Sindh Univ. Res. Jour. (Sci. Ser.) Vol.45 (3) 634-639 (2013)

SINDH UNIVERSITY RESEARCH JOURNAL (SCIENCE SERIES)

Speech Recognition based web scripting from predefined Context Free Grammar (Language Model & Grammar) programmed in Visual Programming and Text Editor

NADEEM AHMED. KANASRO, H.U. ABBASI, M.R. MAREE, A.G. MEMON

Institute of Mathematics and Computer Science, University of Sindh, Jamshoro, Pakistan
Email: nadeemahmedkanasro@yahoo.com;habibullah.abbasi@gmail.com;mujeeb@usindh.edu.pk;ghafoor@usindhedu.pk

Corresponding author: Nadeem Ahmed Kanasro, Email: nadeemahmedkanasro@yajhoo.com Ph. No. +92-0333-2753560

Abstract: This paper presents implementation of our designed Language Model & Grammar and its analysis, designing, development and implementation in an application. In this research article we proposed 'Speech Recognition Application' named 'Text Editor Through Voice', which is operated as 'Speech Recognition (Speaker Dependent) System'. The approach is based on experiencing the praxis using 'Hidden Morkov Model' and application is designed in Visual Basic 6.0 using 'Visual Programming' and 'Object Oriented Programming' methods. In 'Text Editor Through Voice' the use of Speech Recognition engine translates spoken input before finding the specified syntax and tags stored in database. After finding and matching recognized input from database it put that in document area of text editor just like typing on keyboard and pressing a key on the phone keypad, in this application microphone able to do this. For example one might say a word like "HTML", to which application replies by inserting said word in the document area. Furthermore, we show you list of words and phrases in tables with figures that are successfully implemented and executed in our developed application.

Keywords: Language Model & Grammar, Text Editor Through Voice, Speech Recognition Engine, HMMs, DTW

1. INTRODUCTION

The designing and developing a computer application for a machine that mimics person activities, mostly the ability of talking naturally and responding appropriately to spoken language, has intrigued engineers and scientists for centuries. Since the 1930s, when Homer Dudley of Bell Laboratories projected a system model for speech study and synthesis (H. Dudley, 1939), the problem of automatic speech recognition has been approached progressively (Goel, V. Byrne, W. J, 2000), from a simple instrument that responds to a small set of sounds to a complicated system that responds to effortlessly spoken natural language and takes into description the varying information of the language in which the speech is produced. Based on major advances in statistical modeling of speech in the 1980s (Hongbing Hu, Stephen A. Zahorian, 2010), automatic speech recognition systems today find extensive application in farm duties that require a human-machine interface.

Speech based applications are developed to perform different tasks such types of applications are given below;

1 **Simple data entry:** These types of applications are used to enter numbers, characters, and phonemes. For example: entering a credit card number

2 **Voice user interfaces:** These types of applications are used to make a call by (VCD) voice command device, these applications fall into different categories like
  - Voice activated dialing
  - Routing of Calls

3 **Domestic appliance control:** These types of applications are used to control home appliances, for example: turn off tube lights, where particular words are spoken.

4 **Preparation of structured documents:** These types of applications are used in medical science to create reports, for example: a radiology report.

5 **Speech-to-text processing**: These types of applications are used to dictate, process spoken words, word processors or emails are examples of these applications.

Speech recognition is the transformation of verbal inputs known as words, phrases or sentences into content. It is also known as 'Speech to Text', 'Computer Speech Recognition' or 'Automatic Speech Recognition'. It is one kind of technology and was first introduced by AT&T Bell Laboratories in the year 1930s.

Some speech recognition based systems use "trainings" where speakers read a chunk of text. These systems examine specific voice of the individual and use it to fine tune the detection of that person's speech, resulting in more correct transcription. Training based systems are called Speaker Dependent systems while non training based systems are called Speaker Independent systems.

The speech recognition process is performed by a software component known as the **speech recognition engine**. The primary function of the speech recognition engine is to process spoken input and translate it into text that an application understands.

Figure# 1 shows that Speech recognition engine requires two types of files to recognize speeches, which are defined below.

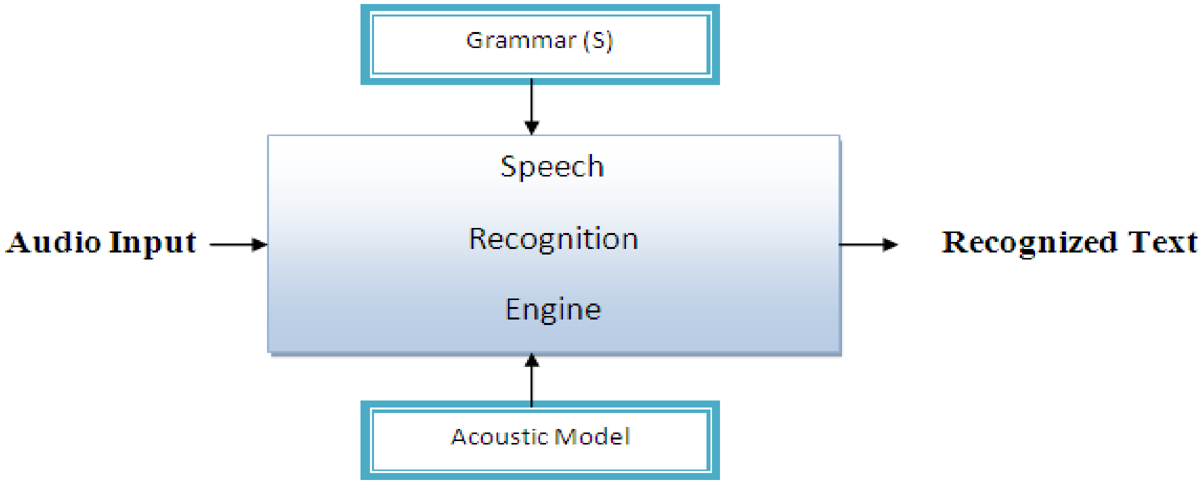

Figure #1: Speech Recognition Engine Components

1- Language Model or Grammar: A Language Model is a file containing the probabilities of sequences of words. A Grammar is a much smaller file containing sets of predefined combinations of words. Language Models are used for 'Dictation' applications, whereas Grammars are used as desktop 'Command and Control' applications.

2- Acoustic Model: Contains a statistical representation of the distinct sounds that make up each word in the Language Model or Grammar. Each distinct sound corresponds to a phoneme.

2, <u>ALGORITHMS AND MODELS</u>

2.1. Dynamic Time Warping:

The Dynamic Time Warping (DTW) is an algorithm, it was introduced in 1960s (R. Bellman and R. Kalaba, 1959). It is an important and aged algorithm was used in speech recognition systems known as Dynamic Time Warping algorithm (Vintsyuk 1971, Itakura 1975, Sakoe and Chiba 1978). It is used to measure the resemblance of objects/Sequences in the form of speed or time. For example similarity would be detected in running pattern where in film one person was running slowly and other person was running fast. This algorithm can be applied to any data; even data is graphics, video or audio. It analyzes data by turning into a linear representation.

This algorithm is used in many areas: Computer animation, Computer vision, Data mining (V. Niennattrakul and C. A. Ratanamahatana, 2007), online signature matching, signal processing (M. Muller, H. Mattes, and F. Kurth, 2006), gesture recognition and speech recognition (C. Myers, L. Rabiner, and A. Rosenberg, 1980).

2.2. Hidden Morkov Model

It is modern general purpose algorithm. It is widely used in speech recognition systems because of that statistical models are used by this algorithm, which creates output in the form of series of quantities or symbols. It is based on statistical models that output a series of symbols or quantities (Goel, V. Byrne, W. J. 2000) & (Mohri, M. 2002).

2.3. Neural Networks

Neural networks were created in the late 1980s. These were emerging and an attractive acoustic modeling approaches used in Automatic Speech Recognition (ASR). From the time then these algorithms have been used in various speech based systems such as phoneme categorization (A. Waibel, T. Hanazawa, G. Hinton, K. Shikano, and K. J. Lang, 1989) speaker adaptation and isolated word recognition (S.A. Zahorian, A. M. Zimmer, and F. Meng, 2002),. These algorithms are attractive recognition models for speech recognition because they formulate no assumptions as compared to Hidden Markov Models regarding feature statistical properties. This algorithm is used as preprocessing i.e; dimensionality reduction (Hongbing Hu, Stephen A. Zahorian, 2010) and feature transformation for Hidden Markov Model based recognition.

2. RESULTS AND DISCUSSION

This research is concentrated on programming of Language Model and Grammar. As discussed in introduction section that Speech Recognition Engine requires two types of files to recognize inputs. First is the Language/Grammar model and the second is Acoustic Model. So we have created one language model and one grammar
Those models/grammars are;

1) **IDE (Integrated Development Environment):** The Grammar designed in this module can be used to operate Text Editor by Microphone.

2) **HTML (Hypertext Markup Language):** The Language Model designed in this module can be used to create web script by microphone in Text Editor.

In the Figure# 2 we have shown the implementation of language model and grammar model in speech recognition engine.

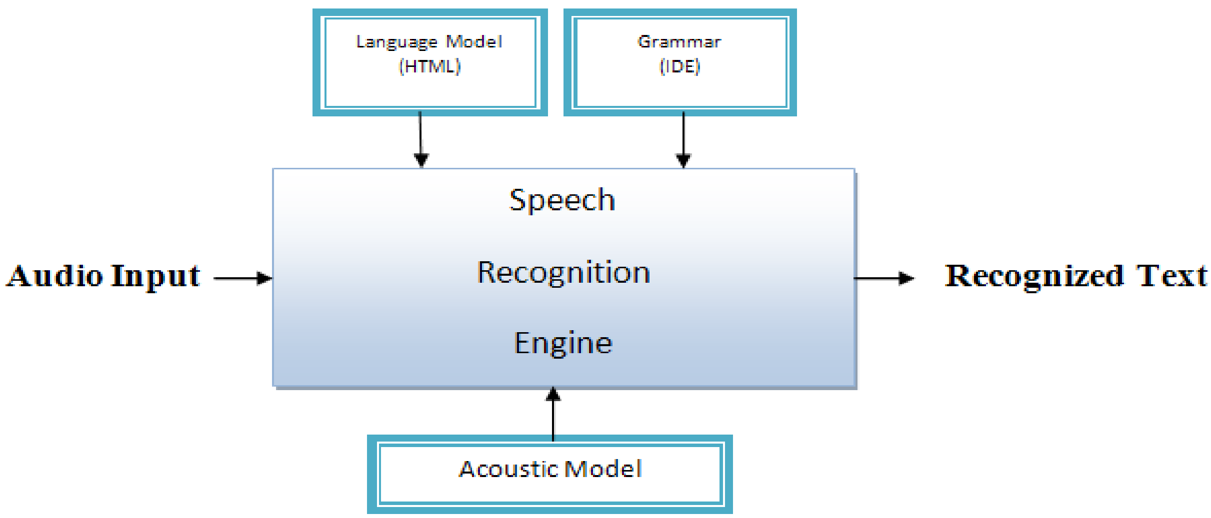

Figure #2: Implementation of Language Model & Grammar

2. APPLICATION SNAPSHOTS AND RESULTS OF PROGRAMMED LANGUAGE MODEL AND GRAMMAR

Figure #3 is GUI (Graphical User Interface) of our designed application. In the left side of image four microphone icons are displayed. Names of these icons are:

- Dictionary
- HTML (Language Model)
- IDE (Grammar)
- S.Characters

The icon namely HTML is linked to the programmed Language Model and the icon namely IDE is linked programmed (Grammar).
In the right side of image four other icons are displayed. Their names are

- Database
- Contents
- About Project
- Programmer

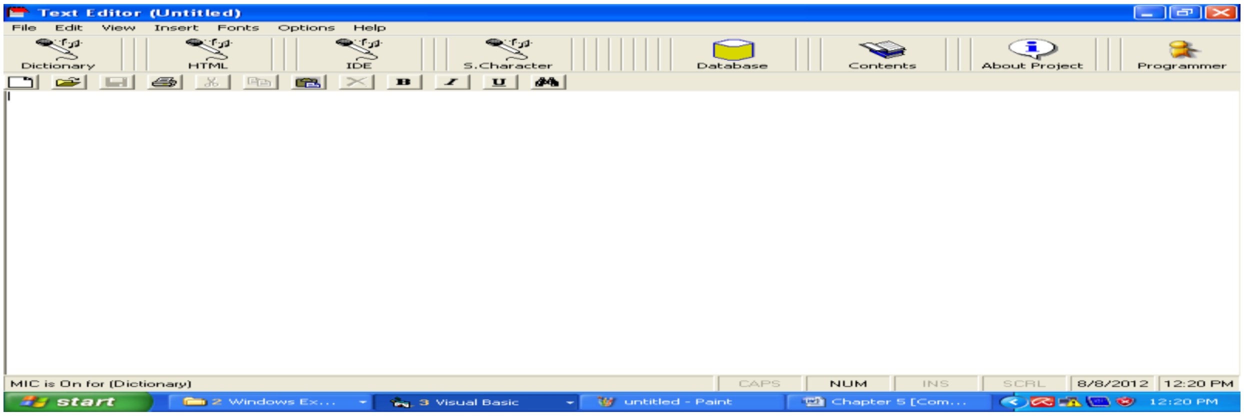

Figure #3: (Text Editor Through Voice) Editor Window

4.1. IDE: This is Grammar used as command and control purpose. Phrases and descriptions are given in Table #1 and Figure #4 shows date and time function is called by speaking corresponding phrase into MIC.

List of Phrases	Description of Phrase
New	To Open new document
Open	To Open saved document
Save	To Save Document
Save As	To Save document with new name
Print	To Print document
Exit	To Exit Text Editor
Delete	To Delete selected text
Cut	To Cut selected text
Copy	To Copy selected text
Paste	To place cut or copied text
Find	To Search text from document
Replace	To Replace document
Select All	To Select All Text
Time	To Insert time in document
Tool Bar	To Call tool bar function
Status Bar	To Call status bar function
Standard Buttons	To Call standard buttons function
Date and Time	To Insert date and time in document
Bold	To change the format of text as Bold
Italic	To change the format of text as Italic
Underline	To change the format of text as underline
Font	To Call font function
Color	To Call color function
Dictionary	To call Dictionary function
HTML	To call HTML function
IDE	To call IDE function
Special Characters	To call special character function
Database	To call database wizard function
De Activate	To Off MIC
Capital Characters	To call capital character function
Small Characters	To call small character function
About Me	To know about Application Developer
About Project	To know about Project Description
Contents	Help and Index

Table #1: List of phrases to control IDE

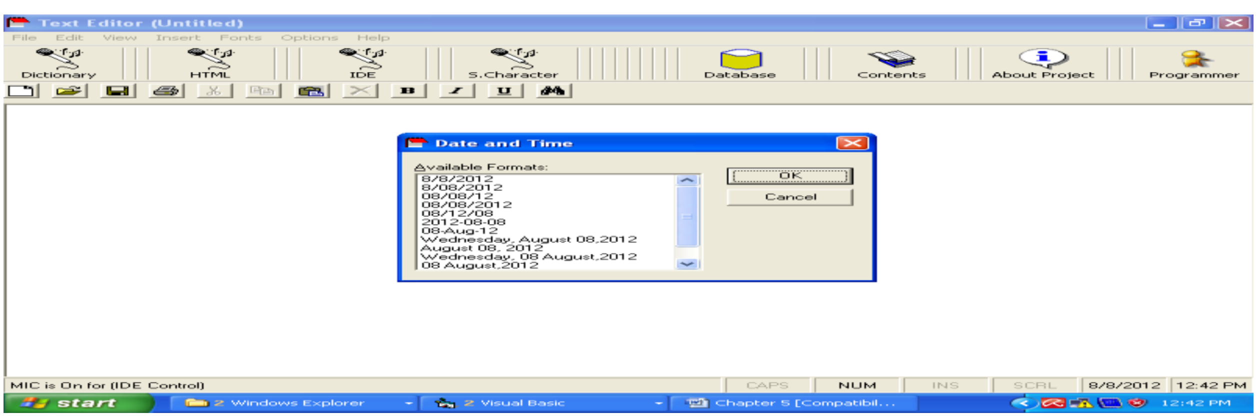

Figure #4: (Editor Window) Editing by MIC (Selected function is IDE)

4.2. HTML: This is Language Model used to create simple web scripts based on dictation. Words and phrases for their corresponding HTML Tags are given in Table #2 and Figure #5 shows simple web script created by speaking their phrases into MIC.

Phrases	Opening Tags	Phrases	Closing Tags
HTML	<HTML>	Close HTML	</HTML>
HEAD	<HEAD>	Close HEAD	</HEAD>
TITLE	<TITLE>	Close TITLE	</TITLE>
Body	<Body>	Close Body	</Body>
Image	<Image>	---	---
B	<B>	Close B	</B>
I	<I>	Close I	</I>
U	<U>	Close U	</U>
Center	<Center>	Close Center	</Center>
Font	<Font>	Close Font	</Font>
HR	<HR>	Close HR	</HR>
BR	 	Close BR	</BR>
P	<P>	Close P	</P>
Table	<Table>	Close Table	</Table>
TH	<TH>	Close TH	</TH>
TR	<TR>	Close TR	</TR>
TD	<TD>	Close TD	</TD>
H1	<H1>	Close H1	</H1>
H2	<H2>	Close H2	</H2>
H3	<H3>	Close H3	</H3>
H4	<H4>	Close H4	</H4>
H5	<H5>	Close H5	</H5>
H6	<H6>	Close H6	</H6>
Sub	_{	Close Sub	}
Sup	^{	Close Sup	}
Marquee	<Marquee>	Close Marquee	</Marquee>
Frame	<Frame>	Close Frame	</Frame>
Frameset	<Frameset>	Close Frameset	</Frameset>
Form	<Form>	Close Form	</Form>
Input	<Input>	---	---
Select	<Select>	Close Select	</Select>
Option	<Option>	---	---
Text Area	<Textarea>	Close Text Area	</Textarea>

Table #2: List of Phrases and HTML Tags

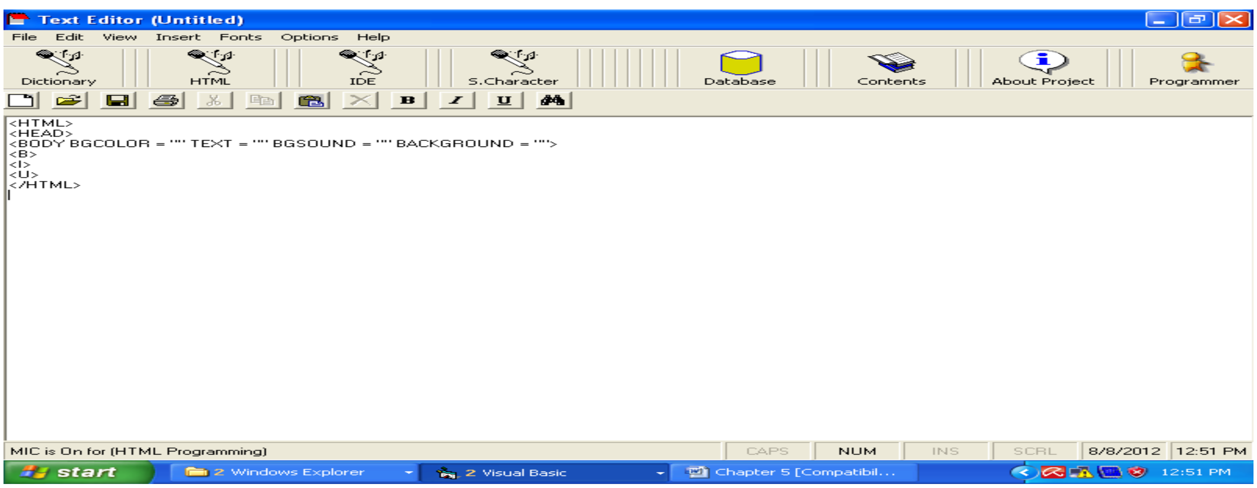

Figure #5: (Editor Window) Editing by MIC (Selected function is HTML)

2. CONCLUSION

It is learnt through this project that the professional work is necessary to be carried out in the software industry. It is proved through study to develop Speech based application named Text Editor Through Voice. Here are designed two types of Language Models/Grammars, and have classified them as dictation and command & control grammars. Further a concept have portrayed that computer programmers can create simple web scripts through the speech base process. It is concluded through Graphical User Interface (GUI) and outputs, to make it possible, to create web scripts via speaking commands. The study is implemented in designed grammar in speech recognition engine in order to prove the solution, which is technically feasible.

The work on this application is oriented to direction as a commercial system. Hidden Markov Model is used in usually speech recognition software, which integrates an acoustic model, a large vocabulary file. One of the software version used for the test phase is equipped with a vocabulary gathering more than 10000 current and specialized words. Ideally, the use of a voice recognition system, really speaker independent, like the Microsoft word processor, we want to improve the robustness of the designed language model and grammar in the whole application. Best results will be selected among different speech recognition engines after implementation according to a framework working at the same time. The software industry is using the machines but it is little tried to create some applications for commercial purposes. Our future work is oriented in this direction as commercial systems. In this scenario it is tried to level best to develop Speech Based Text Editor, which is working properly and needed more improvements as a successful commercial product. For example:

- This application cannot understand any input if was spoken in other than English language. There is acute need to develop an Editor for Sindhi and Urdu languages.
- This application needs perfect pronunciation, sound proof of environment and having no noise.
- There is need to develop the same application in .Net Framework for latest equipments and easy to access on each platform.

REFERENCES

A. Waibel, T. Hanazawa, G. Hinton, K. Shikano, and K. J. Lang, (1989) "Phoneme recognition using time-delay neural networks," IEEE Transactions on Acoustics, Speech and Signal Processing, vol. 37, pp. 328-339.

C. Myers, L. Rabiner, and A. Rosenberg, (1980), Performance trade of in dynamic time warping algorithms for isolated word recognition," Acoustics, Speech, and Signal Processing [see also IEEE Transactions on Signal Processing], IEEE Transactions on, vol. 28, no. 6, pp. 623-635.

Goel, V.; Byrne, W. J. (2000). "Minimum Bayes-risk automatic speech recognition". *Computer Speech & Language* 14 (2): 115–135. doi:10.1006/csla.2000.0138. Retrieved 2011-03-28.

H. Dudley, R. R. Riesz, and S. A. Watkins, (1939) *A Synthetic Speaker*, J. Franklin Institute, Vol. 227, pp. 739-764.

Hongbing Hu, Stephen A. Zahorian, (2010) "Dimensionality Reduction Methods for HMM Phonetic Recognition," ICASSP 2010, Dallas, TX

Itakura, F. (1975). Minimum Prediction Residual Principle Applied to Speech Recognition. *IEEE Trans. on Acoustics, Speech, and Signal Processing*, 23(1):67-72, February 1975. Reprinted in Waibel and Lee (1990).

J. Wu and C. Chan,(1993) "Isolated Word Recognition by Neural Network Models with Cross-Correlation Coefficients for Speech Dynamics," IEEE Trans. Pattern Anal. Mach. Intell., vol. 15, pp. 1174-1185.

M. Muller, H. Mattes, and F. Kurth, (2006) An ancient multiscale approach to audio synchronization," pp. 192-197.

R. Bellman and R. Kalaba, (1959) On adaptive control. pocesses,"Automatic Control, IRE Transactions on, vol. 4, no. 2, pp. 1-9.

S. A. Zahorian, A. M. Zimmer, and F. Meng, (2002) "Vowel Classification for Computer based Visual Feedback for Speech Training for the Hearing Impaired," in ICSLP 2002.

Sakoe, H. and Chiba, S. (1978). Dynamic Programming Algorithm Optimization for Spoken Word Recognition. *IEEE Trans. on Acoustics, Speech, and Signal Processing*, 26(1):43-49, February 1978. Reprinted in Waibel and Lee (1990).

V. Niennattrakul and C. A. Ratanamahatana, (2007) On clustering multimedia time series data using k-means and dynamic time warping," in Multimedia and Ubiquitous Engineering, 2007. MUE '07. International Conference on, pp. 733-738.

Vintsyuk, T. (1971). Element-Wise Recognition of Continuous Speech Composed of Words from a Specified Dictionary. *Kibernetika* 7:133-143.