Voice Enabled Platform Independent System for Switching Among the Applications

Qamar-un-nisa Soomro, Akhtar Hussain Jalbani and Zahid Hussain

***Abstract*—The Speech has been potential mode of the interaction with computers now a days. The recognition of the speech is very much difficult for the computers due to a complexity of the human language. Therefore, in this research paper we have developed a platform and speaker independent Automatic Speech Recognition (ASR) system for English and Sindhi language. The system gets input from the user in the form of voice via microphone and acts according to the voice command. The user can open, close and switch between the applications using voice commands. The open command is used to open an application; close command is used for close an application and switch command is used for switch among the applications. The user pronounce the word switch with particular application i.e. switch notepad, same action will be performed for close and open applications. We have achieved 75% accuracy for Sindhi language and 85% accuracy for English language using our proposed system. The proposed system has been developed using Java, which provides platform independent features and an open-source speech recognition framework called CMU (Carnegie Mellon University) Sphinx4; it is a flexible, open source, the modular and pluggable framework which uses the statistical based approach (Hidden Markov Model) for speech recognition. The proposed system uses the model-based approach. Therefore, in the future, different language models can also be embedded in the system.**

***Index Terms*—CMU Sphinx, Speech Recognition, HMM, Speaker Independent.**

I. INTRODUCTION

Interaction between users and computers can be carried out using different resources: keyboard and mouse, a graphical user interface (GUI), a haptic environment, human conversation, etc. [1]. Speech is a natural mode of communication in human beings; it has been potential to being a fast and convenient mode of interaction with computer [2].

The proposed system developed to provide a natural language voice enabled user interface through which user can interact with the system via his or her voice, based on voice command action will be performed. Possible basic operations of the proposed system are open, close and switch between applications. The proposed system allows to open an application via voice command like **“open notepad”** system will launch the notepad application similarly; the user can close an application via **“close notepad”** command; the system will close the notepad application and **“switch notepad"** command will change the focus of the current window to notepad application. If a user had opened two applications, namely notepad and calculator, at a time only one window will be active, therefore, the focus will be on a recent window. If a user wants to change the focus of calculator or any application to notepad he or she simply say **“switch notepad”**.

II. RELATED WORK

Human-Computer Interaction (HCI) introducing different methods to interact with the system, one of the methods for interaction with computer is the natural language interface. Natural Language Processing (NLP) is the field of Artificial Intelligence (AI); it gives ability to the machine to understand the languages speak by the humans. NLP is further divided into two fields, Speech Recognition and Speech Synthesis.

Speech Recognition is also known as Automatic Speech Recognition (ASR) it is the field of Artificial Intelligence, which is the ability of computer to understand the natural language of humans. In other words, it is a process which converts the spoken words into the text with the help of speech recognizer, shown in Fig 1.

Speech Synthesis is also known as Automatic Speech Synthesis (ASS) it is a process to convert text into the Speech with the help of speech synthesizer, shown in Fig 2.

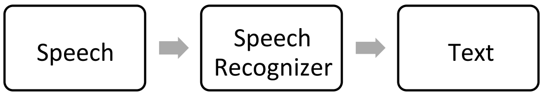


Fig. 1: Speech Recognition Process

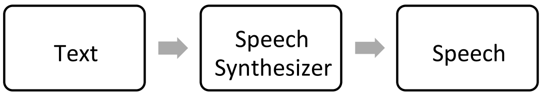


Fig. 2: Speech Synthesis Process

Qamar-un-nisa Soomro, Akhtar Hussain Jalbani, Department of Computer Science, Quaid-e-Awam University of Engineering, Science & Technology Nawabshah, Z. Hussain Department of Information Technology, Quaid-e-Awam University of Engineering, Science & Technology Nawabshah. Email: nisa1207@gmail.com

Speech recognition is classified into following categories [3].

A. Speaker dependent

The system designed for a single speaker, it is more accurate for the particular speakers but less for other speakers. These systems require to train system according to his or her voice. Usually, these systems are easier, cheaper and more accurate, but these systems are not as flexible as a speaker adaptive or speaker independent systems [3].

B. Speaker independent

The system designed for any speaker, and they do not require to train system according to his or her voice. These systems are very flexible but require most expensive and effort less accuracy than Speaker dependent model [3].

C. Isolated words

It is used for individual word recognition, it is easy to implement and have great accuracy of word success rate [3]. It is used for singular words command such as open, close, exit, switch, move, left, right etc.

D. Connected Words

It is used for multiple words' recognitions connected. It is implemented for command control system. The command may consist of multiple words' patterns defined by Context Free Grammar (CFG). The connected words command such as open notepad, close notepad, switch notepad, move left, move right etc.

E. Continuous Speech

It allows users to speak naturally, while the computer will determine the content using speech recognizer. It is basically used for computer dictation.

Speech recognition and Speech synthesis has been used various fields, few of them are: iSign [4], system developed by state university of New York, for aid families and deaf children. In iSign, the user will speak a word that will be recognized by the system, which will then view a video and visuals about that word. Another system is Multi-Purpose Speech Recognition and Speech Synthesis System take input in the form voice or typed, after that query will be processed resultant output can be visual or speech [5].

III. METHODOLOGY

Proposed system consist of the following different components shown in Fig. 3

- Input
- Feature Analysis
- Acoustic Model
- Grammar or Language Model
- Recognizer
- Dictionary or Lexicon
- Probabilistic Model
- Output

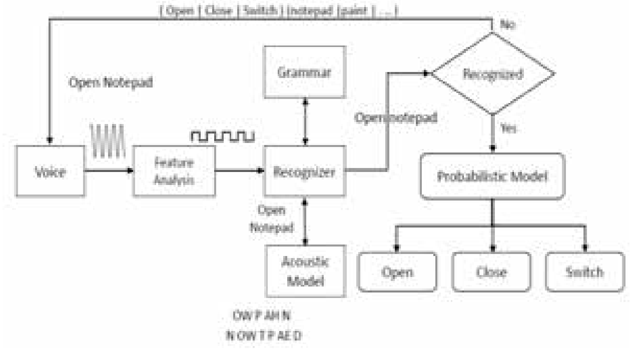


Fig. 3: Proposed system model

A. Input

To give input to the system we required a good quality microphone. The system is initiated by the user by giving the input in the form of vocal commands using a microphone, it receives the sound and then converts the sound waves into the electrical pulse. The sound card is responsible to convert these electrical pulses into the digital signals [6].

B. Feature Analysis

This component is responsible to identifying the linguistic content and discarding all the other stuff, which carries information like background noise, emotion, etc. [3]. The System uses the Sphinx-4 [7] [8] Java's tool kit for speech recognition, which uses the Mel Frequency Cepstral Coefficients (MFCC) [9] for feature extraction purpose.

C. Acoustic Model

The acoustic model is consist of the sub-words are called phonemes, which collectively forms the word. The acoustic model is responsible to convert spoken words into phonemes and from phonemes to words. The English languages uses about 44 phonemes to convey 50,000 or more words it contains, The speech recognition engine work with it to produce best results' [6]. Words will get from the extracted phonemes few of the phonemes are listed in Table 1. Sphinx4 speech recognizer uses the well-known probabilistic model called Hidden Markov Model [10] [11], it works based on probabilities, each word has numerous phonemes , every word has its own HMM represented by graph with different nodes and their transitions from one state to another as shown in Fig.4.

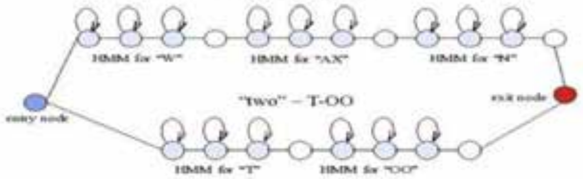


Fig. 4: HMM Word Model [5]

TABLE I. List of phonemes used to form words [6]

Words	Phonemes
Father	Aa
Cat	Ae
Cut	Ah
Foul	Aw
Sing	Ng
Talk	T
Thin	Th
Book	Uh
Too	Uw
Pleasure	Zh

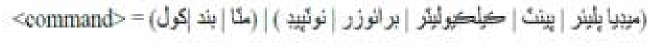
D. Grammar or Language Model

It defines the rules for the language. In other words, it defines the pattern of the sentences or corpus, which defines the sequence of possible voice commands. The proposed system command structure or pattern is shown in Fig. 5.

In other words, the pattern of the sentence or structure of the vocal commands is represented by grammar which shows in which order vocal commands to speak. The command structure of English and Sindhi language of proposed system are defined using context free grammar shown in Figure 5 and Figure 6 respectively.

<command> = (open | start | switch | move | close | exit) |
(notepad | browser | calculator | paint | media player)

Fig. 5: Command structure of English voice command



<command> = (كول|بند|مٿا) | (نوٽپيڊ | برائوزر | ڪيلڪيوليٽر | پينٽ | ميڊيا پليئر)

Fig. 6: Command structure of Sindhi voice command

We can say that the vocal command may consist of two to three words. First word of command shows the action either user wants to open, close or switch between the applications. Therefore, first word can be any of the given words i.e. open, start, switch, move, close, exit. Second word of the vocal command shows application name that can be any of the given application names i.e. notepad, Browser, calculator, paint and media player.

E. Dictionary or Lexicon

It is the component of the speech recognition engine, consists of all possible language words with their pronunciation, will be used in defined language. The dictionary of proposed System consists of English and Sindhi words, listed in Table 2 & Table 3.

TABLE II. List of words in English dictionary

Word	Pronunciation
START	S T AA R T
OPEN	OW P AH N
CLOSE	K L OW S
CLOSE (2)	K L OW Z
EXIT	EH G Z IH T
EXIT (2)	EH K S AH T
SWITCH	S W IH CH
MOVE	M UW V
NOTEPAD	N OW T P AE D
BROWSER	B R AW Z ER
CALCULATOR	K AE L K Y AH L EY T ER
PAINT	P EY N T
MEDIA	M IY D IY AH
PLAYER	P L EY ER

TABLE III. List of words in Sindhi dictionary

Word	Pronunciation
کول	K HH AW L
مٿا	M AA T AH
بند	B AE N D
نوٽپيڊ	N OW T P AE D
برائوزر	B R AW Z ER
ڪيلڪيوليٽر	K AE L K Y AH L EY T ER
پينٽ	P EY N T
ميڊيا	M IY D IY AH
پليئر	P L EY ER

6) Recognizer

The recognizer is responsible for speech to text conversion; the recognizer communicates with the language model and acoustic model for verification either spoken words are present in the dictionary or not and either spoken command is said in specified order (defined using CFG) or not. If user speaks the words which are present in a dictionary and words are spoken in defined order, then according to the spoken command action will be performed.

7) Probabilistic Model

The CMU Sphinx4 uses the Probabilistic model HMM [3] , which handles the uncertain situations using language model, which checks the text command generated by recognizer, which approximately match with the language model and final action will be performed.

8) *Output*

The possible output of the proposed system are to open, close and switch between applications, further detail is present in IV section.

IV. EXPERIMENTAL RESULTS

This section represents that how to access the proposed system using voice command. Table 4 & Table 5 represents the English and Sindhi language voice commands with their corresponding action.

TABLE IV. System commands with corresponding actions

Command	Action
Open notepad	Launch the notepad
Open calculator	Launch the calculator
Open browser	Launch the browser
Open media player	Launch the media player
Open paint	Launch the paint
Close notepad	Exit the notepad
Close calculator	Exit the calculator
Close browser	Exit the browser
Close media player	Exit the media player
Close paint	Exit the paint
Switch notepad	Switch to the notepad
Switch calculator	Switch to the calculator
Switch browser	Switch to the browser
Switch media player	Switch to the media player
Switch paint	Switch to the paint

If a user wants to open an application like Notepad, then it can open by ***"Open Notepad"*** command. Similarly, we can close an application via close command. If user wants to close an application namely notepad, then it can close by ***"Close Notepad"*** command. Whenever user wants to switch between applications, there are will be condition checked before going to switch. First case is whenever user wants to switch an application which is not opened already, then system will launch that particular application. For example initially one application is opened namely notepad and user is commanding to switch to the calculator application which is not opened but system will launch the application calculator and will be the active window as shown in Figure 7.

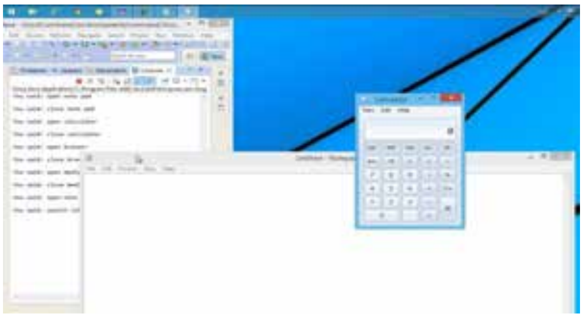


Fig. 7: ***"Switch Calculator"*** application will launch the calculator application because it is not opened

TABLE V. List of Sindhi command list

Action	Command
Launch the notepad	کول نوٽپيڊ
Launch the calculator	کول ڪيلڪيوليٽر
Launch the browser	کول برائوزر
Launch the paint	کول پينٽ
Launch the media player	کول ميڊيا پليئر
Close the notepad	بند نوٽپيڊ
Close the calculator	بند ڪيلڪيوليٽر
Close the browser	بند برائوزر
Close the paint	بند پينٽ
Close the media player	بند ميڊيا پليئر
Switch to the notepad	مٽا نوٽپيڊ
Switch to the calculator	مٽا ڪيلڪيوليٽر
Switch to the browser	مٽا برائوزر
Switch to the paint	مٽا پينٽ
Switch to the media player	مٽا ميڊيا پليئر

Figure 7 shows that two applications are already open namely ***notepad*** and ***calculator****,* at a time only one window can be active so in this case calculator window is active and focused. Figure 8 shows that user is switching from calculator to the notepad using ***"Switch Notepad".***

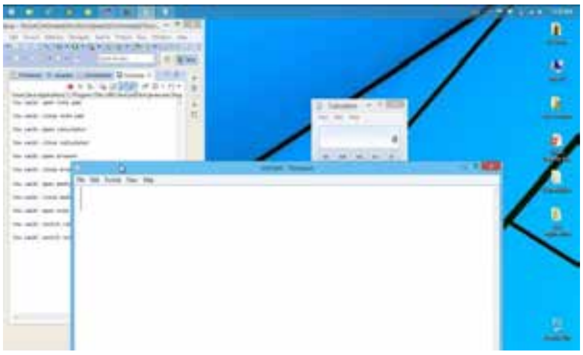


Fig. 8: ***"Switch Notepad"*** focus will change to the notepad

V. SYSTEM PERFORMANCE

System performance of the proposed system evaluated by the Command Success Rate (CSR) using hit and trial method, we had given Twenty Five (25) times command to the system for each operation i.e. open, close and switch. The success ratio of commands shown in Figure 9 & Figure 10 for English and Sindhi respectively. Figure 9 shows that user asked 25 times for open, close and switch command. Take the result for the notepad application, as figure shows that 21 out of 25 commands were successfully recognized by the system as open notepad application and rest of the commands were not recognized by the system. For closing notepad application, 22 commands out of 25 commands were successfully recognized by the system as close notepad application and rest of commands didn't recognized correctly. For switching notepad application, 20 commands out of 25 commands were successfully recognized by the system as switch notepad application and rest of commands recognized wrong.

Similarly, command success rate of defined applications calculated and represented with the help of confusion matrix as shown in Table 6, 7, 8, 9 and 10. The figure 11 shows overall recognition rate of each application using English language voice commands.

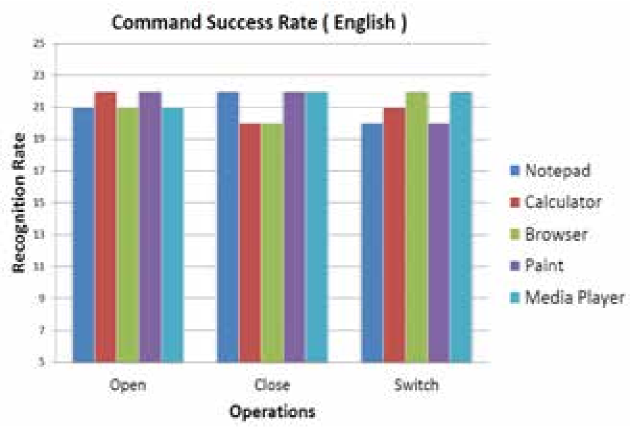


Fig. 9: Command success rate of English commands

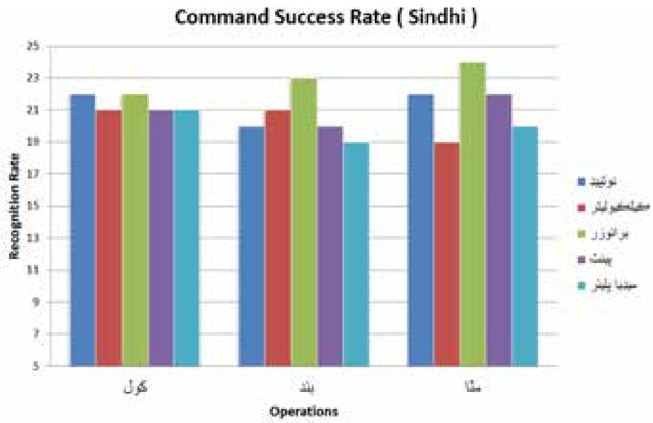


Fig. 10: Command success rate of Sindhi commands

TABLE VI. Overall accuracy of notepad application command

	Open	Close	Switch	Overall Accuracy
Open	21	2	2	84 %
Close	3	22	0	
Switch	3	2	20	

TABLE VII. Overall accuracy of browser application command

	Open	Close	Switch	Overall Accuracy
Open	21	2	2	84%
Close	3	20	2	
Switch	1	2	22	

TABLE VIII. Overall accuracy of calculator application command

	Open	Close	Switch	Overall Accuracy
Open	22	2	1	84 %
Close	3	20	2	
Switch	2	2	21	

TABLE IX. Overall accuracy of paint application command

	Open	Close	Switch	Overall Accuracy
Open	22	1	2	85.33 %
Close	1	22	2	
Switch	2	3	20	

TABLE X. Overall accuracy of media player application command

	Open	Close	Switch	Overall Accuracy
Open	21	1	3	86.66 %
Close	2	23	0	
Switch	1	2	22	

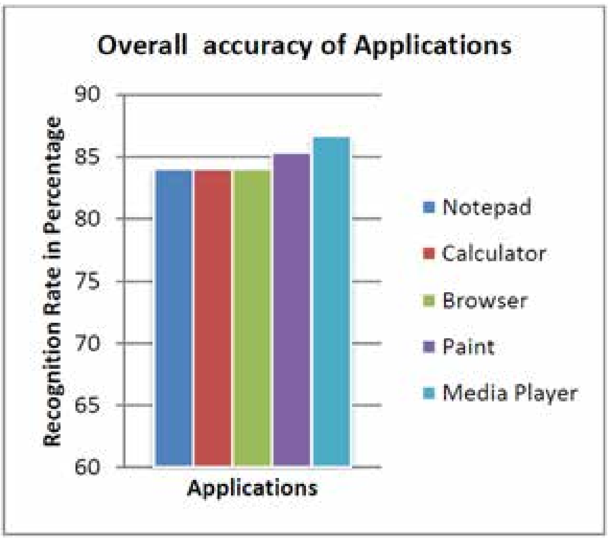


Fig. 11: Overall command success rate of each application

Figure 10 shows the command success rate of each operation in Sindhi language voice command such as ڪہول ا□م and بند The command success rate of each application is calculated and represented with the help of confusion matrix as shown in Table 11, 12, 13, 14 and 15.

TABLE XI. Overall accuracy of notepad application command in Sindhi

	کول	بند	مٽا	Overall Accuracy
کول	22	1	2	85.33 %
بند	3	20	2	
مٽا	0	3	22	

TABLE XII. Overall accuracy of calculator application command in Sindhi

	کول	بند	مٽا	Overall Accuracy
کول	21	4	0	81.33 %
بند	2	21	2	
مٽا	3	3	19	

TABLE XIII. Overall accuracy of browser command in Sindhi

	کول	بند	مٽا	Overall Accuracy
کول	22	0	3	92 %
بند	0	23	2	
مٽا	1	0	24	

TABLE XIV. Overall accuracy of paint command in Sindhi

	کول	بند	مٽا	Overall Accuracy
کول	21	0	4	84 %
بند	3	20	2	
مٽا	2	1	22	

TABLE XV. Overall accuracy of media player application command in Sindhi

	کول	بند	مٽا	Overall Accuracy
کول	21	0	4	80 %
بند	4	19	2	
مٽا	3	2	20	

The figure 12 shows overall recognition rate of each application using Sindhi language voice command.

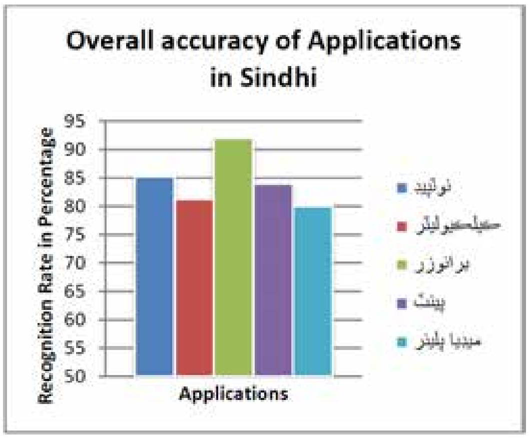


Fig. 12: Overall command success rate of each application

VI. CONCLUSION & FUTURE WORK

The research goal is to develop a voice based interface through which user can interact with the system to manipulate the computer promptly. The developed application is platform independent, which is applicable for multiple operating systems, with the help of java and CMU sphinx4 speech recognition engine we have achieved goal of the research. The proposed system works for English and Sindhi language. The CMU Sphinx provides the trained acoustic model for the English language which gives the accuracy almost more than 85% to 90% while for Sindhi language we obtain 70% to 75% accuracy. If environment is noise free and quality of the microphone is good enough to filter out the background noise or emotions, then this system will perform well, if commands are spoken in noisy environment and quality of a microphone is not enough to filter out the background sound, then system is will perform disorder results. This system is based on the model-based technology, which helps to integrate multiple environments.

In future this research can be carried out for further research: to enhance vocabulary size includes English and Sindhi words for different operations e.g. message, send, receive, up, down, left, right, new, delete, cancel, ok and move. Similarly for Sindhi dictionary e.g. پيغَام، موڪل، وصول ڪر، هيٺ، مٿي، ڳاٻي طرف، ساڄي طرف، نئون، ختم ڪر، بس ڪر، محفوظ ڪر، ٺيڪ آ، and اڳتي etc. for performing different operations on the system. In addition system can be extended to command control with dictation which facilitate to dictate the system which can be used for send email, message and for writing document without use of keyboard, user can directly dictate the system via his or her voice.

ACKNOWLEDGMENT

The authors would like to thank Information Technology department of Quaid-e-awam University of Engineering, Science & Technology Nawabshah for supporting this research work. This research work is taken from the MS (IT) thesis done by Qamar-un-nisa Soomro in 2015.

REFERENCES

[1] R. Justo, O. Saz, A. Miguel, M. I. Torres, and E. Lleida, "Improving language models in speech-based human-machine interaction," *Int J Adv Robotic Sy*, vol. 10, no. 87, 2013.

[2] K. Sharma, T. Suryakanthi, and T. Prasad, "Exploration of speech enabled system for english," *arXiv preprint arXiv:1304.8013*, 2013.

[3] O. P. Prabhakar and N. K. Sahu, "A survey on: Voice command recognition technique," *International Journal of Advanced Research in Computer Science and Software Engineering*, vol. 3, no. 5, 2013.

[4] T. Scarlatos, L. L. Scarlatos, and F. Gallarotti, "isign: Making the benefits of reading aloud accessible to families with deaf children." in *Computer Graphics and Imaging*, 2003, pp. 74–78.

[5] H. M. Karim Jahed, Marwan Fawaz, "Multi-purpose speech recognition and speech synthesis system," *IEEE MULTIDISCIPLINARY ENGINEERING EDUCATION MAGAZINE*, vol. 6, 2011.

[6] A. J. Kadam, P. Deshmukh, A. Kamat, N. Joshi, and R. Doshi, "Speech oriented computer system handling," in *International Conference On Intelligent Computational Systems (ICICS'2012)*, 2012.

[7] W. Walker, P. Lamere, P. Kwok, B. Raj, R. Singh, E. Gouvea, P. Wolf, and J. Woelfel, "Sphinx-4: A flexible open source framework for speech recognition," 2004.

[8] P. Lamere, P. Kwok, E. Gouvea, B. Raj, R. Singh, W. Walker, M. Warmuth, and P. Wolf, "The cmu sphinx-4 speech recognition system," in *IEEE Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP 2003), Hong Kong*. Citeseer, 2003, pp. 2 – 5.

[9] L. Muda, M. Begam, and I. Elamvazuthi, "Voice recognition algorithms using mel frequency cepstral coefficient (mfcc) and dynamic time warping (dtw) techniques," *arXiv preprint arXiv:1003.4083*, 2010.

[10] L. R. Rabiner, "A tutorial on hidden markov models and selected applications in speech recognition," *Proceedings of the IEEE*, vol. 77 , no. 2, pp. 257–286, 1989.

[11] M. LLC. (2003) Shinx4. [Online]. Available: http://cmusphinx.sourceforge.net/wiki/tutorialsphinx4