Isolated Word Recognition In Marathi

Prof.S.A.Paithane
Misba Hudewale, Sukhbir Gill, Sumit Ahirwar
Department of Electronics & Telecommunication
JSPM'S RSCOE, Savitribai Phule Pune University
Pune, Maharashtra, India
mizzy240@gmail.com

***Abstract*— The paper presented here is based on recognition of isolated Marathi words which consists of the numbers and the alphabets. It is a part of Automatic Speech Recognition called as ASR, that is, a technology which uses computer to identify a spoken word of a person through a microphone and converts it to text. The ultimate goal of ASR is to allow computer to recognize in real time, plus the 100% accuracy, the words that are spoken by any person, independent of vocabulary size, the speaker's features, noise or accent. Speech Recognition is the technology which was initially designed for the individuals in disability community. Aim of this paper has a heavy focus on creating proper interfaced and adaptable speech recognition system for Marathi language. Most of present ASRs are made for English language. Most of them do require a lot of changes before they can be used. In the rural areas across India there are people who don't understand English and they are unable to speak proper English. So this available ASR system is of no use for the rural people. There are many regional languages in India, but being Maharashtrian, we have got inspired to think about Speech Recognition system for Marathi language.**

***Keywords*- recognition, isolated, Marathi, ASR**

I. INTRODUCTION

Speech is the most common route for communication. It is a random, natural and convolved signal. It is generated by the human speech production system. When a person speaks, air from lungs is forced through glottis passing from trachea to larynx. The most important part in speech production is the vocal tract which connects larynx to lips. It acts as an acoustic filter. The pitch which is controlled by excitation source due to vocal folds in larynx is convolved with response of vocal tract. The speech is perceived due to change in pressure of air molecules which compress and expand and leads to recognized speech.ASR has greatly influenced in 1990's by automating call handling functions, reducing the operating cost of a call centre. In 1992, AT&T Bell Laboratories in New Jersey, USA had devised Voice Recognition Call Processing Service handles 1.2 billion voice transactions with machines each year under the influence of ASR. In 1988, Apple Inc. created a vision of speech technology & computers for year 2011 called as "Knowledge Navigator", which uses concept of Speech User Interface and a Multimodal User Interface along with smart voice-enabled agents.

During first half of 20th century, Homer Dudley devised Voice Operating Demonstrator using the perception and reception related to the characteristics of a human speech. However, during this time period, Tom Martin, made the use of Linear Predictive Coders his first own speech recognition commercial company called as Threshold Technology. This system was used by FedEx for package sorting on a conveyor belt. It greatly influenced Advanced Research Projects Agency of the U.S. Department to create the fund for Speech Understanding Research program. With this great research, evaluation of speech technology has extended for wide range of vocabularies pursued in twenty-first century. As technology matures, there are new upcoming challenges to develop an ASR[4] which is robust and can recognize speech very efficiently as perceived by human beings.

II. MARATHI LANGUAGE

Since English-speaking population in India is quite low, Marathi is the widely spoken language in the state of Maharashtra, it is an Indo-Aryan language spoken in western and central India. The fluency of Marathi speakers all over world enhances speech recognition. The system would be developed as speaker dependent, isolated word recognizer. Marathi language makes use of Devanagari- a character based script. All in total, Marathi language has 12 vowels and 36 consonants. It has been influenced by its neighbouring Southern states. The alphabets are, however, the phoneme set which is also called as "Barakhadi". A phoneme is a smallest part of speech. Even the numbers, also contribute to phoneme, since the pronunciation of the numbers also make use of Marathi alphabets. The script currently in use for Marathi is called as "Balbodh", also a modified version of Devanagari script. This script is easier to read but it does not have advantage of faster writing. Marathi is also said to be descendant of Maharashtri which was Prakrit spoken by people, residing in Maharashtrian region. Earlier, "Modi", script was in use till the time of Peshawas in eighteenth century. This was developed by Hemadpanta, who was minister of Yadava kings in Devgiri in thirteenth century. All Indo-Aryan languages originate from Sanskrit. Population of Marathi speakers in world is approximately 72 million.

III.SPEECH RECOGNITION

Before proceeding to speaker recognition [7], there are factors that affect speech signal and these are: speaker gender, speaker identity and speaker language, psychological conditions, speaking style and environmental conditions. There are two speaker models[1] in speech recognition:

1) *Speaker Dependent:* Recognizes speech patterns from only one person. This is the model which is chosen for our project.

2) *Speaker Independent:* Recognizes speech patterns from many speakers.

Now, there are two more classifications in speech recognition, they are:

1) Isolated Word Recognition: Simplest speech type because it requires the user to pause between each word.

2) Connected Word Recognition: Is capable of analyzing a string of words spoken together, but not at normal speech rate.

Certain difficulties arise while dealing with ASR:

1) Digitization: Converting analogue signal into digital representation.

2) Signal processing: Separating speech from background noise.

3) Phonetics: Variability in human speech.

4) Continuity: Natural speech is continuous.

IV. METHODOLOGY

This method to be implemented here for isolated word recognition is described in the block diagram below-

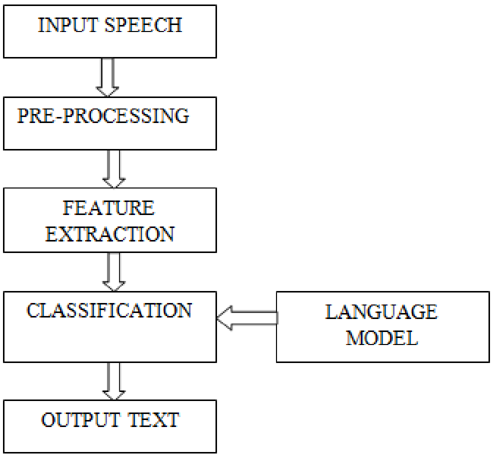

Fig. 1 Block Diagram of IWR

A)Input Speech: Here, the speech will be given via microphone connected to PC which is going to be an isolated word which has to be recognized and displayed on PC.

*B) Pre-*Processing: It consists of pre-emphasis, reducing the noise. This acoustic speech need to be digitally converted for further processing. The sampling rate of this should be such that it avoids aliasing effect. During speech production, lower frequencies are boosted while higher ones are suppressed, which cause loss information in signal. Hence, high pass FIR filter is

C) Feature Extraction: There are various techniques to extract features from the spoken speech like MFCC, PLP (Perceptual Linear Prediction), Rasta-PLP, etc. but we use MFCC (Mel Frequency Cepstral Coefficients. This technique works on Mel-Scale, which approximates human auditory system response. It is a representation of short term power spectrum in speech. Cepstral means non-linear spectrum of spectrum. There are steps to calculate MFCC-

1) Take Fourier Transform of windowed speech signal.
2) Map the values onto Mel Scale using triangular windows.
3) Take log of values at each Mel frequency.
4) Take Discrete Cosine Transform of all Mel values.
5) Hence, MFCC coefficients are obtained.

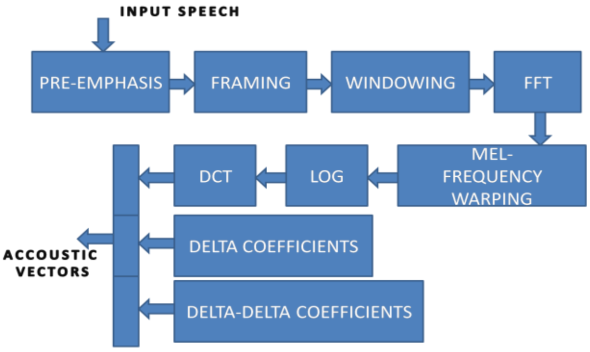

Fig.2 MFCC Block Diagram

The steps to obtain acoustic vectors in MFCC are-

1)Framing: Speech signal is divided in frames of 10 to 20ms since the variation in speech signal is so fast, we can take frame of smaller size and extract features.

2) Windowing: Each frame is multiplied to window function to minimize discontinuities. Hence hamming window is compatible with Mel Scale. The function is given as-

$$w[n] = 0.54 - 0.46(1 - \cos(\frac{2\pi n}{N-1})); 0 \le n \le N-1$$

3) Discrete Fourier Transform: The DFT is applied to each frame but it has more computations. Hence, FFT can speed up computations. Equation of DFT is given by-

$$X(k) = \sum_{n=0}^{N-1} x(n)\, e^{\frac{-i2\pi nk}{N}}, k = 0 \ldots N-1$$

4) Mel Frequency Warping: Based on Mel scale, this equation is devised on basis of perception of Frequencies by human ear. This scale is approximately linear below 1kHz and logarithmic above 1kHz. It is described by-

$$m = 2595 \log_{10}\left(1 + \frac{f}{700}\right) = 1127 \log_e\left(1 + \frac{f}{700}\right)$$

desirable since it preserves features of speech signal.

Where m denotes perceived frequency and f is actual one.

5) Log compression and DCT: The filtered outputs are evaluated with logarithmic process and apply Discrete Cosine Transform to it giving rise to MFCC coefficients. Equation is given by-

$$\sum_{i=1}^{M}\log(Y(i))*\cos((\Pi n/M)*(i-0.5))$$

6) Calculation of delta and its coefficients: These are used to add time evolution information. First order derivative is delta coefficient and second order derivative is delta-delta coefficient. The nth delta is calculated by formula-

$$\Delta f_k(n) = f_{k+M}(n) - f_{k-M}(n)$$

For nth delta-delta coefficient is calculated by-

$$\Delta^2 f_k(n) = f_{k+M}(n) - f_{k-M}(n)$$

M is 2-3 frames.

D) Classification: This is nothing but feature matching method. There are methods like Hidden Markov Models, Neural Networks and DTW. When compared, Dynamic Time Warping (DTW) is a good choice for word spotting and also for short duration phases. This block performs comparison of input sample with stored sample in database. So we use Dynamic Time Warping where it aligns test and stored samples to give average distance. It finds optimal match and displays text which was formally input speech. It is a non linear and efficient technique. DTW works well with small amount of data.

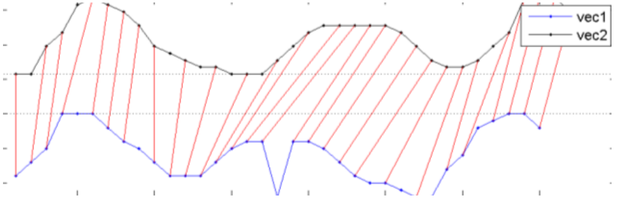

Fig.3 Alignment of input speech and stored sample for an optimal match

Vector 1 is input and Vector 2 is stored sample.

E) Language Model: It is the database for our implementation which has stored voice samples which is most significant block in our project.

F) Output Text: On PC, the text is outputted on screen. This will be nothing but the spoken word given as input from

V. CONCLUSIONS

This paper brings the idea of how speech can be recognized using several transforms and many mathematical computations which are required to bring the output into text format. This implementation will be a boon for disabled people as well as for illiterate people an important means of education. This system can further be made speaker independent, robust and for continuous speech recognition. Most importantly, we have MFCC which extracts most important phonetic features and behaves as an approximate model of human auditory system. MFCC, in fact, has low robustness to noise. Hence, this pre-processing stage makes signal robust and makes it capable for the next stage for further processing. DTW is best option since it is also a de-noising technique. It is capable of compressing signal without degrading the signal information. This can be also implemented in vehicles to follow suitable commands and make appropriate movement of vehicle. Voice based user interface for making call on mobile phone, entering account no. in ATM during transactions. As speech is a signal which is more noise-immune, we can make further modifications like creating a trained database, making use of proper filters, making use of best microphones, etc. This implementation can be carried out for various languages with known phoneme set, fluent speakers and a robust algorithm to create an efficient ASR system.

VI. REFERENCES

[1] *Digital Speech And Audio Processing,* Dr. Shaila D.Apte.

[2]L. Rabiner, B.H. Juang, *Fundamentals Of Speech Recognition,* Prentice Hall, New Jersey, USA.

[3]Juang. B.H,L. Rabiner, Elsevier Encyclopedia Of Language and Linguistics, 2005, Second Edition, *Automatic Speech Recognition-A Brief History Of Technology.*

[4]Huang.X, Acero.A, Prentice Hall, *Spoken Language Processing-A Guide to theory, algorithm and system development,* 2001.

[5]Rabiner. L, Schafer. R, 1978, Digital *Processing Of Speech Signals,* Prentice Hall, New Jersey.

[6]B.Gawali, S. Gaikwad, 2010, Journal Of Computer Applications, Marathi Isolated Word Recognition System using MFCC and DTW features.

[7]Paithane.A.N, D.S.Bormane and Sneha Dinde,"*Human Emotion Recognition using Electrocardiogram signals".*Vol.2, 2014, pp.194-197, ISSN: 2321-8169.

microphone.