GENDER CLASSIFICATION IN SPEECH RECOGNITION USING A NEURO-FUZZY SYSTEM

A Graduate Project Report submitted to Manipal University in partial fulfilment of the requirement for the award of the degree of

BACHELOR OF TECHNOLOGY
In
Electronics and Communication Engineering

Submitted by

Neha Rawat
Reg. No.:110907090

Srishti Saha
Reg. No.:110907266

Under the guidance of
Prof. T.K. Padma Shri
Associate Professor – Senior Grade

DEPARTMENT OF ELECTRONICS AND COMMUNICATION ENGINEERING
MANIPAL INSTITUTE OF TECHNOLOGY
(A Constituent College of Manipal University)
MANIPAL – 576104, KARNATAKA, INDIA

MAY 2015

DEPARTMENT OF ELECTRONICS AND COMMUNICATION ENGINEERING

MANIPAL INSTITUTE OF TECHNOLOGY

(A Constituent College of Manipal University)

MANIPAL – 576 104 (KARNATAKA), INDIA

Manipal

CERTIFICATE

This is to certify that the project titled **GENDER CLASSIFICATION IN SPEECH RECOGNITION USING A NEURO-FUZZY SYSTEM** is a record of the bonafide work done by **NEHA RAWAT** (*Reg. No. 110907090*) and **SRISHTI SAHA** (*Reg. No. 110907266)* submitted in partial fulfilment of the requirements for the award of the Degree of Bachelor of Technology (BTech) in **ELECTRONICS AND COMMUNICATION ENGINEERING** of Manipal Institute of Technology Manipal, Karnataka, (A Constituent College of Manipal University), during the academic year 2014-15.

Internal Guide
Prof. T.K. Padma Shri
Associate Professor – Senior Grade

Prof. Dr. ----------
HOD, E & C.
M.I.T, MANIPAL

ACKNOWLEDGEMENT

We would like to express our gratitude to the Director, Dr. Vinod Thomas and our Head of Department, Dr. Somashekara Bhat for providing us the opportunity and necessary facilities to successfully complete the project. We would like to thank our guide Prof. T.K. Padma Shri, who was a constant support during the tenure of our project. Her wisdom and encouragement were indispensable for the completion of this project. We would also like to thank Prof. Muralikrishna H. for his technical guidance during the course of the project. We are highly indebted to our knowledgeable and diligent panel of professors – our guide Prof. T.K. Padma Shri, Prof. Muralikrishna H., Prof. Shruthi K., Prof. Ujjwal Verma and Prof. Vishnumurthy Kedlaya K., for their constant motivation and unbiased evaluation. Last but not the least; we extend our thanks to the non-teaching staff without whose help this project could not have been successful.

ABSTRACT

Speech is often considered as the most natural and efficient manner of exchanging information. The goal of speech recognition technology is, in a broad sense, to create machines that can receive spoken information and act appropriately upon that information. Speech recognition algorithms can be either speaker-dependent or speaker-independent. A gender discrimination system comes of use in a speaker-independent recognition system where speaker attributed variability is undesirable. The gender of a speaker is one of the influential sources of this variability. Hence, such a system can be used in a myriad of applications such as speaker identification, speaker indexing, annotation and retrieval of multimedia database in financial transactions and telephone banking, smart human computer interaction for biometrics, social robots, security systems for speaker verification, etc. The given gender classification system uses a neuro-fuzzy based approach to achieve its purpose.

The *Adaptive Neuro-Fuzzy Inference System (ANFIS)* being used is a kind of artificial neural network that is based on Takagi-Sugeno fuzzy inference system. Since it integrates both neural networks and fuzzy logic principles, it has the potential to capture the benefits of both in a single framework. It uses a dataset of five frequency and energy-based features extracted from various male and female voice samples to train the classifier. The features that we extracted for this purpose are Pitch, Zero Crossing Rate, Short Time Energy, Energy Entropy and Formants. *Pitch* is a perceptual property that allows the ordering of sounds on a frequency-related scale. It serves as a close proxy for the fundamental frequency of a speech signal. *Zero Crossing Rate* gives the rate of sign changes along the speech signal. *Short Time Energy* gives the energy associated with a short region of speech while *Energy Entropy* gives the sudden changes in the energy levels of the signal. *Formant frequencies* are concentrations of acoustic energy around a particular frequency in the speech wave and represent resonance in the vocal tract. The above features were fed to the classifier which after being trained was tested on some test voice samples.

The classifier was also tested for its accuracy as well as its efficiency as compared to other existing classifiers for gender classification. The objective of the project was to prove the effectiveness of using a neuro-fuzzy based system as opposed to standard methods like neural network systems and Support Vector Machine (SVM).

In conclusion, the proposed gender classification system was seen to provide a much more accurate and efficient discrimination due to the use of a multi-feature neuro-fuzzy based approach. The software MATLAB 2015a was used for the design of this system.

ABBREVIATIONS

ZCR- Zero Crossing Rate
STE- Short-Time Energy
EE- Energy Entropy
ANN- Artificial Neural Network
SVM- Support Vector Machine
ANFIS- Adaptive Neuro-Fuzzy Inference System
TIMIT- Texas Instruments-Massachusetts Institute of Technology

LIST OF TABLES

Table No	Table Title	Page No
4.1	Feature Table for TIDIGITS Dataset	25
4.2	Efficiency Table	41

LIST OF FIGURES

Figure No	Figure Title	Page No
1.1	General Functionality block diagram of a classifier	2
2.1	A Fuzzy Inference System	8
2.2	An Adaptive Network	10
3.1	Position of the bulk of the tongue in the oral cavity during the production of vowels	18
3.2	ANFIS Architecture	22
3.3	Process Summary	24
4.1	Pitch – Male vs Female	29
4.2	ZCR (Normalized) – Male vs Female	30
4.3	STE (Normalized) – Male vs Female	30
4.4	EE (Normalized) – Male vs Female	30
4.5	First Formant – Male vs Female	31
4.6	Output displayed in command window for ANN using the 5 features	32
4.7	Input Membership Functions – Pitch	34
4.8	Output in command window for ANFIS using only Pitch as a feature	35
4.9	Input Membership Functions – ZCR	36
4.10	Input Membership Functions – STE	36
4.11	Input Membership Functions – EE	37
4.12	Input Membership Functions – Formants	37
4.13	Output in command window for ANFIS using the 5 features	37
4.14	Output in command window for SVM using the 5 features	39

Contents

			Page No
Acknowledgement			i
Abstract			ii
List Of Figures			iii
List Of Tables			vi
Chapter 1	**INTRODUCTION**		
	1.1	Introduction	1
	1.2	Speech Recognition-The Dimensions of Difficulty	1
	1.3	Need for Gender Classification in Speech Recognition	2
	1.4	Classifier Systems	2
	1.5	Motivation	4
	1.6	Objective	5
	1.7	Organization of the Project Report	5
Chapter 2	**BACKGROUND THEORY**		
	2.1	Introduction	7
	2.2	Theory	8
	2.3	Literature Survey	10
Chapter 3	**METHODOLOGY**		
	3.1	Introduction	13
	3.2	Pre-Processing of the Speech Signal	13
	3.3	Pitch	14
	3.4	Zero Crossing Rate	15
	3.5	Short-Time Energy	16
	3.6	Energy Entropy	16
	3.7	Formants	17
	3.8	The Adaptive Neuro-Fuzzy Inference System (ANFIS)	19
	3.9	Artificial Neural Networks (ANNs) and Support Vector Machines (SVMs)	23
	3.10	Process Summary	24
Chapter 4	**RESULT ANALYSIS**		
	4.1	Introduction	25
	4.2	Feature Table	25
	4.3	Feature Analysis	30
	4.4	Implementation and Results	32
	4.5	Significance of the Result	41

Chapter 5		**CONCLUSION AND FUTURE SCOPE**	
	5.1	Introduction	42
	5.2	Summary	42
	5.3	Conclusions	42
	5.4	Future Scope of Work	43
REFERENCES			44
PROJECT DETAILS			45

CHAPTER 1
INTRODUCTION

1.1 INTRODUCTION

Speech recognition has been a major topic of research for the past decade. The study of speech recognition is part of a quest for 'artificially intelligent' machines that can 'hear', 'understand' and 'act upon' spoken information. Science fiction gives vivid accounts of the unbounded possibilities of this field. However, our imagination far surpasses our current technical abilities in this domain. In this Chapter, we will tread upon one of the initial, yet important part of the speech recognition problem – gender classification.

1.2 SPEECH RECOGNITION- THE DIMENSIONS OF DIFFICULTY

There are a variety of factors that influence the success or failure of a speech recognition system. These factors can be enumerated in the form of the following questions:

- Is the system required to recognize a specific individual or multiple speakers (including, perhaps, all speakers)?
- What is the size of the vocabulary?
- Is the speech to be entered in discrete units (usually words) with distinct pauses among them, or as a continuous utterance?
- What is the extent of ambiguity and acoustic confusability in the vocabulary?
- Is the system to be operated in a quiet or noisy environment, and what is the nature of the environmental noise if it exists?
- What are the linguistic constraints placed upon the speech, and what linguistic knowledge is built into the recognizer?

All these questions influence the thought that goes into building a useful speech recognition system. For the purpose of our project though, we need to look only into the first question – speaker recognition.

1.2.1 Speaker-Dependent versus Speaker-Independent Recognition

Most speech recognition algorithms are used either in *speaker-dependent* or *speaker-independent* mode. A speaker-dependent recognizer uses the utterances of a single speaker to learn the parameters (or models) that characterize the system's internal model of the speech process. The system is then used specifically for recognizing the speech of its trainer. A speaker-independent recognizer on the other hand, is trained by multiple speakers and used to recognize many speakers (who may be outside the training population). Though, a speaker-dependent system gives higher accuracy, using it for multiple speakers renders it highly inconvenient as one is required to retrain the system each time it is used with a new speaker.

Therefore, for most practical applications a speaker-independent system is used, whereas for some specialized areas a speaker-dependent system is employed.

1.3 NEED FOR GENDER CLASSIFICATION IN SPEECH RECOGNITION

The need for a gender classification system arises in the case of speaker-independent speech recognition systems. It forms a part of automatic speech recognition system to enhance speaker adaptability. It helps in removing one of the major causes of speech variability among speakers i.e. their gender. Thus, it is used in various applications such as:

- Speaker diarization for speaker identification
- Speaker indexing, annotation and retrieval of multimedia database in financial transactions and telephone banking.
- Smart human computer interaction for biometrics, social robots, etc.
- Security systems like modern voice password technologies for speaker verification
- Voice over Internet Protocol (VoIP)

1.4 CLASSIFIER SYSTEMS

Fig 1.1 General Functionality block diagram of a classifier

As shown in Fig 1.1, a *classifier* can be viewed as a function of block. A classifier assigns one class to each point of the input space. The input space is thus partitioned into disjoint subsets, called *decision regions*, each associated with a class. The way a classifier classifies inputs is defined by its decision regions. The borderlines between decision regions are called *decision-region boundaries* or simply *decision boundaries*. In practice, input vectors of different classes are rarely so neatly distinguishable. Samples of different classes may have same input vectors. Due to such *uncertainty*, areas of input space can be clouded by a mixture of samples of different classes. The optimal classifier is the one expected to produce the least number of misclassifications. Such misclassifications are due to uncertainty in the problem rather than a deficiency in the decision regions. A designed classifier is said to generalize well if the classifier achieves similar classification accuracy to both training samples and real world (testing) samples. Classifiers can be broadly classified into two main categories [14] as:

- Hard-computing Algorithms

- Soft-computing Algorithms

1.4.1 Hard Computing Techniques

The attributes of a conventional hard computing algorithm is as follows:

- Hard computing, i.e. conventional computing requires a precisely stated analytical model and often a lot of computation time. It is intolerant of imprecision, uncertainty, and approximation. For example: Bayes Classifier– It always produces an optimal result with minimum error rate or minimum risk; but requires exact knowledge of class prior probabilities and class conditional probabilities of features. This makes it seldom possible because precise knowledge does not exist in practical applications.
- Hard computing is based on binary logic, crisp systems, numerical analysis and crisp software. It thus has the characteristics of precision and precise classification.
- Hard computing requires programs to be written.
- Hard computing uses two-valued logic.
- Hard computing is deterministic.
- Hard computing requires exact input data.
- Hard computing is strictly sequential.
- Hard computing produces precise answers.

1.4.2 Soft Computing Techniques

Contrary to the conventional hard computing techniques, the prime inherent advantage associated with the soft computing techniques is the non-requirement of a mathematical model and hence are becoming increasingly popular as system identification methodology. In effect, the role model for soft computing is the human mind.

The attributes of a soft-computing algorithm are:

- It is liberal and takes in account any probable imprecision, uncertainty, partial truth, and approximation.
- Soft computing is predominantly based on fuzzy logic, neural nets and probabilistic reasoning.
- Soft computing allows parallel computations.
- Soft computing can use multivalued or fuzzy logic. It can thus take in account the fuzzy boundaries and probabilistic values related to the input set of data.
- Soft computing can deal with ambiguous and noisy data.
- Soft computing can yield approximate answers.

Although in hard computing, imprecision and uncertainty are undesirable properties, in soft computing the tolerance for imprecision and uncertainty is exploited to achieve tractability, lower cost, high Machine Intelligence Quotient (MIQ) and economy of communication.

In soft computing based system identification, instead of a single standard method, a collection of techniques has been put forward as possible solutions to the identification problem. The algorithms can be broadly grouped as the following:

- Neural Network based algorithm
- Fuzzy Logic based algorithm
- Genetic algorithm.

These attributes make a soft-computing algorithm a clear choice for the purpose of gender clustering on the basis of features extracted from the speech samples.

1.4.3 Selection of a hybrid system for Gender Classification

The neural network has the inherent advantage of being able to adapt itself and also in its learning capabilities. Similarly the salient feature that is associated with the fuzzy logic is the distinct ability to take into account the prevailing uncertainty and imprecision of real systems with the help of the fuzzy if-then rules.

In order to exploit the advantage of the self-adaptability and learning capability of the neural network and the capability of the fuzzy system to take into account of the prevailing uncertainty and imprecision of real systems with the help of the fuzzy if-then rules, an integrated forecasting approach comprising of both the fuzzy logic and the neural network has been considered.

The hybrid system, so conceptualized is called the Adaptive network based fuzzy inference system (ANFIS). Here the fuzzy system with its expert knowledge stands as a front end preprocessor for the neural network input and output layers. Based on the historical data, the neural network learning algorithms are used to determine the parameters of the expert knowledge based fuzzy system. The use of this hybrid system ANFIS helps to complement the weakness of the respective systems.

1.5 MOTIVATION

The design of a gender classification system in present day scenario is generally developed using the pitch of the speakers as the classifying feature. The pitch gives the relative highness or lowness of a voice sample and is seen to be higher for females than for males. Hence, it generally gives a clear distinction between male and female speakers.

However, there are unusual cases where the pitch of some male speakers has been observed to be in the range of female speakers and vice-versa. Speech being a complex and variable signal, using a single feature for gender discernment would not be a very efficient method.

Therefore, other features like formants, short-time energy, energy entropy, zero crossing rate, MFCCs are also commonly used as an alternative to pitch for this purpose. But even in such alternative systems we generally find the use of only the frequency based features or only the energy based features. Hence, the motivation for this project which uses a multi-feature classification system involving the frequency as well as the energy based features.

The use of a multi-feature based system is expected to give a more accurate classification as even if there is an aberration in the case of pitch or formants, it can be nullified by the values

of the energy-based features. This would then provide a balanced and flexible estimate for comparison between the speakers.
The classifier used for this purpose will be a neuro-fuzzy based classifier, which gives the advantage of combining the flexibility of fuzzy logic with the learning capabilities of neural networks. Therefore, these hybrid systems are expected to be much more efficient that standard minimum distance techniques or neural networks and Support Vector Machines (SVMs).
The finally designed classifier is expected to confirm the above stated expectations in terms of accuracy as well as efficiency.

1.6 OBJECTIVE

Our primary objective is to design an efficient neuro-fuzzy based classifier for gender-based speaker classification using frequency as well as energy based features.
Our secondary objective is to analyze the features extracted and first observe the rough distinction apparent through just the values. Also, we will analyze the methodologies and tools used for feature extraction and the changes observed on modifying them if required.
We also plan to compare the efficiency of the designed classifier with other standard classifiers and neural network based techniques like Support Vector Machines (SVM).

1.7 ORGANIZATION OF THE PROJECT REPORT

This project report has been divided into the five following chapters:

The First Chapter is the Introduction which gives a brief description of the area of the project work and the aspects covered in the project work. It gives the need for gender classification and the difficulties in speech recognition that have given rise to this need. It is followed by a brief insight into the existing and possible classifier systems that can be employed for the purpose of the project and a brief comparative outlook on the options so mentioned. This is followed by the motivation for the design of the stated classifier, its unique features and the significance of the results obtained. Finally, we define the primary and secondary objective of the project as a whole.

The Second Chapter is an introduction to the background theory of the project and a literature survey. It includes definitions of the terms involved in the process and a brief theoretical explanation accompanying each feature. It also succinctly explains the need, architecture and the basic learning rule on the neuro-fuzzy inference system that shall be employed to accomplish the primary objective of the project. This is followed by the literature survey which consists of some of the research works related to the project topic.

The Third Chapter deals with the methodology of the feature extraction techniques and algorithms. It contains a description of the various features extracted, the formulae used and the methods employed for extraction. It includes step-wise algorithms describing the methodology employed for each feature and a diagrammatic representation summarizing the entire process at the end. This is followed by a theoretical explanation of an Adaptive Neuro-Fuzzy Inference System including detailed explanation of the architecture and step-wise algorithm of the implementation of ANFIS on our database. It also encompasses a brief description of Artificial Neural Networks (ANN) and Support Vector Machines (SVMs).

The Fourth Chapter deals with the result analysis of the values obtained through the implementation of the methodologies discussed in the previous chapter. The values obtained for all the speech samples in the dataset are tabulated and graphs are plotted to indicate the distinction between the values of each feature for male and female speakers. This is followed by an analysis of the observed behavior for each given feature. It then covers the results obtained from the implementation of the ANN (multi-feature), ANFIS (using only pitch as a feature), ANFIS (multi-feature) and SVM (multi-feature) and simultaneously compares the efficiency of each of these systems. The significance of the results so obtained have been summarized in the last section of this chapter.

The Fifth Chapter consists of the conclusion and future work. It includes a brief summary of all the work carried in the project including the objectives accomplished. This is followed by the conclusion derived from the analysis of the results obtained. Finally, we end with a brief description of the future scope of work.

CHAPTER 2
BACKGROUND THEORY

2.1 INTRODUCTION

Feature extraction in speech recognition entails the process of reducing data while retaining speaker discriminative information. A variety of techniques are being employed currently for this purpose. The amount of data, generated during the speech production, is quite large while the essential characteristics of the speech process change relatively slowly and therefore, they require less data. This essential data is extracted from the speech signal using short-term energy based approaches and frequency based approaches. This Chapter consists of a brief survey of the methodologies that have generally been used for this purpose.

When sound is emitted from the human mouth, it passes through two different systems before it takes its final form. The first system is vocal cords and the next system is vocal tract. Scientists call the first system the laryngeal tract and the second system the supralaryngeal. Each system contributes specific attribute to the speech coding. Some of the features which can be extracted to aid in speaker discernment, and have also been used in this project, are as follows:

- Pitch
- Formants
- Short-Time Energy
- Energy Entropy
- Zero Crossing Rate

All of the above given features have often been used individually and in groups for the process of gender discrimination for speaker recognition.
To implement the primary motive of gender clustering, the most optimum classifier had to be chosen for the purpose. As discussed in Chapter 1, a neuro fuzzy inference system viz. ANFIS seems to be the best option. To verify this, efficiencies of other classifiers i.e. SVMs (Support Vector Machines) and ANN (Artificial Neural Network) have been calculated and compared to that of the ANFIS (Adaptive-Network-Based Fuzzy Inference System). An adaptive neuro-fuzzy inference system or adaptive network-based fuzzy inference system (ANFIS) is a kind of artificial neural network that is based on Takagi–Sugeno fuzzy inference system. The technique was developed in the early 1990s.

2.2 THEORY

Before proceeding with the literature survey, let us have a brief look on the above mentioned features:

- *Pitch*: The rate at which the vocal cords vibrate gives us an estimation of the pitch of the speech signal. Pitch is essentially a perceptual property but is generally related to the fundamental frequency of the speech signal so as to quantify it as a frequency. Pitch is useful to differentiate speaker genres. Adult females tend to speak at around 200 Hz on average and adult males around 125 Hz.
- *Short-Time Energy (STE):* The amplitude of speech signal varies appreciably with time. In particular, the amplitude of the unvoiced segments is generally much lower than the amplitude of the voiced segments. The short-time energy of the speech signal provides a convenient representation that reflects these amplitude variations.
- *Zero Crossing Rate (ZCR)*: In the context of discrete-time signals, a zero crossing is said to have occurred if successive samples have different algebraic signs. The rate at which zero crossings occur is a simple measure of the frequency content of a signal.
- *Energy Entropy (EE)*: Energy entropy is defined as the sudden different changes in the energy level of a speech signal. It has been observed to be low and distributed for males while for females it is high and remains for a short period of time.
- *Formants*: When the air flows through the vocal tract, it begins to reverberate at particular frequencies determined by the diameter and length of the cavities in the vocal tract. These reverberations are called “resonances” which are represented by the formant frequencies.

  Since, an ANFIS has been implemented to use these features to classify the speech samples according to the gender of the speakers, we should briefly study about the background theory of these Adaptive Networks.

  Before we proceed to the details of ANFIS as a unit, let us introduce ourselves to the concept of Fuzzy Inference systems. Fuzzy inference systems are also known as fuzzy-rule-based systems, fuzzy models, fuzzy associative memories (FAM), or fuzzy controllers occasionally [1].

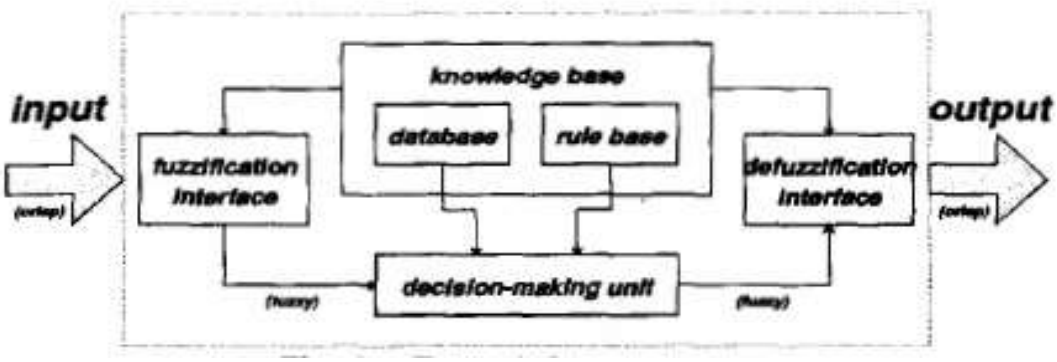

Fig 2.1 A Fuzzy Inference System

Basically a fuzzy inference system is composed of five functional blocks as shown in Fig 2.1. :

- a *rule base* containing a number of fuzzy if-then rules;
- a *database* which defines the membership functions of the fuzzy sets used in the fuzzy rules;
- a *decision-making unit* which performs the inference operations on the rules;
- a *fuzzification interface* which transforms the crisp inputs into degrees of match with linguistic values;
- a *defuzzification interface* which transform the fuzzy results of the inference into a crisp output;

Usually, the rule base and the database are jointly referred to as the *knowledge base.*

The steps of fuzzy reasoning (inference operations upon fuzzy if-then rules) performed by fuzzy inference systems are:

- Compare the input variables with the membership functions on the premise part to obtain the membership values (or compatibility measures) of each linguistic label. (This step is often called *fuzzification*).
- Combine (through a specific T-norm operator, usually multiplication or min.) the membership values on the premise part to get firing strength (weight) of each rule.
- Generate the qualified consequent (either fuzzy or crisp) of each rule depending on the firing strength.
- Aggregate the qualified consequents to produce a crisp output. (This step is called *defuzzification.*)

Now that the concept of a Fuzzy-rule based system has been introduced, we shall have a brief look on what an Adaptive Network means. An adaptive network is a superset of all kinds of feedforward neural networks with supervised learning capability. An adaptive network, as its name implies, is a network structure consisting of nodes and directional links through which the nodes are connected. Moreover, part or all of the nodes are adaptive, which means their outputs depend on the parameter(s) pertaining to these nodes, and the learning rule specifies how these parameters should be changed to minimize a prescribed error measure. The basic learning rule of adaptive networks is based on the gradient descent and the chain rule, which was proposed by Werbos in the 1970's. An adaptive network as indicated in Fig 2.2 is a multilayer feedforward network in which each node performs a particular function (node function) on incoming signals as well as a set of parameters pertaining to this node. The formulas for the node functions may vary from node to node, and the choice of each node function depends on the overall input-output function which the adaptive network is required to carry out. The links in an adaptive network only indicate the flow direction of signals between nodes and no weights are associated with the links.

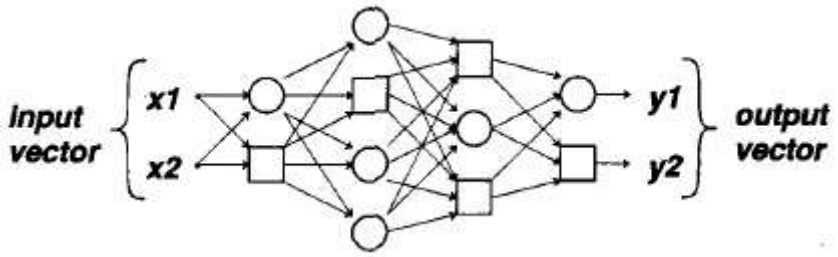

Fig 2.2 An Adaptive Network

To reflect different adaptive capabilities, we use both circle and square nodes in an adaptive network. A square node (adaptive node) has parameters while a circle node (fixed node) has none. The parameter set of an adaptive network is the union of the parameter sets of each adaptive node. In order to achieve a desired input-output mapping, these parameters are updated according to given training data and a gradient-based learning procedure.
By embedding the fuzzy inference system into the framework of adaptive networks, we obtain the ANFIS architecture, an application of which has been implemented in the project.

2.3 LITERATURE SURVEY

Some of the recent research works related to gender-based speaker classification are discussed as follows:

Nandyala et al. [2] have proposed a technique to create a real time isolated word speech recognition system for human-computer communication. Their major task has been to identify the list of words said by the speaker via the microphone. The Mel-frequency cepstral coefficients (MFCCs) that provide good discrimination of the speech signal have been used as features. Using the Dynamic Programming Algorithm, the similarity between the stored template and the test template has been measured for the speech recognition, which provides the optimum distance. The proposed system has achieved a recognition accuracy of 88.0%. They have prepared an elementary list containing ten words of cities names in India and when a particular city name is spoken, the image corresponding to that city name has been displayed.
Lakra et al. [3] have classified male and female speakers using an Adaptive Neuro-Fuzzy Inference System implemented on Matlab using the pitch of the speakers as a discriminative measure. The experiment has successfully been performed over the voice samples of 10 male and 10 female speakers in the age group 24-45 years.
Kumar et al. [4] have used pitch, calculated using the autocorrelation method, Average Magnitude Difference (AMDF) method as well as Cepstrum method, as well as formants calculated using the LP model pole extraction method, and used both of them in combination to design a gender discrimination algorithm using the minimum distance criteria.
Rao et al. [5] have utilized the time-varying glottal excitation component of speech for text independent gender recognition studies. Linear prediction (LP) residual has been used as a

representation of excitation information in speech. A Hidden Markov Models (HMMs) has been used for capturing the gender-specific information in the excitation of voiced speech. The reduce in the error during training and identifying genders during testing phase near to 100 % accuracy have illustrated that the continuous Ergodic HMM can capably capture the gender-specific information in the excitation component of speech. In their gender identification study, they have calculated the size of testing data on the gender recognition performance by using gender specific features in various HMM states, and mixture components. They have used Texas Instruments and Massachusetts Institute of Technology (TIMIT) database for performing the gender recognition studies.

Meena et al. [6] have performed gender classification using the energy based features- short-time energy, zero-crossing rate and energy entropy. They have further used the data obtained to classify the speakers using simple neural networks as well as fuzzy logic and have used the combined results of the two methods to compute various performance parameters like false positive rate, false negative rate, sensitivity, specificity, likelihood ratio positive, likelihood ratio negative, accuracy and precision.

Yune-Sang Lee et al. [7] have proposed although much effort had been directed on the way to understand the neural basis of speech processing, the neural processes concerned in the definite perception of speech had been comparatively less studied, and numerous questions stay open. In this functional magnetic resonance imaging (FMRI), the cortical regions mediating categorical speech perception using an advanced brain-mapping technique, whole-brain multivariate pattern-based analysis (MVPA). Usual healthy human subjects (native English speakers) were scanned as they listened to 10 consonant–vowel syllables along the /ba/–/da/ continuum. To partition the FMRI data into /ba/ and /da/ conditions per subject outside of the scanner, individuals' own category boundaries were calculated. The whole-brain MVPA revealed that area and the left pre-supplementary motor area evoked distinct neural activity patterns between the two perceptual categories (/ba/ vs /da/) area was also found when the same analysis was applied to another dataset, which previously yielded the supramarginal gyrus using a univariate adaptation–FMRI paradigm. The consistent MVPA findings from two independent datasets strongly indicate that area participates in categorical speech perception, with a possible role of translating speech signals into articulatory codes. The difference in results between univariate and multivariate pattern-based analyses of the same data suggest that processes in different cortical areas along the dorsal speech perception stream were distributed on different spatial scales.

As stated in [8] and [9], Most of the work in Automatic speech recognition (ASR) is generally supported by commercial corporations (Kadambe & Srinivasan, 1994). Opinions of the researchers, who study in speech/speaker recognition area, as summarized in Long and Datta (1996); Sarikaya and Hansen (2000) appear to ignore the benefits that can be gained by proper transformations of the input signal. The main task in Automatic speaker recognition (ASR) is to separate various speaker classes (Tufekci&Gowdy, 2000; Long & Datta, 1996; Sarikaya & Hansen, 2000). In literature, some researchers have explored the use of wavelets to provide a richer feature space (Sarikaya, Pellom, & Hansen, 1998; Mallat, 1998; Erzin, Cetin, & Yardimci, 1995; Rabiner, Cheng, Rosenberg,&McGonegal, 1976; Petrovska, Hennebert, Mellin & Genoud, 2000). Nevertheless, there is little evidence of widespread use

of this technique (Long & Datta, 1996). In Petrovska et al. (2000), preprocessing the data allows easier subsequent feature extraction and increased resolution. The signal was transformed from a time domain to a frequency domain using the Fourier transform by engineers (Saito, 1994; Buckheit & Donoho, 1995).

CHAPTER 3
METHODOLOGY

3.1 INTRODUCTION

The time domain waveform of a signal carries most of the auditory information. In order to obtain some relevant information from the signal, it is essential to have mechanisms that reduce information from each segment of the signal into some parameter or feature. These features are then further used for classifying the speakers as male or female. This Chapter shall deal with the methodologies generally used and those used in the project for feature extraction followed by a description of the ANFIS structure used to classify the speakers. Along with the above, we shall also discuss briefly the SVM and artificial neural network based classification systems which will be used for the purpose of comparison with the ANFIS system.

3.2 PRE-PROCESSING OF THE SPEECH SIGNAL

Before implementing the feature extraction methods, we have divided our speech signal into voiced, unvoiced and silence segments.

Voiced Excitation

Voiced sounds are produced by forcing air through the glottis or an opening between the vocal folds. The tension of the vocal cords is adjusted so that they vibrate in oscillatory fashion. The periodic interruption of the subglottal airflow results in quasi-periodic puffs of air that excite the vocal tract. This sound produced by the larynx is called *voice* or *phonation* [12].

Unvoiced Excitation

Unvoiced sounds are generated by forming a constriction at some point along the vocal tract, and forcing air through the constriction to produce turbulence [12].

Silence

Whether silence should be called a form of excitation is debatable, but it is generally included for modelling purposes [12].

3.2.1 Methodology

In our project, we have used a threshold value obtained through observation to divide the speech signal into silence, voiced and unvoiced portions.

Since the amplitude of the silence portion is zero and that of the unvoiced portions are substantially less than the voiced ones, the following steps have been followed to obtain only the voiced segments of speech:

- Analyzing the speech signal frame-by-frame.

- Comparing the maximum amplitude of the frame with the threshold value.
- Rejecting the frame, if the amplitude is lower and accepting it if it is higher.
- Concatenating all the accepted frames to produce a continuation of only the voiced segments.

Since, all the voice samples in our dataset have been recorded in a noise-free environment, additional processing for removal of noise is not required. The sampling frequency for two out of the three datasets is 16 kHz and for the remaining one is 8 kHz.

3.3 PITCH

The Pitch of a speech signal is arguably the most important feature needed for the purpose of gender classification. The methods that are generally used for pitch extraction are as follows:

Autocorrelation Method

The autocorrelation function [13] of a discrete time deterministic signal is defined as:

$$Ø(k) = \Sigma_m x(m).x(m+k) \; ; \; m=(-\infty,+\infty)$$

where x(m) is the speech signal.

The autocorrelation function is a convenient way of displaying certain properties of the signal. The autocorrelation function of a periodic signal is also periodic with the same period. It is an even function and attains its maximum value at k=0.

For a non-stationary signal, the concept of long-time autocorrelation measurement is not really meaningful. Thus a short-time autocorrelation function is defined which operates on segments. The pitch period of a speech signal is determined by obtaining the most prominent peak in the voiced segment of the signal.

Average Magnitude Difference Method (AMDF)

This method is another type of autocorrelation analysis that is often used for pitch estimation. The average magnitude difference function [13] is given as:

$$M(\tau) = (1/N) \Sigma_n |s(n) - s(n+\tau)| \; ; \; n=0,1,2\ldots N-1$$

where s(n) is the signal and s(n+τ) is the time-shifted sample. N is the length of the signal.

The difference function is expected to have a strong local minimum if the shift (or the valley) is equal to or very close to the fundamental period for each frame and the lag for which the AMDF is a global minimum is a strong candidate for the pitch period of that frame.

Cepstrum Method

In speech processing, the pitch is often determined using the cepstrum method. The Cepstrum is formed by taking the FFT (or IFFT) of log magnitude spectrum of a signal. The reason for using the FFT or IFFT interchangeably is because one will just give you a reversed version of the other, so each is equally valid for the processing we wish to do. Once in the cepstral domain, the pitch can be estimated by picking the peak of the resulting signal within a certain range. The lag at which there is most energy represents the dominant frequency in the log magnitude spectrum thereby giving the pitch.

3.3.1 Methodology

In our project, we have used the autocorrelation method to estimate the pitch of the speech signal. The steps followed are as given:

- A non-overlapping rectangular window is used for the short-time analysis of the speech signal because it exactly preserves the temporal characteristics of the signal over the desired range.
- Autocorrelation is then performed for the signal by keeping a lag of 160.
- The region of interest for detection of the pitch is kept between periods corresponding to 100 and 400 Hz as this has been observed to be the standard range for most male and female pitches.
- Within this region, the maximum peak is detected (which is actually the second largest peak as the largest peak is the central peak) and the corresponding time lag gives the pitch period which is used to obtain the pitch frequency.

3.4 ZERO CROSSING RATE

The ratio of the number of time-domain zero crossings occurred to the frame length is generally defined as the zero crossing rate. In our project, we have calculated the normalized ZCR as a means of comparison. The basic expression [13] which has been followed is as given:

$$Z= (1/2N) \Sigma_i \{sgn(x(i)) – sgn(x(i-1))\} \ ; i=1,2,\ldots N-1$$

where x(n) is the speech sample and N is the length of the speech sample. A modified signum function ‘sgn’ has been used which is given as:

$$sgn\{x(i)\}= 1 \ ; x(i) >= 0$$
$$= -1 \ ; x(i) < 0$$

This function is used to detect the zero-crossings which are then counted along the length of the frame to obtain the rate.

3.4.1 Methodology

The steps of calculation of the normalized ZCR can be summarized as follows:

- A rectangular window is used for windowing the speech signal as what we intend to find is the short-time ZCR.
- The signum function is used as per the above given formula to calculate the number of zero crossings in the given frame.
- The obtained value is divided by the length of the speech sample to obtain a normalized value for comparison.

3.5 SHORT-TIME ENERGY

The energy associated with a speech signal is time-varying in nature. Hence, for any useful purpose we require energy associated with a short term region of speech. Further, the energy associated with the voiced portion is large compared to unvoiced. For the purpose of our project, we have calculated the normalized short term energy for a short region of speech. The following formula [13] has been used for calculation:

$$e(n) = \Sigma_m (s(m).w(n-m))^2 \; ; \; m = (-\infty,+\infty)$$

3.5.1. Methodology

The normalized STE has been calculated as follows:

- A rectangular window has been used for short-term analysis of the speech signal.
- The above given formula has been implemented where s(n) is the speech sample and $\Sigma(s(n))^2$ would give the energy for the entire signal. The term w(n-m) is the rectangular window stated above which has been used to confine the energy calculation to only a specific short-time region of speech.
- The value calculated above has been divided by the length of the speech sample to obtain a normalized value.

3.6 ENERGY ENTROPY

Energy entropy of a speech signal indicates the abrupt changes in the energy level of the speech signal. For computing the energy entropy, first the speech signal is divided into k frames and then the normalized energy for each frame is calculated. Then the energy entropy is calculated using the equation given below:

$$E = -\Sigma_i \sigma^2 \log_2 (\sigma^2) \; ; \; i= 0,1,\ldots k-1$$

The normalized energy entropy has been calculated for the different voice samples for the purpose of this project.

3.6.1 Methodology

The steps for calculation of the normalized energy entropy are as given:

- A rectangular window has been used for the short-term analysis of the speech sample which is carried out frame-by-frame.
- Each windowed portion is further divided into sub windows and the normalized energy for each sub frame is calculated.
- The values are summed up to obtain the value for each frame and the obtained values are summed over the total number of frames to get the energy entropy for the entire speech sample using the above stated formula.
- The obtained energy entropy value is divided by the length of the speech sample to get its normalized value.

3.7 FORMANTS

When we study the magnitude spectra of a speech signal, we observe some well-defined regions of emphasis (‘resonances’) and some regions of de-emphasis (‘antiresonances’). These resonances are a consequence of the articulators having formed various acoustical cavities and subcavities out of the vocal tract cavities, much like concatenating different lengths of organ pipe in various orders. So the locations of these resonances in the frequency domain depend upon the shape and physical dimensions of the vocal tract. Conversely, each vocal tract shape is characterized by a set of resonant frequencies. Since these resonances tend to ‘form’ the overall spectrum, speech scientists refer to them as formants or formant frequencies.

In principle, there are an infinite number of formants in a given sound, but in practice, we usually find 3-5 in the Nyquist band after sampling.

Vowels

There are 12 principal vowels in American English. Phoneticians often recognize a thirteenth vowel called a *schwa* vowel, which is a sort of ‘degenerate vowel’ to which many others gravitate when articulated hastily in the course of flowing speech. The phonetic symbols generally used for this are /x/ and /ə/.

Vowels are differentiated by the tongue-hump position and the degree of constriction at that position as shown in Fig 3.1. The position of the hump portion of the tongue (front, central, back) divides the vowels into three groups. The degree to which the hump portion of the tongue is raised towards the palate further delineates each vowel group. Formant frequency locations for vowels are affected by three factors:

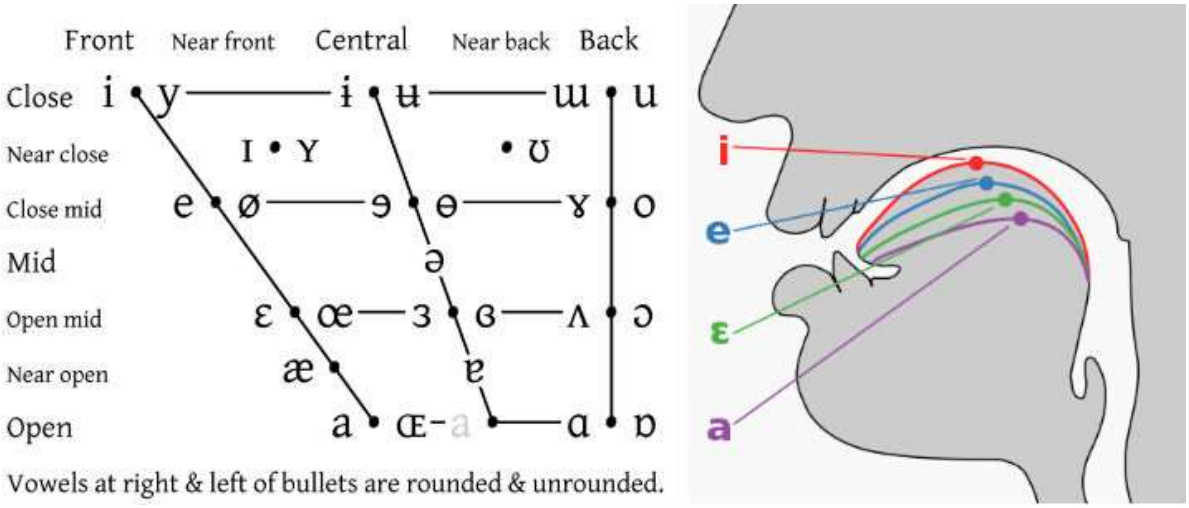

Fig 3.1 Position of the bulk of the tongue in the oral cavity during the production of vowels

- The overall length of the pharyngeal-oral tract
- The location of the constrictions along the tract
- The narrowness of the constrictions

For example, the formant frequencies for a male speaker for the neutral vowel occur near 500, 1500, 2500, 3500 Hz and so on. F1 and F2 are closely tied to the shape of the vocal-tract articulators. The frequency location of the third formant, F3, is significant only to a few specific sounds. The fourth and higher formants remain relatively constant in frequency regardless of changes in articulation.

For the purpose of this project, we have extracted the first two formant frequencies for the vowel /i/ from the word 'cheese' for 12 male and female speakers and the neutral vowel /ə/ from the word 'the' for 9 male and female speakers.

3.7.1 Methodology

The method we have used for the extraction of the respective formants is known as the *Pole Extraction Method of Formant Estimation.*

This method involves the use of *Linear Prediction (LP) parameters*. The Linear Prediction (LP) model is an estimated speech production model which uses an all-pole, minimum phase system. The use of an all-pole model is primarily a matter of analytical necessity. In a practical sense, it has been observed that phase relationships among components of speech have virtually no effect on speech perception i.e. the human ear is fundamentally 'phase deaf'. Therefore, whatever information is aurally gleaned from the speech is extracted from its magnitude spectrum. Hence, the spectrum can be exactly modelled with stable poles.

The all-pole model is given as:

$$\Theta(z) = 1/(1 - \Sigma_i\, a(i)\, z^{-i}) \; ; \; I = 1,2,\ldots M$$

where the a(i)'s form the predictor equation coefficients.

The steps followed for the extraction of the formant frequencies using the above stated method are:

- The desired vowel is extracted from the given word.
- For that vowel, the resonant pole pairs of the representative LP model (two, three or four depending on the Nyquist frequency) are extracted.
- These are selected as representative of formants.
- The formant frequency is deduced from the angle of the pole pair and the bandwidth is related to the pole pair's magnitude.
- This algorithm is generally used as a research tool and not in real-time systems.

3.8 THE ADAPTIVE NEURO-FUZZY INFERENCE SYSTEM (ANFIS)

An adaptive neuro-fuzzy inference system or an adaptive network-based fuzzy inference system is a kind of artificial neural network that is based on the Takagi-Sugeno fuzzy inference system. Since it integrates both neural network and fuzzy logic principles, it has the potential to capture the benefits of both in a single framework. Given an input/output dataset, the ANFIS system constructs a fuzzy inference system whose membership functions are then tuned using a back propagation algorithm alone or in combination with a least squares method. This allows the fuzzy system to learn from the data it is modelling. Hence, it is often seen as a universal estimator.

3.8.1 Neuro-Fuzzy Systems

The fuzzy sets theory initiated by Lofti Zadeh in 1965 provided systems that could manipulate practical knowledge with some level of uncertainty. The fuzzy inference mechanism consists of the three following steps:

- The values of the numerical inputs are mapped by a function according to a degree of compatibility of the respective fuzzy sets, the operation known as fuzzification.
- The fuzzy system processes some rules in accordance with the firing strengths of the inputs.
- The resultant fuzzy values are transformed again into numerical values, this operation known as defuzzification.

These fuzzy systems provided the following advantages over classic techniques:

- Capacity to represent inherent uncertainties of the human knowledge with linguistic variables.
- Simple interaction of the expert of the domain with the engineer designer of the system.
- Easy interpretation of the results, because of the natural rules representation.
- Easy extension of the base of knowledge through the addition of new rules.
- Robustness in relation of the possible disturbances in the system.

But they also included disadvantages like:

- Incapable of generalizing.
- Not robust in relation to topological changes of the system, such changes would demand alterations in the rule base.
- Depended on the existence of an expert to determine the inference logical rules.

At the same time, artificial neural networks existed which tried to shape the biological functions of the human brain. They modified their internal structure and the weights of the connections between the artificial neurons to map the input/output relations. But there was difficulty in determining the number of neurons and layers needed. Also, the system did not provide the flexibility that fuzzy systems offered. However, these systems provided the following advantages which could counteract the disadvantages of fuzzy systems:

- Learning capacity
- Generalization capacity
- Robustness in relation to disturbances

This led to the idea of combining the above two and using neural networks, with their efficient learning algorithms, to automate or support the development of tuning fuzzy systems. Neural systems introduced their computational characteristics of learning in the fuzzy systems and received from them the interpretation and clarity of systems representation. As a result, neuro-fuzzy systems were born.

3.8.2 Types of Neuro-Fuzzy Systems

The different combinations of neuro-fuzzy systems are divided into the following three classes [10]:

Co-operative Neuro-Fuzzy Systems

In a co-operative system, the neural networks are used only in the pre-processing initial phase to determine the sub-blocks of the fuzzy system using the training data. After this the neural networks are removed and only the fuzzy system is executed.

Concurrent Neuro-Fuzzy Systems

In a concurrent system, the neural network works together with the fuzzy system. Here, the input enters the fuzzy system, is pre-processed and then the neural network processes the outputs of the system.

Hybrid Neuro-Fuzzy Systems

A hybrid neuro-fuzzy system is a fuzzy system that uses a learning algorithm based on gradients or inspired by the neural networks (heuristical learning strategies) to determine its parameters (fuzzy sets and fuzzy rules) through the patterns processing (input and output) [10].It can be interpreted as a set of fuzzy rules and has the advantages of learning through patterns and easy interpretation of functionality.

3.8.3 ANFIS Architecture

The ANFIS system is a hybrid neuro-fuzzy system which implements a Takagi-Sugeno fuzzy inference system [11]. It has five layers, given as follows:

- Layer 0-Input Layer (Passive Layer)
- Layer 1-Membership Function Layer
- Layer 2-Rule Layer
- Layer 3-Norm Layer
- Layer 4-Output Layer
- Layer 5-Final Output Layer

Layer 0 consists of all the inputs that are mapped into their respective membership functions. Being a passive layer, it is not counted into the main five layers.
Layer 1 is responsible for mapping the input variable relatively to each of its membership functions.
Layer 2 uses the T-Norm operator to calculate the antecedents of the fuzzy rules.
Layer 3 normalizes the rule strengths to get membership degrees less than or equal to 1.
Layer 4 consists of the determination of the consequents of the fuzzy rules.
Layer 5 calculates the global output as the summation of all the signals that arrive to this layer.

ANFIS uses the error backpropagation training algorithm to determine its input membership function parameters and the consequent parameters. It does so in two steps – forward propagation and back propagation. The error parameter is obtained at the end of the forward propagation and the weights are updated accordingly in the back propagation step.
It also uses the hybrid training method where backpropagation is used to determine the input membership function parameters and least mean squared method is used to determine the consequent parameters. Each step of this iterative learning algorithm has two parts. In the first part, the input patterns are propagated and the parameters of the consequents are calculated using the least mean squared algorithm, while the parameters of the premises are considered fixed. In the second part, the input patterns are propagated again and in each iteration the backpropagation learning algorithm is used to modify the parameters of the premises, while the consequents remain fixed.

The complete architecture is shown below in Fig 3.2.

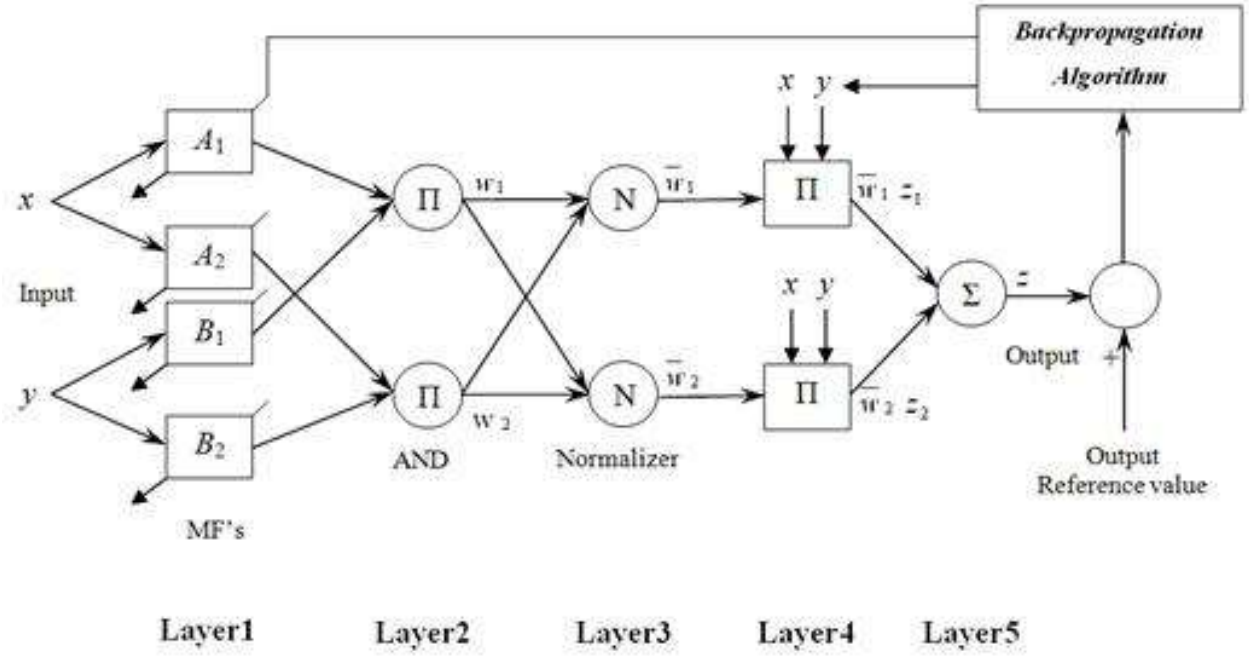

Fig 3.2 ANFIS Architecture

3.8.4 Methodology

For the purpose of our project, we have implemented the multi-feature ANFIS system using the neuro-fuzzy toolbox available in Matlab 2015a. The steps are as follows:

- The training data is consolidated in the form of an input/output matrix.
- This training matrix is passed as an input parameter to the ANFIS to generate the input fuzzy inference structure.
- This is generated using the function *genfis1*.
- Each input parameter for the above is divided into three input and corresponding output membership functions – low, medium and high. The input membership functions are of the bell shape type and the output membership functions are linear (values ranging from 0 to 1).
- Training options like the maximum number of epochs, error goal, step size, step size decrease rate and step size increase rate are specified.
- Display options are provided according to preference.
- The optimization method is specified as hybrid or backpropagation. In this case, backpropagation method has been used.
- The *anfis* function then returns the output fuzzy inference system into which the test data is passed to obtain the corresponding outputs as given by the ANFIS system.
- The above steps are carried on for each of the five features and the results obtained from the five parallel ANFIS structures is then summed up and evaluated using a neuron or processing unit.
- The neuron checks the values and if found to be above the given threshold, the speaker is classified as ‘female’, otherwise ‘male’.
- The results for the given test set are displayed on the command window.

3.9 ARTIFICIAL NEURAL NETWORKS (ANNs) AND SUPPORT VECTOR MACHINES (SVMs)

ANNs and SVMs are the most commonly used systems for classification purposes. Therefore, we compare the performance of these two techniques with the ANFIS technique as a final step.

3.9.1 Artificial Neural Networks (ANNs)

Artificial neural networks are a family of statistical learning algorithms inspired by biological neural networks and are generally used to approximate functions that can depend on a large number of inputs and are generally unknown. They are represented as systems of interconnected neurons (processing units) which can compute values from inputs and are adaptive in nature.

They have often been used to solve a wide variety of tasks, optimization or classification, which are hard to solve using ordinary rule-based programming, including computer vision and speech recognition.

3.9.2 Support Vector Machines (SVMs)

Support Vector Machines are systems generally used for classification and regression analysis. They use a 'safety cushion' while separating the data. Given a set of training examples, each marked as belonging to one of two categories, an SVM training algorithm builds a model that assigns new examples into one category or the other, making it a non-probabilistic binary linear classifier. The SVM model maps points in space so that the examples of the separate categories are divided by a clear gap as wide as possible. Using this, it constructs a hyper plane or a set of hyper planes which are then used for classification or regression purposes.

3.9.3 Methodology

The ANN and SVM systems are also implemented in the multi-feature format as the ANFIS structure. The toolboxes used for the purpose are the Neural Network Toolbox and Support Vector Machine Toolbox. The functions used are *trainlm* for Neural Networks (which uses the Levenberg-Marquardt training algorithm) and *fitcsvm* for modelling Support Vector Machines.

3.10 PROCESS SUMMARY

The entire process can be summarized in form of the block diagram displayed below in Fig 3.3:

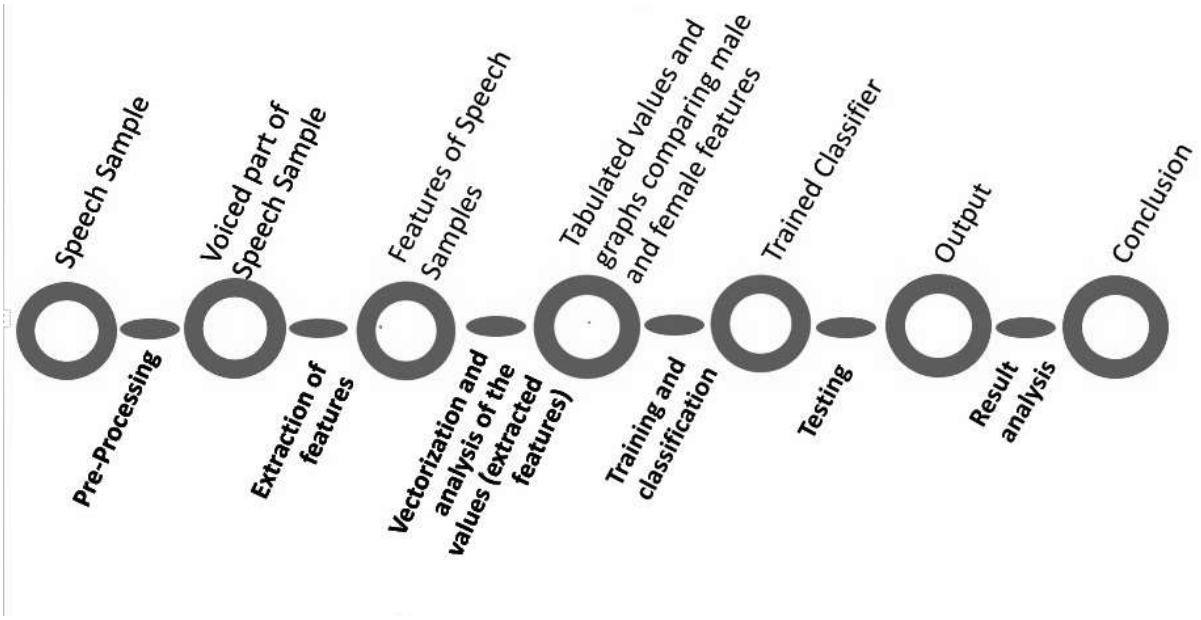

Fig 3.3 Process Summary

CHAPTER 4
RESULT ANALYSIS

4.1 INTRODUCTION

This Chapter deals with the results obtained by the implementation of the methodologies discussed in the previous chapter. The values for the pitch, short-time energy, zero crossing rate, energy entropy and formants for the given data sets are tabulated and graphs are plotted for comparison. An analysis of the data obtained is carried out to determine the relation between the genders and features of different speakers. The ANFIS classifier is trained and tested and the efficiency obtained is compared with the efficiencies of the SVM and ANN classifiers. Finally, the significant observations made during the complete analysis are stated.

4.2 FEATURE TABLE

Database: Modified TIMIT database (derived from TIDIGITS)
Words uttered: ‘oh’, ‘zero’, ‘two’, ‘three’, ‘four’, ‘five’, ‘six’, ‘seven’, ‘eight’, ‘nine’
Sampling Frequency: 8 kHz

Table 4.1 Feature Table for TIDIGITS Dataset

SPEAKERS	PITCH (Hz)	ZCR (norm)	STE (norm)	EE (norm)	FORMANT (Hz)
Male 01	138.54	0.0182	0.000102	0.000106	433.20
Male 02	99.84	0.0175	0.0000454	0.000132	417.30
Male 03	101.45	0.0347	0.000252	0.000113	605.70
Male 04	92.20	0.0219	0.0000414	0.000173	584.80
Male 05	110.79	0.0053	0.0000104	0.0000139	447.90
Male 06	177.58	0.0089	0.0000271	0.00000918	469.10
Male 07	95.68	0.0122	0.0000498	0.000110	562.10
Male 08	95.23	0.0137	0.0000538	0.000147	573.70
Male 09	100.85	0.0156	0.0000407	0.000135	472.10
Male 10	104.02	0.0094	0.0000226	0.000122	423.20
Male 11	93.07	0.0236	0.0000394	0.000231	529.90
Male 12	97.55	0.0201	0.000030	0.000230	519.10
Male 13	102.62	0.0215	0.0000427	0.000264	437.50
Male 14	101.37	0.0141	0.0000416	0.000101	407.70
Male 15	103.30	0.0195	0.0000445	0.000163	494.60
Male 16	100.05	0.0203	0.0000609	0.000188	461.40
Male 17	115.80	0.0272	0.000249	0.000231	302.90
Male 18	164.93	0.0409	0.000421	0.000257	333.10
Male 19	148.78	0.0358	0.000474	0.000212	319.20

Male 20	154.17	0.0275	0.000405	0.000166	325.40
Male 21	151.31	0.0352	0.000743	0.000247	396.60
Male 22	148.26	0.0313	0.000704	0.000211	384.90
Male 23	140.49	0.0568	0.000446	0.000235	592.20
Male 24	150.56	0.0562	0.000641	0.000201	586.60
Male 25	135.45	0.0108	0.000179	0.000115	487.00
Male 26	156.62	0.0126	0.000262	0.000118	456.30
Male 27	153.75	0.0226	0.000315	0.000658	572.10
Male 28	160.43	0.0255	0.000340	0.0000859	554.40
Male 29	136.97	0.0378	0.000301	0.000179	450.00
Male 30	139.70	0.0420	0.000319	0.000198	437.90
Male 31	110.49	0.0388	0.0000971	0.000284	535.10
Male 32	159.63	0.0421	0.000361	0.000255	520.90
Male 33	146.27	0.0391	0.0011	0.000209	436.20
Male 34	152.32	0.0406	0.0011	0.000259	410.30
Male 35	143.06	0.0361	0.000505	0.000242	482.10
Male 36	154.12	0.0380	0.0010	0.000236	492.60
Male 37	129.11	0.0232	0.000245	0.000153	313.30
Male 38	126.92	0.0205	0.000329	0.000196	306.90
Male 39	128.06	0.0229	0.000121	0.000102	328.70
Male 40	125.80	0.0264	0.000169	0.000170	310.30
Male 41	130.33	0.0376	0.000624	0.000252	388.30
Male 42	124.76	0.0269	0.000392	0.000161	396.60
Male 43	118.07	0.0456	0.000247	0.000215	586.80
Male 44	123.52	0.0477	0.000302	0.000202	602.20
Male 45	137.40	0.0162	0.000103	0.0000713	492.30
Male 46	125.76	0.0114	0.0000817	0.0000912	433.30
Male 47	128.03	0.0328	0.000476	0.000162	568.10
Male 48	129.47	0.0263	0.000329	0.000173	576.80
Male 49	132.80	0.0224	0.0000934	0.0000497	415.00
Male 50	125.99	0.0148	0.0000343	0.000176	392.00
Male 51	120.01	0.0507	0.000125	0.000243	509.90
Male 52	117.45	0.0415	0.0000942	0.000232	527.30
Male 53	117.25	0.0333	0.000273	0.000165	444.60
Male 54	118.90	0.0268	0.000187	0.000109	420.70
Male 55	127.27	0.0385	0.000385	0.000212	470.40
Male 56	122.00	0.0351	0.000219	0.000229	484.10
Male 57	100.00	0.0218	0.000056	0.000175	300.00
Male 58	99.83	0.0272	0.0000618	0.000151	327.90
Male 59	101.75	0.0214	0.000504	0.000115	333.30

Male 60	116.16	0.0205	0.000108	0.000154	365.00
Male 61	93.18	0.0262	0.000186	0.000222	583.90
Male 62	95.58	0.0417	0.000157	0.000237	579.10
Male 63	105.73	0.0131	0.000043	0.000052	461.10
Male 64	104.90	0.0195	0.000056	0.0000439	424.80
Male 65	107.60	0.0178	0.0000817	0.000112	580.00
Male 66	99.02	0.0195	0.0000958	0.000134	573.10
Male 67	103.83	0.0299	0.0000583	0.000164	463.90
Male 68	89.80	0.0240	0.0000371	0.000215	480.60
Male 69	93.87	0.0276	0.0000763	0.000140	520.10
Male 70	92.26	0.0346	0.0000904	0.000190	549.10
Male 71	98.96	0.0361	0.000262	0.000232	409.10
Male 72	92.26	0.0316	0.000191	0.000233	433.20
Male 73	99.68	0.0260	0.0000717	0.000156	496.60
Male 74	114.32	0.0240	0.000717	0.000207	501.10
Male 75	108.49	0.0210	0.0000273	0.000235	301.90
Male 76	105.29	0.0143	0.0000110	0.000234	327.50
Male 77	108.83	0.0218	0.0000464	0.000195	333.30
Male 78	102.23	0.0312	0.00000973	0.000166	349.30
Male 79	103.49	0.0178	0.0000242	0.000221	371.30
Male 80	101.38	0.0177	0.0000191	0.000183	390.00
Male 81	98.19	0.0303	0.0000627	0.000185	583.10
Male 82	96.98	0.0207	0.0000163	0.000231	573.70
Male 83	107.75	0.0109	0.0000148	0.0000488	464.70
Male 84	111.12	0.0148	0.0000158	0.000102	430.00
Male 85	101.77	0.0167	0.0000318	0.000165	583.10
Male 86	102.17	0.0213	0.0000478	0.000159	554.90
Male 87	139.15	0.0419	0.0000502	0.000255	378.80
Male 88	105.65	0.0344	0.0000471	0.000228	386.10
Male 89	101.00	0.0289	0.0000413	0.000227	510.10
Male 90	100.98	0.0315	0.000034	0.000195	534.30
Male 91	100.69	0.0279	0.0000528	0.00025	399.70
Male 92	100.29	0.0224	0.0000386	0.0000251	414.40
Male 93	110.61	0.0346	0.000130	0.000235	484.10
Male 94	100.96	0.0198	0.0000117	0.000245	490.30
Male 95	107.46	0.0168	0.0000447	0.000220	317.70
Male 96	99.86	0.0161	0.0000305	0.000247	311.10
Male 97	104.97	0.0173	0.0000379	0.000229	328.10
Male 98	100.78	0.0174	0.0000314	0.0000204	336.20
Male 99	116.25	0.0193	0.000102	0.000127	380.10

Male 100	99.53	0.0229	0.00000845	0.000173	310.00
Female 01	179.25	0.0250	0.000170	0.000108	496.20
Female 02	237.09	0.0301	0.0000633	0.000175	411.60
Female 03	235.17	0.0503	0.000201	0.000243	368.40
Female 04	235.59	0.0282	0.0000543	0.000200	370.10
Female 05	232.91	0.0315	0.000109	0.000201	358.40
Female 06	220.93	0.0364	0.000177	0.000246	461.10
Female 07	211.75	0.0566	0.0000838	0.000271	493.80
Female 08	220.65	0.0595	0.000150	0.000278	700.20
Female 09	219.11	0.0107	0.0000256	0.0000862	698.70
Female 10	230.23	0.0111	0.0000434	0.0000155	558.60
Female 11	232.63	0.0244	0.0000688	0.000172	507.30
Female 12	223.83	0.0240	0.0000829	0.000174	655.40
Female 13	219.95	0.0274	0.0000322	0.000135	643.80
Female 14	219.24	0.0304	0.000113	0.000203	584.10
Female 15	226.49	0.0436	0.0000458	0.000303	592.30
Female 16	225.38	0.0619	0.0000484	0.000318	596.80
Female 17	217.99	0.0341	0.0000698	0.000246	605.50
Female 18	224.14	0.0411	0.000115	0.000242	486.10
Female 19	220.34	0.0320	0.0000929	0.000182	490.00
Female 20	215.04	0.0269	0.0000314	0.000204	496.70
Female 21	236.16	0.0327	0.000123	0.000198	462.60
Female 22	210.74	0.0406	0.000334	0.000206	370.40
Female 23	204.25	0.0362	0.000355	0.000217	359.20
Female 24	188.45	0.0388	0.000269	0.000161	363.30
Female 25	183.94	0.0342	0.000122	0.000221	487.00
Female 26	194.99	0.0446	0.000327	0.000195	490.10
Female 27	195.50	0.0424	0.000223	0.000252	507.30
Female 28	179.10	0.0671	0.000348	0.000160	793.00
Female 29	182.62	0.0641	0.000415	0.000166	684.10
Female 30	197.82	0.0187	0.000103	0.0000928	542.90
Female 31	196.15	0.0188	0.000153	0.0000489	523.10
Female 32	198.57	0.0418	0.000238	0.000175	678.00
Female 33	186.13	0.0466	0.000262	0.000159	693.10
Female 34	195.62	0.0333	0.000118	0.000131	550.00
Female 35	196.36	0.0267	0.000126	0.000127	538.60
Female 36	187.52	0.0670	0.000116	0.000174	551.10
Female 37	191.56	0.0601	0.000131	0.000225	527.50
Female 38	187.15	0.0333	0.000228	0.000160	511.10
Female 39	183.51	0.0424	0.000283	0.000188	479.20

Female 40	182.75	0.0521	0.000205	0.000149	483.20
Female 41	163.09	0.0549	0.000183	0.000173	470.00
Female 42	212.44	0.0371	0.000431	0.000131	400.00
Female 43	237.91	0.0234	0.000162	0.000208	393.80
Female 44	189.95	0.0503	0.000276	0.000143	371.90
Female 45	199.31	0.0450	0.000249	0.000147	396.60
Female 46	195.73	0.0332	0.000398	0.000241	505.00
Female 47	203.69	0.0367	0.000932	0.000206	503.90
Female 48	179.89	0.0687	0.000639	0.000141	666.60
Female 49	201.06	0.0760	0.00160	0.000150	677.50
Female 50	193.63	0.0250	0.000596	0.0000344	517.10
Female 51	192.04	0.0191	0.000199	0.0000560	521.10
Female 52	190.49	0.0327	0.000336	0.000151	634.40
Female 53	192.17	0.0318	0.000301	0.000137	625.20
Female 54	234.64	0.0422	0.000330	0.000114	542.20
Female 55	251.47	0.0668	0.000705	0.000253	563.10
Female 56	192.41	0.0584	0.000249	0.000190	604.30
Female 57	169.79	0.0821	0.000255	0.000186	579.10
Female 58	190.07	0.0495	0.00260	0.000228	493.10
Female 59	189.60	0.0367	0.00130	0.000158	510.00
Female 60	167.02	0.0628	0.000518	0.000157	483.70
Female 61	223.62	0.0551	0.00110	0.000207	471.10
Female 62	226.96	0.0320	0.000217	0.000252	434.80
Female 63	206.09	0.0302	0.000105	0.000246	421.70
Female 64	196.13	0.0324	0.000596	0.000288	354.30
Female 65	192.54	0.0322	0.000115	0.000265	381.10
Female 66	211.46	0.0321	0.0000601	0.000247	473.90
Female 67	197.54	0.0267	0.000110	0.000180	489.90
Female 68	196.59	0.0418	0.0000380	0.000295	663.00
Female 69	210.44	0.0544	0.000130	0.000294	696.60
Female 70	219.35	0.0103	0.000357	0.000127	521.10
Female 71	216.54	0.0101	0.000384	0.000123	530.00
Female 72	216.22	0.0254	0.0000515	0.000212	660.60
Female 73	204.67	0.0305	0.000810	0.000249	626.10
Female 74	203.30	0.0260	0.0000755	0.000226	579.10
Female 75	226.72	0.0280	0.0000845	0.000237	591.30
Female 76	201.26	0.0275	0.0000440	0.000299	600.70
Female 77	186.55	0.0399	0.0000870	0.000297	585.50
Female 78	203.43	0.0456	0.0000830	0.000208	486.30
Female 79	202.07	0.0482	0.0001550	0.000259	447.50

Female 80	191.16	0.0327	0.0000813	0.000180	490.00
Female 81	202.07	0.0369	0.000129	0.000185	483.00
Female 82	191.08	0.0178	0.000275	0.000173	407.90
Female 83	196.98	0.0201	0.000135	0.000131	393.30
Female 84	160.29	0.0310	0.000351	0.000164	354.60
Female 85	182.55	0.0271	0.0000555	0.000233	360.00
Female 86	177.68	0.0288	0.0000550	0.000183	493.10
Female 87	173.01	0.0359	0.000115	0.000226	471.10
Female 88	158.16	0.0598	0.000188	0.000161	706.70
Female 89	163.01	0.0460	0.0000959	0.000164	681.90
Female 90	196.62	0.0082	0.0000954	0.0000549	545.50
Female 91	173.93	0.0080	0.0000239	0.0000791	563.10
Female 92	169.93	0.0246	0.000102	0.000156	635.20
Female 93	174.66	0.0309	0.0000846	0.000162	650.00
Female 94	180.68	0.0281	0.0000446	0.000173	588.10
Female 95	162.50	0.0391	0.000107	0.000227	579.30
Female 96	152.38	0.0635	0.0000994	0.000245	573.30
Female 97	164.40	0.0572	0.0000967	0.000258	551.40
Female 98	162.89	0.0267	0.0000501	0.000149	494.40
Female 99	181.67	0.0364	0.000226	0.000226	500.00
Female 100	166.79	0.0337	0.0000551	0.000204	487.10

4.3 FEATURE ANALYSIS

4.3.1 Pitch

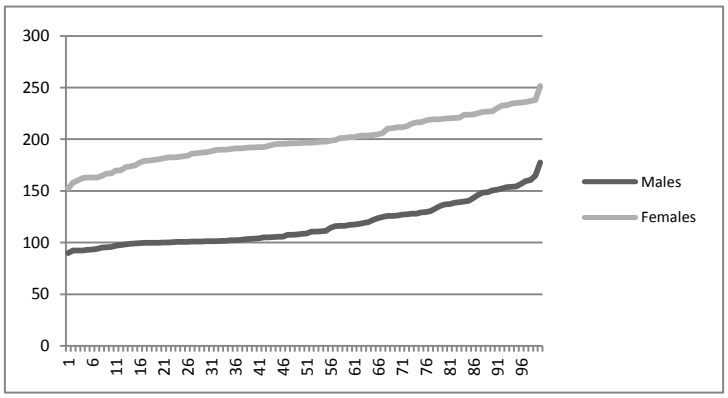

Fig 4.1 Pitch – Male vs Female

As seen in Fig 4.1, the pitch frequencies of males when plotted against the female speakers show greater values for females. The minimum as well as the maximum values are higher for the female speakers and a clear distinction is evident between the two classes of speakers

4.3.2 Zero Crossing Rate

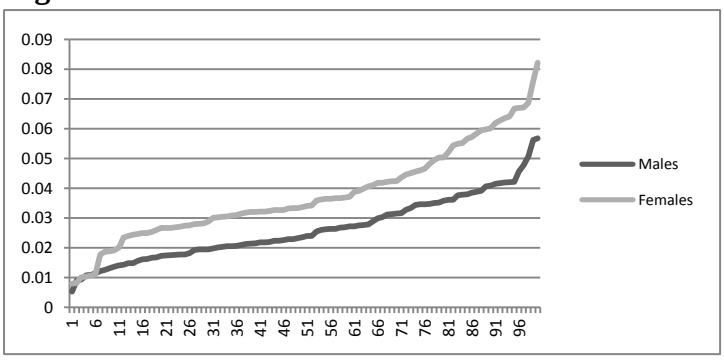

Fig 4.2 ZCR (Normalized) – Male vs Female

As seen in Fig. 4.2, the Zero Crossing Rate values for female speakers are higher than those of their male counterparts. Hence, a clear distinction is evident in this case as well between the two classes of speakers.

4.3.3 Short-Time Energy

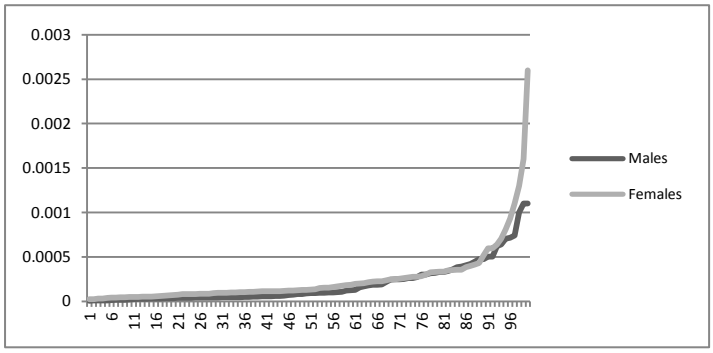

Fig 4.3 STE (Normalized) – Male vs Female

As seen in Fig. 4.3, the Short-Time Energy values for female speakers are slightly higher than those of their male counterparts. The minimum value is slightly higher and the maximum value significantly higher with only a small overlapping segment. Hence, a distinction can be made between the two classes.

4.3.4 Energy Entropy

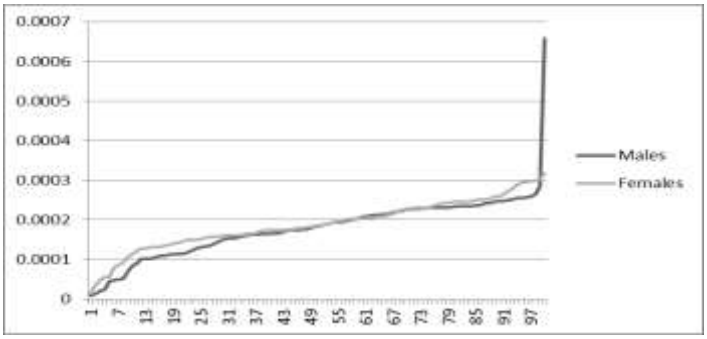

Fig 4.4 EE (Normalized) – Male vs Female

As seen in Fig 4.4, the Energy Entropy values for females are generally slightly higher than those for males except at the end, where the maximum value for males is higher than that of females. Overall, including the minimum value, the values for females are seen to be either higher or at the same level as that of the males, leading to the possibility of distinction in all mid-value cases and some aberrations in the exceptional cases.

4.3.5 Formants

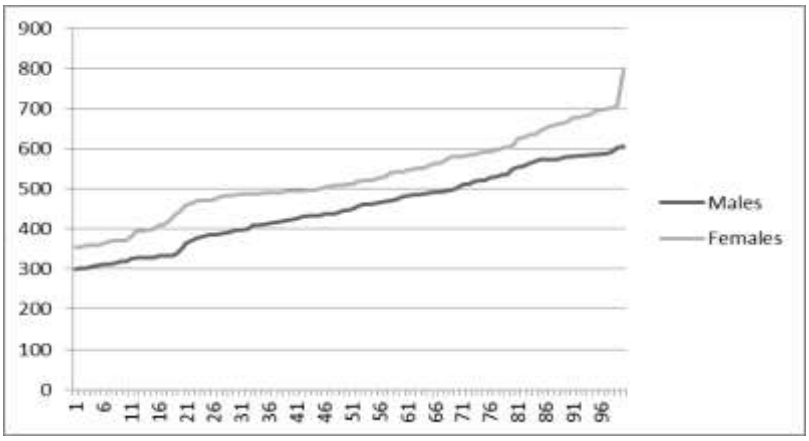

Fig 4.5 First Formant – Male vs Female

In Fig 4.5, the values of the First Formants are plotted for the male and female speakers. The values are seen to be significantly higher for females as compared to males. Hence, a clear distinction is seen in this case too between the two classes of speakers.

Finally, we observe from the feature analysis that at least four out of the five features provide suitable distinction between the male and female speakers. Hence, we proceed towards the implementation of the classifier using Matlab 2015a and the results obtained.

4.4 IMPLEMENTATION AND RESULTS

4.4.1 Artificial Neural Network (Multi-Feature)

No. of training samples: 180
No. of testing samples: 20
Threshold Value: 2.65
Training method used: Levenberg-Marquardt
Function used: *trainlm*
Toolbox used: Neural Network Toolbox, Matlab 2015a

Output obtained:

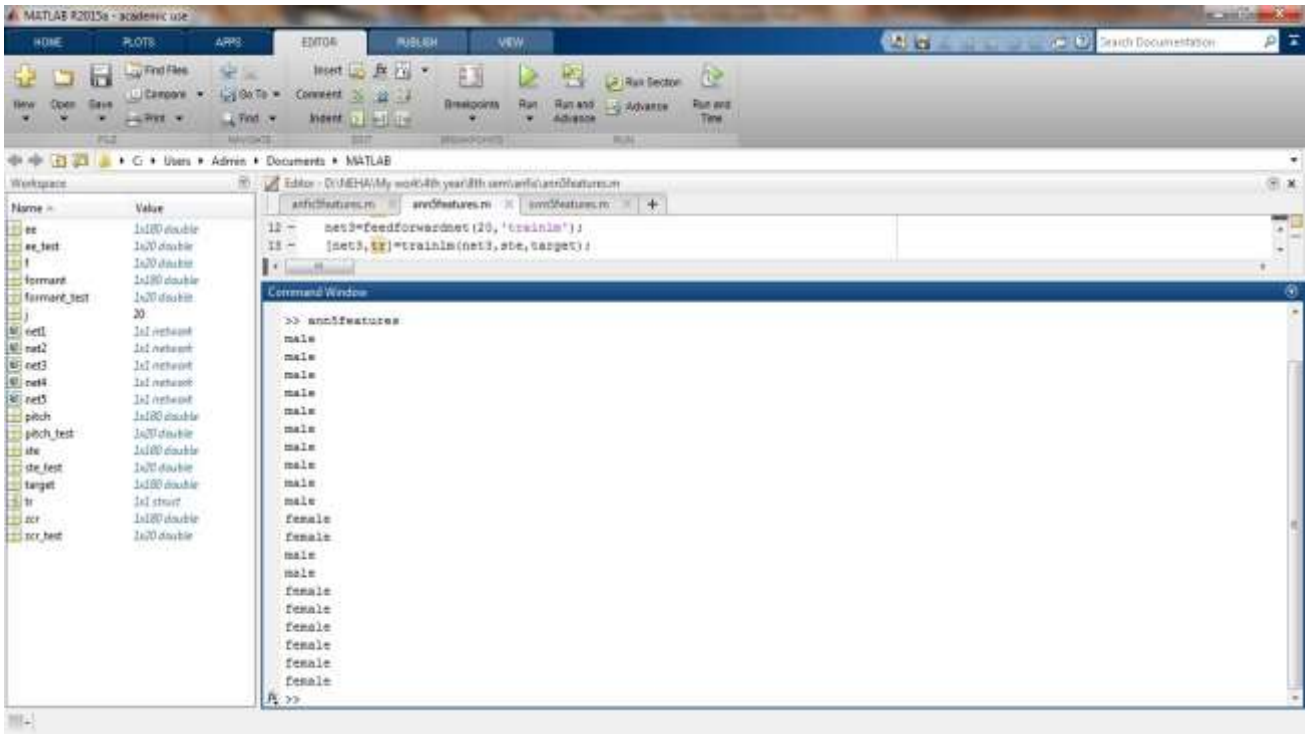

Fig 4.6 Output displayed in command window for ANN using the 5 features

Out of the 20 testing samples for the ANN structure, the first 10 are males while the last 10 are females. As seen in Fig 4.6, the first 10 males have been correctly classified whereas out of the last 10 females, 2 have been incorrectly labelled as males. Hence, the efficiency of the neural network is obtained using the following formula:

$$Efficiency = (TP+TN) / (TP+FP+TN+FN)$$

Where Females – Positive Class, Males – Negative Class
TP – True Positive i.e. no. of correct positive cases
FP – False Positive i.e. no. of incorrect positive cases
TN – True Negative i.e. no. of correct negative cases
FN – False Negative i.e. no. of incorrect negative cases

Therefore, Efficiency = ((8+10) / 20)*100 %
= **90.00 %**

4.4.2 Adaptive Neuro-Fuzzy Inference System (Only Pitch)

No. of training samples: 180
No. of testing samples: 20
No. of training epochs: 150
Threshold value: 0.55
Training method used: Error Backpropagation Training Algorithm (EBPTA)
Function used: *anfis*
Toolbox used: Neuro-Fuzzy Toolbox, Matlab 2015a

ANFIS Input Arguments:

- ***trnData***: Training data, specified as a matrix. For an FIS with N inputs, trnData has N+1 columns, where the first N columns contain input data and the final column contains output data.
- ***initFis***: FIS structure used to provide an initial set of membership functions for training, specified as one of the following:
  - Positive integer — Specifies the number of membership functions for all inputs and generates an initial FIS using genfis1.
  - Vector of positive integers — Specifies the number of membership functions for each input individually and generates an initial FIS using genfis1.
- ***trnOpt***: Training options, specified as a vector of scalars that represent the following settings:
  - trnOpt(1) — Training epoch number (default: 10)
  - trnOpt(2) — Training error goal (default: 0)
  - trnOpt(3) — Initial step size (default: 0.01)
  - trnOpt(4) — Step size decrease rate (default: 0.9)
  - trnOpt(5) — Step size increase rate (default: 1.1)
- ***dispOpt*** : Display options that specify information to display in the Command Window during training, specified as a vector of integers that represent these settings:
  - dispOpt(1) — ANFIS information, such as numbers of input and output membership functions
  - dispOpt(2) — Error values
  - dispOpt(3) — Step size at each parameter update
  - dispOpt(4) — Final results

  Each display option is specified as:
  1 (default) — Display the corresponding information.
  0 — Do not display the corresponding information.
  NaN — The default option is used.
- ***optMethod*** : Optimization method used in membership function parameter training, specified as an integer with the following values:

- 1 (default) — Hybrid method. This method is a combination of least-squares estimation and back-propagation.
- 0 — Back-propagation method

If any other value is specified, the default method is used.

ANFIS Output Arguments:

- ***fis:*** FIS structure whose parameters are tuned using the training data, returned as a structure.
- ***error:*** Root mean squared training data errors at each training epoch, returned as an array of scalars.
- ***stepsize:*** Step sizes at each training epoch, returned as an array of scalars.

No. of membership functions used: 3

- Low
- Medium
- High

Type of input membership function: *gbellmf (bell function) (as shown in Fig 4.7)*
Type of output membership function: linear

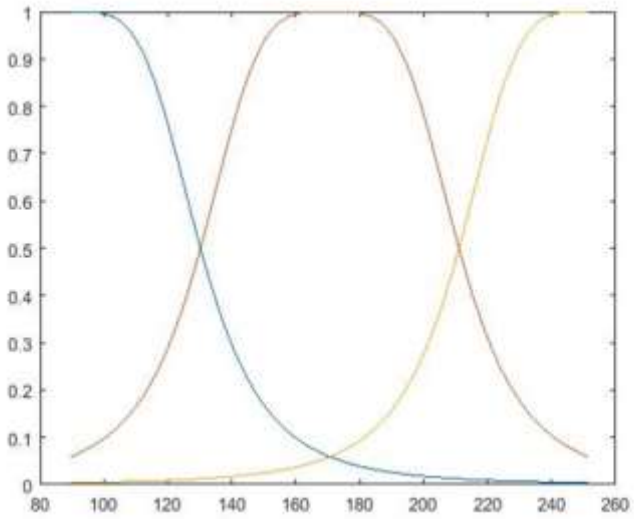

Fig 4.7 Input Membership Functions – Pitch

Rule Base:

- If (input1 is in1mf1) then (output is out1mf1)
- If (input1 is in1mf2) then (output is out1mf2)
- If (input1 is in1mf3) then (output is out1mf3)

Output Obtained:

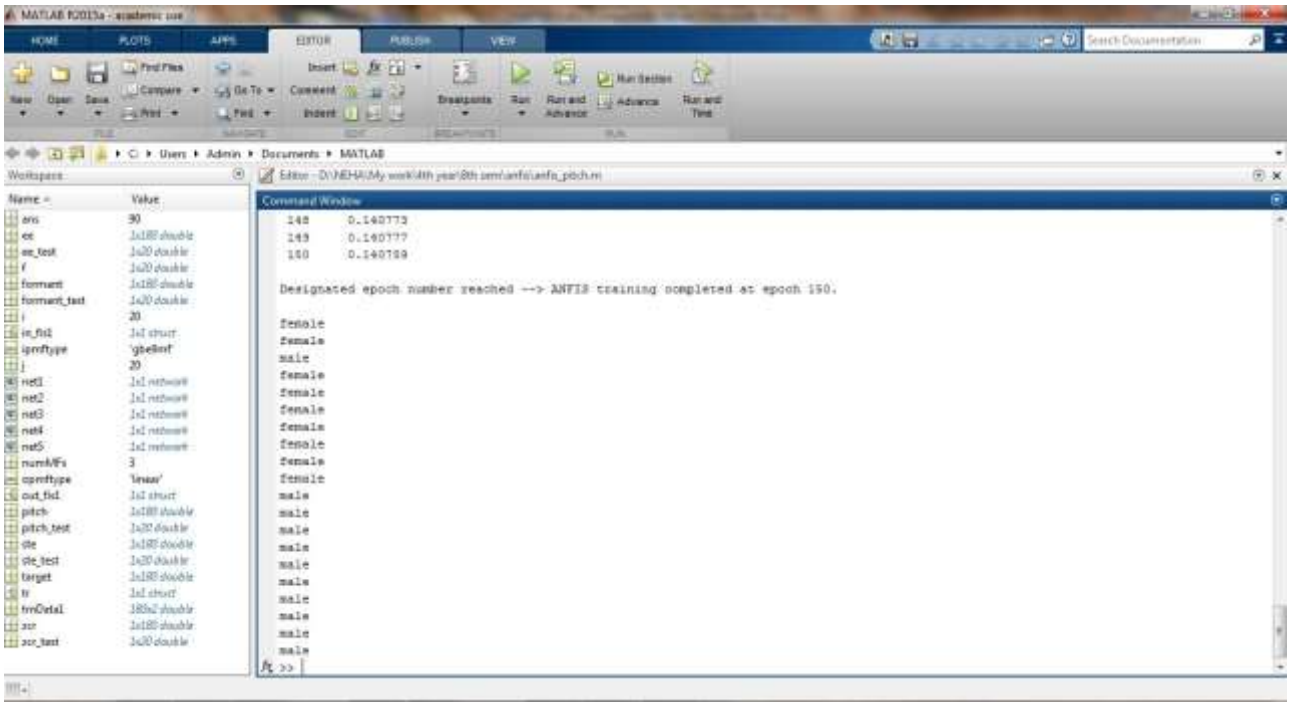

Fig 4.8 Output in command window for ANFIS using only Pitch as a feature

For the ANFIS structure, the testing dataset has 20 samples out of which the first 10 are females whereas the last 10 are males. As shown in Fig 4.8, this ANFIS structure classifies all the males correctly and 9 out of the 10 females correctly. Hence, its efficiency is given as:

Efficiency = ((9+10) / 20) * 100 %

= **95.00 %**

This is greater than that for neural networks and is close to the efficiency obtained in the base paper which classified speakers using a Pitch based ANFIS structure.

4.4.3 Adaptive Neuro-Fuzzy Inference System (Multi-Feature)

No. of training samples used: 180
No. of testing samples used: 20
No. of training epochs: 150
Threshold value: 2.65
Training method used: Error Backpropagation Training Algorithm (EBPTA)
Function used: *anfis*
Toolbox used: Neuro-Fuzzy Toolbox, Matlab 2015a

*Remaining parameters same as described in previous section.

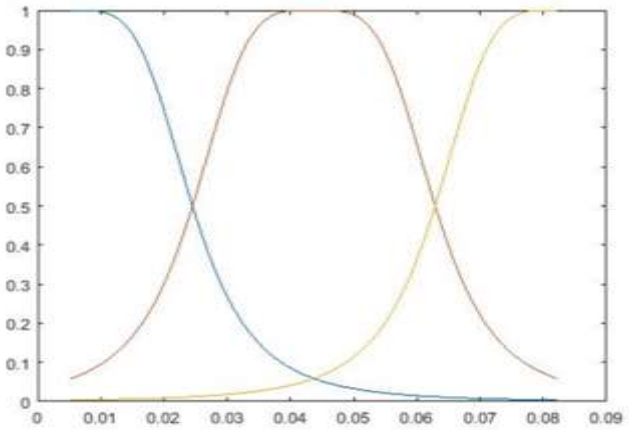

Fig 4.9 Input Membership Functions – ZCR

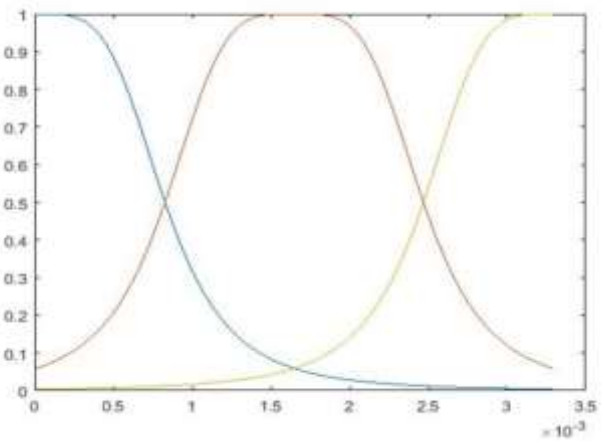

Fig 4.10 Input Membership Functions – STE

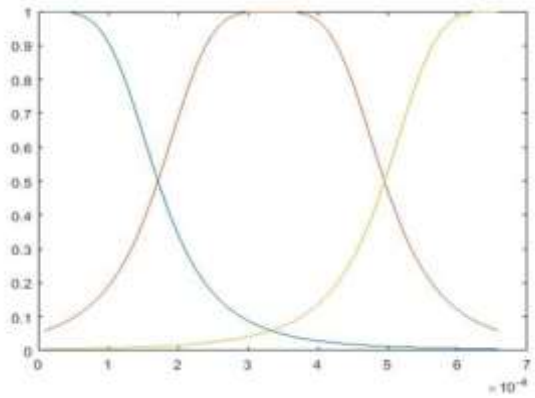

Fig 4.11 Input Membership Functions – EE

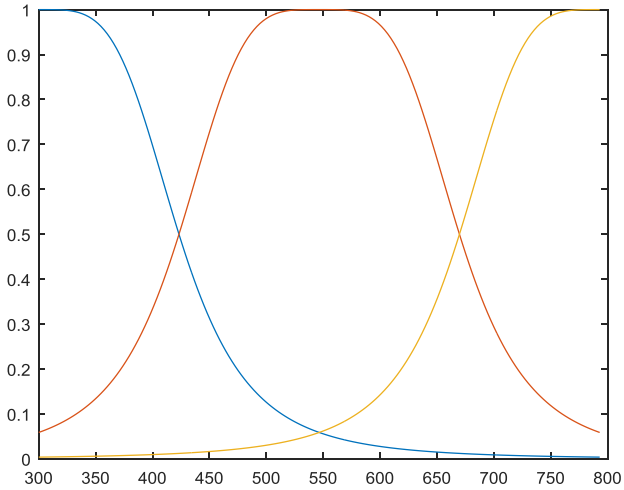

Fig 4.12 Input Membership Functions – Formants

*****Rule Base and Membership Functions of same type as in previous section.

Output Obtained:

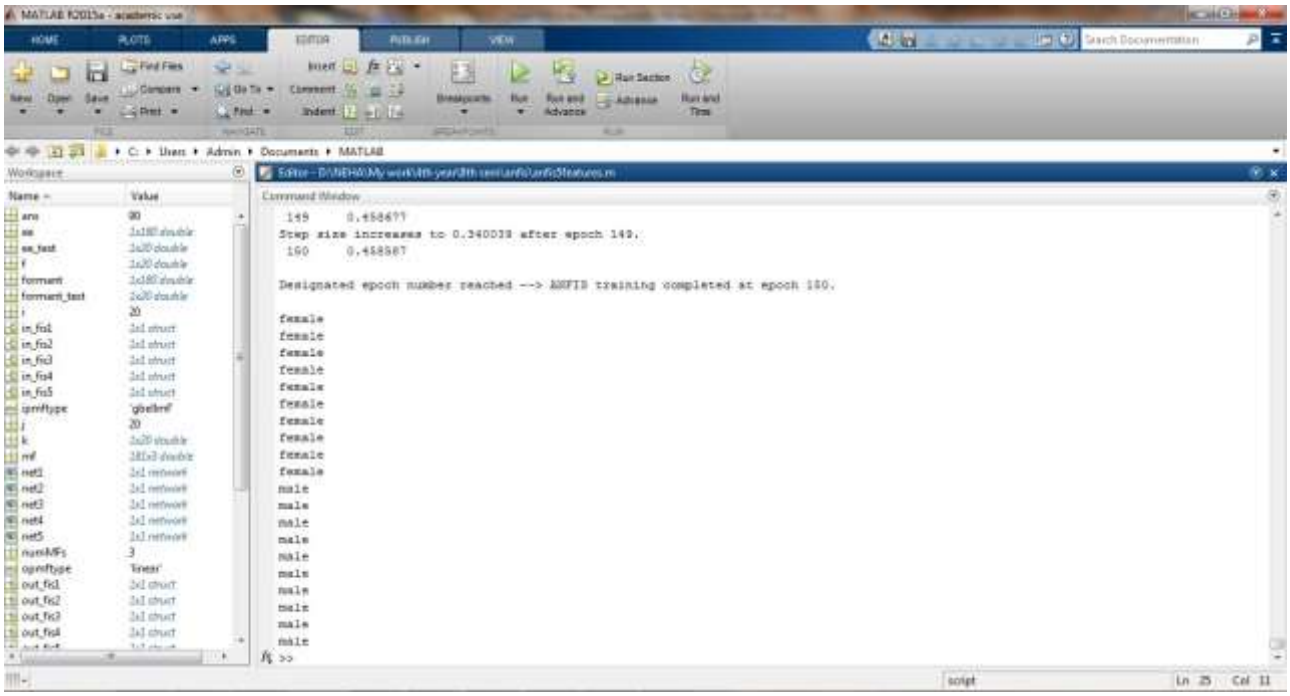

Fig 4.13 Output in command window for ANFIS using the 5 features

Out of the 20 testing samples for the multi-feature ANFIS structure, the first 10 are females while the last 10 are males. As seen in Fig 4.13, the ANFIS system classifies all of them correctly, given a threshold value of 2.65. On increasing the threshold value to 3.00 though, the second female speaker is incorrectly classified as male. Hence, we take the optimal threshold value as 2.65 for our operation. In this case, the efficiency obtained is as follows:

Efficiency = ((10+10) / 20) * 100 %
= **100.00 %**

This indicates that a multi-feature ANFIS classifier works to counter the defects of neural networks as well as a single-feature system. Probable reasons for this seem to be the combined benefits of a good learning algorithm as well as the concept of membership functions provided due to the fuzzy systems incorporated in the system. Neural networks on the other hand set hard limits for classification and thus do not provide enough scope for classification of uncertain cases.
Finally, we compare it with a Support Vector Machine classifier, which is one of the popular methods used for classification purposes aside from neural networks.

4.4.4 Support Vector Machine (Multi-Feature)

No. of training samples: 180
No. of testing samples: 20
Negative Class '0'- Male
Positive Class '1'-Female
Threshold Value: 2.00
Training method used: Support Vector Machines
Function used: *fitcsvm*
Toolbox used: Statistics and Machine Learning Toolbox, Matlab 2015a

Output Obtained:

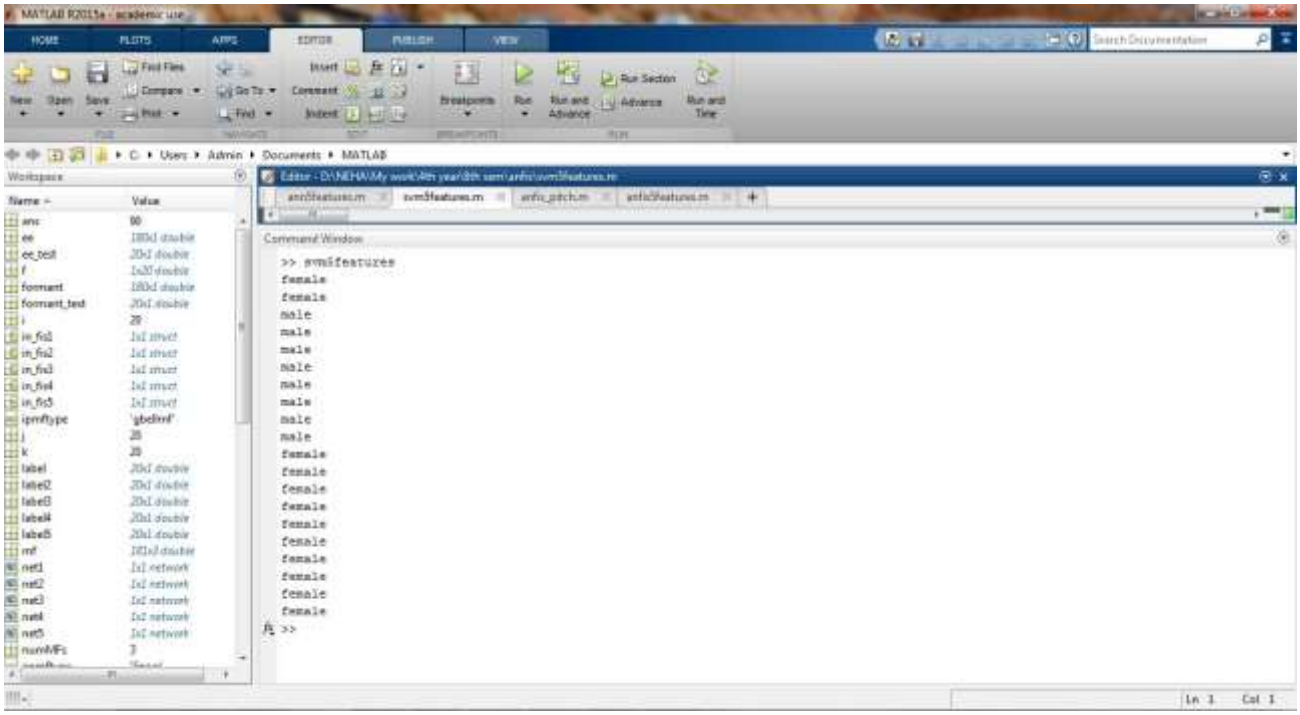

Fig 4.14 Output in command window for SVM using the 5 features

Out of the 20 testing samples for the SVM structure, the first 10 are males while the last 10 are females. As seen in Fig 4.14, the system classifies all the females correctly whereas only 8 males correctly. Hence, we can deduce the efficiency as:

Efficiency = ((8+10) / 20) * 100 %
= **90.00 %**

Thus, we see that SVM though an efficient classifier, makes errors in exceptional cases. The multi-feature ANFIS however manages to counteract these defects. A possible reason for this could be that the SVM is a non-probabilistic binary classifier and therefore provides hard definite values for the speakers instead of soft probabilistic values and so is less flexible than the ANFIS system.

The efficiencies obtained from the above four techniques are summarized in Table 4.2 given below:

Table 4.2 Efficiency Table

METHOD USED	EFFICIENCY OBTAINED
ANN (Multi-Feature)	90.00 %
SVM (Multi-Feature)	90.00 %
ANFIS (using only Pitch as a feature)	95.00 %
ANFIS (Multi-Feature)	100.00 %

4.5 SIGNIFICANCE OF THE RESULT

The following points can be concluded from the results obtained:

- The pitch of the female speakers is significantly higher than the male speakers.
- The normalized zero crossing rate values are also higher for the female speakers than the male speakers.
- The normalized short-time energy values are slightly higher for the female speakers than the male speakers.
- The normalized energy entropy values are also in general slightly higher for the female speakers than the male speakers.
- The first formant frequencies are higher for the female speakers when compared to male speakers.
- The Neural Network and Support Vector Machine classifiers have an efficiency of 90.00% in the multi-feature format.
- The ANFIS network using only Pitch as a feature has an efficiency of 95.00%, showing improvement over the standard methods.
- The multi-feature ANFIS gives an efficiency of 100.00% for the given dataset, hence rectifying the error arising from the exceptional case when using only Pitch as a discriminating factor.

In conclusion, we see that the multi-feature ANFIS proposed in this project does achieve its objective of accurate and efficient classification of male and female speakers. Also, in comparison with other standard classifiers it gives much better accuracy.

CHAPTER 5
CONCLUSION AND FUTURE SCOPE OF WORK

5.1 INTRODUCTION

This Chapter deals with the finishing details of our project report. It summarizes the basic objectives of the entire project which is followed by a brief explanation of the significance of the results obtained up till now. Finally, we conclude with a look at the future scope of work.

5.2 SUMMARY

The main objective of the project is the design of an efficient gender classification system. Towards this purpose, we have extracted some important features of the speech samples and vectorized them to create our training data set. Also, we have compared the values obtained to get a rough idea of whether a distinction between the two genders is possible using these features and the previously stated methodologies. We have observed that a distinction is very much evident between the respective values for female and male speakers. The pitch, zero crossing rate, short-time energy, energy entropy and the first formant frequency show a general trend of higher values for female speakers and comparatively lower values for male speakers. We have designed four classifiers for the purpose of classification based on these features: A multi-feature ANN classifier, a multi-feature SVM classifier, a Pitch-based ANFIS classifier and a multi-feature ANFIS classifier. On comparison of the results obtained, we have observed the multi-feature ANFIS to have the highest efficiency for the given dataset.

5.3 CONCLUSIONS

The results confirm our expectations of the possibility of achieving a clear distinction between the male and female speakers using a multi-feature ANFIS system. In addition, it shows a higher accuracy and efficiency as compared to the neural network and SVM classifiers.
This provides us with two important conclusions:

- A multi-feature system is much more efficient than a system that relies on a single feature such as pitch, zero crossing rate, etc.

- A neuro-fuzzy system, which incorporates the uncertainty handling benefit of fuzzy systems with the learning capability of neural networks, is a much better option than a standalone neural network or support vector machine classifier.

5.4 FUTURE SCOPE OF WORK

The given system is only an initial block in what may be a sophisticated speaker and speech recognition engine. It may be used in a variety of security systems for domestic as well as industrial purposes. Especially in financial transactions and maintaining databases in financial sectors, such a system could prove to be of great use. The system can also be further modified to produce a gender recognition system which takes age also as a factor or an emotion recognition system which takes gender into consideration. Such an emotion recognition system can be used for detection of pathological speech conditions for detection of diseases like Parkinson's disease, etc.

REFERENCES

Journal / Conference Papers

[1] Jyh-Shing Roger Jang, "ANFIS: Adaptive-Network-Based Fuzzy Inference System", IEEE transactions on systems, man, and cybernetics, vol. 23, no. 3, May June 1993.

[2] Siva Prasad Nandyala and T. Kishore Kumar, "Real Time Isolated Word Speech Recognition System for Human Computer Interaction", International Journal of Computer Applications, Vol. 12, No. 2, pp. 1-7, Nov 2010.

[3] Sachin Lakra, Juhi Singh and Arun Kumar Singh , "Automated Pitch-Based Gender Recognition using an Adaptive Neuro-Fuzzy Inference System", IEEE International Conference on Intelligent Systems and Signal Processing (ISSP), 2013

[4] Pawan Kumar, Nitika Jakhanwal, Anirban Bhowmick, and Mahesh Chandra, "Gender Classification Using Pitch and Formants", ICCCS, February, 2011

[5] R. Rajeshwara Rao, A. Prasad, "Glottal Excitation Feature based Gender Identification System using Ergodic HMM", International Journal of Computer Applications, Vol. 17, No. 3, pp. 31-36, March 2011.

[6] Dr. K. Meena, Dr. K. R. Subramaniam, M. Gomathy, "Performance Analysis of Gender Clustering and Classification Algorithms", International Journal on Computer Science and Engineering (IJCSE), Vol.4, 03 March 2012

[7] Yune-Sang Lee1, Peter Turkeltaub, Richard Granger, and Rajeev D. S. Raizada, "Categorical Speech Processing in Broca's Area: An fMRI Study Using Multivariate Pattern-Based Analysis", The Journal of Neuroscience, Vol.32, No.24, JAN 2012.

[8] Kadambe, S., & Srinivasan, P. (1994). Applications of adaptive wavelets for Speech. Optical Engineering, 33(7), 2204–2211.

[9] Sarikaya, R., & Hansen, H. L. (2000). High resolution speech feature parameterization for monophone-based stressed speech recognition. IEEE Signal Processing Letters, 7(7), 182–185.

[10] D. Nauck, F. Klawon; R. Kruse, "Foundations of Neuro-Fuzzy Systems", J. Wiley & Sons, 1997.

[11] R. Jang, "Neuro-Fuzzy Modelling:Architectures, Analysis and Applications", PhD Thesis, University of California, Berkley, July 1992.

Reference / Hand Books

[12] John R. Deller, Jr., John H.L. Hansen, John G. Proakis, "Discrete-Time Processing of Speech Signals", IEEE Press, ISBN 0-7803-5386-2

[13] Lawrence R. Rabiner, Ronald W. Schafer, "Digital Processing of Speech Signals", Prentice-Hall Signal Processing Series, ISBN 0-13-213603-1

Web

[14] www.wordpress.com/2011/04/20/hard-computing-vs-soft-computing

PROJECT DETAILS

Student Details			
Student Name	**Neha Rawat**		
Register Number	110907090	Section / Roll No	A/10
Email Address	nea.rawat@gmail.com	Phone No (M)	9008746304
Student Name	**Srishti Saha**		
Register Number	110907266	Section / Roll No	D/27
Email Address	srishti280992@yahoo.com	Phone No (M)	8197346936
Project Details			
Project Title	**Gender Classification in Speech Recognition using a Neuro-Fuzzy System**		
Project Duration	4 months	Date of reporting	12th January 2015
Organization Details			
Organization Name	**Dept. of Electronics and Communication, Manipal Institute of Technology**		
Full postal address with pin code	Manipal Institute of Technology, Manipal University, Manipal Karnataka- 576104		
Website address	www.manipal.edu		
Internal Guide Details			
Faculty Name	**Prof. T.K. Padma Shri**		
Full contact address with pin code	Dept of E & C Engg, Manipal Institute of Technology, Manipal – 576 104 (Karnataka State), INDIA		
Email address	padma.shri@manipal.edu		