Scientific Journal of Impact Factor(SJIF): 3.134
e-ISSN(O): 2348-4470
p-ISSN(P): 2348-6406

International Journal of Advance Engineering and Research Development

Volume 2,Issue 5, May -2015

Improve accuracy of Speech Recognition for different Indian Accents using MFCC, LPC, Zero Crossing and Power Spectrum

Pahini A. Trivedi[1], Sagar H. Virani[2]

[1]*Computer Engineering Department, V.V.P. Engineering College-Rajkot, Gujarat Technological University,pahinitrivedi21@gmail.com*
[2]*Ass. Prof. Computer Engineering Department,V.V.P. Engineering College-Rajkot,sagarvirani@gmail.com*

Abstract-*Nowadays, smart phones and its novel applications are widely used by different users. Among those applications, speech recognition has been the most fascinating and useful application. In smart phones user gives speech/spoken words as input, command given is interpreted and relevant task is carried out. The recognition of the speech for different living in different countries is a challenging job. Observations show that most smart phones provide more accurate and efficient results to American users rather than users from other part of the world. When these smart phones are used in foreign countries like India, Japan etc. it gives less accurate results. This is due to the fact that the speech recognition system used in smart phones uses majority of data from American origin in order to training it. According to survey more and more people are using smart phones in India in recent times.The work undertaken uses an accent base approach to improve the accuracy. Speech recognition system is developed for each accent and users can acquire them based on their origin. For training purpose 3 accents are considered – Gujarati,Bengali and Malayalam.* ***Speech Recognition*** *Features used in this are MFCC, LPC, Zero Crossing and Power Spectrum.SVM is used for classification purpose.*

Keywords- MFCC, LPC, Zero Crossing, Power Spectrum. Support Vector Machine (SVM)

I. INTRODUCTION

In computer science and electrical engineering, **speech recognition** (SR) is the translation of spoken words into text. It is also known as "automatic speech recognition" (ASR), "computer speech recognition", or just "speech to text" (STT).Some SR systems use "speaker-independent speech recognition" while others use "training" where an individual speaker reads sections of text into the SR system. These systems analyse the person's specific voice and use it to fine-tune the recognition of that person's speech, resulting in more accurate transcription. Systems that do not use training are called "speaker-independent" systems. Systems that use training are called "speaker-dependent" systems [3].

1.1 History of Speech Recognition

The first speech recognizer appeared in 1952 and consisted of a device for the recognition of single spoken digits [1] another early device was the IBM Shoebox, exhibited at the 1964 New York World's Fair. Lately there have been numerous improvements like a high speed mass transcription capability on a single system like Sonic Extractor [2] One of the most notable domains for the commercial application of speech recognition in the United States has been health care and in particular the work of the medical transcriptionist (MT).

According to industry experts, at its inception, speech recognition (SR) was sold as a way to completely eliminate transcription rather than make the transcription process more efficient, hence it was not accepted. It was also the case that SR at that time was often technically deficient.

A distinction in ASR is often made between "artificial syntax systems" which are usually domain-specific and "natural language processing" which is usually language-specific. Each of these types of application presents its own particular goals and challenges [4].

1.2 Structure of a standard speech recognition system

The structure of a standard speech recognition system is illustrated in Figure 1. The elements are as follows:
• Raw speech. Speech is typically sampled at a high frequency, e.g., 16 KHz over a
Microphone or 8 KHz over a telephone. This yields a sequence of amplitude values over time.

• Signal analysis**.** Raw speech should be initially transformed and compressed, in order to simplify subsequent processing.

Many signal analysis techniques are available which can extract useful features and compress the data by a factor of ten without losing any important information. Among the most popular:

➢ Fourier analysis (FFT) yields discrete frequencies over time, which can be interpreted visually. Frequencies are often distributed using a *Mel* scale, which is linear in the low range but logarithmic in the high range, corresponding to physiological characteristics of the human ear.

➢ Perceptual Linear Prediction (PLP) is also physiologically motivated, but yields coefficients that cannot be interpreted visually.

• Linear Predictive Coding (LPC) yields coefficients of a linear equation that approximate the recent history of the raw speech values.

• Cepstral analysiscalculates the inverse Fourier transform of the logarithm of the power spectrum of the signal.

In practice, it makes little difference which technique is used. Afterwards, procedures such as Linear Discriminant Analysis (LDA) may optionally be applied to further reduce the dimensionality of any representation, and to decor relate the coefficients [4].

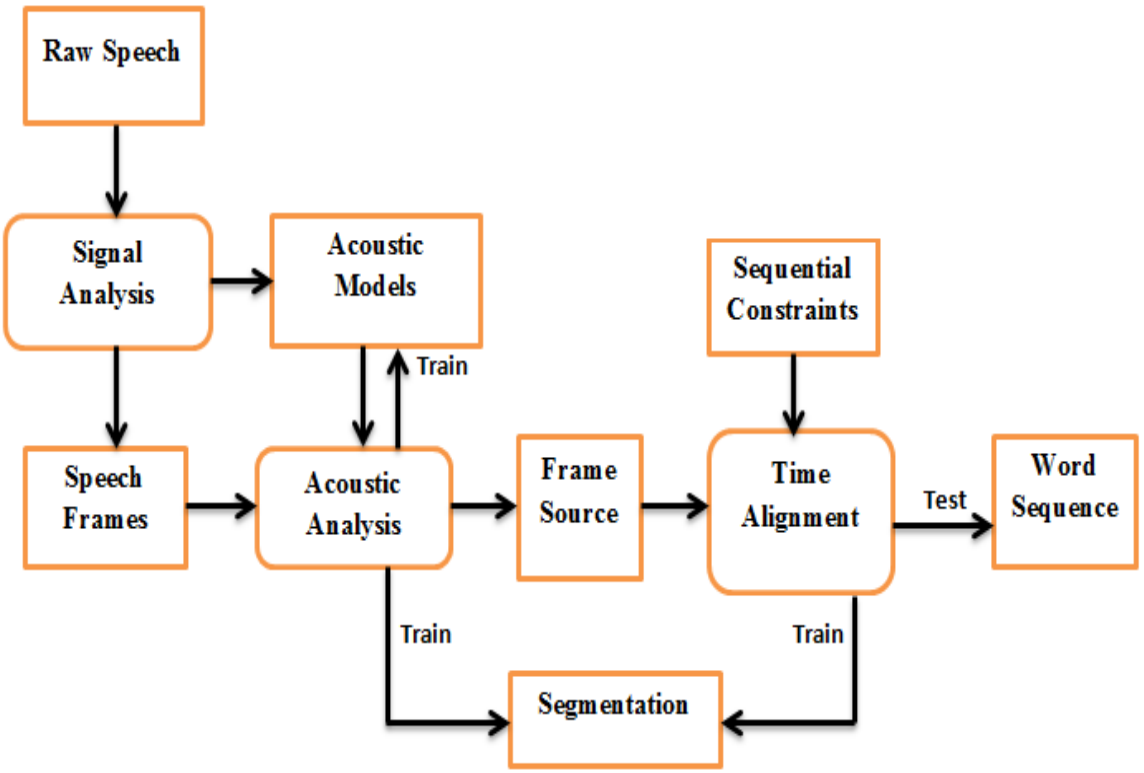

Figure 1: Structure of a standard speech recognition system[4]

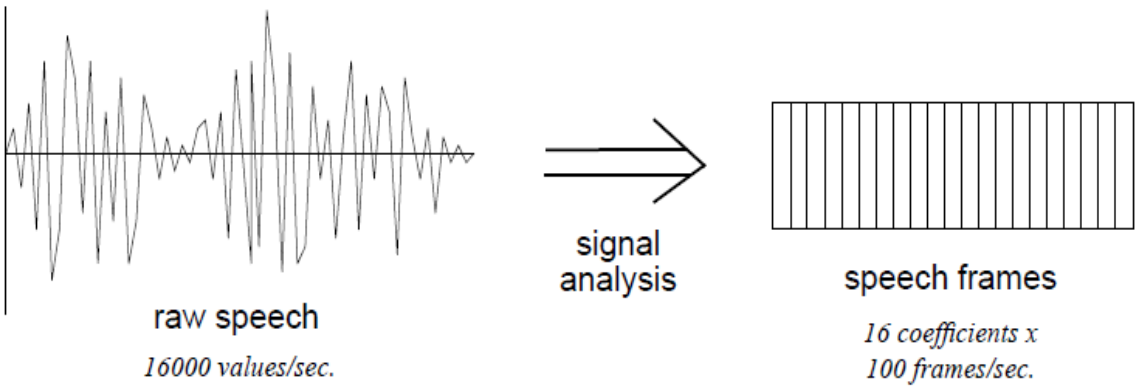

Figure 2: Signal analysis converts raw speech to speech frames [4]

II. PROBLEM IN EXISTING SYSTEM

The problem in existing System (Particular in Smart Phone) is that there is only one SR system for different accents and trained through multi accents. So, it is not identifying single accent properly, that's why we get less accurate result. Smart phone support American English well rather than other accent (e.g. British or Indian English).So when this phone used in America it gives more accuracy compare to the same phone used in India, Japan or any other foreign country.

The main problem occurs when users want to use smart phone in their own accent. For Example, if there is a Gujarati user and he/she want to know the information about town, then he/she gives spoken words as an input and it will return information about spoken words into form of frequencies, this is a main purpose of Speech Recognition. As Speech Recognition is a conversion of Speech to Text.

Now come to the problem, when Guajarati user speak any particular word as he/she want information about it, Smart phone misinterpreted the word and give different result.

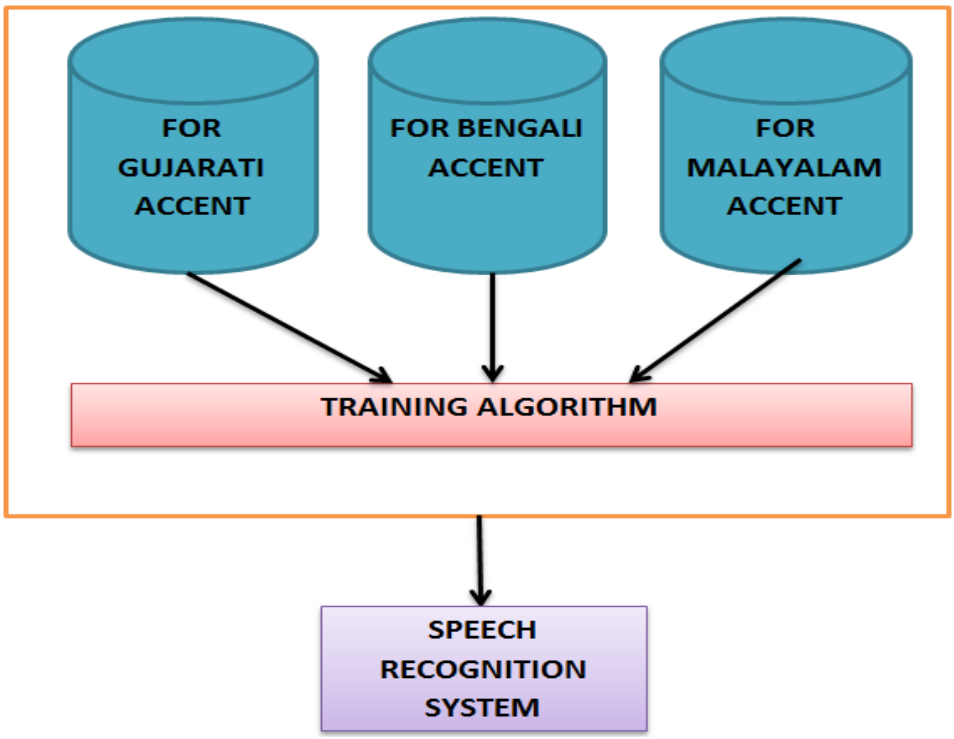

Figure 3: Problem in Existing System

This is one of the major problem user faced due to the fact that the algorithm uses the training set which contains majority of data form American accents. So when user speaks any type of English, No matter whether it is Indian/American/British English, all are trained using single training algorithm.

2.1 Motivation

Now-a-days as new technologies are introducing in mobile phone, People want more accurate result, which is really useful in means of Speech Recognition.

Speech recognition is a challenging problem of smart phone. Perhaps in smart phone SR is an exciting feature. Smart phones provide more accurate and efficient output for American Accents rather than other accents. So when this smart phone used in America it gives more accuracy compare to the same phone used in India, Japan or any other foreign country. This is due to the fact that the algorithm uses the training set which contains majority of data form American accents. Many approaches used for targeting this problem yet there is only 50-60% accuracy achieved, comparatively it is very low.

Prime goal of thesis is to increase accuracy of speech recognition in smart phone. To improve accuracy of Speech Recognition by Proposed System Architecture. Novel Approach is given which improve accuracy of Speech Recognition by Proposed System Architecture. There is different SR system for different accents (Gujarati, Bengali, and Malayalam) trained within training algorithm. These all things deploy within one server. So when users want to use smart phone in their own accent they can easily select it by two scenarios.

(i) Scenario-1 : Automatic request for SRS through GPS
(ii) Scenario-2 : Manually request for SRS through GPS

Above mentioned scenarios gives effective output which increases accuracy of speech recognition.

2.2 Objectives

The major objectives of this dissertation work could be summed up as follows:

- To improve accuracy of Speech Recognition by Proposed System Architecture.
- Generate Training–set of different accents.
- Develop Server for Proposed System Architecture.
- Develop mobile interface which communicate with server.
- To study features of Speech Recognition.
- Combine derived features to improve accuracy.
- Select proper classifier for spoken word recognition.

III. PROPOSED SYSTEM ARCHITECTURE

In proposed system architecture each database for Gujarati, Bengali and Malayalam accent having their own speech recognition system.

As shown in above figure of proposed system architecture is given. There are different SR systems for different accents (Gujarati, Bengali, and Malayalam) trained within training algorithm. These all things deploy within one server. Suppose Gujarati user uses the smart phone then he/she can do automatic request for their own accent through GPS that request placed in database.

In database speech is given as an input and frequency of a corresponding word is given as an output. In ANN spoken words frequencies are between 0-1.For examples, “Hello” word spoken by user and it stored at 0.5 frequencies in database. User by user accent is change little bit so when user speaks hello, suppose frequency is achieved 0.489, which is nearer to 0.5, database give output of “Hello” as training algorithm.

User can automatic as well as manually request for SRS through GPS.

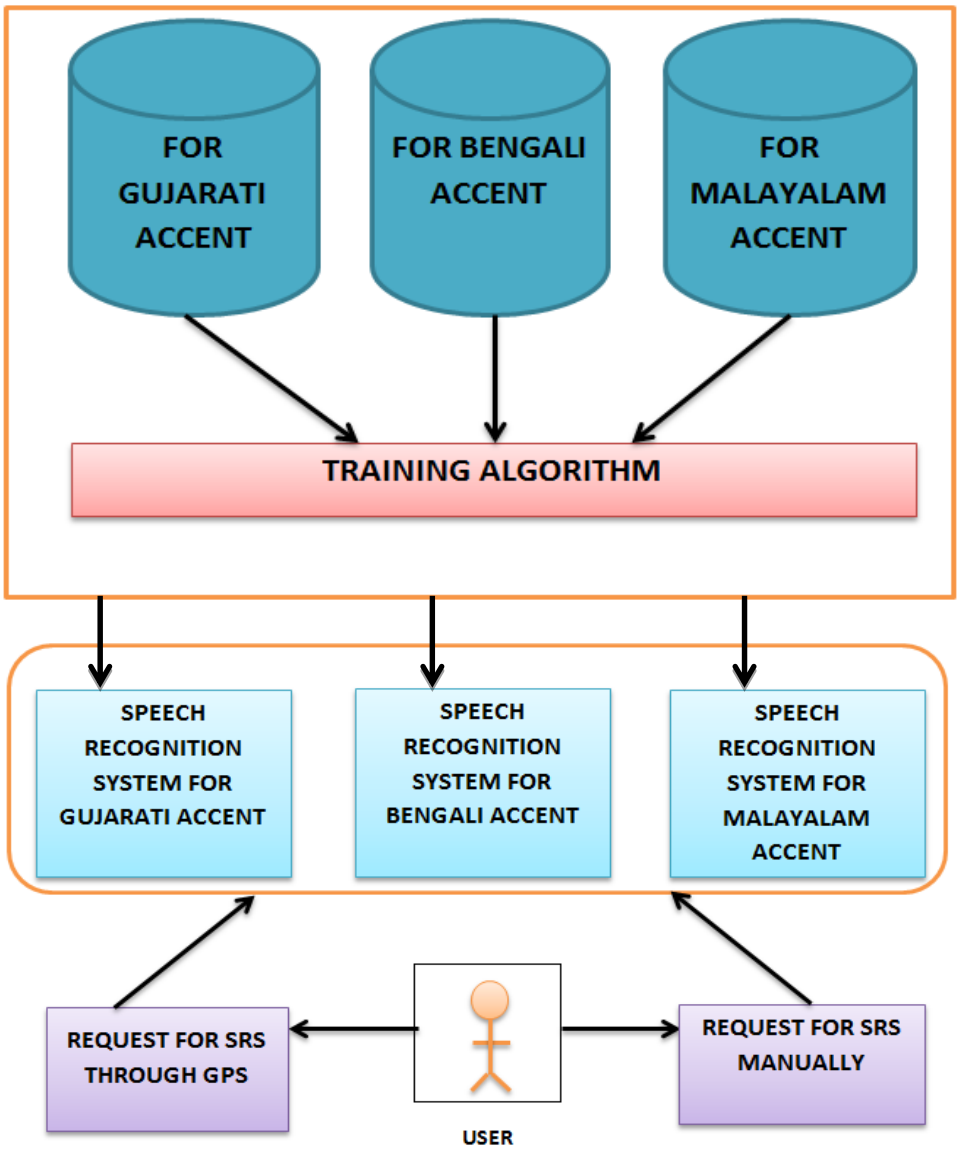

Figure 4: Proposed System Architecture

IV. IMPORTANT FEATURES OF SPEECH RECOGNITION

Features of SRS which are important for further implementation are as follows:

- MFCC
- LPC
- Zero Crossing
- Power Spectrum

1) MFCC

This is a representation of the short-term power spectrum of a sound.MFCCs are the amplitudes of the resulting spectrum and the cepstral representation of the audio clip [5][7].

2) LPC

LPC provides an accurate estimate of the speech parameters. Speech sample can be approximated as a linear combination of past speech samples in LPC. It is performed to provide observation vectors of speech in form of frequency [6].

3) Zero Crossing

This is a good measure of the pitch as well as the noisiness of a signal. Zero crossings are calculated by finding the number of times the signal changes sign from one sample to another (or touches the zero axis).Counting zero-crossing is a method used in speech processing to estimate the fundamental frequency of speech [5] [8].

4) Power Spectrum

This is a good measure of the power of different frequency components within a window. The power spectrum is found by first calculating the FFT with a Hamming window. It is a part of signal processing. Power spectral analysis consist of Perceptual Linear Predictive (PLP) Coefficients and Relative spectra filtering of log domain coefficients (RASTA).[5][6].

V. TOOLS

Implementation of this will be done in malab and android base application run on java .

- Mat lab R2011a
- jAudio 1.0.4.jar
- Total Video Converter 3-71 win.exe

In matlab there are some tools that can use for speech recognition.

1) ASR (Automatic Speech Recognition) Toolbox [11]

This toolbox provides functions for ASR (Automatic Speech Recognition) based on HTK (Hidden Markov Model Toolkit.Before using the toolbox, following toolboxes are needed:

- Utility Toolbox
- SAP Toolbox

2) VOICEBOX: Speech Processing Toolbox for MATLAB [10]

VOICEBOX is a speech processing toolbox consists of MATLAB routines. The routines are available as a zip archive and are made available under the terms of the GNU Public License. The routine VOICEBOX.M contains various installation-dependent parameters which may need to be altered before using the toolbox.

In particular it contains a number of default directory paths indicating where temporary files should be created, where speech data normally resides, etc.

3) HDecode - HTK Speech Recognition Toolkit [9]

The HMM-based Speech Synthesis System (HTS) for HMM-based synthesis. This toolkit is released as a patch code to HTK. Modifications needed to HTK are listed below:

- Context clustering based on MDL criterion
- Stream-dependent context clustering
- Multi-space probability distribution as state output probability
- State duration modelling and clustering
- Speech parameter generation from continuous density HMMs

VI. CONCLUSION AND ROADMAP TO FUTURE WORK

Among all features of Speech Recognition -MFCC, LPC, Zero Crossing and Power Spectrum are useful features to be implemented. So first that all features will be implement. Combine that features to improve accuracy. At last select proper classifier for spoken word recognition that will improve accuracy of Speech Recognition.

REFERENCES

[1] George E. Dahl, Dong Yu, Senior Member, IEEE, Li Deng, Fellow, IEEE, and Alex Acero, Fellow, IEEE,"Context-Dependent Pre-Trained Deep Neural Networks for Large-Vocabulary Speech Recognition"IEEE TRANSACTIONS ON AUDIO, SPEECH, AND LANGUAGE PROCESSING,VOL. 20, NO. 1, JANUARY 2012.

[2] G. Zweig and P. Nguyen,"A segmental CRF approach to large vocabulary continuous speech recognition", in Proc. ASRU, 2009, pp.152–155.

[3] http://en.wikipedia.org/wiki/Speech_recognition

[4] Joe Tebelskis,"Speech Recognition using Neural Networks",May 1995, CMU-CS-95-142, School of Computer Science, Carnegie Mellon University Pittsburgh, Pennsylvania 15213-3890.

[5]Mrs G.M.Bhandari1, Dr. R.S.Kawitkar2," Audio Segmentation for Speech Recognition Using Segment Features ", Research scholar of JJTU , Rajasthan , India ,Department of Electronics Engineering, IPASJ International Journal of Computer Science (IIJCS), Volume 2, Issue 3, March 2014.

[6]Vimala.C , Radha.V ," Suitable Feature Extraction and Speech Recognition Technique for Isolated Tamil Spoken Words" , Department of Computer Science,Avinashilingam Institute for Home Science and Higher Education for Women,Coimbatore, Tamil Nadu, India, (IJCSIT) International Journal of Computer Science and Information Technologies, Vol. 5 (1) , 2014, 378-383.

[7]"MFCC", http://en.wikipedia.org/wiki/Mel frequency_cepstrum

[8]"Zero-Crossing", http://en.wikipedia.org/wiki/Zero_crossing

[9] http://htk.eng.cam.ac.uk/extensions

[10] http://www.ee.ic.ac.uk/hp/staff/dmb/voicebox/voicebox.html

[11] http://mirlab.org/jang/matlab/toolbox/ASR

