Robust Speech Recognition Technique Using Mat lab

B.Siva Ayyapa Kumar[1], [2]Shaik.Yasmine

[1]*Assistant Professor, Dept. of computer science& Engineering, Vignan's lara Institute of Technology&science , Andhra Pradesh, India*
[2]*Student, Dept. of computer science& Engineering, Vignan's lara Institute of Technology&science , Andhra Pradesh, India*

---***---

Abstract – *In our project we represent a new methodology for robust speech recognition. Speech recognition is a method by which operator is able to identify the natural language. In present speech recognition is most commonly used due to increase in the digitalisation. In the first step we are taking audio signal as the input. And in the second step we are eliminating the noise. And in the third step match the patterns to recognize the input. Voice signal is the main interface for the acceptance of devices. Speech recognition is the branch of science known as Natural Language processing (NLP). NLP is the field of computer sciences, linguistics and artificial sciences which is bothered to be the interaction between the natural language (spoken by humans) and computer systems. NLP connects the computer system to drive the meaning out of the input language thus it makes the user easy to operate the device. Recent research proposal is a general method, that recognize the voice by decoding it to patterns, These patterns are used for the indexation of weighted automaton which produce a value that decides the probability of a particular word is spoken or not. In our proposed method we have used LRU technique which required for penetrating pattern in the database. In this method we stored the pattern instead of the voice which again reduces the space complexity of the system. The structure of paper is as follows: In section 2 we offer a brief introduction of speech recognition system and the database use for the system. In section 3 we will discuss how voice signals, the noise is started and active methods. In section 4 we will propose our method and evaluate results. The results are summarized in section 5.*

Keywords: Voice Recognition; Natural Language Processing; Fast Fourier Transform; Hidden Markov Model

1. INTRODUCTION

Speech identification (SI) is the inter-disciplinary sub-field of computational linguistics that develop methodologies and technologies that enable the identification and translation of spoken language into text by computers. It is also known as "automatic speech identification" (ASI), "computer speech identification", or just "speech to text" (STT). It incorporates knowledge and research in the linguistics, computer science, and electrical engineering fields.

Some SI systems use "training" (also called "staffing") where an individual speaker reads text or secluded vocabulary into the system. The systems analyze the person's specific voice and use it to fine-tune the identification of that person's speech, resulting in bigger accurateness. Systems that do not use training are called "speaker independent"[1] system. System that use training are called "speaker reliant".

Speech recognition applications include voice user interfaces such as voice dial (e.g. "Call home"), call direction-finding (e.g. "I would like to make a collect call"), domotic appliance control, search (e.g. find a podcast where particular words were spoken), simple data entry (e.g., entering a credit card number), preparation of structured documents (e.g. a radiology report), speech-to-text processing (e.g., word processors or emails), and aircraft (usually termed Direct Voice Input).

The term voice identification[2][3][4] or speaker identification[5][6] refers to identifying the speaker, rather than what they are saying. Identifying 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 confirm or verify the identity of a speaker as part of a safekeeping process.
From the technology perspective, speech identification has a long history with several waves of major innovation. Most recently, the field has benefited from advances in deep learning and big data. The advances are evidence not only by the surge of academic papers published in the field, but more importantly by the worldwide industry adoption of a variety of deep learning methods in designing and deploying speech recognition systems

2. Speech Identification System And Database

The speech identification system is continuous mixture density hidden Markov model (HMM) system whose parameters are estimated by Vertebra training [2]. In our Speech identifying technique first the normalization is done, which is to find Cepstral coefficient by taking the

Fourier transform of a short time window speech followed by decorating the spectrum using inverse Fourier transform and then recognize the pattern for different speakers and recording conditions. In a second step phoneme are recognized on the basis of the pattern. The database will be consisting of a different set of patterns which will represent the states required for the successful completion for a given phoneme. This will have ascendancy over the other methods which use warehouse of training audio, for more than one time. The multilevel checking of audio input from the desired audio is done. The database also contents weight, assign to each state for the successful completion, input voice has to complete and cross through the threshold limit. If any input audio does not successfully complete stages of any phoneme in the database, then its output is directed to the database phoneme whose threshold limit is close to threshold limit of the input phoneme (audio).

Searching of the phoneme in the database system on the technique in which most recently used voice patterns are kept at the top in the database. This technique is popularly known as LRU (Least Recently Used). The best part of using this technique is to decrease the time complexity at the time of searching the desired phoneme. This technique makes it more efficient than others. The drawback of conventional voice indexing based on phonemes is that sub-word units selected for the purpose of indexing are not reliable due to relatively low phoneme recognition rate [3].

3. NOTATIONS AND ASSUMPTIONS

Several methods combine the enframing the composition with a design of the pattern based on the FFT of the signal [6]. While the energy of the whole signal is preserved by such a transformation, the computed energy on each interval may be drastically changed [7]. We use the notation F (x) denoting Fourier transform and Denoting inverse Fourier transform of an audio signal. We first took the Fourier transform F (x) of the input audio signal. Its data is extracted using Fourier transform, that data is discretised using inverse Fourier transform. N (⍰) is considered as a normalization function holding noise eliminated values. Energy level data is calculated using the equation.

$$f(x) = \int_{-x}^{x} F(K) e^{2\pi/RX}\, dK$$

$$F(X) \blacksquare \int_{-X}^{X} f(x)\, e^{-2\pi RX} \,|\, dX$$

Here,

$$F(k) \blacksquare R[f(x)]K$$

$$= \int_{-x}^{x} f(x) e^{-2\pi/RX} dx$$

Is called the forward (-i) Fourier transform, and

$$f(x) \blacksquare F^{-1}_{R}[F(K)](x)$$

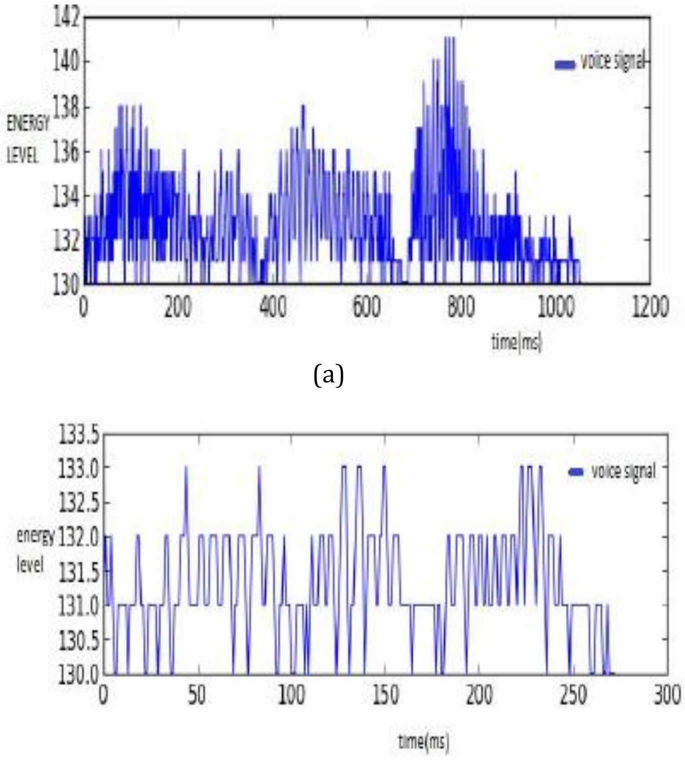
$$= \int_{-x}^{x} F(K) e^{2\pi/RX} dK$$

Is called the inverse (+i) Fourier transform.

A.NOISE ELIMINATION

We assume that the observed signal is a realization of wide sense stationary process [4]. If the signal to noise ratio is not too low, a simple method to detect speech is based on signal energy [2]. As the level of energy in the speech signal is higher than the level of noise energy. On this basis a threshold limit can be imposed on the energy signal, all the values above it will be considered as utterance, information and all the values below the threshold value will be discarded as probability those values will be nice. The equation for the noise elimination is given as follows

$$N(\omega) = \textstyle\sum_{-x}^{x} F(x) \qquad \text{In } F(x) > \text{threshold limit}$$



(a)

(b)

Fig 1. (a) Graph showing energy level of noise elimination (b) energy level after noise elimination

3. Proposed Method

The motivation behind this method is to give voice commands to the system by making a speech utterance. As now in the digital world using voice command to operate a device will be more appreciated and more convenient than using mouse. As mentioned in Fig 2, first audio is taken from the microphone which is a natural language with the environmental noise. In the second step, audio file generated by the sound recorded from the microphone, is used to generate Fourier transform of it which gives the sound energy level data along the time. In the third step, the noise elimination is done on the data generated by the Fourier transform. This is to make ensure that the audio is now free from noise, which intern increase the chances of decoding the main voice command from the audio file recorded for the recognition. In the fourth step, noise eliminated data is converted into a string by finding the ASCII transformation for each data generated which is used to encode the voice to string pattern for extraction of the required patterns. In the fifth step, the patterns are matched from the patterns already stored in the database. The percentage of pattern matched gives a value which is calculated by summing up the weight associated with the states of the automata, which must cross threshold value for any word in the database for its successful recognition. After the threshold value is crossed successfully the task associated with that command is done. The task is also stored in the database corresponding to the word spoken.

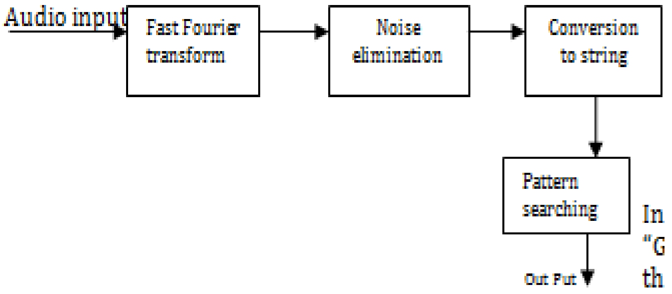


Fig 2. Framework of proposed method

5. Results

We first took the convolution between the speeches of the same kind and the graphs generated by convolution gave us the result, that there are some common areas where the data of two or more utterances of the same word is same. For example, in Fig 3 we get common region 0-400 (approx taken along time axis having a same corresponding energy level), it ensures that if the pattern of the waves is generated in this region can be used as a state in the automata of the word "open". Fig 3 and fig 4 convey the same.

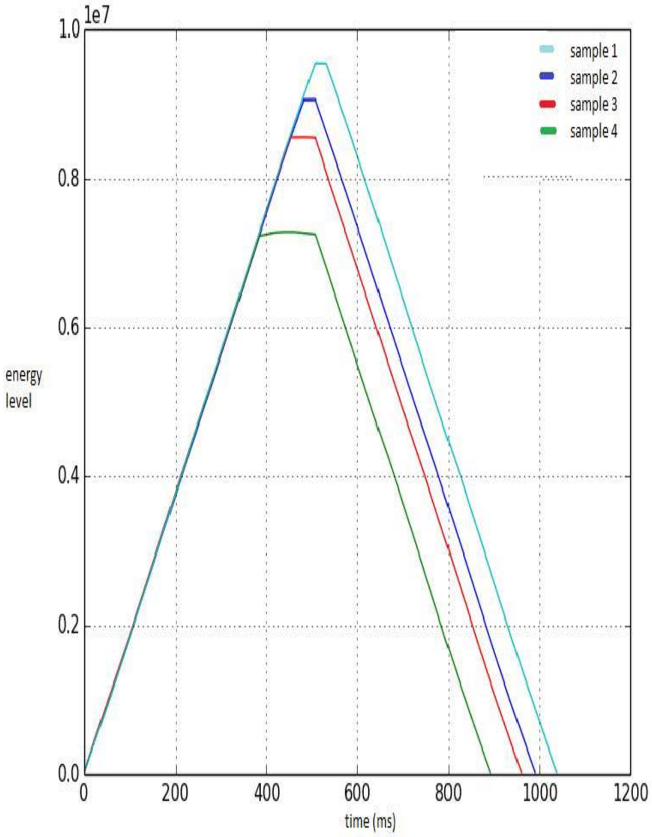


Fig.3. Graph representing utterance of the word "Open"

In fig 4 there are plots of 4 utterances of the word "Google". It Indicates the utterance level of the word is at the same level on a time scale of 0-400 ms, corresponding with unique energy values, which indicates that patterns for "Google" will be found unique between these intervals compared to other words.

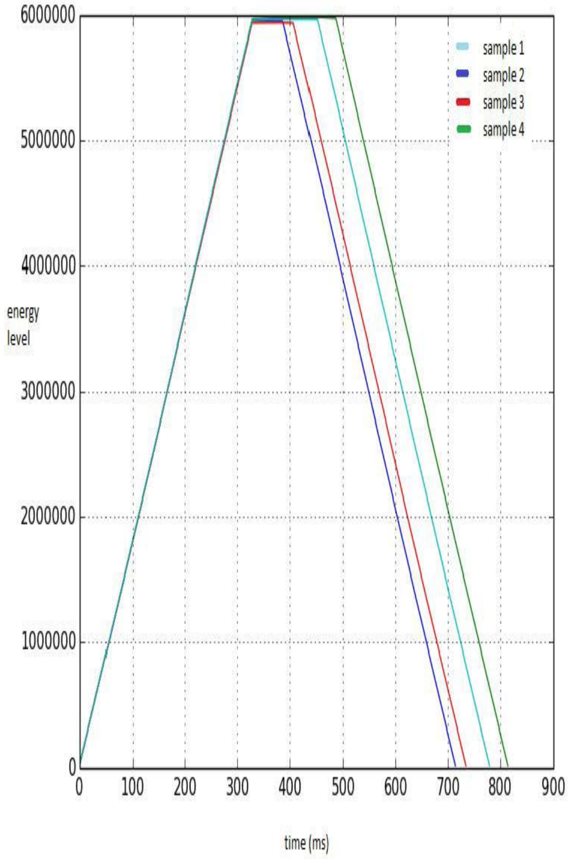


Fig4. Graph representing utterance of the word "Google".

6. CONCLUSIONS

In this paper, we presented the technique for the robust recognition of the word (voice command) utterance. It is a procedural approach in which voice has to pass through the automata, gain sum value which must be greater than equal to the threshold value for successful recognition and applies error tolerance to achieve robustness against human errors. This technique is fit for designing voice commanding system. Using pattern for the complete recognition of a word, followed by threshold value checking makes it more robust to identify the utterance of the word accurately. 90%of the utterance can be recognized correctly. Further work will be focusing on making a system more efficient and user friendly. We are implementing this method on a voice commanding system.

References

[1] Changxue Ma, Uniterm Voice indexing and search for mobile devices, applications & software research center Motorola Inc 1295 e. Algonquin IL 60196.

[2] Volker Stahl, Alexander Fischer and Rolf Bippus, quantile based noise estimation for spectral subtraction and wiener filtering, Philips research laboratories.

[3] Olivier Siohan, Michiel Bacchiani, Fast Vocabulary-Independent Audio Search Using Path-Based Graph Index. INTERSPEECH, 2005.

[4] M. H. Hayes, "Statistical Digital Signal Processing and Modeling" John Wiley & Sons, Inc., 1996.

[5] C. Allauzen, M. Mohri & M. Saraclar. General Indexation Of Weighted Autometa – Application to Spoken Utterance Retrieval ACL, HLT, 2004.

[6] Jerome Lebose, Luc Brun, Jean Cluade Pailles, "A Robust Audio Fingerprint Extraction Algorithm".

[7] S. Mallat. A Wavelet tour of signal processing. Academic press, 1999. Chapter VIII p.363.

[8] S. Furui, "Recent advances in robust speech recognition," in Proc. ESCA-NATO Tutorial and Research Workshop on Robust Speech Recog- nition for Unknown Communication Channels, Pont-a-Mousson, France, Apr. 1997, pp. 11–20.

[9] Y.-F. Gong, "Speech recognition in noisy environments: A survey," Speech Commun., vol. 16, pp. 261–291, 1995.

[10] B.-H. Juang, "Speech recognition in adverse environments," Comput. Speech Lang., vol. 5, pp. 275–294, 1991.

[11] C.-H. Lee, "On feature and model compensation approach to robust speech recognition," in Proc. ESCA-NATO Tutorial and Research Workshop on Robust Speech Recognition for Unknown Communication Channels, Pont-a-Mousson, France, Apr. 1997, pp. 45–54.