IJNRD.ORG ISSN : 2456-4184
INTERNATIONAL JOURNAL OF NOVEL RESEARCH AND DEVELOPMENT (IJNRD) | IJNRD.ORG
An International Open Access, Peer-reviewed, Refereed Journal

VOICE DETECTOR FOR SECURITY PURPOSE IN WEB APPLICATION

Tapadipta Roy
Department of computer Science
Dr.M.G.R Educational And Research Institute
Maduravoyal,Chennai,600095,
Tamilnadu,India
Tapadiptaroy90@gmail.com

Rohinth Rex A
Department of Computer Science
Dr.M.G.R Educational And Research Institute
Maduravoyal,Chennai,600095,
Tamilnadu,India
rohitrex33@gmail.com

Sakaniesh Ram K
Department of Computer Science
Dr.M.G.R Educational And Research Institute
Maduravoyal,Chennai,600095,
Tamilnadu,India
neymarsaka12@gmail.com

Prof. Shobarani A
Faculty of Computer Science
Department of computer Science
Dr.M.G.R Educational And Research Institute
Maduravoyal,Chennai,600095,
Tamilnadu,India
shobarani.cse@drmgrdu.ac.in

Trisha Ghosh
Department of Architecture & Design
Saveetha College of Architecture & Design
Saveetha Nagar,Chennai,602105,
Tamilnadu,India
ghoshtrisha008@gmail.com

Abstract— **Online voice detector is a web application where we will develop detecting of voice of real owner. The Voice is a signal of many information. Digital processing of voice signal is very important for high & low speed and precise automatic voice recognition technology. Nowadays it is being used for many health care services, telephony military and people with disabilities therefore the digital signal processes such as Feature Extraction and Feature Matching are the latest many issues for study of voice signal. In order to extract valuable information from the speech signal, make decisions on the process, and obtain results, the data needs to be manipulated and analyzed. Basic method used for extracting the features of the voice signal is to find the Mel frequency cepstral coefficients. This paper is divided into two modules. Mel-frequency cepstral coefficients (MFCCs) are the coefficients that collectively represent the short-term power spectrum of a sound, based on a linear cosine transform of a log power spectrum on a nonlinear mel scale of frequency. Under the first module feature of the speech signal are extracted in the form of MFCC algorithm coefficients and in another module the non linear sequence alignment known as Dynamic Time Warping (DTW). Since it's obvious that the voice signal have different rate, the alignment is important to produce the better performance. This paper presents the feasibility of MFCC to extract features and DTW to compare the test patterns.**

Keywords— **Recording, Analyzing Voice, Application Programming Interface, Communication , Speech to Text;**

I. INTRODUCTION

Voice detector is built into most devices. For example, smart phones, laptops and tablets include good microphones which will support voice input and commands. Similarly, computers often come with inbuilt cameras, microphones, and speakers. Voice detector can provide an alternative to typing and copy pasting. At its simplest form, it provides a super fast method of writing on a computer, tablet or Smartphones. You can speak to text into an external microphone, headset or built-in microphone, and your words appear as text on the screen exactly what is detected. This might be in the text bar of a search engine, in a chat or a messenger application, or in an email or document.

A. Why use voice recognition?

Voice detector offers significant benefits to a wide variety of users. Most probably, it is extremely useful for those with a physical disability who finds typing very difficult, painful or impossible. similarly, it can help to reduce the risk of getting a repetitive strain injury (RSI) or to manage any wrong writing more effectively.

Voice recognition program can also greatly benefit people with many problems who would otherwise struggle with spelling and structuring sentences correctly.

II. LITERATURE SURVEY

[1].Thing, et al. (2011) presented speech recognition using Linear Predictive Coding (LPC) and Artificial Neural Network (ANN) for controlling movement of mobile robot. Input signals were sampled directly from the microphone and then the extraction was done by LPC and ANN.

[2]. Ms.Vimala.C and Dr.V.Radha (2012) proposed a speaker-independent isolated speech recognition system for the Tamil language. Feature extraction, acoustic model, pronunciation dictionary, and language model were implemented using HMM which produced 88% of accuracy in 2500 words.

[3]. Cini Kurian and Kannan Balakrishnan (2012) found the development and evaluation of different acoustic models for Malayalam continuous speech recognition. In this paper HMM is used to compare and evaluate the Context-Dependent (CD), Context Independent (CI) models and Context-Dependent tied (CD tied) models from this CI model 21%. The database consists of 21 speakers including 10 males and 11 females.

[4]. Annu Choudhary et al. (2013) proposed an automatic speech recognition system for isolated and connected words of Hindi language by using Hidden Markov Model Toolkit (HTK). Hindi words are used for the dataset extracted by MFCC and the recognition system achieved 95% accuracy in isolated words and 90% in connected words.

[5]. Preeti Saini et al. (2013) recommended Hindi automatic speech recognition using HTK. Isolated words are used to recognize the speech with 10 states in HMM topology which produced 96.61%.

[6]. Md. Akkas Ali et al. (2013) presented an automatic speech recognition technique for Bangla words. Feature extraction was done by, Linear Predictive Coding (LPC) and Gaussian Mixture Model (GMM). Totally 100 words were recorded 1000 times which gave 84% accuracy

[7]. Maya Moneykumar, et al. (2014) developed Malayalam word identification for speech recognition system. The proposed work was done with syllable-based segmentation using HMM on MFCC for feature extraction . Jitendra Singh Pokhariya and Dr. Sanjay Mathur (2014) introduced Sanskrit speech recognition using HTK. MFCC and two states of HMM were used for extraction which produces 95.2% to 97.2%.

III. PROBLEMS AND SOLUTIONS

A. *PROBLEM IDENTIFICATION*

The current big problems of voice detectors are caused by two major factors – reach and loud environments. Voice detecting at a loud family time or meetings with various voice from background. These are the upcoming challenges to be solved for next-gen voice detection.

Beyond this, voice detection needs to be made available for more languages and cover each and every topic. Because as of now, ASR needs a lot of data to work well and some of them just hasn't been collected as well certain languages and topics in the database. Without adding these, ASR systems will remain noticeably helpless.

Voice detecting systems isn't always able to interpret spoken words properly. This is due to computers not being on par with humans in understanding the relation of words and series of sentences, causing interpretations of what the person trying to say or achieve.

B. *PROPOSED SYSTEM*

1 On-premises deployment

On-premises deployment of voice technology enables users to keep their data secure within their own premises with not much need for data to go into the cloud. It is often done by using virtual appliances or containers, so they can be deploy into existing technology stacks of a system. This is particularly important for some sort of industries such as banking, financial, and insurance where compliance and regulatory issues mean customer data and voice data cannot leave the industrial premises .

2 Cloud deployment

All Private cloud deployments are secure enough to keep data safe for many applications. If cloud deployment security is good enough for the business and use case needs, cloud deployment is often the most prefered option due to low operational cost and less complexity.

IV. OVER ALL ARCHITECTURE

- Speech Capturing Device
- Pre-processed signal
- Reference Speech Pattern

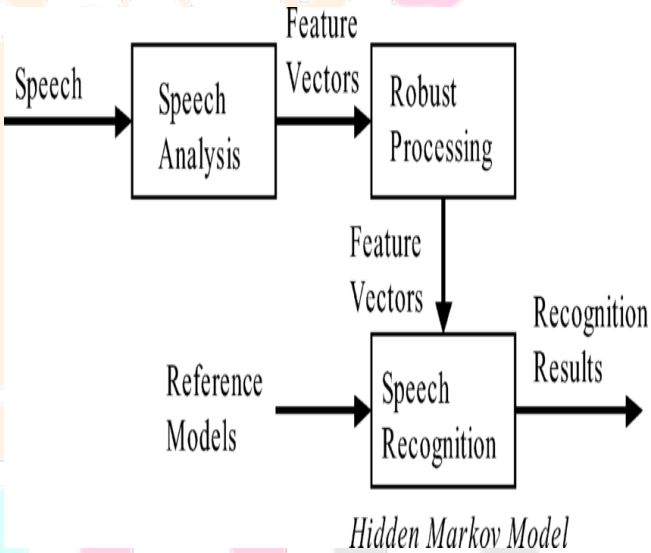

Fig:1:Over all Architecture

V. MODULE DESCRIPTION

A. *Continuous listening*

The speech detector system can either stop listening after the user stops speaking or it can keep listening until the user stops it. If you only want to detect a phrase or a word, you can set this to false. For this tutorial, let's set it to true.

B. *Attend and Spell*

We now describe the Attend And Spell function. The function is computed using the algorithm LSTM transducer .At every output step, the transducer produces a probability distribution over the next character conditioned on all the characters mentioned previously.

C. *Learning*

The Listen and check And rewrite functions can be trained jointly for end-to-end speech detector. The sequence to sequence methods condition the next step prediction and maximizes the log probability.

D. Decoding and Restoring

During inference we want to find the most probable character sequence given the input Decoding is performed with a simple left-to-right beam search algorithm At each timespan, each partial hypothesis in the beam is expanded with every possible character When the token is encountered, it is removed from the beam and added to the set of complete hypothesis. A dictionary can optionally be added to constrain the search space to valid words, however we found that this was not necessary since the model learns to spell real words almost all the time. We have lot quantities of text data compared to the amount of transcribed speech utterances. We can use language models trained on text corpora alone similar to conventional speech systems.

E. Interim results

Interim results are results that are not yet final. If you enable this property, the speech detector object will also return the interim results along with the final results. Let's set it to true.

VI. BLOCK DIAGRAM

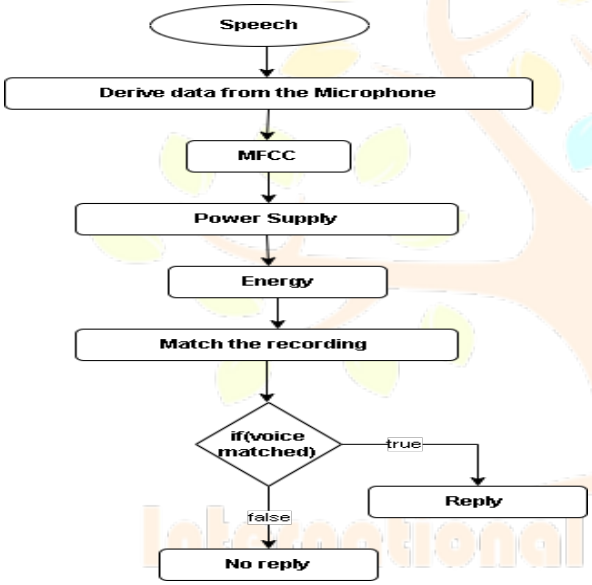

Fig:2:Decryption

VII. DATA FLOW DIAGRAM

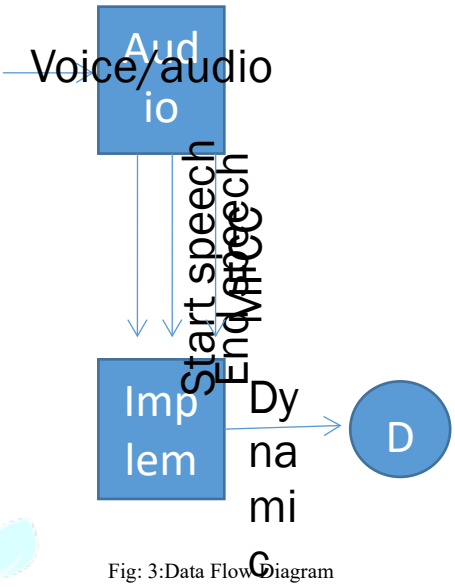

Fig: 3:Data Flow Diagram

VIII. USE CASE DIAGRAM

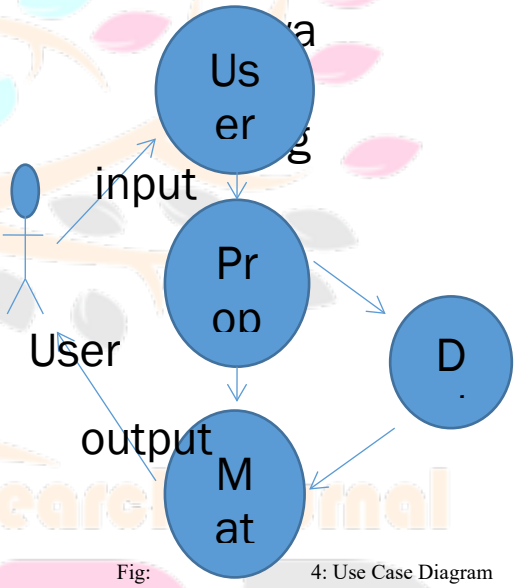

Fig: 4: Use Case Diagram

IX. CLASS SEQUENCE DIAGRAM

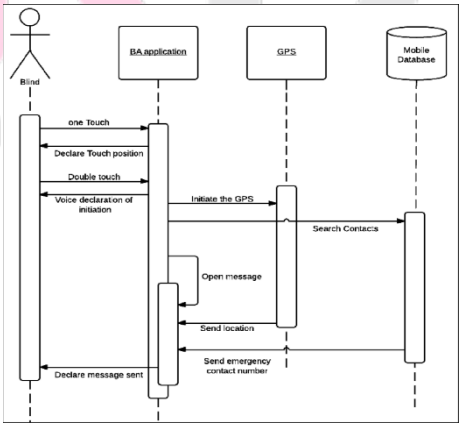

Fig: 5: Class Sequence Diagram

X. ALGORITHM/METHOD SPECIFICATION

A. *Long short-term memory (LSTM)*

It is an artificial recurrent neural network (RNN) architecture used in the field of deep learning method. Unlike good feed forward neural networks, LSTM has connections. It can process not only single data points (such as images), but also entire sequences of data (such as speech or video and text). For example, LSTM is applicable to tasks such as unsegmented, connected handwriting recognition, speech detector and oddity detection in network traffic or IDSs (intrusion detection systems).A common LSTM unit is composed of a cell, an input gate, an output gate and a forget gate. The cell remembers values over arbitrary time intervals and the three gates regulate the flow of information into and out of the cell.

B. ***Automatic Speech Recognition(ASR)***

It's known in short, is the technology that allows human beings to use their voices to speak with a computer interface in a way that, in its most sophisticated variations, resembles normal human conversation.

The most advanced version of currently developed ASR technologies revolves around what is called **Natural Language Processing**, or **NLP** in short. This variant of ASR comes the closest to allowing real conversation between people and machine intelligence and though it still has a long way to go before reaching an apex of development, we're already seeing some remarkable results in the form of intelligent smart phone interfaces like the Siri program on the iPhone and other systems used in business and advanced technology contexts.

However, even these NLP programs, despite and "accuracy" of roughly 96 to 99% can only achieve these kinds of results under ideal conditions in which the questions directed at them by humans are of a simple yes or no type or have only a limited number of possible response options based on selected keywords (more on this shortly).

The basic sequence of events that makes any Automatic Speech Recognition software, regardless of its sophistication, pick up and break down your words for analysis and response goes as follows:

You speak to the software via an audio feed

1. The device you're speaking to creates a wave file of your words
2. The wave file is cleaned by removing background noise and normalizing volume
3. The resulting filtered wave form is then broken down into what are called phonemes. (Phonemes are the basic building block sounds of language and words. English has 44 of them, consisting of sound blocks such as "wh", "th", "ka" and "t".

C. *MEL-FREQUENCY CEPSTRAL COEFFICIENTS (MFCCS)*

1. Take the Fourier transform of (a windowed excerpt of) a signal.
2. Map the powers of the spectrum obtained above onto the mel scale, using triangular overlapping windows or alternatively, cosine overlapping windows.
3. Take the logs of the powers at each of the mel frequencies.
4. Take the discrete cosine transform of the list of mel log powers, as if it were a signal.
5. The MFCCs are the amplitudes of the resulting spectrum.

D. *DYNAMIC TIME WARPING (DTW)*

1. Every index from the first sequence must be matched with one or more indices from the other sequence, and vice versa
2. The first index from the first sequence must be matched with the first index from the other sequence (but it does not have to be its only match)
3. The last index from the first sequence must be matched with the last index from the other sequence (but it does not have to be its only match)
4. The mapping of the indices from the first sequence to indices from the other sequence must be monotonically increasing, and vice versa, i.e. if {displaystyle j>i} are indices from the first sequence, then there must not be two indices {displaystyle l>k} in the other sequence, such that index {displaystyle i} is matched with index {displaystyle l} and index {displaystyle j} is matched with index {displaystyle k}, and vice versa.

CONCLUSION AND RESULT

The goal of this project was to create a voice detector web application, and apply it to a speech of a known speaker. By investigating the extracted features of the known speech and then compare them to the stored extracted features for each different speaker in order to identify and follow the order that speaker. The feature extraction is done by using MFCC (Mel Frequency Central Coefficients). MFCC and methods were used as primitive methods for recognizing the speaker uttering the same word in the training and testing phases, are giving good results to detect the speaker MFCC method which recognizes the speaker either utters the same or different word is generated by clustering the training feature of each speaker and then stored in the speaker database. In the recognition stage, a distortion measure which based on the minimizing the Euclidean distance was used when matching an unknown speaker with the speaker database. During this project, we have found out that the MFCC based clustering approach provides us with the faster speaker identification process than only FFT approach, hence it is the best choice to build high efficient speaker detector system.

For the future, Voice recognition assistants are capable of more than just answering your queries for Google. Thanks to technical due diligence investment, software engineers and developers can continue growing the technology. More applications are being made specifically to be compatible with smart devices, such as smart household appliances.

REFERENCES

[1]. Abdulla, W. H., Chow, D., & Sin, G. (2019, October). Cross-words reference template for DTW-based speech recognition systems. In TENCON 2019. Conference on Convergent Technologies for the Asia-Pacific Region (Vol. 4,pp.1576-1579).IEEE.

[2]. Al-Qatab, B. A., & Ainon, R. N. (2015, June). Arabic speech recognition using hidden Markov model toolkit (HTK).In Information Technology (ITSim), 2015 International Symposium in (Vol. 2, pp. 557-562). IEEE.

[3]. Bahl, L. R., Brown, P. F., De Souza, P. V., & Mercer, R. L. (2021). Estimating hidden Markov model parameters to maximize speech recognition accuracy. Speech and Audio Processing, IEEE Transactions on, 1(1),77-83.

[4]. Bahl, L. R., Brown, P. F., de Souza, P. V., Mercer, R. L., & Picheny, M. A. (2020). A method for the construction of acoustic Markov models for words. Speech and Audio Processing, IEEE Transactions on, 1(4), 443-452.

[5]. Butzberger, J., Murveit, H., Shriberg, E., & Price, P. (2020, February). Spontaneous speech effects in large vocabulary speech recognition applications. In Proceedings of the workshop on Speech and Natural Language (pp. 339-343). Association for Computational Linguistics.

[6]. Charles, A. H., & Devaraj, G. (2018). Alaigal-A Tamil Speech Recognition. Tamil Internet.

[7]. Dumitru, C. O., & Gavat, I. (2019, June). A Comparative Study of Feature Extraction Methods Applied to Continuous Speech Recognition in Romanian Language. In Multimedia Signal Processing and Communications, 48th International Symposium
ELMAR-2019 focused on (pp. 115-118). IEEE.

[8]. Furui, S., Ichiba, T., Shinozaki, T., Whittaker, E. W., & Iwano, K. (2015). Cluster-based modeling for ubiquitous speech recognition. Interspeech2015, 2865-2868.

[9]. Gaikwad, S. K., Gawali, B. W., & Yannawar, P. (2016). A review on speech recognition technique. International Journal of Computer Applications, 10(3), 16-24.

[10]. Ghai, W., & Singh, N. (2017). Literature review on automatic speech recognition. International Journal of Computer Applications, 41(8).