***MID TERM REPORT OF***

**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 ENGINEERING**

**In**

**Electronics and Communication Engineering**

*Submitted by*

**Neha Rawat Srishti Saha**

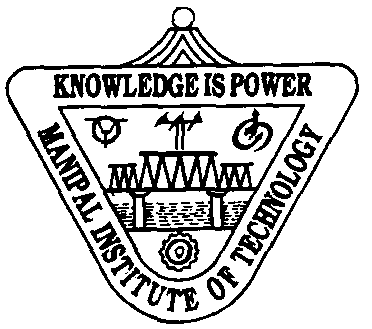
**Reg. No.:110907090 Reg. No.:110907266**

*Under the guidance of*

**Prof. T.K. Padma Shri**

**Associate Professor**

**DEPARTMENT OF ELECTRONICS AND COMMUNICATION ENGINEERING**



**MANIPAL INSTITUTE OF TECHNOLOGY**

(A Constituent College of Manipal University)

MANIPAL – 576104, KARNATAKA, INDIA

**MARCH/APRIL 2015**

**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 six frequency and energy-based features extracted from various male and female voice samples to train the classifier. The features extracted for this purpose are Pitch, Zero Crossing Rate, Short Time Energy, Energy Entropy, Formants and Mel-frequency Cepstral Coefficients (MFCCs). *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. A *Mel-frequency cepstrum (MFC)* is a representation of the short-term power spectrum of a sound, based on a linear cosine transform of a log power spectrum on a nonlinear Mel scale of frequency. MFCCs are the coefficients that collectively make an MFC. The above features will be fed to the classifier which after being trained will be tested on some test voice samples.

The classifier will be tested for its accuracy as well as its efficiency as compared to other existing classifiers for gender classification. The objective would be to prove the effectiveness of using a neuro-fuzzy based system as opposed to standard methods like decision tree and minimum distance classification as well as simple neural network systems like Support Vector Machines (SVM).

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

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Contents** | | | | | |
|  | | | | | Page No |
| Abstract | | | |  | i |
|  | | | | | |
| **Chapter 1** | | | **INTRODUCTION** | |  |
|  | **1.1** | Introduction | | | 1 |
|  | **1.2** | Speech Recognition- The Dimensions of Difficulty | | | 1-2 |
|  | **1.3** | Need for Gender Classification in Speech Recognition | | | 2 |
|  | **1.4** | Motivation | | | 2 |
|  | **1.5** | Objective | | | 3 |
|  | **1.6** | Organization of the Project Report | | | 3-4 |
|  | | | | | |
| **Chapter 2** | | | BACKGROUND THEORY | |  |
|  | **2.1** | Introduction | | | 5 |
|  | **2.2** | Theory | | | 5-6 |
|  | **2.3** | Literature Survey | | | 6-7 |
|  | | | | | |
| **Chapter 3** | | | **METHODOLOGY** | |  |
|  | **3.1** | Introduction | | | 8 |
|  | **3.2** | Pre-Processing of the Speech Signal | | | 8-9 |
|  | **3.3** | Pitch | | | 9-10 |
|  | **3.4** | Zero Crossing Rate | | | 10-11 |
|  | **3.5** | Short-Time Energy | | | 11 |
|  | **3.6** | Energy Entropy | | | 11-12 |
|  | **3.7** | Formants | | | 12-14 |
|  | | | | | |
| **Chapter 4** | | | **RESULT ANALYSIS** | |  |
|  | **4.1** | Introduction | | | 15 |
|  | **4.2** | Feature Tables | | | 15-18 |
|  | **4.3** | Result Analysis | | | 18-22 |
|  | **4.4** | Significance of the Result | | | 22 |
|  | | | | | |
| **Chapter 5** | | | **CONCLUSION AND FUTURE SCOPE** | |  |
|  | **5.1** | Introduction | | | 23 |
|  | **5.2** | Summary | | | 23 |
|  | **5.3** | Conclusion | | | 23 |
|  | **5.4** | Future Scope of Work | | | 23-24 |
|  | | | | | |
| **REFERENCES** | | | | | 25 |
| **PROJECT DETAILS** | | | | | 26 |

**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 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.

**1.5 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 Network Machines (SVM).

**1.6 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 present day scenario regarding the project work. It gives the need for gender classification and the difficulties in speech recognition that have given rise to this need. 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. 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 work carried up till now. 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.

The Forth 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 datasets 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.

The Fifth Chapter consists of the conclusion and future work. It includes a brief summary of all the work carried up till now 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

Another class of features which is often used in the field of speaker recognition is *Mel-Frequency Cepstral Coefficients (MFCCs)*. They basically represent the short-term power spectrum of the speech signal. All of the above given features have often been used individually and in groups for the process of gender discrimination for speaker recognition.

**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.

**2.3 LITERATURE SURVEY**

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

Nandyala et al. [1] 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. [2] 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. [3] 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. [4] 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. [5] 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.

**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 methods must extract the features of the various segments such that it makes them viable for grouping and comparison. A variety of methods exist for different features. This Chapter shall deal with the different methodologies available as well as the ones used in the project.

**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* [6].

***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 [6].

***Silence***

Whether silence should be called a form of excitation is debatable, but it is generally included for modelling purposes [6].

**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 [7] of a discrete time deterministic signal is defined as:

Ø(k) =Σm x(m).x(m+k) ; m=(-∞,+∞)

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 [7] is given as:

M(τ) = (1/N) Σ n |s(n) – s(n+τ)| ; n=0,1,2…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 [7] which has been followed is as given:

Z= (1/2N) Σi {sgn(x(i)) – sgn(x(i-1))} ; i=1,2,…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 [7] has been used for calculation:

e(n) = Σm (s(m).w(n-m))2 ; m = (-∞,+∞)

**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 Σ(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 = -Σi σ2 log2 (σ2) ; i= 0,1,…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**

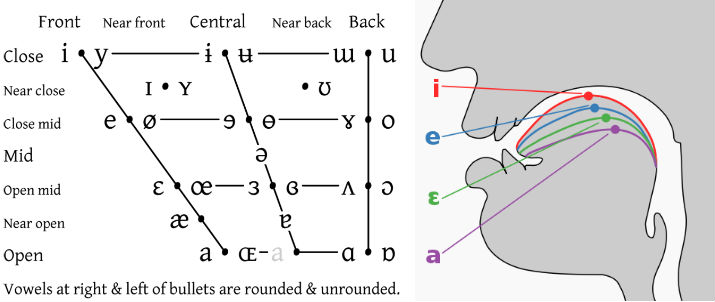
When we study the magnitude spectra of a speech signal, we observe some well-defined regions of emphasis (‘resonances’) and some regions of de-emphasis (‘antiresonances’). These resonances are a consequence of the articulators having formed various acoustical cavities and subcavities out of the vocal 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 /ә/.

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 /ә/ 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 [8]. 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:

Ө(z) = 1/(1 – Σi a(i) z-i) ; I = 1,2,…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.

**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.

**4.2 FEATURE TABLES**

**4.2.1 Database 1- NOIZEUS (University Of Texas, Dallas)**

Word extracted- ‘the’

Sampling Frequency- 8 kHz

Table 4.1 gives the features extracted for Database 1.

**Table 4.1 Feature table for Database 1**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **SPEAKERS** | **PITCH (Hz)** | **NORMALIZ-ED STE** | **NORMALIZ-ED ZCR** | **NORMALIZ-ED EE** | **FORMANTS (Hz) (‘e’-the)** |
| **Male Speaker 1** | 380.00 | 0.000324 | 0.0275 | 0.00096 | F1=384.4  F2=1158.7 |
| **Male Speaker 2** | 186.04 | 0.000066 | 0.0123 | 0.00047 | F1=375.4  F2=1418.1 |
| **Male Speaker 3** | 123.07 | 0.000805 | 0.0616 | 0.00170 | F1=328.7  F2=1831.2 |
| **Female Speaker 1** | 400.00 | 0.003100 | 0.0521 | 0.00150 | F1=456.8  F2=1622.3 |
| **Female Speaker 2** | 160.00 | 0.001800 | 0.0986 | 0.00190 | F1=523.7  F2=1397.4 |
| **Female Speaker 3** | 205.12 | 0.000471 | 0.0381 | 0.00063 | F1=451.1  F2=1187.6 |

**4.2.2 Database 2- American English Database (International Telecommunication Union)**

Word extracted- ‘the’

Sampling frequency- 16 kHz

Table 4.2 gives the features extracted for Database 2.

**Table 4.2 Feature table for Database 2**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **SPEAKERS** | **PITCH (Hz)** | **NORMALIZ-ED STE** | **NORMALIZ-ED ZCR** | **NORMALIZ-ED EE** | **FORMANTS (Hz) (‘e’-the)** |
| **Male Speaker 1** | 177.70 | 0.000297 | 0.0097 | 0.000093 | F1=216.30  F2=3385.50 |
| **Male Speaker 2** | 195.12 | 0.00028 | 0.0263 | 0.000853 | F1=203.10  F2=3569.80 |
| **Male Speaker 3** | 137.93 | 0.00130 | 0.0590 | 0.000965 | F1=230.80  F2=2506.20 |
| **Male Speaker 4** | 145.45 | 0.00190 | 0.0428 | 0.000704 | F1=224.80  F2=2225.60 |
| **Male Speaker 5** | 137.93 | 0.00170 | 0.0158 | 0.000585 | F1=194.30  F2=2333.10 |
| **Male Speaker 6** | 124.03 | 0.00140 | 0.0200 | 0.000317 | F1=232.10  F2=2931.70 |
| **Female Speaker 1** | 134.45 | 0.00059 | 0.0297 | 0.000949 | F1=364.00  F2=2910.00 |
| **Female Speaker 2** | 326.53 | 0.00039 | 0.0145 | 0.000716 | F1=349.20  F2=2422.60 |
| **Female Speaker 3** | 380.95 | 0.00014 | 0.0652 | 0.000580 | F1=371.20  F2=1421.00 |
| **Female Speaker 4** | 170.21 | 0.00077 | 0.0232 | 0.000783 | F1=495.00  F2=2174.00 |
| **Female Speaker 5** | 340.42 | 0.00092 | 0.0125 | 0.000322 | F1=225.00  F2=2360.80 |
| **Female Speaker 6** | 280.70 | 0.00210 | 0.0183 | 0.000944 | F1=232.50  F2=1919.40 |

**4.2.3 Database 3- Modified TIMIT Database (Prof. Dan Ellis, Electrical Engineering, Columbia University)**

Words extracted- ‘cottage’, ‘cheese’, ‘chives’, ‘delicious’

Sampling frequency- 16 kHz

The mean values of the following features (average of the values calculated for each extracted word) have been given in Table 4.3.

**Table 4.3 Feature table (mean values) for Database 3**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **SPEAKERS** | **PITCH (Hz)** | **NORMAL-IZED STE** | **NORMALIZE-D ZCR** | **NORMALI-ZED EE** | **FORMANTS (Hz)**  **(‘e’-cheese)** |
| **Male Speaker 1** | 190.89 | 0.0031 | 0.019 | 0.00012 | F1=287.20  F2=2923.90 |
| **Male Speaker 2** | 176.57 | 0.0024 | 0.208 | 0.00015 | F1=300.60  F2=2668.20 |
| **Male Speaker 3** | 236.84 | 0.0038 | 0.024 | 0.00013 | F1=261.20  F2=2309.20 |
| **Male Speaker 4** | 172.07 | 0.0016 | 0.053 | 0.00011 | F1=226.10  F2=2955.60 |
| **Male Speaker 5** | 233.72 | 0.0030 | 0.016 | 0.00012 | F1=300.00  F2=2496.40 |
| **Male Speaker 6** | 161.71 | 0.0013 | 0.218 | 0.00022 | F1=257.47  F2=2937.00 |
| **Male Speaker 7** | 230.52 | 0.0031 | 0.019 | 0.00013 | F1=297.80  F2=2329.40 |
| **Male Speaker 8** | 230.70 | 0.0014 | 0.027 | 0.00009 | F1=221.70  F2=2456.10 |
| **Male Speaker 9** | 186.93 | 0.0018 | 0.019 | 0.00013 | F1=245.70  F2=2595.80 |
| **Male Speaker 10** | 302.05 | 0.0019 | 0.200 | 0.00015 | F1=294.90  F2=3171.70 |
| **Male Speaker 11** | 148.53 | 0.0026 | 0.024 | 0.00017 | F1=334.80  F2=2303.80 |
| **Male Speaker 12** | 222.18 | 0.0018 | 0.118 | 0.00036 | F1=292.70  F2=2699.20 |
| **Female Speaker 1** | 235.34 | 0.0115 | 0.072 | 0.00650 | F1=481.60  F2=2819.50 |
| **Female Speaker 2** | 228.15 | 0.0058 | 0.517 | 0.00055 | F1=393.30  F2=4282.10 |
| **Female Speaker 3** | 257.88 | 0.0066 | 0.066 | 0.00100 | F1=413.40  F2=2134.30 |
| **Female Speaker 4** | 245.81 | 0.0086 | 0.144 | 0.00092 | F1=383.30  F2=2830.30 |
| **Female Speaker 5** | 253.80 | 0.0177 | 0.078 | 0.00186 | F1=565.50  F2=2418.70 |
| **Female Speaker 6** | 136.31 | 0.0027 | 0.258 | 0.00029 | F1=442.50  F2=3742.50 |
| **Female Speaker 7** | 269.07 | 0.0065 | 0.028 | 0.00022 | F1=385.70  F2=2292.20 |
| **Female Speaker 8** | 225.09 | 0.0039 | 0.038 | 0.00023 | F1=457.90  F2=2336.10 |
| **Female Speaker 9** | 200.58 | 0.0036 | 0.018 | 0.00015 | F1=446.10  F2=2658.20 |
| **Female Speaker 10** | 251.57 | 0.0020 | 0.220 | 0.00019 | F1=272.90  F2=4154.40 |
| **Female Speaker 11** | 217.26 | 0.0038 | 0.080 | 0.00019 | F1=403.10  F2=2335.10 |
| **Female Speaker 12** | 188.49 | 0.0021 | 0.049 | 0.00042 | F1=466.40  F2=2670.10 |

**4.3 RESULT ANALYSIS**

**4.3.1 Pitch**

Word- ‘the’

**Fig. 4.1 Pitch – Male vs Female**

Words- ‘cottage’, ’cheese’, ’chives’, ’delicious’

**Fig. 4.2 Pitch (mean value) – Male vs Female**

Fig. 4.1 and Fig. 4.2 show a comparison between the pitch values of the female and male speakers in the three datasets given above. In the first figure, we see a comparison between the pitch values of the speakers for the word ‘the’. It is observed that the pitch value for female speakers is higher than that of the male speakers. A similar result, except for a few deviations, is seen in the second figure, which shows a comparison for the speakers of dataset 3 on the basis of their average pitch values. The female speakers have an average pitch value of 243.22 Hz while the male speakers have an average pitch value of 195.23 Hz.

**4.3.2 Zero Crossing Rate**

Word- ‘the

**Fig. 4.3 Normalized ZCR – Male vs Female**

Words- ‘cottage’, ’cheese’, ’chives’, ’delicious’

**Fig. 4.4 Normalized ZCR (mean value) - Male vs Female**

Fig. 4.3 and Fig. 4.4 compare the values of the normalized zero crossing rate obtained for the male and female speakers for the word ‘the’ and the mean value obtained for the words ‘cottage’, ‘cheese’, ‘chives’ and ‘delicious’. It can be observed from the plots that the normalized ZCR has an increasing trend for females while for males it is remains at lower values.

**4.3.3 Short-Time Energy**

Word – ‘the’

**Fig. 4.5 Normalized STE – Male vs Female**

Words - ‘cottage’, ’cheese’, ’chives’, ’delicious’

**Fig. 4.6 Normalized STE (mean value) – Male vs Female**

Fig. 4.5 and Fig. 4.6 compare the values of the normalized short-time energies obtained for male and female speakers of the three datasets. Except for one deviation, the values for short-time energy are observed to be higher for female speakers than for male speakers.

**4.3.4 Energy Entropy**

Word – ‘the’

**Fig. 4.7 Normalized EE – Male vs Female**

Words – ‘cottage’, ‘cheese’, ‘chives’, ‘delicious’

**Fig. 4.8 Normalized EE (mean value) – Male vs Female**

Fig. 4.7 and Fig. 4.8 compare the values of the normalized energy entropy of the male and female speakers of the three datasets. It is observed from the plots that the energy entropy for females is higher than that for males. In few cases, some male speakers have values of energy entropy that are very close or equal to those of females, but in general the females have values sufficiently higher than their male counterparts

**4.3.5 Formants**

Vowel – ‘e’ (neutral vowel: /ә/ or vowel: /i/) in ‘the’

**Fig. 4.9 First Formant-Male vs Female Fig. 4.10 Second Formant-Male vs**

**Female**

Fig. 4.9 and Fig. 4.10 show the first and second formant frequencies for male and female speakers for the vowel ‘e’ in the word ‘the’. The first formant generally depends on gender while the second depends on the age of the speaker along with gender [9]. As per our observations, we have found higher values of the first formant for female speakers as compared to male speakers, but for the second formant in this case, the male speakers have been observed to attain higher values. A possible reason for this may be the age difference between the speakers as the age of the selected speakers was not taken into consideration while forming the datasets.

Vowel – ‘e’ (vowel: /i/) in ‘cheese’

**Fig. 4.11 First Formant-Male vs Female Fig. 4.12 Second Formant-Male vs**

**Female**

Fig. 4.11 and Fig. 4.12 show the first and second formant frequencies for the male and female speakers of dataset 3. The vowel extracted is ‘e’ from the word ‘cheese’ and the results are similar to the ones obtained for the previous datasets. The first formant frequencies are higher for the female speakers while the second formant frequencies, in this case, are almost equal for both speakers except at the end where the females attain higher values than the male speakers.

**4.4 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 higher for the female speakers than the male speakers.
* The normalized energy entropy values are also in general higher for the female speakers than the male speakers.
* The first formant frequencies are higher for the female speakers while the second formant frequencies show discrepancies as they are sometimes higher for males, sometimes equal for both speakers and in some cases higher for females. A possible reason could be the age difference between the speakers.

In conclusion, we can observe that features like pitch, zero crossing rate, short-time energy and energy entropy can be used to obtain an effective discernment between male and female speakers. The first formant frequency, as it majorly depends on the gender of the speaker, can also be used as a differentiating factor.

**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 as well as the work done up till now. That is followed by a brief explanation of the significance of the results obtained up till now and finally, 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 as well as values of the two formants show a general trend of higher values for female speakers and comparatively lower values for male speakers.

**5.3 CONCLUSIONS**

The results confirm our expectations of a clear distinction between the male and female speakers.

Thus, we can conclude from the observations that the given features can be used as stepping stones for the design of the classification system. The pitch values obtained are higher for females than males. The normalized zero crossing rates, short-time energies and energy entropy are also higher for females than for males. The first formant frequencies are higher for the female speakers than the male speakers while discrepancies due to possible age differences can be observed in the case of second formant frequencies. Overall, the results provide a strong possibility for the design of an efficient gender classification system.

**5.4 FUTURE SCOPE OF WORK**

The extraction of the above stated features will be followed by the extraction of another important feature – the Mel-frequency Cepstral Coefficients (MFCCs) which will serve as an additional tool in the design of the classification system.

After gathering this final tool, we will proceed with the design of an Adaptive Neuro-Fuzzy Inference System (ANFIS) which will be trained on the vectorized data we have computed to provide us with an accurate and efficient gender classification unit.

The system will also be tested further to compare its efficiency with existing classifiers designed using algorithms like minimum distance techniques, decision trees and Support Vector Machines (SVMs). The results obtained will be analyzed to ascertain the truth of our research objective i.e. the efficiency of a multi-feature neuro-fuzzy system over existing standard technologies.

**REFERENCES**

*Journal / Conference Papers*

[1] 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.

[2] 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

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

[4] 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.

[5] 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

*Reference / Hand Books*

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

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

*Web*

[8] Wikipedia

[9] [www.ncbi.nlm.nih.gov](http://www.ncbi.nlm.nih.gov) – “Formants of children, women and men: the effects of vocal intensity variation”, J. Acoust. Soc. Am., 1999

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 |
| Expected date of completion of project |  |  |  |
|  |  | | |
| *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 | | |