International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 4 Issue 10, October-2015

Sinhala Speech Recognition

N. A. C Sandasarani
Department of Physical Sciences,
Rajarata University
Mihintale, Sri Lanka.

***Abstract* - Speech recognition stands to convert the human voice into the text that similar to the information being conveyed by the speaker. This paper aims to find a suitable method for Sinhala Speech Recognition. For that the paper goes through 2 approaches, the isolated word recognition and continuous speech recognition. The paper uses Artificial Neural Networks to go through the Continuous Speech Recognition approach, and Mel Frequency Cepstral Coefficients (MFCC) and Dynamic Time Warping (DTW) as feature matching and feature extraction techniques under isolated word recognition. By using above techniques, experimental results show that overall accuracy is higher in isolated word recognition than continuous speech recognition.**

Keywords – Speech Recognition, Feature Matching, Feature Extraction

I. INTRODUCTION

Basically Automatic Speech Recognition is a process by which a computer or any other equivalent device correctly recognized what was spoken. Most of the work in speech recognition is done regarding to the English or other European languages. For Sinhala language still any well-known Speech Recognizer does not exist.

Sinhala is the native language of the island nation of Sri Lanka. It belongs to the Indo-Aryan branch of the Indo-European languages. Sinhala is the mother tongue of about 15 million Sinhalese, while it is spoken by about 19 million people in total [1].

Research into the concept of speech technology began as early as 1936 at Bell Labs. Since that period to today, speech technology shows significant additional improvements [2].

The objective of this paper is to find a path for Sinhala Speech Recognition, through two different approaches. The isolated word recognition and continuous speech recognition are those two approaches. Here the main attention was given to the acoustic phonetic modeling under the continuous speech recognition approach. Although I got some acceptable results in isolated word recognition but due to its limitations of having short vocabulary, this technique was only acceptable for short vocabularies.

II. ISOLATED WORD RECOGNITION

First approach I go through is based on the matching of incoming word with one of a number of stored acoustic pattern templates that exist in the recognition vocabulary. This process is named as Isolated Word Recognition or Pattern Matching Approach. Here the incoming speech consists of isolated words. The task is to match the incoming word against each word that stored.

A. Feature Extraction for the isolated word recognition

In this approach there must be a feature set to store and match with each other. In speech recognition feature extraction techniques are used for this purpose. Several feature extraction algorithms can be used to do this task, such as Linear Predictive Coefficients (LPC), Linear Predictive Cepstral Coefficients (LPCC), Mel Frequency Cepstral Coefficients (MFCC) and Human Factor Cepstral Coefficients (HFCC) [2]. Here I used MFCC algorithm to extract the features. I have done the implementation using Matlab.

MFCC are chosen because MFCC are the most important features, which are required among various kind of applications, and it gives high accuracy results especially for clean speech and also MFCC can be regarded as the standard features in speakers as well as speech recognition.

In MFCC algorithm the speech is first decomposed into frames of the size which are usually chosen as a power of two to fit the FFT algorithm. In the next block, Hamming window is applied to those frames. Next FFT algorithm is applied to get the magnitude spectrum of the windowed speech data. The next block is mel-filtering. It provides a model of hearing realized by the bank of triangular filters.

B. Feature mapping for isolated speech recognition

I have used Dynamic Time Warping (DTW) to match the incoming feature set against to the stored feature set. DTW algorithm is based on Dynamic Programming techniques as described in [3]. This algorithm is for measuring similarity between two time series which may vary in time or speed. This technique also used to find the optimal alignment between two times series if one time series may be “warped” non-linearly by stretching or shrinking it along its time axis. Then this warping between two time series

can be used to find corresponding regions between the two time series or to determine the similarity between the two time series [3].

III. CONTINUOUS SPEECH RECOGNITION

There exist finite, distinctive units in spoken language called phonetics. These phonetic units are probably characterized by a set of properties that are manifested in the speech signal or the spectrum over time. Usually English and other European languages use these phonetic units in Continuous Speech Recognition. But in Sinhala language there's no exact phonetic symbols for some Sinhala letters. So here I have used another method (as described in next part) to extract set of properties that are manifested in the Sinhala speech. After extracted feature set for the Sinhala letters I have used Artificial Neural Networks for trained those features. Figure 1 showed parts I have gone through in Continuous Speech Recognition.

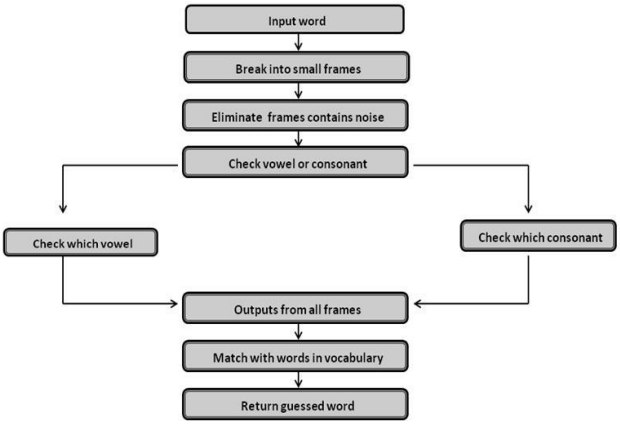

Fig. 1. Process that followed for continuous speech recognition

A. *Check which vowel*

First I have digitized the speech signal, break those signal into small frames (0.05 sec) and plot time domain graphs of those small frames. Those graphs showed a special characteristic for all vowels in Sinhala language. There exist a repeating pattern throughout the graph and this pattern is different from vowel to vowel. Letters "ෙ" and "ෘ" also act like same as vowels. So in my further developments I use those characters as vowels. Figure 2 showed a 0.05sec graph for letter "ඇ".

For extracted features for letter, "ඇ" I have broken the speech signal that contained the letter "ඇ" into 0.05 second frames and use FFT function for those frames. Then I have used threshold value to cut frequencies that are not in human voice ranges and get the resulted feature set to identify letter "ඇ".

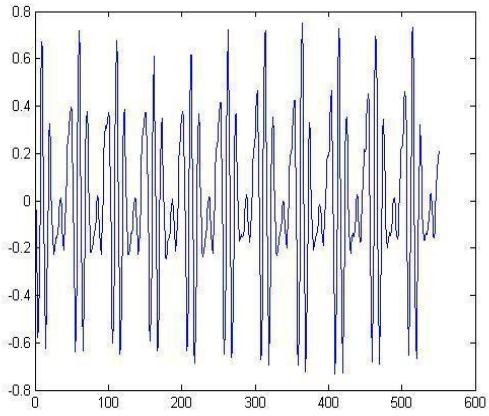

Fig. 2. 0.05 Second frame of time domain signal of letter 'ඇ'

B. *Check Vowel or Consonant*

When analyzing signals it revealed that there's no any repeating unit in consonant's wave pattern. Figure 3 shows a time domain signal of a sound sample that contains consonant and a vowel. I have used ANN again to separate consonants from vowels. For that I have used frequency domain signal and break signals into 0.05 sec frames. Then I have found the maximum y-axis value and get 65 y- axis values from that value onwards, along the x-axis. If frame is not enough to get 65 samples, took samples from next frame. I have taken those 65 samples as feature set for one letter.

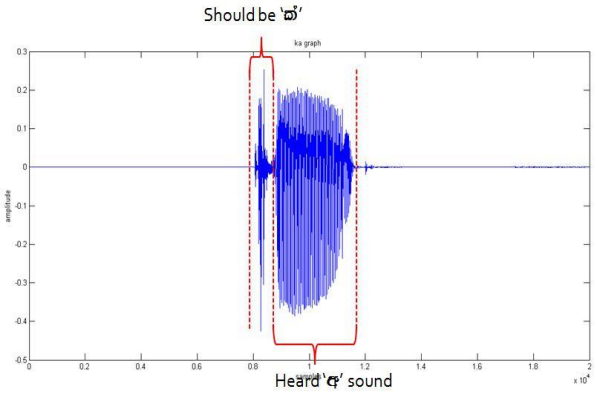

Fig 3. Graph for letter 'ක'

C. *Check which Consonant*

To identify consonants separately, I have extracted the part that represent the consonant and got Fourier transformation of those parts. Then got highest 20 values (in y-axis and relevant x-axis values) from each of that frames. Got those 40 features to represent identity of one letter and have given that feature set as the input to ANN.

D. *Train Artificial Neural Networks*

I go through 2 types of training methods. To check which vowel and to check vowel or a consonant I have to use one method. Here I have trained one train file for one letter and make set of train files. That means when create a train file for a particular letter; output has defined as "1", if the output is that letter. Otherwise output is "0".

Then I have given test sample to every train file, and find which train file gives the maximum output. The letter relevant to that train file becomes the output. I use 300 samples from each letter when I train ANN.

To identify consonants separately I have used another training method. Here also I create one train file for one letter. But I used 300 samples from the particular letter, that relevant to the train file. And use 50 samples from each other letters. I have created 6 sets of train files from one letter. When gained the output, first retrieve the result for the input frame using sets of train files and selected the most frequent output as the final output. This method increase accuracy when check which consonants. But increase the time to give the output.

I stored words in a structure array and create a "mat" file in Matlab. After I have collected output for all frames of the input word, I check which word in the vocabulary match with the final output. Find that best match word and return that word as the guessed word.

IV. RESULTS

I used 4 Sinhala words to check the accuracy.

A. *Isolated word recognition*

Word	Accuracy
එක	37%
තුන	80%
නිවන්න	28%
දාන්න	21%

Fig. 4. Accuracy for voices that not in stored database

Word	Accuracy
එක	95%
තුන	90%
නිවන්න	90%
දාන්න	90%

Fig. 5. Accuracy for voices that exist in database

B. *Continuous speech recognition*

Letter	Accuracy
ක්	80%
ත්	50%
ප්	10%
ද්	53%
න්	20%
ට	97%
ච	67%

Fig. 6. Highest accuracy gained when check for consonants

Letter	Accuracy
අ	60%
ඇ	96%
එ	96%
ඉ	73%
ඔ	90%
උ	96%
ල්	90%
මි	66%

Fig. 7. Highest accuracy gained when check for vowels

Word	Accuracy
එක	75%
තුන	30%
නිවන්න	40%
දාන්න	80%

Fig. 8. Highest Accuracy for male voices

Word	Accuracy
එක	10%
තුන	60%
නිවන්න	13%
දාන්න	70%

Fig. 9. Highest accuracy for female voices

V. CONCLUSION

According to the results gained, isolated word recognition for words that exists in the database gives the highest accuracy. But it can use only for a limited vocabulary. Accuracy will be decrease and the time to return the output will be increase when number of words in the vocabulary increasing.

Accuracy gained for the continuous speech recognition won't be decrease with the vocabulary. If we could use high performances for the training process, accuracy will be increased than this. Then this can be use for any of applications regarding the Sinhala speech to text conversion.

Although the Sinhala language is not very much familiar with this research topic this work prove that there will be a good future for the Sinhala language with respect to speech recognition.

VI. ACKNOWLEDGEMENT

I remain thankful to, my senior supervisor, my supervisor for their useful discussions and suggestions during the preparation of this technical paper. And I thankful all the others for their continuous help and support in all stages of the research.

REFERENCES

[1] SajikaGallege, "Analysis of Sinhala Using Natural Language Processing Techniques", University of Wisconsin-Madison.

[2] M.A.Anusuya, S.K.Katti , "Speech Recognition by Machine: A Review " in (IJCSIS) International Journal of Computer Science and Information Security , Vol. 6, No. 3, 2009.

[3] Lindasalwa Muda, Mumtaj Begam, I.Elamvazuthi, "Voice Recognition Algorithms using Mel Frequency Cepstral Coefficient (MFCC) and Dynamic Time Warping (DTW) Techniques " in Journal of Computing , Volume 2, Issue 3, March 2010.

[4] Ms.Vrinda, Mr. Chander Shekhan , "Speech Recognition System for English Language " in International Journal of Advanced Research in Computer and Communication Engineering, Vol. 2, Issue 1, January 2013.

[5] M.A.Anusuya, S.K.Katti , "Speech Recognition by Machine: A Review " in (IJCSIS) International Journal of Computer Science and Information Security , Vol. 6, No. 3, 2009.

[6] Urmila Shrawankar , Dr. Vilas Thakare, " Techniques For Feature Extraction In Speech Recognition System" , SGB Amravati University, Amravati.

[7] eff Blimes. (2002, January),"What HMMs can do", [Online]. Available:https://www.ee.washington.edu/techsite/papers/documents/uweetr-2002-0003.pdf. Accessed : September 2009.

[8] Representation of Digital Signals.[Online]. Available: http://storage.sk.uni-bonn.de/Milca/ssv/content/ssv_s132_en.xhtml Accessed: July 2012.

[9] Ruwan Weerasinghe, Asanka Wasala, Dulip Herath, Viraji Welgama, "NLP Applications of Sinhala: TTS & OCR" in Language Technology Research Laboratory in University of Colombo School of Computing.

[10] Wouter Gevaert, Georgi Tsenov, Valeri Mladenov, "Neural Networks Used For Speech Recognition", Journal Of Automatic Control, University Of Belgrade, Vol. 20:1-7, 2010.

[11] Ahmed M. Abdelatty Ali, Jan Van Der Spiegel, Paul Mueller, Gavin Haentjens, Jeffrey Berman, "An Acoustic-Phonetic Feature-Based System for Automatic Phoneme Recognition In Continuous Speech", Dept. Of Electrical and Computer Engineering, Carnegie Mellon University, 5000 Forbes Ave, Pittsburgh.

[12] Valery A. Petrushin, "Hidden Markov Models: Fundamentals and Applications", Center for Strategic Technology Research, Northbrook, Illinois.

[13] Manjot Kaur Gill, Reetkamal Kaur, Jagdev Kaur, "Vector Quantization based Speaker Identification", International Journal of Computer Applications (0975 – 8887), Volume 4 , No.2, July 2010.