Decision level data fusion in Speech and Image Recognition Systems

A Project Report

submitted by

Gaurav Sanjay Newalkar (12EC43)

Nithin Rao Koluguri(12EC51)

Nikhil Lunavath (12EC68)

under the guidance of

Dr.Ashvini Chaturvedi

in partial fulfilment of the requirements

for the award of the degree of

BACHELOR OF TECHNOLOGY

DEPARTMENT OF ELECTRONICS AND COMMUNICATION ENGINEERING

NATIONAL INSTITUTE OF TECHNOLOGY KARNATAKA

SURATHKAL, MANGALORE - 575025

November 17, 2015

ABSTRACT

We aim to design and implement Isolated words speech recognition system using Matlab. This work was based on the Hidden Markov Model (HMM), which provides a highly reliable way for recognizing speech. The system is able to recognize the speech waveform by translating the speech waveform into a set of feature vectors using Mel Frequency Cepstral Coefficients (MFCC) technique. In this project we focuses on English digits from (One through Four), which is based on isolated words structure. In clean environment and isolated words speech recognition module, we achieved 98.33% when tested on training set and achieved 92.5% when tested on testing set. These recognition rates are relatively successful if compared to similar systems.[10]

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

CHAPTER 1
Introduction

1.1 Problem definition

The idea of human machine interaction led to research in Speech recognition. Automatic speech recognition uses the process and related technology for converting speech signals into a sequence of words or other linguistic units by means of an algorithm implemented as a computer program. Speech understanding systems presently are capable of understanding speech input for vocabularies of thousands of words in operational environments. Speech signal conveys two important types of information: (a) speech content and (b) The speaker identity. Speech recognizers aim to extract the lexical information from the speech signal independently of the speaker by reducing the inter-speaker variability. Speaker recognition is concerned with extracting the identity of the person. [1] speech recognition concerns the study of signals of a spoken language where an analysis using the feature extraction method is used to produce a sequence of words. In short, this field attempts to develop machines that mimic human speech communication capabilities. There are several approaches to automatic speech recognition, namely acoustic phonetic, statistical pattern recognition and artificial intelligence approaches.[2]
In speech applications such as dictation software, the application's response to hearing a recognized word may be to write it in a word processor. In an interactive voice response system, the speech application might recognize a person's name and route a caller to that person's phone. Speech recognition is also different from voice recognition, though many people use the terms interchangeably. In a technical sense, voice recognition is strictly about trying to recognize individual voices, not what the speaker said. It is a form of biometrics, the process of identifying a specific individual, often used for security applications.
Because we all have distinct speaking styles this is why you can tell your mom's voice from your favorite radio talk show host's computers can take a sample of speech and analyze it for distinct characteristics, creating a "voice print" that is unique to an individual in the same way a fingerprint is. A common voice recognition system might make the user speak a password. It would then compare the speaker's voice print to a stored voice print and authenticate the user if they matched.
Though speech recognition uses some of the same fundamental technology as voice recognition, it is different because it does not try to identify individuals. Rather it tries to recognize what individuals say. It's the difference between knowing who is speaking and what is said. We are currently more concerned about what is said than who is.

1.2 Previous work

Peeling and Moore (1987) [3] applied MLPs to digit recognition with excellent results. They used a static input buffer of 60 frames (1.2 seconds) of spectral coefficients, long enough for the longest spoken word; briefer words were padded with zeros and positioned randomly in the 60-frame buffer. Evaluating a variety of MLP topologies, they obtained the best performance with a single hidden layer with 50 units. This network achieved accuracy near that of an advanced HMM system: error rates were 0.25% versus

0.2% in speaker-dependent experiments, or 1.9% versus 0.6% for multi-speaker experiments, using a 40-speaker database of digits from RSRE. In addition, the MLP was typically five times faster than theHMM system.

Kammerer and Kupper (1988) [4] applied a variety of networks to the TI 20-word database, finding that a single-layer perceptron outperformed both multi-layer perceptrons and a DTW template-based recognizer in many cases. They used a static input buffer of 16 frames, into which each word was linearly normalized, with 16 2-bit coefficients per frame; performance improved slightly when the training data was augmented by temporally distorted tokens. Error rates for the SLP versus DTW were 0.4% versus 0.7% in speaker-dependent experiments, or 2.7% versus 2.5% for speaker-independent experiments.

Lippmann (1989) [5] points out that while the above results seem impressive, they are mitigated by evidence that these small-vocabulary tasks are not really very difficult. Burton et al (1985) demonstrated that a simple recognizer based on whole-word vector quantization, without time alignment, can achieve speaker-dependent error rates as low as 0.8% for the TI 20-word database, or 0.3 for digits. Thus it is not surprising that simple networks can achieve good results on these tasks, in which temporal information is not very important.

Burr (1988) [6] applied MLPs to the more difficult task of alphabet recognition. He used a static input buffer of 20 frames, into which each spoken letter was linearly normalized, with 8 spectral coefficients per frame. Training on three sets of the 26 spoken letters and testing on a fourth set, an MLP achieved an error rate of 15% in speaker-dependent experiments, matching the accuracy of a DTW template-based approach.

1.3 Objectives

1. To recognize 4 different words through automatic speech recognition and convert to 3 digit binary word.
2. Based on the 3 digit binary word, A Image Recognition system to be developed to recognize the object pertaining to the command given.

CHAPTER 2
Description

Speech recognition systems can be separated in several different classes by describing what types of utterances they have the ability to recognize. These classes are classified as the following: Isolated Words, Connected Words, Continuous Speech, Spontaneous Speech. In this project we work on Isolated words since our objective here is to recognize only 4 words that are isolated from each other.

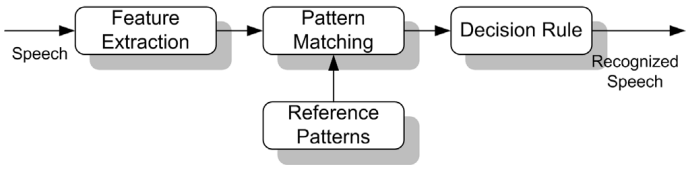

Figure 2.1: Basic Model of Speech recognition

2.1 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 accepts single words or single utterance at a time. 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.

2.2 Approaches to speech recognition

Basically there exist three approaches to speech recognition. They are
a)Acoustic Phonetic Approach
b)Pattern Recognition Approach
c)Artificial Intelligence Approach.
The pattern-matching approach [8] involves two essential steps namely, pattern training and pattern comparison. We use HMM as a tool here to perform both the steps.
Recognizers generally use a similar process to figure out what a speaker said:
1.The Automatic Speaker Recognition(ASR) loads a list of words to be recognized. This list of words is called a grammar.
2.The ASR loads audio from the speaker. This audio is represented as a waveform, essentially the mathematical representation of sound.
3.The ASR compares the waveform to its own acoustic models. These are databases that contain information about the waveforms of individual sounds and are what allow the engine to recognize speech.
4.The ASR compares the words in the grammar to the results it obtained from searching its acoustic models.
5.It then determines which words in the grammar the audio most closely matches and returns a result.

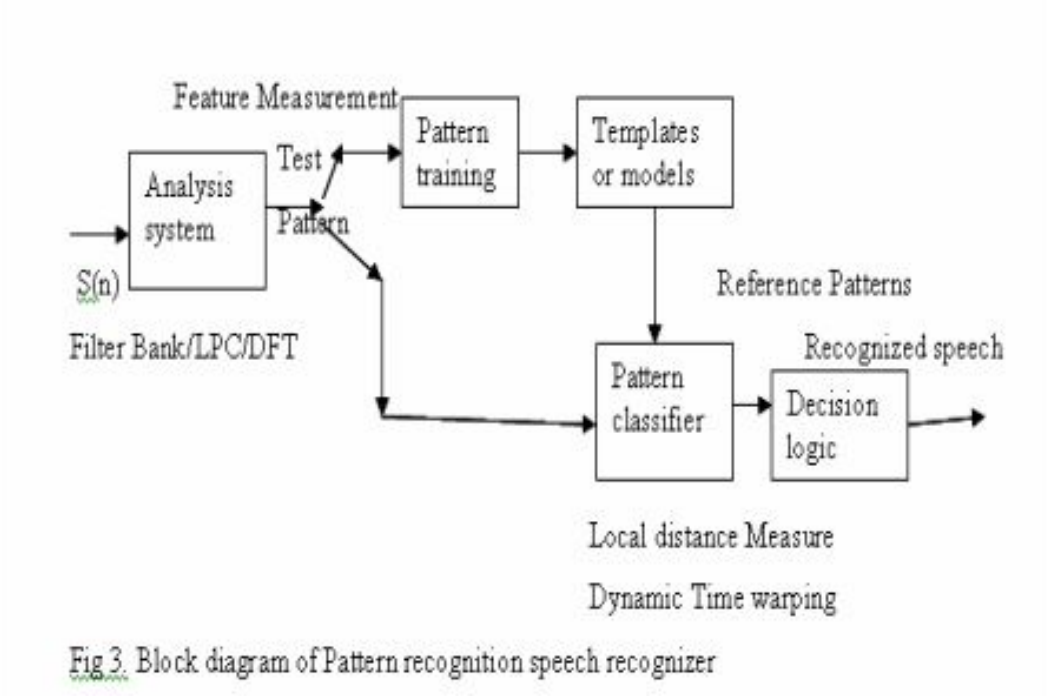

Figure 2.2: Block Diagram of Pattern Recognition and speech Recogniser

In the pattern-comparison stage of the approach, a direct comparison is made between the unknown speeches (the speech to be recognized) with each possible pattern learned in the training stage in order to determine the identity of the unknown according to the goodness of match of the patterns. The pattern-matching approach has become the predominant method for speech recognition in the last six decades. A block schematic diagram of pattern recognition is presented in fig.2.2. It has been noticed that the success of any automatic speech recognition system requires a combination of various techniques and algorithms, each of which performs a specific task for achieving the main goal of the system. Therefore, a combination of related algorithms improves the accuracy or the recognition rate of such applications. Figure 3 shows the architecture of the HMM based English digits speech recognition system.

2.3 Feature Extraction :

Figure 2.3 had just shown the main steps to perform the HMM based speech recognition system as follows:
1. Receiving and digitizing the input speech signal.
2. Extracting features for all input speech signals using MFCC algorithm, where its computational steps are shown in Fig.2.4 and Fig. 2.5, then converting and storing each signals features into a feature vector.
3. Training of feature vectors is done by Baun-Welch Algorithm
4. Finally, performing a Viterbi search which is an algorithm to compute the optimal (most likely) state sequence in HMM given a sequence of observed outputs.

Hidden Markov Model (HMM) is one of the most powerful and dominating statistical approaches, which has been applied for many years. The basic theory of HMM was

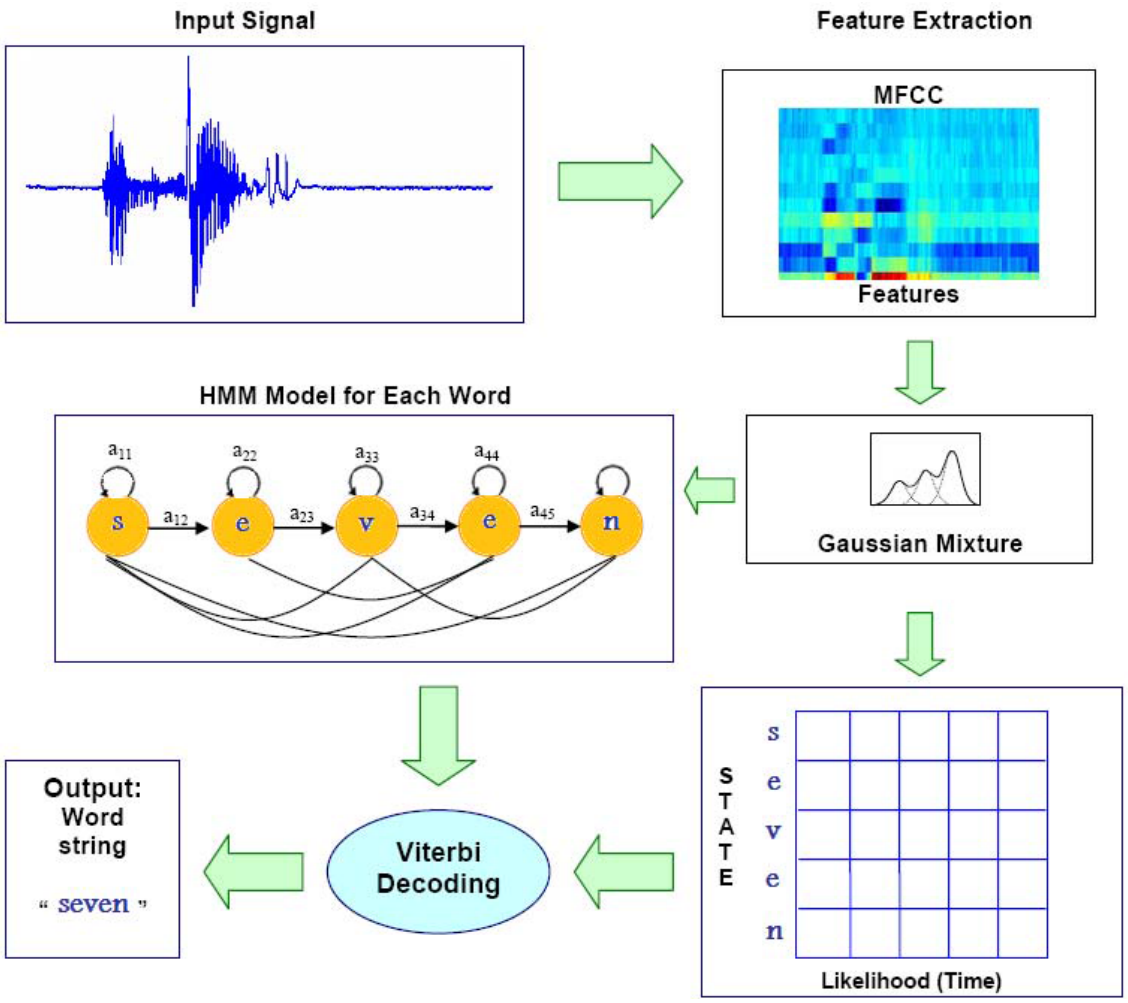

Figure 2.3: Techniques for speech recognition

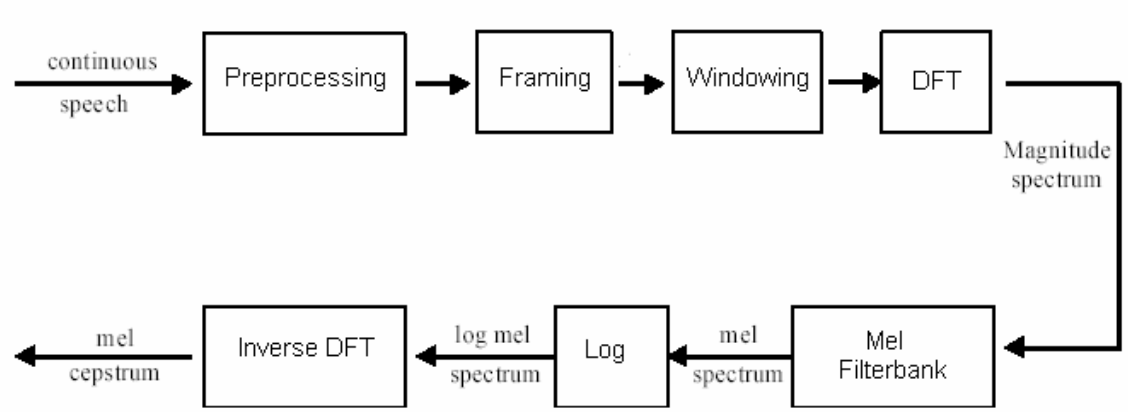

Figure 2.4: Steps performed for MFCC extraction

published in a series of classic papers by Baum and his colleagues in the late 1960s and early 1970s which was then implemented for speech recognition applications by Baker at Carnegie Mellon University (CMU) and by Jelinek and his colleagues at IBM in the 1970s [8].

An HMMs are specified by a set of states Q, a set of transition probabilities A, a set of observation likelihoods B, a defined start state and end state(s), and a set of observation

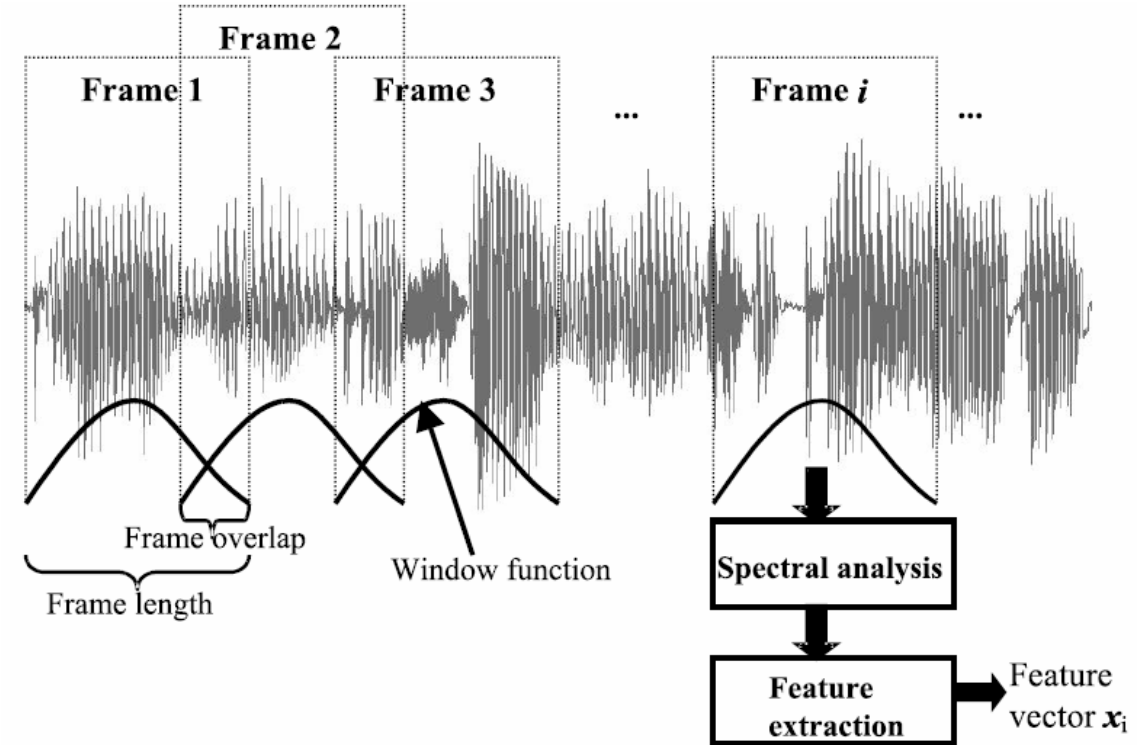

Figure 2.5: Frame Blocking

symbols O, which is not drawn from the same alphabet as the state set Q [11].

2.4 Project

In our project we collected samples from various males and females as well. For each word we have taken 20 male voice samples and 20 female voice samples. Hence a total of 4*(20+20)=160 samples have been collected. Out of 160 samples two third of them i.e., 120 samples are given for training and the remaining 40 samples for testing. Spectrogram for one of the voice sample(word : two) is shown below Fig 2.6 and also melfrequency cepstrum coefficients also in different stages of speech signal two. Fig 2.7

2.5 Results

When performed testing on the test data set we accurately classified 37 out of 40 samples which is accounting to 92.5% accuracy. We also tested on training data set as well and accurately classified 118 out of 120 samples which accounts to 98.33%accuracy. We also present the confusion matrix of our results.

2.6 Future Work

There are copious future applications for speech recognition. The output evaluated after speech recognition are used in various programs and applications to make physical

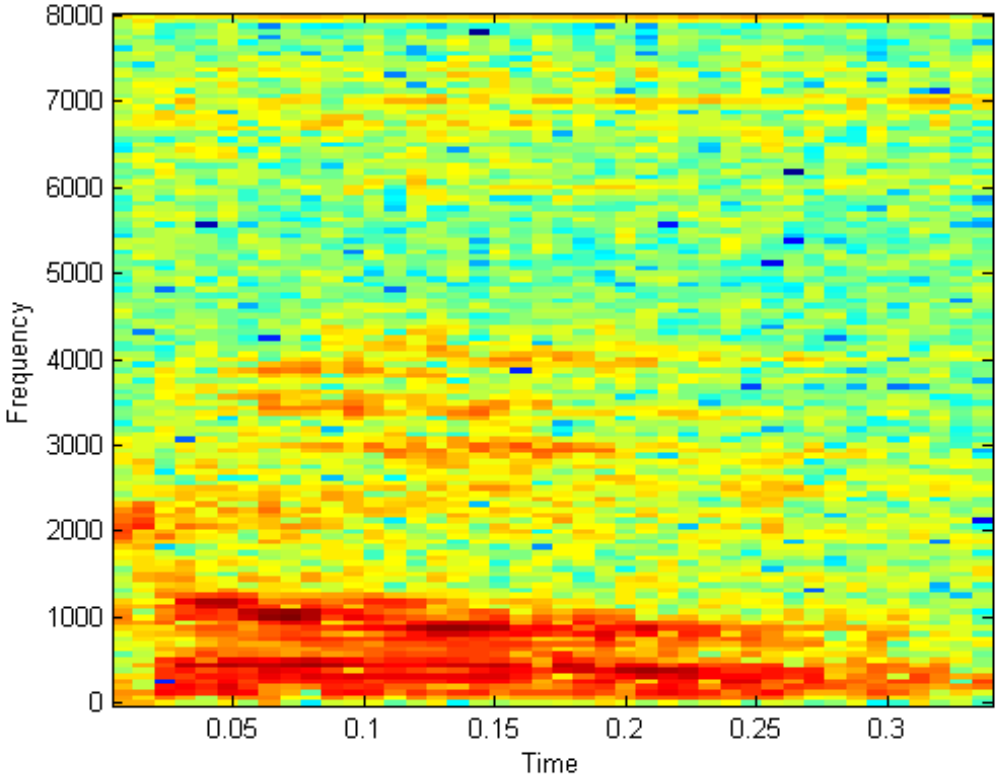

Figure 2.6: Techniques for speech recognition

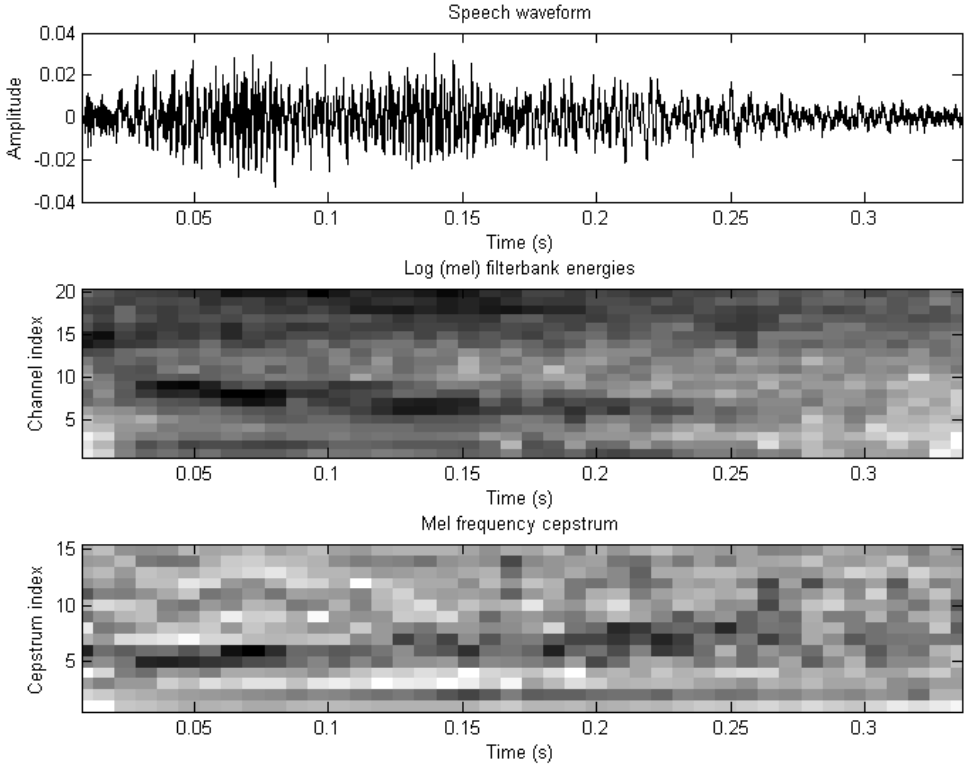

Figure 2.7: Feature Extraction Methods

Words/Digits	1	2	3	4
1	40	0	0	0
2	0	38	2	0
3	0	1	38	1
4	0	1	0	39

Table 2.1: Confusion Matrix

interaction less systems. One such future application is our second objective where we use the O/P we got in first objective to be given to image recognition system where it recognizes the object placed in front of it based on the command given.

REFERENCES

[1] Suma Swamy and K.V Ramakrishnan, *"AN EFFICIENT SPEECH RECOGNITION SYSTEM",*An International Journal (CSEIJ), Vol. 3, No. 4, August 2013 .

[2] Bassam A.Q.Al-Qatab and Raja.N.Aninon, *"Arabic Speech Recognition using Hidden Markov Model ToolKit (HTK)",*IEEE Information Technology (ITSim), 2010,page 557-562.

[3] S.M. Peeling, R.K. Moore, *" Isolated digit recognition experiments using the multi-layer perceptron",*1987

[4] Bernhard R. Kmmerer, Wolfgang A. Kpper *"Experiments for isolated-word recognition with single- and two-layer perceptrons",* 1988

[5] Richard P Lippmann, *" Review of Neural Networks for Speech Recognition"*1989

[6] Burr, D.J., *""Experiments on neural net recognition of spoken and written text",* IEEE Transactions on , vol.36, no.7, pp.1162-1168, Jul 1988.

[7] . M.A.Anusuya ,S.K.Katti , *"Speech recognition by machine: A review",* in Proceedings of the IEEE , vol.64, no.4, pp.501-531, April 2009.

[8] Rabiner and Juang *"Fundamentals of Speech Recognition".*1993

[9] R.K.Moore, *Twenty things we still don t know about speech , Proc.CRIM/ FORWISS Workshop on Progress and Prospects of speech Research an Technology"* 1994.

[10] *English Digits Speech Recognition System Based on Hidden Markov Models"* 2010 May .

[11] D. S., Jurafsky, and J. H., Martin, *"Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition. Prentice Hall, Upper Saddle River, NJ, 2nd edition, "* 2008 .