Testing Speech to Text Application on Android Platform

Theint Zarni Myint
University of Computer Studies Yangon
chutpit@gmail.com

Abstract

Speech recognition reduces the overhead caused by alternate communication methods. Speech has not been used much in the field of electronics and computers due to the complexity and variety of speech signals and sounds. However, with modern processes, algorithms, and methods we can process speech signals easily and recognize the text. In this paper, we are going to develop an on-line speech-to-text engine. The system acquires speech at run time through a microphone and processes the sampled speech to recognize the uttered text. An application like voice search is developed in this work that allows a user to record and convert spoken input into Bing Search Engine that will be search the user's data. This application will find the famous places in Myanmar. Speech recognition is done via the Internet, connecting to Google's server. The application is adapted to input messages in English. Speech recognition for Voice uses a technique based on Hidden Markov Model (HMM). It is currently the most successful and most flexible approach to speech recognition.

1. Introduction

Mobile phones have become an integral part of our Everyday life, causing higher demands for content that can be used on them. Smart phones offer customer enhanced methods to interact with their phones but the most natural way of interaction remains speech. Market for smart mobile phones provides a number of applications with speech recognition implementation. Voice recognition technology is the process of identifying and understanding the voice signals of a user, which is converted into text or commands for a program.

Google's Voice Actions and recently phone's Siri are applications that enable control of a mobile phone using voice, such as calling businesses and contacts, sending texts and email, listening to music, browsing the web, and completing common tasks. Both Siri and Voice Actions require an active connection to a network in order to process requests and most of Android phones can run on a 4G network which is faster than the 3G network that the Note II runs on.

In this work we have developed an application for searching the desired user's query on Bing Search Engine by using speech input which uses Google's speech recognition engine. The main goal of application Voice search which will search only the famous places in Myanmar is to allow user to input spoken information and search s desired data on Internet using this android application. The user is able to manipulate text message fast and easy without using keyboard, reducing spent time and effort. In this case speech recognition provides alternative to standard use of key board for text input, creating another dimension in modern communications.

2. Android Architecture

Android is a software environment for mobile devices that includes an operating system, middleware and key applications [1]. In 2005 Google took over company Android Inc., and two years later, in collaboration with the group the Open Handset Alliance, presented Android operating system (OS).

Main features of Android operating system are:

- Enables free download of development environment for application development.
- Free use and adaptation of operating system to manufacturers of mobile devices.
- Equality of basic core applications and additional applications in access to resources.
- Optimized use of memory and automatic control of applications which are being executed.
- Quick and easy development of applications using development tools and rich database of software libraries.
- High quality of audiovisual content, it is possible to use vector graphics, and most audio and video formats.
- Ability to test applications on most computing platforms, including Windows, Linux…

The application layer of Android OS is visible to end user, and consists of user applications. The application layer includes basic applications which come with the operating system and applications which user subsequently takes. All applications are written in the Java programming language. Framework is extensible set of software components used by all applications in the operating system. The next layer represents the libraries, written in the C and C + + programming languages, and OS accesses them via framework. The Android operating system (OS) architecture is divided into 5 layers as shown in the following figure.

Dalvik Virtual Machine (DVM), forms the main part of the executive system environment. Virtual machine is used to start the core libraries written in the Java programming language. Unlike Java's virtual machine, which is based on the stack, DVM bases on registry structure and it is intended for mobile devices. The last architecture layer of Android operating system is kernel based on Linux OS, which serves as a hardware abstraction layer.

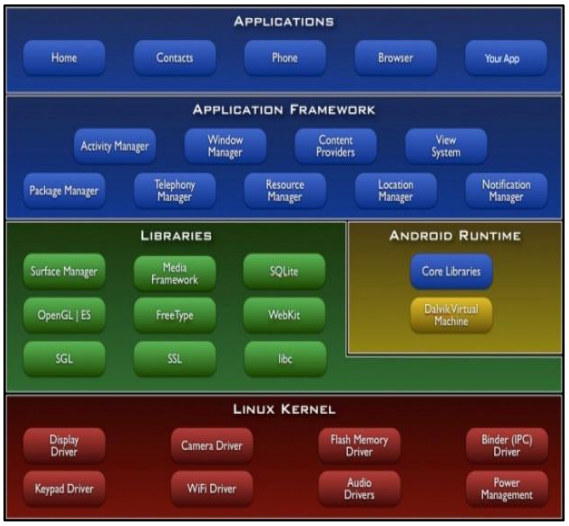

Figure 1. Android Architecture

2.1. History of Speech Recognizer

Speech Recognition research has been ongoing for more than 80 years. Over that period there have been at least 4 generations of approaches, and a 5th generation is being formulated based on current research themes. To cover the complete history of speech recognition is beyond the scope of this paper.

By 2001, computer speech recognition had reached 80% accuracy and no further progress was reported till 2010. Speech recognition technology development began to edge back into the forefront with one major event: the arrival of the "Google Voice Search app for the iPhone". In2010, Google added "personalized recognition" to Voice Search on Android phones, so that the software could record users' voice searches and produce a more accurate speech model. The company also added Voice Search to its Chrome Browser in mid-2011. Like Google's Voice Search, Siri relies on cloud-based processing. It draws on its knowledge about the speaker to generate a contextual reply and responds to voice input. [2]

Parallel processing methods using combinations of HMMs and acoustic- phonetic approaches to detect and correct linguistic irregularities are used to increase recognition decision reliability and increase robustness for recognition of speech in noisy environment.

3. Speech Recognition on Android

Speech recognition is the process of converting voice signal into corresponding text or commands. In this research ,the user speaks the interesting places in Myanmar through the microphone and the converted text is inserted in the corresponding result field and stored in the database. The speech recognition application in smart phones is incorporated to implement this speech to text conversion operation. Figure. 2 depict the basic system architecture of the speech recognizer.

Speech recognition for application Voice SMS is done on Google server, using the Hidden Markov Model (HMM) algorithm. HMM algorithm is briefly described in this part. Process involves the conversion of acoustic speech into a set of words and is performed by software component. Accuracy of speech recognition systems differ in vocabulary size and confusability, speaker dependence vs. independence, modality of speech task and language constraints.

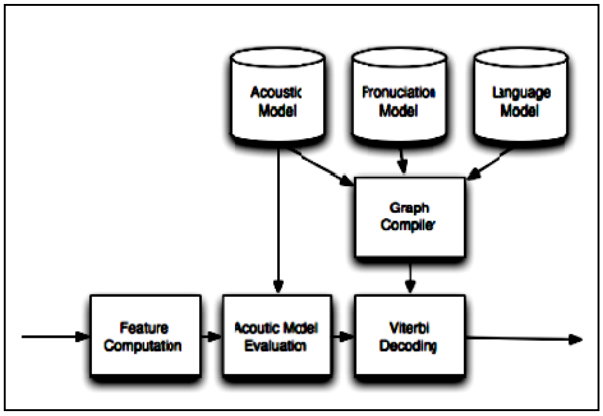

Figure 2.Block diagram of Speech Recognition

Speech recognition system can be divided into several blocks: feature extraction, acoustic models database which is built based on the training data, dictionary, language model and the speech recognition algorithm. Analog speech signal must first be sampled

on time and amplitude axes, or digitized. Samples of speech signal are analyzed in even intervals. This period is usually 20 ms because signal in this interval is considered stationary.

Speech feature extraction involves the formation of equally spaced discrete vectors of speech characteristics. Feature vectors from training database are used to estimate the parameters of acoustic models. Acoustic model describes properties of the basic elements that can be recognized. The basic element can be a phoneme for continuous speech or word for isolated words recognition.

These vectors of speech characteristics are called observations and used in further calculations. To develop an acoustic model, it is necessary to define states. Continuous speech recognition, each state represents one phoneme. Under the concept of training we mean the determination of probabilities of transition from one state to another and probabilities of observations. Iterative Baum-Welch procedure is used for training. The process is repeated until a certain convergence criterion is reached, for example good accuracy in terms of small changes of estimated parameters, in two successive iterations. In continuous speech the procedure is performed for each word in complex HMM model.

Once states, observations and transition matrix for HMM are defined, the decoding (or recognition) can be performed. Decoding represents finding of most likely sequence of hidden states using Viterbi algorithm, according to the observed output sequence. It is defined by recursive relation. During the search, n-best word sequences are generated using acoustic models and a language model.

3.1. Process of Hidden Markov Model

Hidden Markov Processes are the statistical models in which one tries to characterize the statistical properties of the signal with the underlying assumption that a signal can be characterized as a random parametric signal of which the parameters can be estimated in a precise and well-defined manner.

In order to implement an isolated word recognition system using HMM, the following steps must be taken

(1) For each uttered word, a Markov model must be built using parameters that optimize the observations of the word.
(2) Maximum likelihood model is calculated for the uttered word.

The above section has already explained about the speech recognition system on android using HMM to translate the speech to text. This system applied this process to develop the application for searching the famous places in Myanmar like google voice search. This paper will present about this application on section 4.

4. Implementation

4.1. Application functionality Principle

This application integrates direct speech input enabling user to record spoken information as text message, and send it to Bing Search Engine. After application has been started display on mobile phone shows button which initiate voice recognition process. When speech has been detected application opens connection with Google's server and starts to communicate with it by sending blocks of speech signal. If we use this kind of speech recognizer it is very likely that our voice is stored on Google's servers. This fact provides continuous increase of data used for training, thus improving accuracy of the system. When process of recognition is over, user can see the list of possible statements. Process can be repeated clicking on the button Image Button. Pressing the most accurate option, selected result is entered into interface for searching the data.

4.2 Results over application

While speech is both convenient and effective as an input method, especially as an alternative to typing on tiny mobile keyboards, spoken output is very limited given its sequential nature. During most mobile speech apps require the user to press a button to initiate recording, only some require the user to manually stop the recording after speaking by pressing it again, or pressing another button.

Firstly, the voice recognition application that handles the intent processes the voice input, then passes the recognized string back to our application. The start activity of our application has a button to speak, a edit text editor for accepting the output and a web view which is applying the Bing Search Engine for searching this query. The start activity of the application is shown in figure 3.

Secondly, after the user press the speak button, it is displayed the screen to talk about the query that is shown as in figure 4. This application only can search the famous places in Myanmar like pagoda, lake, city etc.

This application is its adaption only for English language, and the need for permanent Internet connection. If it has no connection , it will be

present on the screen like the dialog box “Connection Error” including “Cancle” and Try again”. Lastly, we consider the screens in Figure 5 which shows the results displayed for the same ”Shwedagon Pagoda in Yangon” query using Bing Search feature on Android.

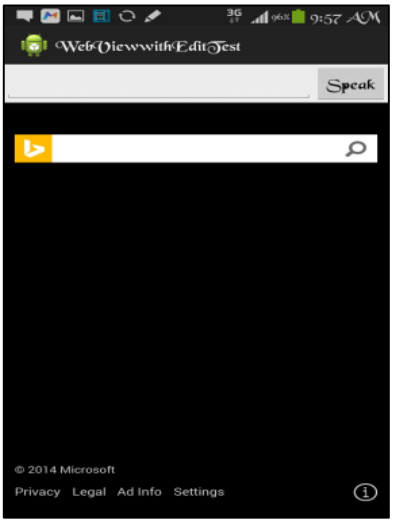

Figure 3. Start Activity of Application

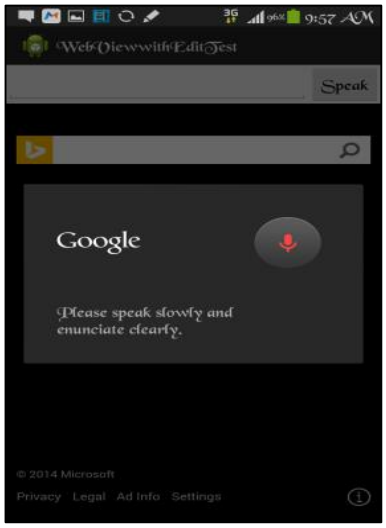

Figure.4. Enable process after clicking speak button

Figure 5. Final output of the application

Making sure the recognizer at least starts with a clean and complete recording of what the user actually said is key for any speech application.As we’ve seen, this is far from automatic and many strategies are currently in use. It may be that some are equally effective and address different user preferences. However, we are also likely to discover that some are simply more effective.

5. Experimental Results on Application

Application Voice Search that will find the most famous places in Myanmar integrates direct speech input enabling user to record spoken information as text data, and send it into Bing Search Engine to be searched for user’s desired query. This system is tested on four different (2 male and 2 female) speakers are asked to speak the same word ten times from the given list of words. The speakers are then asked to utter the same words in a random order and the recognition results noted. In speech recognition phase, the experiment is repeated ten times for each of the above words. The overall efficiency of a speech recognition system obtained is 92.5~ 93% on android.

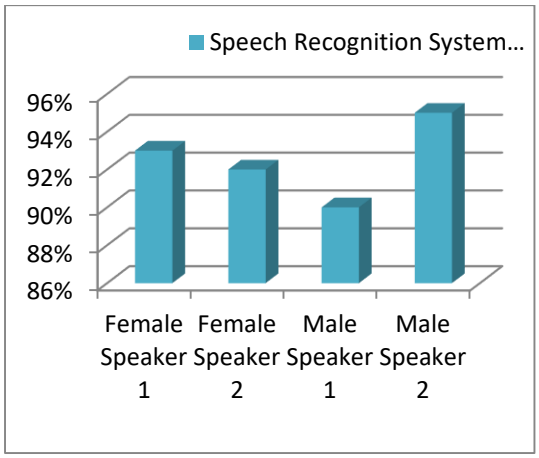

Figure.6. Efficiency Chart for Speech Recognition System

6. Conclusion

With the development of software and hardware capabilities of mobile devices, there is an increased need for device-specific content, what resulted in market changes. Speech recognition technology is of particular interest due to the direct support of communications between human and computers.

The objective for the future is development of models and databases for multiple languages which could create a foundation for everyday use of this technology worldwide. The main goal of application is to allow searching the famous places in Myanmar of text output based on uttered voice messages. Testing has shown simplicity of use and high accuracy of data processing. Results of recognition give user option to select the most accurate result. Further work is planned to implement the model of speech recognition for different language. This project is just only applied the speech recognizer that is automatic transform from speech to text in android. The accuracy is calculated on this recognizer. This is just only testing for speech to text application on android.

References

[1] B. Harb, C. Chelba, J. Dean, and G. Ghemawhat. Back-Language Model Compression. 2009.

[2] C. Allauzen, M. Riley, J. Schalkwyk, W. Skut, and M. Mohri. OpenFst: A general and efficient weighted _nite-state transducer library. Lecture Notes in Computer Science, 4783:11, 2007.

[3] C. Van Heerden, J. Schalkwyk, and B. Strope. Language modeling for what-withwhere on GOOG-411. 2009.

[4] D. Povey, D. Kanevsky, B. Kingsbury, B. Ramabhadran, G. Saon, and K. Visweswariah. Boosted MMI for model and feature-space discriminative training. In Proc. of the Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP), 2008.

[5] H. Hermansky. Perceptual linear predictive (PLP) analysis of speech. Journal of the Acoustical Society of America, 87(4):1738{1752, 1990.

[6] J. Tebelskis, Speech Recognition using Neural Networks, Pittsburgh: School of Computer Science,Carnegie Mellon University, 1995.

[7] www.androiddevoloper.com