ANALYZING HUMAN VOICE for IDENTIFYING and PREVENTING the USAGE of UNINTENTIONAL WORDS USING VUS ALGORITHM

A. Suresh Kumar, N. Umamaheshwaran, H. Vasantha Kumar
Department of Computer Science and Engineering
S.K.P Engineering College
Tiruvannamalai, Tamil Nadu
Email: creativegeeksuresh@gmail.com

***Abstract*— Human Voice recognition is an essential fact for modern voice recognition applications and its related products. The Main reason for these kind of problem is the variation in user's communication at different places and at different mental condition. The problems faced by many people today is their lack of improper expression in communication. In order to overcome this problem a better way is to improvise their vocabulary capabilities and control the usage of particular word in a particular place. This paper presents an android based application for effectively improvising user speech and thereby increasing the chances for better recognition of human voice by the voice recognition applications. Mobile device record audio input and process them by using Variation of Uniquely Spoken words algorithm.**

***Keywords*— Android, Voice recognition, ,Vocabulary ,Acostics, Unintentional words.**

I. INTRODUCTION

Voice recognition also referred as Speech recognition plays a vital role in modern world technology which gives the user a better and easy to use products. Speech recognition (SR) is the inter-disciplinary sub-field of computational linguistics that develops methodologies and technologies that enables the recognition and translation of spoken language into text by computers. It is also known as "automatic speech recognition" (ASR), "computer speech recognition", or just "speech to text" (STT). It incorporates knowledge and research in the linguistics, computer science, and electrical engineering fields. Some SR systems use "training" also called "enrollment" where an individual speaker reads text or isolated vocabulary into the system. The system analyzes the person's specific voice and uses it to fine-tune the recognition of that person's speech, resulting in increased accuracy. Systems that do not use training are called "speaker independent" systems. Systems that use training are called "speaker dependent". Speech recognition applications include voice user interfaces such as voice dialing, call routing, search, simple data entry, preparation of structured documents, speech-to-text processing, and aircraft which is usually termed as Direct Voice Input.

The term *voice recognition* or *speaker identification* refers to identifying the speaker, rather than what they are saying. Recognizing the speaker can simplify the task of translating speech in systems that have been trained on a specific person's voice or it can be used to authenticate or verify the identity of a speaker as part of a security process. There are various factors which plays a vital role in obtaining is a better view for voice recognizing. The primary factor is Pattern recognition which is the word flow of user's speech or which kind of words the user frequently use following by what words. The second factor is Emotion recognizing which is the user's state of mind. The external factor like Acoustics which is the branch of science that deals with sound is also playing a key role as the sound absorption or finding the presence of external noise. The removing of unwanted noise is a key in data pre-processing as an error free voice given as an input will be used for efficient calculation of result.

The existing system uses Sentimental analysis for improvising user communication in terms of vocabulary by training user with a pre-defined set of words. Sentiment analysis also known as opinion mining refers to the use of natural language processing, text analysis and computational linguistics to identify and extract subjective information in source materials. Sentiment analysis is widely applied to reviews and social media for a variety of applications, ranging from marketing to customer service. Generally speaking, sentiment analysis aims to determine the attitude of a speaker or a writer with respect to some topic or the overall contextual polarity of a document. The attitude may be his or her judgment or evaluation, affective state, or the intended emotional communication.

II. RELATED WORK

In this section, we are about to survey the recent work related to our approach. Firstly, we review some approaches based on Gesture recognition [4]. Then, we review the often utilized Hidden Markov Model (HMM) for voice recognition. Also, the sentimental analysis and applications are provided in detail.

A. Gesture Recognition

Gesture recognition is a part in computer science and language technology with the goal of interpreting human gestures via mathematical algorithms. Gestures can originate from any bodily motion or state but commonly originate from the face or hand. Current focuses in the field include emotion recognition from face and hand gesture recognition. Users can use

simple gestures to control or interact with devices without physically touching them. This paper describes an application created on an Android mobile phone that recognizes gestures using the smartphone's orientation sensor. These gestures can be used to trigger events in another program running on a remote computer. We present a prototype application that generates different events for controlling a PowerPoint presentation, namely starting and stopping, as well as displaying the next and previous slide. In [5] it's working is explained clearly.

B. Voice Recognition

1) *Hidden Markov Model (HMM)*

Hidden Markov Model is used to model speech recognition application. We start with mathematical understanding of HMM followed by it and its solution. Real-world processes generally outputs which can be characterized as signals. The signals can be pure or can be corrupted from other signal sources or by transmission distortions.

One can classify signal model in to two types

1) Deterministic Model: It exploit some known property of signal (like Amplitude of wave).
2) Statistical Model: It takes statistical property of signal in to account. Example of this type of model is Gaussian Model, Poisson Model, Markov Model and Hidden Markov model.

Various algorithms used in various steps:

1) Forward Algorithm for Evaluation Problem.
2) Viterbi Algorithm for Decoding Hidden State Sequence.
3) Baum-Welch Algorithm for Learning.

2) *Isolated Word Recognizer*

The purpose of the Front-End is to parameterize an Input signal such as audio into a sequence of output Features. Voice samples are taken every 10-25msec. This sample data is feed to the Front-End module for further processing. Output of Front-End is list of feature vector. This feature vector are then mapped to symbol using vector quantization.

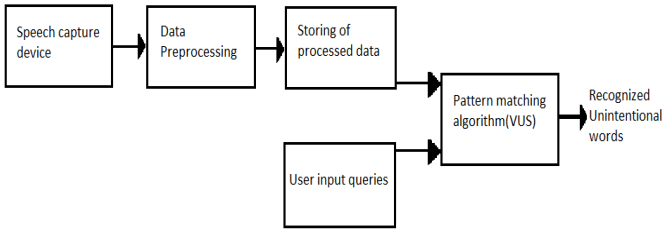
Vector Quantization: It maps Feature vector to symbol. This is also knows as acoustic modeling. This symbols represent HMM state. During recognition process this symbol are matched against unknown symbols. This gives us way to map complex vector in to manageable symbol set.

HMM model creation: Depending on implementation, HMMs are created for every basic sound unit, in our case it is Phoneme. Further all HMM are linked together to represent the vocabulary under consideration. This linked representation is known as search space for given problem. During recognition phase this graph is searched for finding occurrence of given word.

3. *Neural Network*

Neural Network is another method, which uses gradient decent method with back propagation algorithm. Study shows that this method works well in presence of large amount of training data also the recognition accuracy is high if the word under consideration is from training data set. While in HMM, the recognition ability is good for unknown word. HMM is generic concept and is used in many area of research. In whole architecture of speech recognition, HMM is just one block which help in creating search graph. It work in tandem with other block such as front-end, language model, lexicon to achieve desired goal. The purpose of HMM is to map feature vector to some representable state and emit symbol, concatenation of which gives desired phoneme sequence. Problem with continuous speech recognition is, to determine word boundary. It requires knowledge of language construct and regional accent understanding. Other problem is presence of noise in sample data. Efficient solution to this two problem is required for accurate continuous speech recognition. What we discussed in the above section is the recognition of English sentence, we can extend this model to recognize multi-lingual sentence.

III. System Architecture



IV. Proposed Approach

The purpose of our system is to find the usage of user's unintentional words by continuously observing the user's speech and determining his word flow. In this paper, we first obtain the user queries that is the unintentional words which the user wishes to forget while having an conversation. The user queries is first obtained and is stored which is then compared with the real time data obtained from the microphone present on the android device. The obtained voice input is then preprocessed for cleaning unwanted noise.

Then we introduce the VUS algorithm which is the variation of uniquely spoken words for determining the user query in the real time data obtained from microphone. Unlike, the existing system the proposed system gives user the choice to choose his own input rather than forcing him to choose from the set of input which is already given. Based on analysis done on user's data of previous conversion it determines all the possible predecessor and successor words that are used.

Finally, we determine the best set of successor and predecessor word of that particular unintentional word and also by continuously monitoring the user speech we also determine the user stress level. In order to simplify the work various API are available just like Google API which is offered by Google for the developer to create application integrating google voice recognition and google text to speech engine. In [1] they use the speech recognition for improving children's language learning support. The various process involved in our approach are as follows:

A. Data Pre-processing

IJSER

Data preprocessing is a technique that involves transforming raw data into an understandable format. Real-world data is often incomplete, inconsistent, and/or lacking in certain behaviors or trends, and is likely to contain many errors. Data preprocessing is a proven method of resolving such issues.

After raw data passed to data preprocessing step various kinds of unwanted noise are removed.

The real time data late obtained are then processed in this step before going for further step in which the comparison is made.

B. Speech Recognition

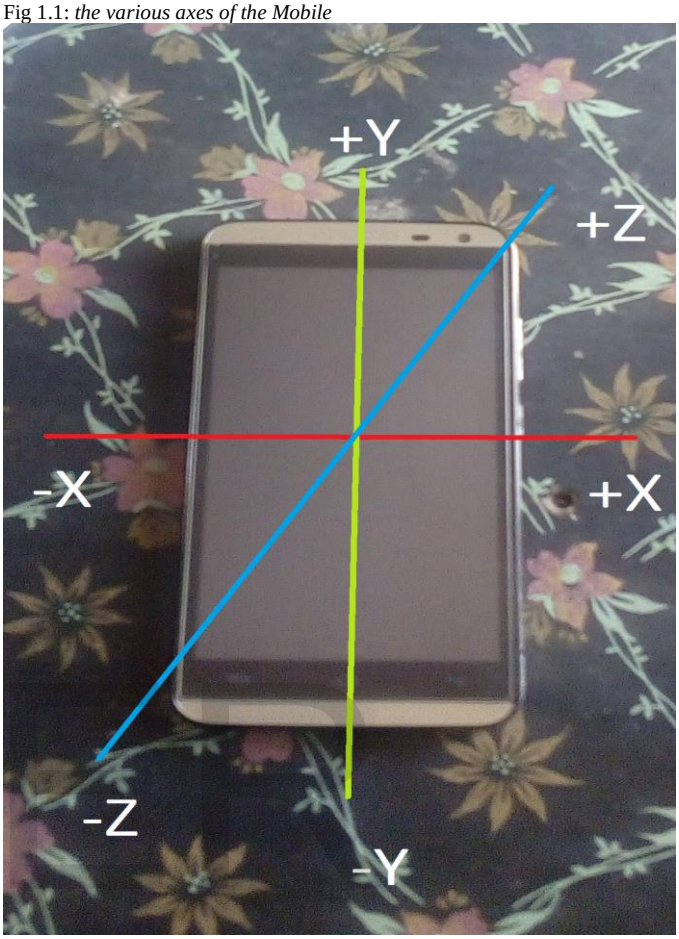
After the data is processed in the data preprocessing step it is processed by using various mechanism for identifying the user queries that is the user input from the real time data obtained from the microphone in the android device. Google provides various API which allows to process the voice.

1) Google cloud speech API

Google Cloud Speech API enables developers to convert audio to text by applying powerful neural network models in an easy to use API. The API recognizes over 80 languages and variants, to support your global user base. You can transcribe the text of users dictating to an application's microphone, enable command-and-control through voice, or transcribe audio files, among many other use cases**.** Recognize audio uploaded in the request**,** and integrate with your audio storage on Google Cloud Storage, by using the same technology Google uses to power its own products. Applying the most advanced deep learning neural network algorithms to your user's audio for speech recognition with unparalleled accuracy. Speech API accuracy improves over time as Google improves the internal speech recognition technology used by Google products. Speech API can stream text results, returning partial recognition results as they become available, with the recognized text appearing immediately while speaking. Alternatively, Speech API can return recognized text from audio stored in a file. You don't need advanced signal processing or noise cancellation before sending audio to Speech API. The service can successfully handle noisy audio from a variety of environments. Automatic Speech Recognition (ASR) powered by deep learning neural networking to power your applications like voice search or speech transcription. Returns recognition results while the user is still speaking. Speech recognition can be customized to a specific context by providing a set of words and phrases that are likely to be spoken. Especially useful for adding custom words and names to the vocabulary and in voice-control use cases. Audio input can be captured by an application's microphone or sent from a pre-recorded audio file. Multiple audio encodings are supported, including FLAC, AMR, PCMU and Linear-16. Handles noisy audio from many environments without requiring additional noise cancellation. Filter inappropriate content in text results for some languages. Audio files can be uploaded in the request or integrated with Google Cloud Storage.

Mobile application developer can use the resource offered by google by using the API which is available to all developer to integrate it in their application and along with the various other API available for integrating with their application in order to perform any functionality.

Fig 1.1: *the various axes of the Mobile*



2) Sensor

Various kinds of sensor are available in mobile devices which can be used by android application. In Android platform we use two kinds of sensors for this project:

1) Motion sensors
2) Position sensors

Motion sensors: These sensors measure acceleration forces and rotational forces along three axes. This category includes accelerometers, gravity sensors, gyroscopes, and rotational vector sensors.

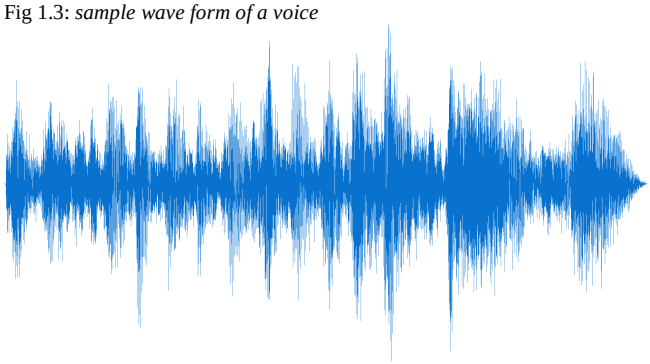
Position sensors: These sensors measure the physical position of a device. This category includes orientation sensors and magnetometers.

The Linear acceleration sensors help in detecting whether the user is in movement while having a conversation or in a static position. The gyroscope sensor helps in detecting the angle in which the phone is in while receiving the real time data using the microphone present on the mobile device. Fig 1.1 illustrates what are the various axes than can be possible. Sensor framework can be used to determine which sensors are available in the device and to determine an individual sensor's capabilities such as its maximum range, manufacturer,

power requirements. It can also be used to acquire raw sensor data and define the minimum rate at which we can acquire a sensor data.

Fig 1.2: Application[OLI] *storing raw data for processing*

Fig 1.3: *sample wave form of a voice*



C. VUS algorithm

Variation of Uniquely Spoken words is the algorithms that is to be deployed in this project for classifying and categorizing the user unintentional words which is spoken accidently during his conversation. The algorithms not only determines the total count of the usage of that particular word but also its predecessor and successor word. Once the predecessor and successor word set is obtained it is then processed for obtaining the best/mostly used set of predecessor and successor word of that particular unintentional word.

Algorithm:

Array of User input → A
Array Index → i , j
Raw data obtained → R
Predecessor word → p1
Successor word → s1
Predecessor and Successor word set → P, S
Best set of predecessor and successor word → B
Database → D : Database for storing processed data
Word count → w

INITIALIZE
i, j ← 0
W ← 0
D ← null
P, S ← null
B ← null

START
Read A[i] : user input
Store A[i] in the D
Loop:
Read R
if (A[i]==R)
  w++; increment word count
  Find p1, s1 and store it in P,S set.
  Goto Loop;
End if

for(p1←P[i],s1←S[i]; (P,S); i++)
  for(j←sizeof(P); (P,S); j++)
    if(p1[j]==p1[i] && s1[j]==s1[i])
      Store p1[j] and s1[j] in B
    End if
  End for
End for

The above algorithm will be used for categorizing user unintentional word and its mostly used successor and predecessor word set which is present in B and also the possible data sets which is present in P, S. The total number of the times the user used a particular unexpected word can be obtained from the word count value stored in w.

V. EXPERIMENTS

We conducted a set of texts in our experiment which is the our project deployed as application in android with a random number of user's. In the previous approach they experimented on analyzing user smoking behavior and its effect based on the user voice as in [3]. In our approach we are analyzing user word flow by using his verbal/voice conversations. The Experimental result can be used to support the factor of increasing the possibility of user communication capabilities.
The user queries and the raw data obtained are being processed and converted to text with the help of google voice recognition which is freely available for the developer to use. Unlike other systems as in [2] this system does not forces user to do certain actions, it only insist on the actions to be done giving user the total freedom to choose by himself.

IJSER

During the testing phase a sample set of data obtained from a random user are processed and its result which is obtained by using the variation of uniquely spoken words is listed in detail in the Table 2.1.

Table 2.1: sample data set obtained from user using VUS

UNINTENTIONAL WORDS	PREDECESSOR WORDS	SUCCESSOR WORDS
Hello	Hi	How
Must	You	Do
Dumb	Like	Man
Worst	You are	Student

Among the array of user input a ranking is done based on the word count factor "w" in the VUS algorithm which helps the user to focus on that word instead of focusing on the correction of all the word which saves time and provides more efficiency. The ranking done on the mostly used unintentional words among all the unintentional words or the user queries given is listed in Table 2.2.

Table 2.2: ranking based on the mostly used words

Unintentional words	Ranking based on the count
Dumb	13
Wow	10
Worst	8

The experimental results also allows an another additional factor called user stress to be predicted based on the high frequency range continuously obtained in the raw data processing and classification done based on the various frequency levels used by the user at different emotional level during his conversation. The Table 2.3 gives a clear perception of the variation in high frequencies and classification done based on those frequencies.

Table 2.3: Classification of user stress based on their voice

User	Frequency level	Stress level
Person 1	Low	Normal
Person 2	High	High
Person 3	Moderate	Normal

VI. CONCLUSION

In this paper, a voice recognition mechanism was formulated for determining user's unintentional words. We highlighted the concept of obtaining user pattern for word flow using VUS method. The experiment results show significant improvement in user communication capabilities based on the real time dataset and slighter improvement in their usage of other voice recognition applications.

REFERENCES

[1] Takahiro Nakadai, Motohiko Hano and Hiroshi Mizoguchi, "Novel application of microphone array sensor for children's language learning support — Intelligent Interactive Design using voice separation and recognition system" in 2015 9th International Conference on Sensing Technology (ICST).

[2] Changhai Wang,Yuwei Xu, Jianzhong Zhang and Wenping Yu, "SW-HMM: A Method for Evaluating Confidence of Smartphone-Based Activity Recognition" in 2016 IEEE Trustcom/BigDataSE/ISPA, Pages: 2086 2091, DOI: 10.1109/TrustCom.2016.0320

[3] Abdurrahman Yildiz and Umut Arioz, "Analysing human voice and classification of voice frequencies according to smoking effect" in 2015 International Conference on Pervasive Computing (ICPC), Pages: 1 - 4, DOI: 10.1109/PERVASIVE.2015.7087119

[4] Shraddha Uddhav Khadilkar and Narendra Wagdarikar, "Android phone controlled voice, gesture and touch screen operated smart wheelchair" in 2015 International Conference on Pervasive Computing (ICPC),Pages: 1 - 4, DOI: 10.1109/PERVASIVE.2015.7087119.

[5] Eric Torunski; Abdulmotaleb El Saddik; Emil Petriu, "Gesture recognition on a mobile device for remote event generation" in 2011 IEEE International Conference on Multimedia and Expo, Pages: 1 - 6, DOI: 10.1109/ICME.2011.6012188

[6] A.Suresh (2016), "Speech Stress Analysis based on Lie Detector for Loyalty Test", in International Journal of Printing, Packaging & Allied Sciences,(IJPPAS) ISSN: 2320- 4387, *Vol. 04, No.01, December 2016*, pp.*631* – 638.

IJSER