Asian Journal of Information Technology 13 (11): 720-724, 2014
ISSN: 1682-3915

An Interactive Speech Web Site in Arabic and English

[1]O. Al-Dakkak, [1]N. Ghneim and [2]I. Salman
[1]Higher Institute of Applied Science and Technology (HIAST),
P.O. Box 31983, Damascus, Syria
[2]Alshiekh Saad, Tartous, Syria

Abstract: In this study, we propose the approach to implement an interactive speech web site in Arabic and English; this method relies on the Integration between open source software systems of speech recognition, Text to Speech (TTS) and dialogue systems to build the application. the application is targeted to enable blind people and children to browse a news web site with short stories. It can help to enter the digital world in spite of their difficulties.

Key words: Speech recognition, speech synthesis, JavaScript API, natural language processing, interactive web site

INTRODUCTION

The field of voice interaction between man and machine involves more and more researchers. The aim of this research is to enable people with written language interaction difficulties; like blind people or young children to interact with the machine and have access to the contents arranged in a web site with their voices only. Many open source packages are given to help building such applications. We will scan some of these packages; choose best ones in terms of portability, performance and quality to build the application, together with the help and remarks from the targeted people to ensure ease of use in both Arabic and English.

This study presents the different components used in such systems (Automatic Speech Recognition (ASR Systems), Text-To-Speech Systems (TTS) and Dialogue Systems), then we introduce the implementation method and obtained results. Finally, we give the conclusion.

ASR SYSTEM

The process of speech recognition, means receiving computerized speech, when processing the speech signal to determine the words spoken (Herscher and Cox, 1972). A further step of understanding the meaning of these words is needed to implement voice command associated with it. In the application, we will use voice commands to navigate between the pages of the site. Over the years a number of different methodologies have been proposed for isolated word and continuous speech recognition. These can usually be grouped in two classes: speaker-dependent and speaker-independent. Speaker dependent methods usually involve training a system to recognize each of the vocabulary words uttered single or multiple times by a specific set of speakers (Herscher and Cox, 1972; Itakura, 1975) while for speaker independent systems such training methods are generally not applicable and words are recognized by analyzing their inherent acoustical properties (Gupta *et al*., 1978). Hidden Markov Models (HMM) have been proven to be highly reliable classifiers for speech recognition applications and have been extensively used with varying amounts of success (Rabiner, 1989; Betkowska *et al*., 2007). In this technique, we build an HMM for each word to recognize, based on the probabilities of being in the first state and the transition probabilities of between states and the inherent probabilities of being in each states. We usually use five states models for isolated words recognition, supposing each word consisting roughly of five voices in average. Artificial Neural Networks (ANN) technique has also been demonstrated to be an acceptable classifier for speech recognition (Thiang and Wijoyo, 2011), trying to imitate the recognition process in the brain of living creatures.

TTS SYSTEM

Speech synthesis is the artificial production of human speech. A computer system used for this purpose is

Corresponding Author: O. Al-Dakkak, Higher Institute of Applied Science and Technology (HIAST), P.O. Box 31983, Damascus, Syria called a speech synthesizer and can be implemented in software or hardware products. A Text-To-Speech (TTS) system converts normal language text into speech; other systems render symbolic linguistic representations like phonetic transcriptions into speech (Al-Dakkak *et al*., 2005; Allen *et al*., 1987). Synthesized speech can be created by concatenating pieces of recorded speech that are stored in a database. Systems differ in the size of the stored speech units; a system that stores phones or diphones provides the largest output range but may lack clarity. For specific usage domains, the storage of entire words or sentences allows for high-quality output. Alternatively, a synthesizer can incorporate a model of the vocal tract and other human voice characteristics to create a completely "synthetic" voice output (Rubin *et al*., 1981). The quality of a speech synthesizer is judged by its similarity to the human voice and by its ability to be understood clearly. An intelligible Text-To-Speech Program allows people with visual impairments or reading disabilities to listen to written works on a home computer. Many computer operating systems have included speech synthesizers since the early 1990s, for English and some other languages, however, we don't have many free Arabic TTS available.

DIALOGUE SYSTEMS

A dialog system or Conversational Agent (CA) is a computer system intended to converse with a human with a coherent structure. Dialog systems have employed text, speech, graphics, haptic, gestures and other modes for communication on both the input and output channels.

There is much different architecture for dialog systems. What sets of components are included in a dialog system and how those components divide up responsibilities differs from system to system. The dialog manager is a principal component to any dialog system. It is the component that manages the state of the dialog and dialog strategy. A typical activity cycle in a dialog system contains the following phases (Jurafsky and Martin, 2009):

I The user speaks and the input is converted to plain text by the system's input recognizer/decoder which may include:
   - Automatic Speech Recognizer (ASR)
   - Gesture recognizer
   - Handwriting recognizer

II The text is analyzed by a Natural Language Understanding Unit (NLU) which may include:
   - Proper name identification
   - Part of speech tagging
   - Syntactic/semantic parser

III The semantic information is analyzed by the dialog manager that keeps the history and state of the dialog and manages the general flow of the conversation

IV Usually, the dialog manager contacts one or more task managers that have knowledge of the specific task domain

V The dialog manager produces output using an output generator which may include:
   - Natural language generator
   - gesture generator
   - layout engine

VI Finally, the output is rendered using an output renderer which may include:
   - Text-To-Speech engine (TTS)
   - Talking head
   - Robot or avatar

Dialog systems that are based on a text-only interface (e.g., text-based chat) contain only stages II-V.

IMPLEMENTATION

In this study, we will present the approach used to build the system with its three main components, Text-To-Speech, Spoken Command Recognition and Web site Dialogue. The JAD web site was designed, to be a simple speech interactive web site in Arabic and English. Figure 1 presents the structure of the system.

Text-To-Speech System: The system synthesizes English and Arabic texts. To perform the synthesis we have used:

- Speech SDK 5.1 from Microsoft to synthesize English Texts
- Things where more complicated for Arabic. First we have to input diacritized Arabic texts. These texts are entered to a grapheme to phoneme systems we built for this purpose. The phonemes are then introduced to MBROLA System from TCTS Laboratory (Theorie des Circuits et Traitement du Signal laboratory (TCTS Labs) which concatenate diphones based on the list of phonemes and produces speech. In fact, research has been done to synthesize rough prosody based on punctuation (Al-Dakkak *et al*., 2005)

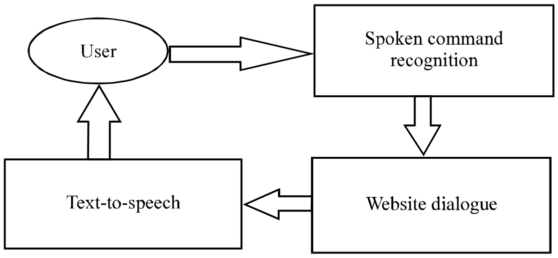

Fig. 1: The structure of the system

Spoken command recognition: As for the speech recognition researchers have used JavaScript library called Annyang which is a free library, performing Multi-Speaker recognition. Annyang is a tiny JavaScript library that lets the visitors of a site control this site with voice commands. Annyang supports multiple languages but does not support Arabic language. It has no dependencies, weighs just 2 KB and is free to use. As this library does not support Arabic Language, we have built an adapter from Arabic into Latin characters. For example, when the user gives the command meaning "link" this word is converted to "Rabit" and then executes the command voice. The fact that the application, an interactive site that performs specific voice commands, we will be using the individual sounds (phonemes) for the English-language most suited to the Arabic language, as shown in Table 1.

In fact, we just have few words to recognize in both English and Arabic versions. The list of the commands is given in Table 2. In addition, a correction strategy is adopted as follows: when the user utters a word not listed in the commands the system chooses the nearest command to it and asks the user if the command is the correct one.

Web site dialogue: This component handles the process of interaction between the user and the site. It begins by a welcoming phrase (welcome to Jad dialogue web site") and says the available options, either he recognizes the command or recognizes a command similar in pronunciation to an available uttered word or he doesn't recognize the command at all. Below are some sections of a dialogue between the user and the system (named "JAD"):

Case 1: The system does not recognize the voice command. Jad: Welcome to Jad dialogue web site; User: About France; Jad: Sorry, your command has not been found. Please enter another command or say help.

Table 1: Arabic phonemes used in the system

Phoneme	No. crafts contrast	Phoneme	No. crafts contrast
B	2	F	20
T	3	Q	21
T	4	K	22
Z	5	L	23
X	6	M	24
X	7	N	25
D	8	H	26
D	9	W	27
R	10	J	28
Z	11	?	Hamza
S	12	Sonic characters short	
S	13	A	Fatha
s.	14	I	Kasra
d.	15	U	Alddama
t.	16	Sonic characters long	
z.	17	a:	"1" in the last word
H	18	i:	"28" in the last word
G	19	u:	"27" in the last word

Table 2: Recognition rates of Arabic words

Words	Rates (%)
Like meaning "Story"	75
Like meaning "Political"	90
Like meaning "Summary"	85
Like meaning "Link"	95
Like meaning "First"	90
Like meaning "second"	90
Like meaning "Third"	95
Like meaning "yes"	90

Case 2: The system recognizes the voice command. Jad: Welcome to Jad dialogue web site. User: SUMMARY. Jad: Ok. Jad: Converting text on the page to speech.

Case 3: The system predicts the command, when it is not properly spoken. Jad: Welcome to Jad dialogue web site. User: New. Jad: if you mean "News", say "ok" or enter another command or say help.

And we have a similar cases in Arabic/. We can move on to the Arabic language through the voice command ("Go to Arabic") and we can go back to English language by voice command ("Rabir English").

RESULTS

The evaluation of the system was done by conducting a survey on ten blind people. The survey is composed of the following questions:

1. Is the idea of the application suits you?
2. Do you find it comfortable to use the application?
3. Do you react with the application easily?
4. Is the quality of Arabic synthesized Speech understandable?
5. Is the quality of English synthesized speech understandable?
6. Do you have a problem in executing the command you want?

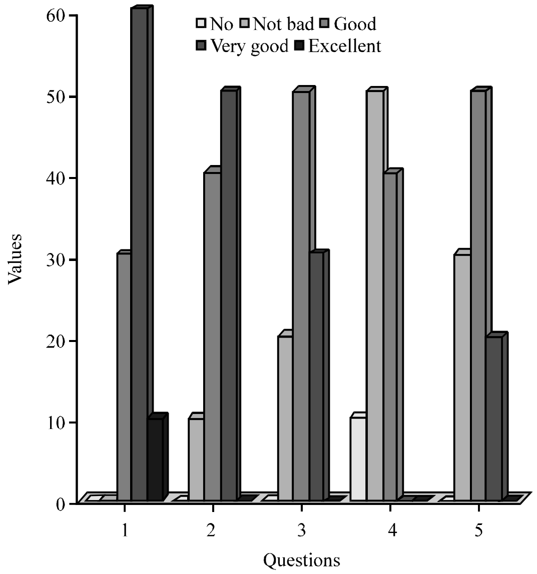

Fig. 2: Results of questions 1-5

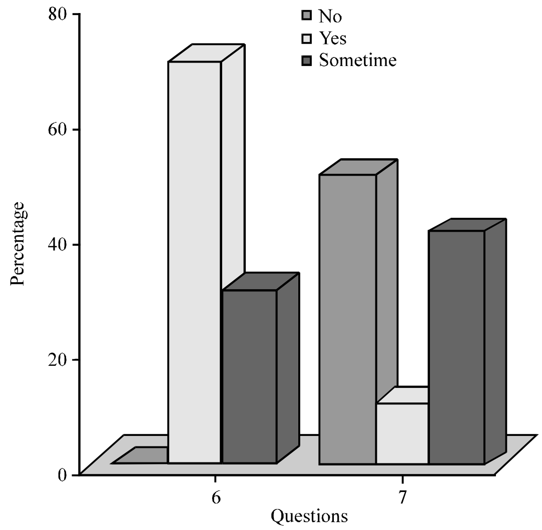

Fig. 3: Results of question 6 and 7

7. Do you feel bothered repeating the command back again?
8. What kind of web site content would you like to be in interactive speech?

And we got the results in Fig. 2-4. We notice from the previous figures that the general idea of the application has suited most of the blind people as they seemed satisfied using the application and they found that the way the application responds to them is convenient and easy. However, they all expressed that "Arabic speech synthesis process needs to be improved". And we also notice that there is no problem neither with the voice commands implementation, nor with repeating them. Moreover, most of them would like all the web sites to be interactive.

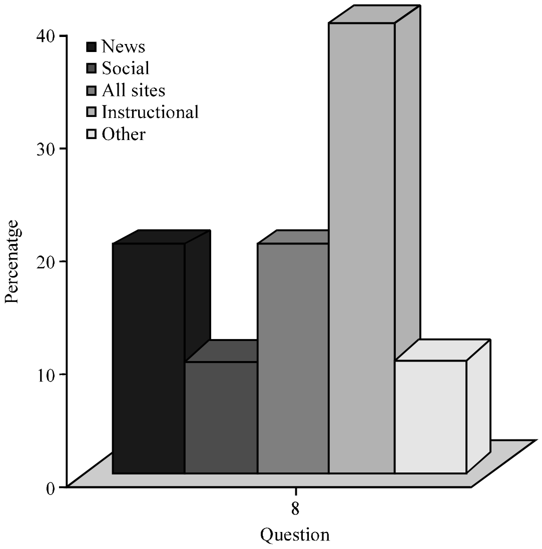

Fig. 4: Results of question 8

On the other hand and as the system is required to be a Multi-user Speech Recognition System and in order to evaluate the recognition rate, we have performed an evaluation test where we asked 25 people to pronounce a set of Arabic words used in the interface and calculated the recognition rate of each word. The results for the Arabic words are shown in Table 2. We remark that the recognition rates of the system are very good and acceptable.

CONCLUSION

In this study, we presented the approach to use Arabic speech synthesis and speech recognition techniques to build a verbal interactive web site of great significance to the community and especially to the blind and visually impaired computer users providing them with the capability of browsing web sites through speech. However, this research is a first step to make the web available to the blind and further research has to be done to integrate the research on automatic prosody generation for Arabic which is taking place in HIAST (Higher Institute for Applied Sciences and Technology) in order to enhance synthesized Arabic speech quality.

REFERENCES

Al-Dakkak, O., N. Ghneim, M.A. Zliekha and S. Al-Moubayed, 2005. Emotion inclusion in an Arabic text-to-speech. Proceedings of 13th European Signal Processing Conference EUSIPCO 2005, September 4-8, 2005, Antalya, Turkey, pp: 1-4.

Allen, J., M.S. Hunnicutt, D.H. Klatt, R.C. Armstrong and D.B. Pisoni, 1987. From Text to Speech: The MITalk System. Cambridge University Press, New York, USA.

Betkowska, A., K. Shinoda and S. Furui, 2007. Robust speech recognition using factorial HMMs for home environments. EURASIP J. Adv. Signal Process. 10.1155/2007/20593.

Gupta, V., J. Bryan and J.N. Gowdy, 1978. A speaker-independent speech recognition system based on linear prediction. IEEE Trans. Acoustics Speech Signal Process., 26: 27-33.

Herscher, M.B. and R.B. Cox, 1972. An adaptive isolated word speech recognition system. Proceedings of the Conference on Speech Communication and Processing, April 1972, Newton, MA., pp: 89-92.

Itakura, F., 1975. Minimum prediction residual principle applied to speech recognition. IEEE Trans. Speech Signal Process., 23: 67-72.

Jurafsky, D. and J.H. Martin, 2009. Speech and Language Processing. 2nd Edn., Chapter 24, Prentice Hall, New York.

Rabiner, L.R., 1989. A tutorial on hidden Markov models and selected applications in speech recognition. Proc. IEEE, 77: 257-286.

Rubin, P., T. Baer and P. Mermelstein, 1981. An articulatory synthesizer for perceptual research. J. Acoust. Soc. Am., 70: 321-328.

Thiang and S. Wijoyo, 2011. Speech recognition using linear predictive coding and artificial neural network for controlling movement of mobile robot. Proceedings of 2011 International Conference on Information and Electronics Engineering, May 28-29, 2011, Bangkok, pp: 179-183.