IOSR Journal of Electronics and Communication Engineering (IOSR-JECE)
e-ISSN: 2278-2834,p- ISSN: 2278-8735.Volume 12, Issue 4, Ver. III (Jul.-Aug. 2017), PP 62-68
www.iosrjournals.org

Automatic Speech Recognition system for class room database management in Fixed – C Language

*MounikaJammula
Assistant Professor, Chaitanya Bharathi Institute of Technology, Hyderabad.
*Corresponding Author: *MounikaJammula*

Abstract: *Speech is a natural mode of communication for people. People are more comfortable with speech to interact with computers than using of interfaces such as keyboards and pointing devices. This type of speech interfaces finds applications in many areas like telephone directory assistance, hand busy applications in medicine or field work or even automatic voice translation into foreign languages. The speech recognition involves many real time challenges. These challenges arise as the system may need to deal with different people with different accents and dialects. Moreover a person does not speak the same word alike twice and the biggest challenge of all is that the speech recognition system should be capable of working with continuous speech. There shall also be the concern of invasion of channel noise. In the current paper, the automatic speech recognition system is implemented for Isolated words, which in turn can be made part of any standalone application like class room data base management. The training and the implementation is carried out for speaker independent isolated word speech recognition. Furthermore this implementation is intended to be extended towards the usage in real time in the Digital signal processors. Once the ASR is realized on DSP's they can be brought into open market.*

Index terms: *ASR, Mel scale, MFCC, DTW, Baum-Welch re-estimation.*

Date of Submission: 25-05-2017 Date of acceptance: 10-08-2017

I. Introduction

Automatic Speech Recognition (ASR) is the process of converting an acoustic signal, captured by a microphone or a telephone, to a set of words. The recognised words can be used for applications such as commands & control, data entry, and document preparation. They can also serve as the input to further linguistic processing in order to achieve speech understanding. Speech recognition is different from speaker verification and speaker identification. Speech recognition is detecting what a person says, where as speaker verification authenticates a person as who she/he claims to be and speaker identification assigns an identity to the voice of an unknown person.

The steps involved in ASR systems are speech data acquisition, feature extraction, training the features of the data for obtaining the acoustic models for words or phonemes, language modeling and finally recognition of the speech with developed models.

ASR finds huge number of applications in diversified fields, such as Healthcare, Training Air Traffic Controllers, Hands-Free Computing, but not just limited to these areas and the possibilities are enormous.

Speech recognition is generally considered advantageous for users with limited mobility, such as those that are partially paralyzed. It allows them to type documents and use commands at nearly the same speed as those without this particular type of disadvantage. It is also useful for the general public, particularly for those that are less computer literate or those that are not able to type quickly (such as those that cannot "touch type").

A. Challenges in Speech Recognition

Humans use more than their ears while listening and they utilize the knowledge they have about the speaker and the subject. Words are not arbitrarily sequenced together, there is a grammatical structure and redundancy that humans use to predictwords not yet spoken. Furthermore, regular phrases that are used usually that is, the sentences more often expressed make prediction even easier. In ASR there is only speech signal. In recent times, model for the grammatical structure and also some kind of statistical model to improve prediction are being constructed, but there is still the problem prevailing in terms of modeling i.e. world knowledge, the knowledge of the speaker. This is encyclopedic in nature and still beyond the understanding. The main challenge in here may not be to model world knowledge exhaustively, but as how much ASR needs can to scale nearer to human comprehension.

Outstanding work in speech recognition and computing has produced the commercial ASR systems for voice-driven computing and word-processing systems in English and European languages. Though significant

research work is being carried out in Indian languages too, like Hindi, Tamil, Bengali and Telugu, ASR systems are not yet launched into the Indian market at full level. In such a scenario, we designed a simple isolated word speech recognition system which can used as a part of class room data base management. There will be 50-100 students in a class. Marks and attendance of every student and subject should be entered manually by the class teacher. This is laborious and time consuming one. If there is a method to enter the data through voice instead of using keyboard for typing, the process will be efficient and fast.

The main aim of this paper is to develop, train and practically implement speaker independent isolated word speech recognition in C-language by taking the ASR system developed already in MATLAB. MATLAB needs a license and very costly affair to use it by every teacher. If it is developed in C, we can use it without any license fee.Further, it was converted to fixed point C to be more relevant when tryingto implement on DSP's.

The next sections in this paper are organized as follows: section 2 gives an overview of ASR and also includes discussion about its different steps, section 3 discusses about dynamic time warping algorithm and its implementation in C, Section 4 discusses the modeling process for isolated word recognition and finally section 5 deals with testing process, results and analysis in detail.

II. Overview of Implementation of ASR

In an isolated word speech recognition system, there are two phases: training and recognition. During the training phase, a training vector is generated for each word spoken by the user. The training vectors extract the spectral features for separating different classes of words. Each training vector can serve as a template for a single word or a word class. These training vectors (patterns) are stored in a database for subsequent use in the recognition phase [2](Talal Bin Amin*et al*., and Iftekhar Mahmood, 2008).

During the recognition phase, the user speaks any word for which the system was trained. A test pattern is generated for that word and the corresponding text string is displayed as the output using a pattern comparison technique. The two key issues limiting the performance of any speech recognition system are:

- How well the system was trained
- How to model the statistical variations of different variants of the same word.

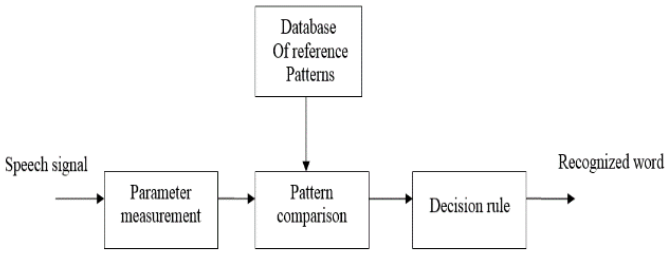

Figure 1: Block diagram of an ASR System

The Automatic speech recognition system as seen in the figure 1 includes two main stages. The first stage comprises the feature extraction and also storage of these extracted features as training data, called the Models. The second stage is testing, where the features of the new speech input are extracted and these features are used in order to match with stored features thus to recognize the word.

MFCC algorithm is used for feature extraction and dynamic time warping algorithm is used to reduce amount of achieved data in form of models. These data are saved as acoustic vectors. In the matching stage, features of speech input are compared with each model using the Euclidean distance, Mahalanobis distance or Normal distribution. The speech recognition system may be viewed as working in four stages,

a) Speech data acquisition
b) Feature extraction
c) Training the system to build the reference models
d) Testing the system with the built in models (pattern comparison)

A. Data Acquisition:

Speech data acquisition is a process of collecting real time speech data in the form of sentences (with silence between words) pronounced by various speakers. The acoustic waveform is converted into electrical waveform by a microphone. This analog waveform should be converted to digital form using ADC. The speech analysis stage involves framing which is used for segmenting speech signal where the converter samples the speech signal at the rate of 16 kHz and stores the speech for further digital processing.

B. Feature Extraction

The aim of the feature extraction process is to translate the information contained in acoustic signals into a data representation that is suitable for statistical modeling and likelihood calculation.There are different methods that can be applied for parametrically representing the speech signal in the speech recognition task, such as

- Linear Prediction Coding Coefficients (LPCC)
- Perceptual Linear Prediction Coefficients(PLPC)
- Mel Frequency Cepstral Coefficients (MFCC)

We chose MFCC for the project because of its advantages. The mel scale takes the human perception of sound frequency into account and is based on the subjective pitch of pure tones. Thus working better than its counter parts.

C. C. Isolated Word Recognition Models

During training, set of typical speech sounds are obtained from large corpora of training data. However, there are numerous factors that cause variation in the acoustic properties of typical speech sounds, which turn to be great hindrances in ASR. Hence for speech sounds to be recognized correctly, it is therefore necessary to model their typical acoustic characteristics (average value) as well as the variations observed in these typical properties (variance). There are two major types of models for classification: stochastic models (parametric) and template models (non-parametric). We chose template models, as they are the simplest ones. Moreover, it includes, Dynamic Time Warping (DTW) for model buildingwhich is appropriate for text dependent recognition.

D. D.Recognition and Testing:

The simplest way to recognize an isolated word sample is to compare it against a number of stored templates and determining the best match. This can be done by optimal alignment between templates and input speech, which is performed on employing the Dynamic Time Warping algorithm. Simply speaking, in the speech recognition technique the data is converted to templates and the incoming speech is matched with these stored templates. The template with the lowest distance measure from the input pattern is the recognized word. The best match (lowest distance measure) is based upon dynamic programming. This is called a Dynamic Time Warping (DTW) word recognizer.

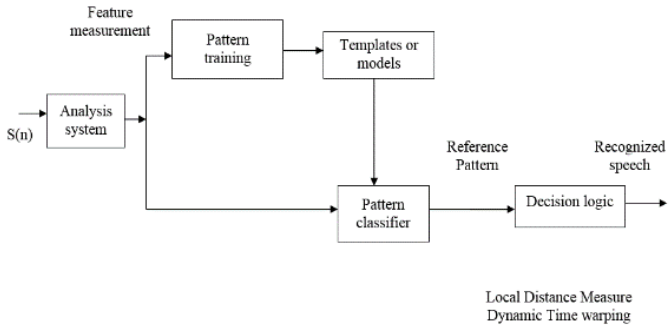

Figure 2: Pattern matching shown with respect to the DTW algorithm

Thus recognition phase involves the comparison of existing trained acoustic models with the processed voice input. A decoder (Viterbi algorithm, Baum-Welch) is used to match the voice input with the most likely acoustic models as the path is made through the network. The decoder transcribes the continuous speech input into a sequence of textual symbols, which an application can directly process. The goal is to match up the symbols into recognizable groups by comparing them with the acoustic speech models.

III. Implementation of Dynamic Time Warping

In order to understand Dynamic Time Warping [1], two concepts need to be dealt with, they are namely

- Features: the information in which each signal is represented.
- Distances: A metric used in order to obtain a match path.

Distance again is categorized as two types:

- Local: a computational difference between a feature of one signal and the other.
- Global: the overall computational difference between an entire signal and another signal of possibly different length.

Since the feature vectors could possibly have multiple elements, a means of calculating the local distance is required. In the current paper, the distance measure between two feature vectors is calculated using probability distribution. Therefore the local distance between feature vector 'x' (test feature frames) of signal 1 and feature vector 'y' (reference feature frames) of signal 2 is given by relation below,
Node cost of the "Reference Feature Frame" and "Test feature Frame" can be given as

1. Probability: (Normal distribution)

$$\text{trelly(row,column)} = \frac{1}{\sqrt{2\pi\sigma}} e^{-0.5*\left(\frac{(Referenceframes)-(Testframes)}{(Varianceframes)}\right)^2} \quad (1)$$

$\boldsymbol{\sigma}2$ = Variance frames (39 features)
μ = Reference frames (39 features)
Speech waveform is time-dependent, hence the utterances of the same word can have different durations, and utterances of the same word with the same duration may differ due to the words being spoken at different rates. To obtain a global distance between two speech patterns (represented as a sequence of vectors) a time alignment must be performed between the reference pattern (template) and the input pattern. The best matching template is the one for which there is the lowest distance path aligning the input pattern to the template. A simple global distance score for a path is simply the sum of local distances that go to make up the path. To make the algorithm reduce excessive computation certain restriction is implied on the direction of propagation, where the matching paths cannot go backwards in time.

A. Implementation of Forward Algorithm

- Two binary files with MFCC features extracted (.mfc) for test and reference signals are taken and the DTW algorithm is performed. One file is taken as the test feature file (input), while the other is applied as the reference file (reference template).
- In DTW algorithm a trellis matrix is used and initiated with large values (10000). A path is found on employing the DTW algorithm and which signifies the Feature vectors per state.
- Now the Probability distribution is computed among each feature vector of both the test and reference files and the corresponding value is filled in the matrix.
- The minimum of these computed values is checked while the matrix is being filled with the values and traversing from the starting element to the last element in each row and column and a path is assumed to be passing through all these minimum values. This path is called as forward path.

IV. Implementation of Backward Algorithm in DTW

The trellis matrix is filled with all the respective costs and the cost at the matrix element (M×N) is considered. A path from this element is traced back to the start of the matrix using the minimum cost conditionsand this path is retrieved and is called as backward path. This backward pathalso signifies the feature vectors per state.

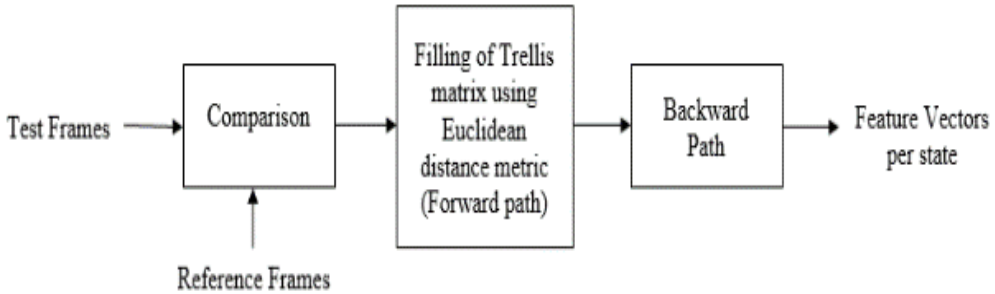

Figure 3: Block Diagram for finding path in DTW algorithm

The backward path is used in modeling as well as in the recognition process.

V. Implementation of DTW Models for IWR

A. Acoustic model is the main component for an ASR and it accounts for most of the computational load and performance of the system. It is used to link the observed features of the speech signals with the expected phonetics of the hypothesis sentence. Language model is the single largest component trained on billions of words developed for detecting the connections between the words in a sentence with the help of pronunciation dictionary.

B. Implementation of Model Building

Dynamic Time Warping algorithm as said in the previous section is utilized for aligning the K patterns and where a Viterbi algorithm is used to decode the K patterns along this aligned path. It finds optimal warping path between K patterns in the multi-dimensional trellis in the K dimensional space. The K patterns are nothing but the same word spoken repeatedly by one speaker, and here it is 10 times.

- The MFCC's are extracted from the audio files (.WAV) recorded by 20 people (speakerA, speakerB, …) and each of them uttering the same sentence 10 times(speakerA1, speakerA2…..etc.). The Silence is also obtained 10 times and features are extracted to be saved in binary format.
- Initially, each pupil's utterance is assigned equal number of feature vectors per state.

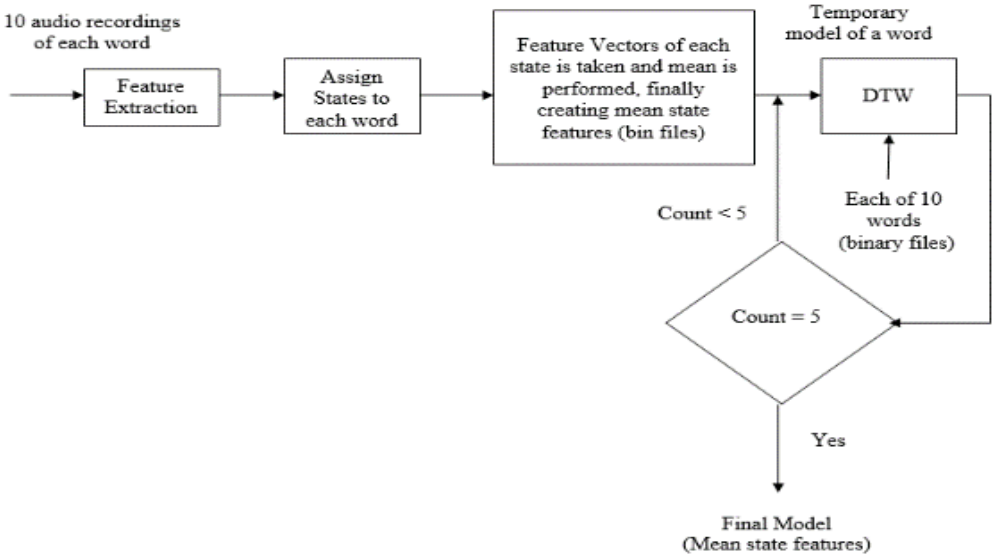

Figure 4: Block diagram explaining the procedure for modeling

- Now 10 binary files of a pupil are considered and foreach state of all these 10 binary files, mean and variance is computed, i.e. first state of each bin file is considered and mean and variance is computed for the state, similarly every state of the 10 binary files is considered to compute mean and variance, thus finally forming mean state features. These mean state features are also referred to as a temporary model.
- The DTW algorithm is applied among each of the 10 binary files and the corresponding temporary model. The path obtained here gives the feature vectors per state which initially were considered as equal for each state. Hence 10 different paths are obtained which form 10 new binary files with new feature vectors per state. With these 10 binary files again a new temporary model is built on following the procedure discussed above. This process is repeated 5 times and ultimately a model is achieved for each of them and saved under the same name which is utilized later in the recognition process.

VI. Template Matching and Performance of the Developed IWR System

In Template matching unknown speech is compared against a set of template words in order to find the best Match. A collection of prototypical speech patterns are stored as reference patterns representing the dictionary of words. Recognition is then carried out by matching an unknown spoken utterance with each of these reference templates and selecting the category of the best matching pattern..

A. Implementation of Recognition process

Training phase of the ASR was dealt in the sections above and now testing process is briefed here in this section. The forward algorithm for DTW is utilized in the pattern matching process. The models that were built as explained in the section preceding are utilized here in the recognition process.

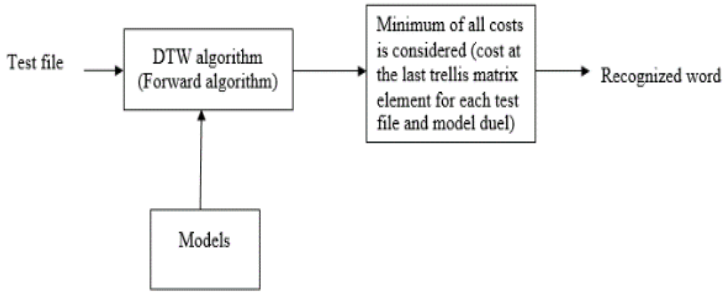

Figure 6: Block diagram of the recognition process

In the recognition process speech (word) is acquired and the corresponding MFCC binary file is obtained applied to the DTW algorithm.

- Each time the speech test file along with the saved models is given to the DTW algorithm to find the minimum cost function.
- The minimum cost among all the models decides the recognized word. Thus finally the name with the corresponding model giving the minimum cost is the word to be recognized.

VII. Results And Analysis

The performance of the developed system is an important task in the process of developing a system.We publish here results obtained on testing the utterances against the same trained models. Originally, the models trained with 10 utterances for each word were tested for the same 10 utterances and while 16 speaker participated in this training and testing process. As publishing all the results is impractical, we put performance details of the first 5 speakers of their 3 utterances while testing.

Table 1: PERFORMANCE OF DEVELOPED ASR SYSTEM

Feature Vector File (in .bin)	Uttered Word	Recognized as	Cost Obtained (Float model)	Cost Obtained (Fixed model)
Silence1	Silence	Silence	-7243.2617	-7235
Silence2	Silence	Silence	-2256.6997	-2248
Silence3	Silence	Silence	-5866.4873	-5861
SpeakerA1	SpeakerA	SpeakerA	-2495.6340	-2487
SpeakerA2	SpeakerA	SpeakerA	-5028.8115	-5023
SpeakerA3	SpeakerA	SpeakerA	-4637.5478	-4632
SpeakerB2	SpeakerB	SpeakerB	-2983.5285	-2978
SpeakerB3	SpeakerB	SpeakerB	-4915.7480	-4910
SpeakerB4	SpeakerB	SpeakerB	-4763.3657	-4754
SpeakerC1	SpeakerC	Silence	5132754944	5132754960
SpeakerC2	SpeakerC	SpeakerC	-4430.2299	-4426
SpeakerC3	SpeakerC	SpeakerC	-4526.9179	-4520
SpeakerD1	SpeakerD	Silence	5673760768	5673760775
SpeakerD2	SpeakerD	SpeakerD	-5259.5161	-5247
SpeakerD3	SpeakerD	SpeakerD	-5202.1879	-5197

VIII. Conclusions

The Speech recognition is limited by the challenges like Speaker variability and Speech variability and many more; these issues are to dealt with to build the ASR.In the project during training the MFCC features are extracted from the speech files in MATLAB and the binary files of the same are taken and applied to the DTW algorithm in C. The models are built in C and stored, thus completing the training phase.

During test phase again the binary files of the extracted features of the speech data to be tested is taken and matched to the stored models in order of the recognition process. In recognition also DTW algorithm is utilized to compare the stored models and the test speech data (word) to find the minimum cost. The minimum cost obtained for a particular pair (model and test speech) reveals the word and this happens to be the recognized word. A more robust class room database management can be incorporated with the following enhancements.

1. Noise cancellation techniques
2. Connected word recognition algorithm
3. Integration of real time data acquisition.
4. Development of automatic segmentation of recorded data in to silent and non-silent words. It will be useful training for more number of words.

IX. Future Scope

The project is implemented for Isolated Word Recognition and it can be extended for continuous words as well. The IWR system is realized in C language initially using the binary files obtained from MATLAB. Now, this beingexecuted to Fixed C, realizing the project in assembly language to be implemented on a Digital signal processor will become quite straightforward. Once implemented on a DSP efficiently, it can lead to the realization on mass scale on the mobile phone technology. One can understand the gigantic scope of the work if the mobile phones are considered to be employed upon.

References

[1]. Dynamic Time warping tutorial, http://www.cnel.ufl.edu/~kkale/dtw.html.
[2]. Mahdi Shaneh, and AzizollahTaheri, "Voice Command Recognition System Based on MFCC and VQ Algorithms", World Academy of Science, Engineering and Technology 57 2009.
[3]. Noelia AlcarazMeseguer, "Speech Analysis for Automatic Speech Recognition", MS Thesis , July 2009.
[4]. Rabiner,L,JuangB,Yegnanarayana B, "Fundamentals of speech recognition",2009.
[5]. Talal Bin Amin, Iftekhar Mahmood, "Speech Recognition Using Dynamic Time Warping", 2nd International Conference on Advances in Space Technologies, Islamabad, Pakistan,2008

IOSR Journal of Electronics and Communication Engineering (IOSR-JECE) is UGC approved Journal with Sl. No. 5016, Journal no. 49082.

MounikaJammula. "Automatic Speech Recognition system for class room database management in Fixed – C Language." IOSR Journal of Electronics and Communication Engineering (IOSR-JECE), vol. 12, no. 4, 2017, pp. 62–68.