and parallel algorithm.

Dr. Saleh Alshomrani is an Associate Professor in the Information Systems Department at King Abdulaziz University. He is also the Vice Dean of the Faculty of Computing and Information Technology, North Jeddah Campus, at King Abdulaziz University. He obtained his bachelor's degree in computer science from King Abdulaziz University, Jeddah, KSA, in 1997. He received his master's degree in computer science from Ohio University, USA, in 2001. He earned his Ph.D. in computer science from Kent State University in Ohio, USA, in 2008 in the field of Internet and Web-based distributed systems and technologies. His research areas include Web data mining, algorithms, e-learning and e-government.

Dr. Imtiaz H. Khan is an Assistant Professor in the Department of Computer Science at King Abdulaziz University, Jeddah, KSA. He received his master's degree in computer science from the University of Essex, UK, in 2005. He earned his Ph.D. in artificial intelligence from the University of Aberdeen, UK, in 2010. His research interests are natural language processing, particularly natural language generation and evolutionary computation.

How to cite this paper: Najwa K. Bakhsh, Saleh Alshomrani, Imtiaz Khan,"A Comparative Study of Arabic Text-to-Speech Synthesis Systems", IJIEEB, vol.6, no.4, pp.27-31, 2014. DOI: 10.5815/ijieeb.2014.04.04