Speech to Text Translator for Document Writing with Speech Recognition for Sri Lankan Accent

W.D.M.R.D. Warnakula¹, M. Kumarakuruparan², P.D. Kurukulasuriya³, G.T.U. Kalyanaratne⁴, M. Tissera⁵, L. Wijeratna⁶

Sri Lanka Institute of Information Technology, Sri Lanka

1. rdilshani.w@hotmail.com, 2. mathurakuma@hotmail.com, 3. praveendilanka@gmail.com, 4.technika1989@gmail.com, 5. muditha.t@sliit.lk, 6. laksri.w@gmail.com

Abstract - The work described in this paper is based on an attempt to implement a speech recognizer and a document typing application for English speaking users with Sri Lankan accent. This speech recognizer and document typing application is not like the other software that is already in the market. The general speech recognizer applications in the market cater generally to the British and the American accents of the English language. This is a disadvantage for Sri Lankan users when using those software applications. Therefore we thought of attempting to build a speech recognizer for the Sri Lankan accent of the English language. This uniqueness of the software is being achieved by generating a voice corpus which uses the vocal data sets containing Sri Lankan accent. Hence higher accuracy can be achieved for Sri Lankan users. The underlying concept of building the speech recognizer is the Hidden Markov Model (HMM) and the system is developed using the Hidden Markov Model Toolkit (HTK) which is open source software owned by the University of Cambridge. The first stages of building the speech recognizer involve the preparation of speech samples for training and the creation of the pronunciation dictionary which lists all the speech samples along with their phonetic representations. A noise reduction method has been applied at the front end of the speech recognizer application to reduce the background noise and filter the speech signal from the initial stage. The middle stages comprise of employing a good feature extraction technique to enhance the speech recognition, and building and training the acoustic model to match a spoken word to the observed input while the latter stages involve the creation of the language model to determine which word has spoken. Other than that we have also provided the user with the functionality to use voice commands. Using voice commands the user can easily interact with the system and give instructions to the system. Also we have provided an inbuilt document writer in which the users can easily edit and format documents. The results indicating the accuracy level of the application show that 40%-50% of the words are correctly recognized by the speech recognizer in a quiet environment, while 30% of the words are correctly recognized in a noisy environment.

Keywords – English Speech Recognition, Information Technology, Sri Lankan Accent, Document Writer, Hidden Markov Models

I. Introduction

Information Technology (IT) plays a major role in today’s development. At present the modern day world has come to a point where we cannot live without technology. Information technology is used immensely to accomplish our day to day activities accurately and efficiently. New technologies have emerged to provide more help and convenient ways to accomplish human activities. That has brought the human-machine interaction to a very sophisticated level. Numerous mechanisms have been invented by humans to interact with the computer. Keyboard invention is one such and still it is the mostly used mechanism among human beings. However, with the busy schedules, interactions with computers become a tedious task if it cannot be done efficiently and easily.

The primary means of communication among humans is speech. If the speech could be used as the media for human-machine interaction, that would provide countless benefits, especially to ease their busy lives. Therefore a need has arisen to provide a voice interface between the user and the computer. Automatic speech recognition is used in order to fulfill this need. The research on automatic speech recognition has a long history, dating back to the 1952 Bell Labs paper describing a technique for digit recognition. Using speech for interaction between human and computer comes under natural language processing (NLP) category. This is because the human computer interactions were enhanced using natural languages such as voice.

Even though English is a global language, the way English is spoken is different when considering different nationalities around the globe. This difference is due to the difference of their accents and the usage of phonetics. Pronunciation of any language is affected by the mixture and the influence of the accents of their mother tongue. This is why the English pronunciation of Sri Lankan community is different than that of the British and the Americans.

Many commercial applications are available for speech recognition. (E.g. Windows Speech recognition, Simon, Dictation Pro) All such software applications have been developed to recognize British or American accents. Using these applications, the accuracy of recognizing the words of a Sri Lankan is lower than that of the British or the American users. That accuracy level also highly depends on the amount of training that the users perform for such an application. This training is tedious and time consuming as well.
Speech recognition can be categorized into two types.

Speaker independent – The software doesn’t need to be trained for different users.

Speaker dependent – The software has to be trained according to the speaker.

Speech to text translator for document typing with speech recognition for Sri Lankan accent is something different than the normal voice recognition tools. This is a voice recognition platform for document typing with speaker independent framework within Sri Lankans. The target of this research was achieved by building software for the Sri Lankan community. In this attempt, we collected datasets from Sri Lankans and the entire system was based on Sri Lankan accent. Target of this framework is to have an accuracy of 50%.

Further, the system that we developed would be a platform for future young researches for their own research work. Speech recognition for English language for Sri Lankans may be further extended by future research groups who are interested in the field of Natural language processing. The system is ultimately capable of reaching a level of around 70% accuracy, with a large number of datasets over a wide range of vocabulary.

II. Methodology

A. Methodology

1) Information Gathering: Initially, we collected speech data from around 400 males and 400 females. The spontaneous speech consisted of free-form monologues where each speaker was asked to discuss a topic of their choice from a small set of predetermined topics.

2) Approach: In order to develop the proposed system, it was very important to identify the best software development methodology. Iterative development was chosen where the application is designed, developed and tested in repeated cycles. Additional features were designed, developed and tested iteratively until it became solid and fully functional software that was ready to release to customers.

3) Design: After gathering the requirements that were needed for the research, we had to understand the requirements properly. Before starting the coding phase it was important to know how the end product looked like. In the Design phase we prepared the architectural design, interface design and the database design as well. The requirements gathered in the requirement gathering phase were studied and understood in the design phase. The architectural design was done according to the system requirements. In this, we should provide a basic idea of the relationships among the modules. Apart from the system design and architectural design, the database and interface designs were also be prepared.

B. Testing and Implementation

Implementation: After the design phase, once the general design was done, then came the implementation phase where it is worked out. In the implementation phase, the implementation of the approved user requirements were taken into account and implemented.

Mapping speech recognition with the in-built document typing interface

We further developed a rich user interface for document typing, which was mapped with the voice recognition system that we have developed. This mapping led to the compatibility of the application with the system. The voice to text translator translates the voice into text in the software application itself. The captured voice is saved as a .wav file in the background every 3 seconds and through a parallel task it creates new thread to run a batch file which contains HCopy and HVite commands which generates files relevant for mapping. Then the output files generated are scanned and the translated text is extracted. This is done through a new process as the existing GUI of the client should stay interactive at all times. Then the final translated text is appended to the end of the document.
Setting voice commands to the software

Voice tags are used in automated speech recognition in a voice command device, allowing the user to "speak" commands. These voice commands most are general while some vary from application to application. These tags help identify the command you want to execute when you say it through the microphone. The system should be able to identify the difference between words spoken to be typed in the document and voice commands to execute an action in the document. An example is to save a file or if you take another example is in the need to underline a word.

In this system, the word “Action” is used to give the execution command to keywords. This will help the system to identify what is a keyword and what is not, thus enhancing the capabilities of voice execution of commands. For example, copy is a keyword, cat is not a keyword. Then consider the following scenario and what should take place.

<table>
<thead>
<tr>
<th>Voice Input</th>
<th>Expected Result</th>
</tr>
</thead>
<tbody>
<tr>
<td>Copy</td>
<td>Print the word copy</td>
</tr>
<tr>
<td>Action copy</td>
<td>Execute copy command</td>
</tr>
<tr>
<td>Action cat</td>
<td>Print the word action cat</td>
</tr>
</tbody>
</table>

Catching the voice and noise cancellation

The voice is captured by an input device either through a microphone or a microphone attached headset. Then the voice captured is sent to the system that we have developed. In that system, the sound waves are processed using audio signal technique. The voices of people may vary from person to person; therefore a clear input of the voice would be best to identify minor changes in the audio signals that are given out when speaking out.

Processing Frequencies

Audio signal processing, sometimes referred to as audio processing, is the intentional alteration of auditory signals, or sound, often through an audio effect or effects unit. As audio signals may be electronically represented in either digital or analog format, signal processing may occur in either domain. Analog processors operate directly on the electrical signal, while digital processors operate mathematically on the digital representation of that audio signal. In general, a majority of audio processing techniques address the following 3 application areas: compression, classification, and security.

Voice datasets

In order to recognize Sri Lankan accent for document typing, it is required to have those audio frequencies for the specific words. For this purpose, in our framework we utilized the concept of speech corpus, also known as spoken corpus. Speech corpus refers to a database of speech audio files and text transcriptions. In Speech technology, speech corpora are used, among other things, to create acoustic models which can then be used with speech recognition engines. The building of large-scale corpora of text for our project is not feasible. What is ideal is to build a small-scale corpus which can be extended and enhanced in the future. This will be an approach to keep the project within the scope and it would benefit to the users to test it out through iteration.

In order to collect data for the database different types of voice samples were required and the amount of data sets will have a clear reflection on the accuracy of the system. The higher the datasets, what is input through voice will be correctly typed. We managed to get around 8,000 record sound file data sets of different voices. Even though we focus only the Sri Lankan accent, that also have a wide variety of voice changes based on age, gender and medical conditions. Therefore the task of creating the corpus has its own complexity, whereas we found ways to compress, identify and retrieve the voices as required.

Process of creating the voice corpus:

Task Grammar creates a regular grammar and converts it to an intermediate form of decoding network. Create the word net file. Create Dictionary is creating pronunciation dictionary covering all words in the grammar. When collecting speech data to train the voice corpus we had to make sure that they were recorded in the correct frequency. The frequency used to record from a computer was 16 KHz. The frequency that we used to record from a phone was 8 KHz. When recording the speech data for the training we had to make sure that all the speech utterances we selected to use for the recordings covered all the phonemes in the English language. Transcription is creates a Master Label File (MLF) describing the phoneme transcriptions of the training data. Feature Extraction with HCopy. Initializing a prototype HMM by flat starting, using speech data to compute all HMM probabilities. Fixing the Silence Models here we create the Master Macro File of all monophones using the prototype HMM. Realigning the training Data by re-estimating each phoneme (monophone) HMM using Baum-Welch algorithm. Make a triphone MLF from monophone MLF and clone triphone HMMs using monophone HMMs. Training Triphone HMMs by re-estimate triphone HMMs and append triphones and biphones.

Converting to text:

The processed and enhanced audio waves of the voice input were compared with the data sets, and then using the Hidden Markov model pick the most appropriate or closest to word that was pronounced. Most modern software recognition frameworks utilize this model as it has a high success rates compared to other techniques such as Dynamic time warping (DTW)-based speech recognition and Neural networks. A hidden Markov model (HMM) is
a statistical Markov model in which the system being modeled is assumed to be a Markov process with unobserved (hidden) states. A HMM can be considered the simplest dynamic Bayesian network. We used this technique to get our framework done. This enabled us to focus on the Sri Lankan accent with the objective of achieving a high accuracy level.

III. Results and Discussions

Even though there are several voice to text translators available in the market, there are limitations in those applications. The main limitation is that they are not aimed towards the Sri Lankan users who speak the general English in their native accent. As a result those existing applications might not give satisfactory result when it is used by a common Sri Lankan user.

So in order to overcome such issues, we incorporated several different features into the application.

- The software is able to clearly identify the accent used by the Sri Lankan English speaking users. Since this is the primary goal of the software, the main focus is given to this.

- The software is capable of identifying the accent uniquely up to a certain level of accuracy. We are aiming on achieving accuracy more than 50%.

- The user is able to use the particular software without any additional training. This will reduce the effort required by the users in order to use the particular software.

- The user is able to handle the software by using simple voice commands. These voice commands are enabled to make document typing faster and easier.

During the process of achieving these objectives we came across several problems. We faced several challenges in overcoming those problems.

- The type of the microphone and the environment in which the voice is captured is dependent on the accuracy of the output received. The type of environment under which a particular voice is captured is dependent on the type of output received. And similarly the quality of the microphone also affects the accuracy of the output. The lesser the amount of noise in the voice input, the more the accuracy of the result will be.

- While capturing voices we had to remove the ambiguity while recognizing some words. There are several words in the English vocabulary which sound the similar but are completely different and are spelt differently. In order to resolve the issue in identifying these type of words. For examples words like ‘there’ and ‘their’. We used a mechanism to give a list of words as suggestions from which the particular user is able to select the appropriate word which the user wants to be appeared in the document.

- We also had problems identifying the end of a sentence while capturing speech. While speech is captured by the software we had to correctly identify the end of a sentence. Due to this concern we captured the speech into a different recording every 3 seconds.

- Even though, it is expected that the user is expected to have a general accent, sometimes we came across instances where some users pronounced some words differently.

In such instances we had trouble identifying the words correctly. To resolve this issue, we have made the document editable, so that user can easily make necessary changes.

Therefore in the near future we hope to make any best possible improvements to our speech to the text translator for Sri Lankan accent in order to produce the best possible results. We have improved our client side in order to make it as user friendly as possible.

IV. Conclusion and Future Work

Document writing with the aid of voice recognition gives an easy way to write the documents efficiently and accurately. “Speech to text translator for document typing with speech recognition for Sri Lankan accent” is a deviation from the general voice recognition tools.

Voice recognition is a directive under the field of Natural Language Processing as voice input is a natural way of human-computer interaction. The specialized voice recognition system will be targeting to a specific scope of people, in this case English speaking people having Sri Lankan accent are targeted.

Few primary objectives of this project were, the software must be capable of identifying the English accent of the Sri Lankans properly, the voice to text translations should be at around 50% accurate and the user should be able to handle the software by using simple voice commands. However, the proposed system is unable to intelligently recognize “Isolated Words”, “Connected Words”, “Continues Speech” and “Spontaneous Speech”. The software has been developed to provide a fast method for document writing on a computer and it can help people with a variety of disabilities. It is very useful to those who are having physical disabilities such as blindness, inability to write due to a pain or completely impossible due to loss of limb, etc. Voice-recognition software can also help those with spelling difficulties, including users with dyslexia. Speech to text translator for English accent of Sri Lankans would be useful to everyone who has a desirable knowledge in English. This user-friendly system allows its users to easily read out pages of documents and let the software type it for them. This will
improve the creation of documents such as reports, notes, etc. with much ease for the Sri Lankan community.

To improve the accuracy of the speech recognizer further with the help of training the speech corpus is one of our future plans. Also developing functionality that let the user to select the correct word whenever the words are misspelled is another such target. Further, a mobile version of the system for smart-phones on different mobile operating systems is part of the development plan for future work.

ACKNOWLEDGMENT

We would like to express our profound gratitude to Ms. Dinusha Thilini for providing us with valuable comments and feedback throughout the project. Also I would like to thank Mr. Amal Siriwardena for his support on this research from the beginning. Also we would like to express our thanks to the people who provided us with the voice samples to be taken as the datasets for the research.

REFERENCES