ISSN:0975-9646

Design and Development of Word Recognition for Marathi Language

Potale Shubham[1], Kharpude Pratik[1], Patil Rahul [1], Ajay Kumar Gupta [2]

Department of Computer Engineering IOK College of Engineering Pune-412208

***Abstract*—– Speech is the most prominent form of communication among human beings. English is an official language at many institutes all over the world and maximum research work is carried out for English language. As far concern with Indian languages, small amount of work is carried out. The work carried out for Indian languages namely Hindi, Bengali, Punjabi, Telugu, Malayalam and Assamese. But very little amount of work is carried out for Marathi language.**

It is said that spoken Marathi Language changes at every 14 miles and has different types of phonetics and intonation. On the other hand Marathi is used as official language at very few places. There are very few institutes which have authority to comment on correctness of language. The authenticity of correctness of language is a major problem in regional languages.

Our project is capable to recognize the isolated Marathi word. At the initial level effort is made to provide help for basic operations. Most of the research is carried out on MATLAB but we go for Sphinx 4 platform as it concerns with JAVA.

Keywords: **Speech Recognition System, Sphinx4, MATLAB etc.**

I. INTRODUCTION

Speech is most prominent way of communication in world wide. Speech is the heart of human activity because it helps human to interact each other in more natural and effective way. Speech recognition is the process to identify words or phrase from spoken language and convert into machine readable format. Speech recognition is the process by which a computer (or other type of machine) identifies spoken words. Basically, it means talking to your computer, AND having it correctly recognizes what you are saying.

Most of the communication in Maharashtra is done in Marathi language .While going through papers we studied that for Marathi language very small amount of work and research is done and is not enough to use in practical application. This project can be extended to other languages in Maharashtra were each and every person can use voice application to command ,to search, to interact with internet computer etc.

II. TYPES OF SPEECH RECOGNITION SYSTEM

There are four types of Speech Recognition System are as follows:

Isolated Words

Isolated word recognizers usually require each utterance to have quiet (lack of an audio signal) on BOTH sides of the sample window. It doesn't mean that it accepts single words, but does require a single utterance at a time. Often, these systems have "Listen/Not-Listen" states, where they require the speaker to wait between utterances (usually doing processing during the pauses). Isolated Utterance might be a better name for this class.

Connected Words

Connect word systems (or more correctly 'connected utterances') are similar to Isolated words, but allow separate utterances to be 'run-together' with a minimal pause between them.

Continuous Speech

Continuous recognition is the next step. Recognizers with continuous speech capabilities are some of the most difficult to create because they must utilize special methods to determine utterance boundaries. Continuous speech recognizers allow users to speak almost naturally, while the computer determines the content. Basically, it's computer dictation.

Spontaneous Speech

There appears to be a variety of definitions for what spontaneous speech actually is. At a basic level, it can be thought of as speech that is natural sounding and not rehearsed. An ASR system with spontaneous speech ability should be able to handle a variety of natural speech features such as words being run together, "ums" and "ahs", and even slight stutters.

III.LITERATURE SURVEY

Speech recognition came into existence during 1920. The first machine i.e. Radio Rex ,a toy to recognize voice was manufactured. Bell Labs developed a speech synthesis machine at the World fair in New York. But later on they discarded efforts based on an incorrect conclusion that the AI is ultimately required for success. In order to develop systems for ASR, attempts were made in 1950s where researchers studied the fundamental concepts of phonetic-acoustic. Most of the systems in 1950 for recognizing speech examine the vowels spectral resonancews of each utterance. At Bell Labs Davis, Biddulph and Balashek(1952) premeditated a isolated digit recognition system for a single speaker using formant frequencies estimated during vowel regions of each digit. At RCA Labs, Olson and Belar (1950) built 10 syllables recognizer of a single speaker and Forgie and Forgie built a speaker-independent 10-vowel recognizer at MIT Lincoln Lab, by measuring spectral resonances for vowels. Fry and Denes (1959) tried to build a phoneme recognizer to recognize four vowels and nine consonants at University College in

England by using a spectrum analyser and a pattern matcher to make the recognition decision. Japanese labs entering recognition field in 1960-70. As computers are not fast enough, they designed special purpose H/W as a part of their system. In Tokyo, Nagata et.al described a system of the Radio Research lab, was a H/W vowel recognizer. Another effort was the work of Sakai and Doshita in 1962, of Kyoto University who built a H/W phoneme recognizer. In 1963, Nagata and co-workers at NEC Labs built a the digit recognizer. This led to a long productive research program. In 1970, the key focus of research was on isolated word recognition. IBM researchers studied in large vocabulary speech recognition. At AT&T Bell Labs, researchers began speaker independent speech recognition experiments. A large number of clustering algorithms were used to find the number of distinct patterns required to represent words to achieve speaker independent speech recognition. This research has been refined so that the techniques for speaker independent patterns are widely used. Carnegie Mellon University's Harphy system recognizes speech with vocabulary size of 1011 words with reasonable accuracy. It was the first to make use of finite state network to reduce computation and determine the closest matching strings efficiently. In 1980, the key focus of research was on connected words speech recognition. In the beginning of 1980, Moshey J. Lasry studied speech spectrogram of letters and digits and developed a feature based speech recognition. There is a change in technology in 1980 from template based approaches to statistical modelling approach specially HMM in speech research. The most significant paradigm shift has been the introduction of statistical methods, especially stochastic processing with HMM (Baker, 1975 & Jelinek, 1976) in the early 1970's (Portiz 1988). More than 30 years later, this methodology still predominates. Despite their simplicity, Ngram language models have proved remarkably powerful. Now days, most practical speech recognition systems are based on statistical approach and their results with additional improvements have been made in 1990s. In 1980, Hidden Markov model (HMM) approach is one of the key technologies developed. IBM, Institute for Defence Analysis (IDA) and Dragon Systems understood HMM , but it was not renowned in the mid-1980s. Neural networks to speech recognition problems is the another technology that was reintroduced in the late 1980s. In 1990, Pattern recognition approach was developed. It followed Bayes's framework traditionally but it has been altered into an optimization problem with minimization of the empirical recognition error. The reason for this alteration is that the distribution functions for the speech signal could not be chosen accurately and under these conditions, Bayes' theory cannot be applied. However, aim is to design recognizer with least recognition error rather than best fitting to given data. The techniques used for error minimization are Minimum Classification error (MCE) and Maximum Mutual Information (MMI). These techniques led to maximum likelihood based approach to speech recognition performance. A weighted HMM algorithm is proposed to address HMM based speech recognition issues of robustness and discrimination. In order to decrease the acoustic mismatch between given set of speech model and test utterance, a maximum likelihood stochastic matching approach was proposed. A narrative approach for HMM speech recognition system is based on the use of a neural network as a vector quantizer which is remarkable innovation in training the neural network. Nam Soo Kim et.al. Described a variety of methods for estimating a robust output probability distribution based on HMM. An extension of the viterbi algorithm made second order HMM efficient as compared to existing viterbi algorithm. In 1990s, the DARPA program continued. After that the centre of attention is Air Travel Information Service (ATIS) task and later focus on transcription of broadcast news (BN). Advances in continuous speech recognition and noisy environment speech recognition, have been explained. In the area of noisy robust speech recognition, minor work has been done.For noisy environment, for robust speech recognition, a new approach to an auditory model was proposed. This approach is computationally efficient as compared with other models. A model based spectral estimation algorithm has been developed. In 2000, a variational Bayesian estimation technique was developed. It is based on posterior distribution of parameters. Giuseppe Richardi has developed a technique to solve the problem of adaptive learning in ASR. In 2005, some improvements have been made on Large Vocabulary Continuous Speech recognition system for performance improvement. A 5-year national project Corpus of Spontaneous Japanese (CSJ) was conducted in Japan. It consists of 7 million of words approximately, corresponding to speech of 700 hours. The techniques used in this project are acoustic modelling, sentence boundary detection, pronunciation modelling, acoustic as well as language model adaptation, and automatic speech summarization]. Utterance verification is being investigated to further increase the robustness of speech recognition systems, especially for spontaneous speech,. When humans speak to each other, they use multimodal communication. It increases the rate of successful transfer of information when the communication takes place in a noisy environment. In speech recognition, the use of the visual face information, especially lip movement, has been examined, and results show that using both mode of information gives better performances than using only the audio or only the visual information, specially, in noisy environment.

IV. Tools For Speech Recognition

Following are the various tools used for ASR

PRAAT: It is free software with latest version 5.3.04 which can run on wide range of OS platforms and meant for recording and analysis of human speech in mono or stereo

AUDACITY: It is free, open source software available with latest version of 1.3.14(Beta) which can run on wide range of OS platforms and meant for recording and editing sounds.

CSL: Computerised Speech Lab is a highly advanced speech and signal processing workstation (software and hardware). It possesses robust hardware for data acquisition and a versatile suite of software for speech analysis.

SPHINX: Sphinx 4 is a latest version of Sphinx series of speech recognizer tools, written completely in Java programming language. It provides a more flexible framework for research in speech recognition.

SCARF: It is a software toolkit designed for doing speech recognition with the help of segmental conditional random fields.

MICROPHONES: They are being used by researchers for recording speech database. Sony and I-ball has developed some microphones which are unidirectional and noiseless.

V. Performance Of Speech Recognition System

The performance of speech recognition is specified in terms of accuracy and speed. Accuracy is measured in terms of performance accuracy which is known as word error rate (WER) whereas speed is measured with the real time factor.

Word Error Rate It is a common metric of the speech recognition performance. As recognized word sequence have a different length from the reference word sequence, there is difficulty in measuring performance.

WER = N IDS □□

Where S is number of substitutions D is number of deletions I is number of insertions and N is number of words in the reference.

Sometimes word recognition rate (WRR) is used instead of WER while describing performance of speech recognition.

WRR = 1- WER
= 1- N IDS □□
= N IDSN □□□

Speed It is measured by real time factor. If it takes time T to process an input of duration D then real time factor is defined by

RTF = D T

RTF ⩽1 implies real time processing

VI. Problem Identification

By doing a research we come to a conclusion that almost all the work for Speech Recognition for isolated word is done in MFCC, LPC, DTW, Matlab, etc. Sphinx Tools is used for Tamil and Telugu Languages and is having best performance accuracy. So we proposed that Sphinx can be used to recognized isolated word for Marathi language.

VII. Proposed System

Sphinx is a simple feasible and open source tool used in speech recognition engine to communicate with smart devices. We are using sphinx tools with java language to identify isolated word for Marathi language. We are using Small vocabulary initially and then we will expand the vocabulary to large amount. Here we are going to create a Database of 50 Speakers (25 Female & 25 Male). And starting from 10 isolated words we will take 5 utterances of each word from thoughts 50 speakers.

VIII. Summary

Research in speech recognition has been carried out intensively for the last 60 years.

A. Conclusions

Hence we studied work done for isolated Marathi language and we come to a conclusion that for isolated word all the work is done in MFCC and LPCC or in Matlab No another work is done using Sphinx tools

Acknowledgment

We thank the anonymous referees for helpful comments and suggestions. We also thank College to provide us a useful guidance and to encourage us for this project.

References

1. *"Speaker Independent Connected Speech Recognition- Fifth Generation Computer Corporation" Fifthgen.com. Retrieved 2013-06-15.*
2. "British English definition of voice recognition". Macmillan Publishers Limited. RetrievedFebruary 21, 2012.
3. *"voice recognition, definition of". WebFinance, Inc. Retrieved February 21, 2012.*
4. **Jump up**^ *"The Mailbag LG #114". Linuxgazette.net. Retrieved 2013-06-15.*
5. *McKean, Kevin (Apr 8, 1980). "When Cole talks, computers listen". Sarasota Journal. AP. Retrieved 23 November 2015.*
6. *Morgan, Nelson; Cohen, Jordan; Krishnan, Sree Hari; Chang, S; Wegmann, S (2013).Final Report: OUCH Project (Outing Unfortunate Characteristics of HMMs). CiteSeerX: 10.1.1.395.7249.*
7. Kincaid, Jason. "The Power Of Voice: A Conversation With The Head Of Google's Speech Technology". *Tech Crunch*. Retrieved 21 July 2015.
8. *Froomkin, Dan. "THE COMPUTERS ARE LISTENING". The Intercept. Retrieved20 June 2015.*
9. *Deng, L.; Hinton, G.; Kingsbury, B. (2013). "New types of deep neural network learning for speech recognition and related applications: An overview".doi:10.1109/ICASSP.2013.6639344.*
10. Markoff, John (November 23, 2012). "Scientists See Promise in Deep-Learning Programs". New York Times. Retrieved 20 January 2015.
11. *"Speech Recognition for Learning". National Center for Technology Innovation. 2010. Retrieved 26 March 2014.*
12. Mohri, M. (2002). "Edit-Distance of Weighted Automata: General Definitions and Algorithms" (PDF). *International Journal of Foundations of Computer Science*. **14** (6): 957–982. doi:10.1142/S0129054103002114. Retrieved 2011-03-28
13. Sadaoki Furui, November 2005, 50 years of Progress in speech and Speaker Recognition Research , ECTI Transactions on Computer and Information Technology,Vol.1. No.2.
14. K.H.Davis, R.Biddulph, and S.Balashek, 1952, Automatic Recognition of spoken Digits, J.Acoust.Soc.Am.,24(6):637-642.
15. H.F.Olson and H.Belar, 1956, Phonetic Typewriter , J.Acoust.Soc.Am.,28(6):1072-1081.
16. J.W.Forgie and C.D.Forgie, 1959, Results obtained from a vowel recognition computer program , J.Acoust.Soc.Am., 31(11),pp.1480-1489.
17. D.B.Fry, 1959, Theoritical Aspects of Mechanical speech Recognition , and P.Denes, The design and Operation of the Mechanical Speech Recognizer at Universtiy College London, J.British Inst. Radio Engr., 19:4,211-299.
18. K.Nagata, Y.Kato, and S.Chiba, 1963, Spoken Digit Recognizer for Japanese Language , NEC Res.Develop., No.6.
19. T.Sakai and S.Doshita, 1962 The phonetic typewriter, information processing 1962 , Proc.IFIP Congress.
20. L.R.Rabiner, S.E.Levinson, A.E.Rosenberg, and J.G.Wilpon, August 1979, Speaker Independent Recognition of Isolated Words Using Clustering Techniques , IEEE Trans. Acoustics, Speech, Signal Proc., ASSP-27:336-349.
21. B.Lowrre, 1990, The HARPY speech understanding system ,Trends in Speech Recognition, W.Lea,Ed., Speech Science Pub., pp.576-586.
22. R.K.Moore, 1994, Twenty things we still don t know about speech , Proc. CRIM/ FORWISS Workshop on Progress and Prospects of speech Research an Technology.
23. J.Ferguson, 1980, Hidden Markov Models for Speech, IDA,Princeton, NJ.

24. L.R.Rabiner, February 1989, A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition , Proc.IEEE,77(2):257-286.
25. B.H.Juang and S.Furui, 2000, Automatic speech recognition and understanding: A first step toward natural human machine communication , Proc.IEEE,88,8,pp.1142-1165.
26. K.P.Li and G.W.Hughes, 1974, Talker differences as they appear in correlation matrices of continuous speech spectra , J.Acoust.Soc.Am. , 55,pp.833-837.
27. Ananth Sankar, May 1996, " A maximum likelihood approach to stochastic matching for Robust Speech recognition", IEEE Transactions on Audio, Speech and Language processing Vol.4,No.3.
28. Gerhard Rigoll, Jan.1994, "Maximum Mutual Information Neural Networks for Hybrid connectionistHMM speech Recognition Systems ", IEEE Transactions on Audio, Speech and Language processing Vol.2,No.1, PartII.