VOCAL EMOTION RECOGNITION

A Project Report submitted in the partial fulfillment of the requirements for the Degree of Bachelor of Engineering in Electronics and Telecommunication Engineering

By

Gilke Mandar Nandkumar
Kachare Pramod Haribhau
Kothalikar Rohit Ravindra
Rodrigues Varun Pius Felix

Under the Guidance of

Ms. Madhavi S. Pednekar

Don Bosco Institute of Technology,
Department of Electronics & Telecommunication Engineering,
Kurla (West), Mumbai- 400 070.
(2011-2012)

CERTIFICATE

This is to certify that following students of final year Engineering in Electronics & Telecommunication discipline have satisfactorily completed the project work entitled "***Vocal Emotion Recognition***'' in the partial fulfillment of the requirements for the Degree of Bachelor of Engineering (B.E.) in Electronics and Telecommunication Engineering during the academic year 2011-2012.

Project Team Members :

1. Gilke Mandar Nandkumar B. E. – 27
2. Kachare Pramod Haribhau B. E. – 28
3. Kothalikar Rohit Ravindra B. E. – 33
4. Rodrigues Varun Pius Felix B. E. – 62

Ms. Madhavi S. Pednekar	Ms. Pratibha Dumane	Dr. N. G. Joag
Project Guide	**Head of the Department**	**Principal**

ACKNOWLEDGEMENT

We were bestowed the golden opportunity to conduct our Project on "Vocal Emotion Recognotion", and hence take this opportunity to express our heartfelt gratitude to all those who have been associated with our project. This project bears an imprint of many people's hard work.

We would like to thank our Lord, The Almighty, whose constant intercession we relied on, whose blessings were constant source of inspiration all through the project.

We would like to thank *Ms. Madhavi Pednekar*, for providing us with endless support and encouragement in all our endeavors at every movement during our project. They say, 'The task of the excellent teacher is to stimulate "apparently ordinary" people to unusual effort'. Ms. Madhavi has been one such teacher who constantly brought winning mentality out of our ordinary souls; and most importantly always provided us with all possible help we needed.

We would also like to express our sincere thanks to Ms. Pratibha Dumane, the Head of the Department of Electronics and Telecommunication, for her constant support.

We are also greatly thankful to our principal *Dr. N. G. Joag*, for his generous guidance and co-operation throughout our project.

Last but not the least; we acknowledge all our fellow students for their co-operation and collaboration. Our entire team would like to thank all those who were remotely associated with the project that we unintentionally have forgotten to mention from the bottom of our heart.

Gilke Mandar Nandkumar B. E. – 27

Kachare Pramod Haribhau B. E. – 28

Kothalikar Rohit Ravindra B. E. – 33

Rodrigues Varun Pius Felix B. E. – 62

TABLE OF CONTENTS

ABSTRACT

Speech interface technology, which includes automatic speech recognition, synthetic speech, and natural language processing, is beginning to have a significant impact on business and personal computer use. Today, powerful and inexpensive microprocessors and improved algorithms are driving commercial applications in computer command, consumer, data entry, speech-to-text, telephone, and voice verification. Robust speaker-independent recognition systems for command and navigation in personal computers are now available; telephone-based transaction and database inquiry systems using both speech synthesis and recognition are coming into use. Large-vocabulary speech interface systems for document creation and read-aloud proofing are expanding beyond niche markets. Today's applications represent a small preview of a rich future for speech interface technology that will eventually replace keyboards with microphones and loudspeakers to give easy accessibility to increasingly intelligent machines.

This report is concerned with emotion recognition based on speech signals. The conventional method of estimation of emotion in speech has three steps. At first, researchers collect a number of speech prototypes, then obtain the speech features using frequency analysis extraction and later compute the statistical value of them. Finally, a classifier is made from the statistical value using a learning algorithm.

Automatic detection of emotion is evaluated using Standard Mel Frequency Cepstral and LPC (Linear Predictive Coding) Coefficients. The techniques used for development of classifier are KNN (K-Nearest Number) and Neural Networks using Back Propagation algorithm. Five basic human emotions including anger, happiness, neutral, fear and sadness are investigated. First, various features related to speech are calculated using MATLAB; they are compared to one another and the best ones are chosen. A database of the selected features for different emotion is created for each and every distinct emotion to be classified. Then the test samples are chosen and classifiers are used in order to detect the best matched emotion contained in the database. The efficiency of the emotion recognition thus depends upon the database used, the number of samples in the database and the classifier accordingly.

LIST OF FIGURES

LIST OF TABLES

1. INTRODUCTION

Introduction

Humans interact with others in several ways such as speech, gesture and eye contact. Among them, speech is the most effective way of communication through which people can readily share information. Emotions color the speech and can make the meaning more complex.

Speech signals apart from carrying just the information and facts also carry paralinguistic information. This may include the age, gender, mental state, emotional state and regional background of the speaker. Emotion can make the listeners react differently, according to what kind of emotion the speaker expresses .Till date robots can only understand commands said but not the "way it is said". Thus to make complete and efficient Human Machine Interface (HMI) it is required that robots understand the emotion in speech and react correspondingly, for example, when the user is happy the robot responds by giving a smile.

'Emotion in speech' is a topic that has received much attention during the last few years, in the context of speech synthesis as well as in Automatic Speech Recognition. It has been observed that the performance of the speech recognition system decreases under stress. Human expression is always affected by their psychological state. Hence speech always reflects the present emotional state of the person. Thus there may be variations in pitch, intensity, formant frequency of the said information. This may lead to error in speech detection. There are recent studies exploring emotional content of speech for call center applications or for developing toys that would advance human-toy interactions one step further by emotionally responding to humans.

In the interpersonal communication partners adapt in their acoustic parameters to show sympathy for each other. A technical system enabled to talk by speech synthesis therefore needs to know the actual user emotion and the according acoustic parameters to adapt instead of staying neutral all the time. Furthermore the communication channels of a speaker interact with each other. The knowledge of the implicit channel is needed to interpret the explicit channel. Irony might be a good example to demonstrate that prosodic features help understand the explicitly uttered intention. An emotion recognition system might also be called in for an objective judgment in psychiatric studies.

2. LITERATURE SURVEY

2.1 Case Studies

1. In studies done to analysis emotion in speech, pitch was selected as the deciding parameter by Shrikanth Narayan, et.al [1]. The pitch control is one of the important properties of speech that is effected emotional modulation. He presented an analysis of emotion derived from speech control.

1. Pitch features from emotional are compared with those from neutral speech using kullback-leibler distance.
2. Emotionally discriminated power of pitch features is quantified by comparing nested power logistic regression module.
3. Parameters such as maximum pitch, minimum pitch and range of pitch provided more information than pitch shift.
4. Analyzing the pitch statistics at the utterance level is more accurate than for shorter speech regions.
5. Best features were selected for distinguishing between neutral speech and emotional speech.
6. The results show that the recognition accuracy of the system is over 77%.
7. When compared to conventional classification schemes, the proposed approach performs better both in accuracy and robustness.

2. In the paper of Mandarin speech, et.al [2] the following was observed:

1. A mandarin speech based emotion classification is used to distinguish five emotions.
2. Emotions being anger, sadness, happy, neutral and boredom.
3. Features such as 16 LPC and 20 MFCC components were extracted from the speech signals.
4. Two classification techniques were used.
    a. Minimum distance method for which an accuracy of 79.1% was obtained.
    B. Nearest class mean for which 89.1% average accuracy was obtained.

3. Noam Amir, et.al [3] uses a corpus that has been studied extensively. The data base is a property of group of technology of habla. He verifies it through subjective listening tests. Best results are obtained using distance measure based classifiers. The recognition rates 70% for neutral, 76% for happy, 83% for sad and 61% for angry utterances .the overall accuracy being 70%.

4. Valery A. Petrushin , et.al [4] performed and developed a computer agent for emotion recognition .the study dealt 700 short utterances .statistical parameters such as pitch, first formant, second formant, energy and speaking rate was selected .the best results were obtained using the ensembles of neural network recognizers .the average accuracy is

about 70%. He also developed a real time emotion recognizer using neural networks for call centers applications with accuracy of 77%.

5. Research done in Spain used the hidden semi continuous Markova model[5]. Both the selection of low level features and design of recognition system were successfully done. Results were obtained using database in Spanish. The accuracy in recognizing seven different emotions –six defined in MPEG-4 plus neutral style was exceeding 80%. It was observed that instantaneous features provide better performance than the syllabic features for both energy and pitch. Thus instantaneous pitch turns out to be a good classifier for emotions.
Study done in Greece provided the following conclusions:

- Automated emotion recognition of real time databases cannot achieve accuracy greater than 50%.
- Natural emotions cannot be easily classified as acted once.
- Emotions in decreasing frequency of appearance are anger sadness happiness fear disgust joy and surprise.

6. Data collection was made in Assamese language and classified using GMM[24]. The highest rose to 74.4% when window size of 23.22ms was chosen. The average mean success score of the experiment (speaker independent) in all cases remained lower than in the other experiment (speaker dependent). The results show that the surprise is the most difficult to differentiate from other emotions, since surprise may be expressed along with any other emotion such as angry- surprise, fear-surprise, happy-surprise, etc.
The study used is HMM classifier with 39 coefficients (13 MFCCs, 13 Delta coefficient, 13 acceleration coefficients). The HMM classifier obtained a better performance in recognizing anger emotion in all cases. Accuracy 80% was obtained in recognizing anger and 60% in remaining emotions.

7. In study done at IIT Kanpur, et.al [7] the use of spectral and prosody features of speech both at per frame level and at utterance level was carried out for this purpose. Measurement of mean and variance of MFCC coefficient was an important feature. Three statistical methods were employed (KNN, GMM and SVM) of which SVM gives the best result of 70%. Further improvement in results was obtained using fused GMM and SVM.
The emotional speech features changed by not only emotional information phoneme information which also should be extracted.

- Evaluation of several features such as pitch energy and formant.
- The features were classified into emotion reflective features and phoneme dominant features.
- Emotion reflective features were extracted based on phoneme information which was classified by phoneme dominant features. This method was more sensitive to emotion, but less sensitive to phoneme.

The performance of the speaker recognition system decreases when the speaker is under stress or emotion. Three features namely pitch, amplitude and duration (together called PAD) were observed. PAD vectors of similar phones in different words of speaker are

closed to each other confirming that the way a speaker stresses syllables in their is unique to them.

8. Ben J. Shannon, et.al [20] has analysed the performance of MFCC and BFCC features are also compared to uniform frequency Cepstral coefficient (UFCC) where it is shown that Bark scale and Mel scale filter banks have equivalent performance in speech recognition task. It has also been shown that these features provide an advantage over uniformly space filter banks only when the training and testing condition are not matched.

9. R.V. Pawar, et.al [21] the author in his paper has extracted features using LPC coefficients, calculating AMDF and DFT. The neural network is trained by applying these features as input parameters. The features are stored in templates for further comparison.

10. K R Aida-Zade, et.al [22] the authors discuss computing algorithms of speech features as the main part speech recognition system, are analyzed. Also the determination algorithms of MFCC and LPC coefficients expressing the basic speech features are developed. The training and recognition process are realized by ANN in the automatic speech recognition system. This results in decrease in error rate. The combined result was 84.6% recognition.

2.2 Books:

"Digital processing of speech signal", et.al [2] by Lawerence Rabiner and Ronald Schafer gives basic understanding of speech signal. the book is helpful for understanding of production of speech signal, its mathematical model, various properties of speech signal and feature extraction methods.

"Introduction to neural networks ", et.al [19] by S.Sivanandan, S.Smathiand and S.Deepa gives idea about various neural network and their application. It describes each neural network with algorithm and example of MATLAB implementation which makes it easy to understand and implement.

2.3 Conclusion:

Thus by studying the above researches we conclude that it is best to work with synthetic databases and use MFCC features to distinguish between various emotions. HMM technique commonly used for this purpose provides the best results. It has been found that LPC coefficient for a sound sample can be extracted, though its accuracy is less as compared as compared to other techniques like MFCC and Neural networks.

3. SPEECH CHARACTERIZATION

3.1 Definition of Speech:

Voice (or vocalization) is the sound produced by humans and other vertebrates using the lungs and the vocal folds in the larynx, or voice box. Voice is not always produced as speech, however. Infants babble and coo; animals bark, moo, whinny, growl, and meow; and adult humans laugh, sing, and cry. Voice is generated by airflow from the lungs as the vocal folds are brought close together. When air is pushed past the vocal folds with sufficient pressure, the vocal folds vibrate. If the vocal folds in the larynx did not vibrate normally, speech could only be produced as a whisper. Your voice is as unique as your fingerprint. It helps define your personality, mood, and health.

3.2 Speech Production:

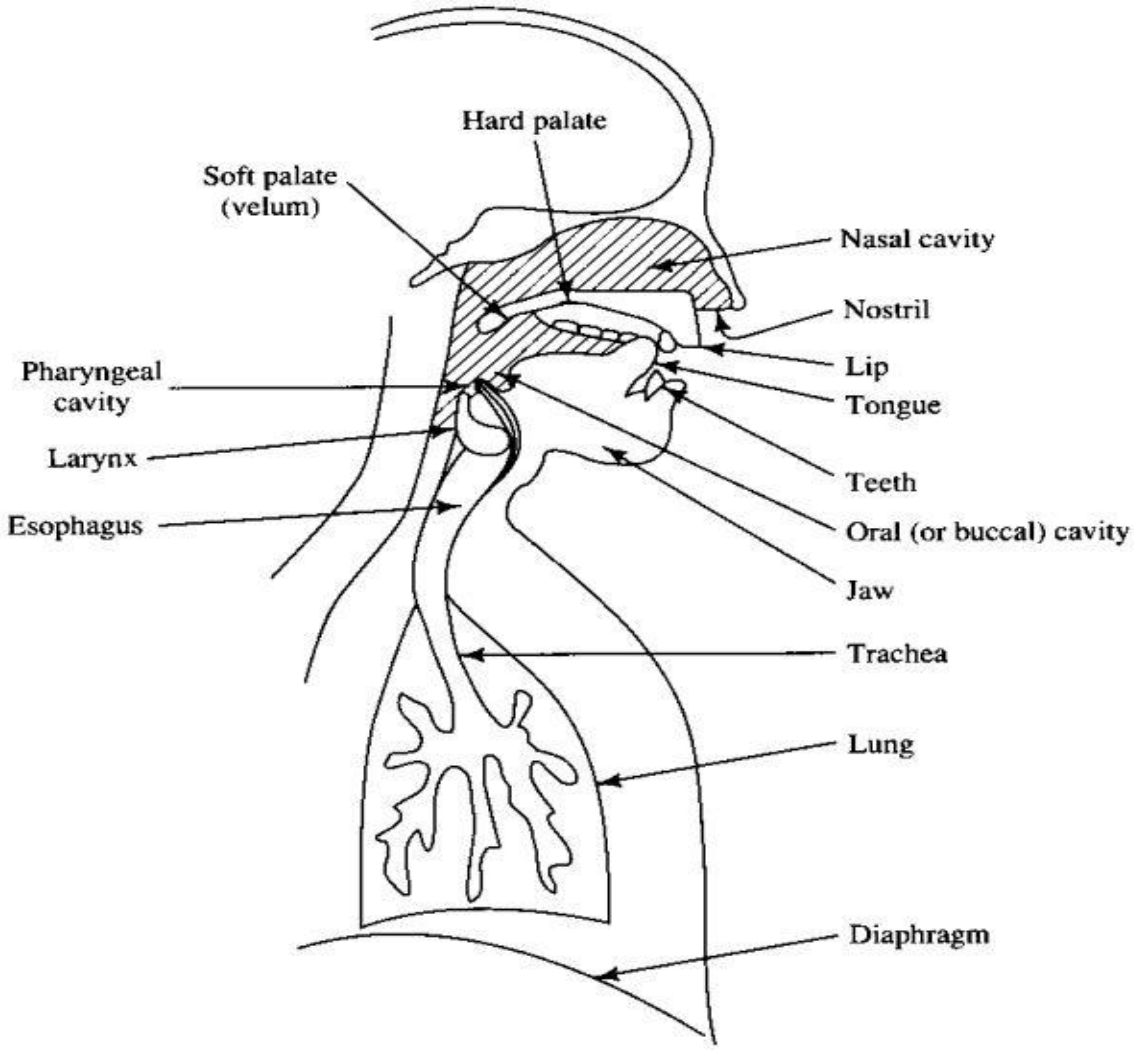

Figure 3.1 Human Vocal System [18]

The human apparatus concerned with speech production and perception is complex and uses many important organs - The lungs, mouth, nose, ears controlling muscles and the brain. It is remarkable that this apparatus has developed to enable not only the speech production but also serves other purposes such as breathing or eating. It was discovered that various specific areas in the brain are regarded to be of prime importance for speech and language.

The vocal tract and vocal cord play a major role in speech production. The vocal tract consists of several organs and muscles which are regularly monitored and carefully controlled by the speech centers. The precise controlling is achieved by internal feedback in the brain. As an example auditory feedback helps us to ensure that we are producing the correct speech sounds and that they are of the correct intensity for the environment. Speech sounds are produced when air is exhaled from the lungs and causes either vibration of vocal cord or turbulence at some point of constriction in the vocal tract. The shape of the vocal tract influences the sound harmonics. The way in which the vocal cord is vibrated and the shape of the vocal tract is varied in order to produce a range of speech sounds.

The vocal cord is situated in larynx called the Adam's apple. The vocal cord is the source for speech production in humans. It generates two kinds of speech sounds these are voiced and unvoiced. The vibration of vocal cords produces the sound called the voicing and the unvoiced sound due to turbulence of flow of air at constriction at all possible sites in the vocal tract. The frequency of vibration of the cord is determined by several factors; the tension exerted by the muscle, its mass and its length. These factors vary between sexes and according to age.

The vibration of vocal cord produces harmonics - the amplitude of the harmonics decrease with increasing frequency illustrated as follows:

- Air pressure from the lungs creates a steady flow of air through the trachea (windpipe), larynx (voice box) and pharynx (back of the throat).
- The vocal folds in the larynx vibrate, creating fluctuations in air pressure that are known as sound waves.
- Resonances in the vocal tract modify these waves according to the position and shape of the lips, jaw, tongue, soft palate, and other speech organs, creating formant regions and thus different qualities of sonorant (voiced) sound.
- Mouth and nose openings radiate the sound waves into the environment.

3.3 Speech and Vocal Tract:

The vocal tract is divided into two parts, first one is called the oral tract which is highly mobile and consists of the tongue, pharynx, plate, lips, and jaw The position of these organs are varied to produce different speech sounds, which we hear as the radiation from the lips or nostrils. The second one is the nasal tract where is immobile but is coupled with oral tract by changing the position of the velum. The shape of the vocal tract responds better for some basic frequency produced by vocal cord than others, this is the essential mechanism for the production of different speech sounds. The lowest resonance frequency for a particular shape of the vocal tract is called the first formant (f1) and next the second formant frequency (f2) and so on.

3.4 Speech and language:

The purpose of speaking is to convey meaningful ideas to the listener. In order to do this, the listener should be able to interpret the meaning of the spoken sounds. One way of doing this is by providing a coding mechanism with set of rules enabling the listener to interpret the meaning of the speech. The human being uses linguistics as the tool for coding the information. The coding mechanism is not straight forward. The new ideas are converted into linguist structure. This requires selection of appropriate words or phrases. These words are ordered in sequence according to grammatical rules.

3.5 Speech and sound:

From the linguistic point of view the smallest speech unit is known as phonemes, which indicates a different in meaning and is normally written between slashes as for example /m/ in hum. In fact the sounds produced for individual phonemes vary depending on where it appears in a word, phonemes sets are different for different languages, as for example about 40 phonemes are sufficient to discriminate between all the sounds made in British English.

Phonemes are characterized in to six different groups. These are the vowels, dipthongs, semi vowels, stop constant, fricative and affricative. The grouping of these phonemes is based on the way these sounds are produced. Each phonemes is a combined version of the first three dominant formant frequency which is originated due to vibration of the vocal cord. However the formant frequency largely varies depending on the speaker.

3.6 Features of speech signal:

1. **Prosody:**

   Prosody may reflect various features of the speaker or the utterance: the emotional state of a speaker; whether an utterance is a statement, a question, or a command; whether the speaker is being ironic or sarcastic; emphasis, contrast, and focus; or other elements of language that may not be encoded by grammar or choice of vocabulary.

2. **Intonation:**

   In linguistics, **intonation** is variation of pitch while speaking which is not used to distinguish words. All languages use pitch semantically, that is, as intonation, for instance for emphasis, to convey surprise or irony, or to pose a question. *Rising intonation* means the pitch of the voice increases over time; *falling intonation* means that the pitch decreases with time. A *dipping intonation* falls and then rises, whereas s*peaking intonation* rises and then falls.

3. **Accent:**

   In linguistics, an **accent** is a manner of pronunciation peculiar to a particular individual, location, or nation. An accent may identify the locality in which its speakers reside (a geographical or regional accent), the socio-economic status of its speakers, their ethnicity, their caste or social class, their first language (when the language in which the accent is heard is not their native language), and so on. Accents typically differ in quality of voice, pronunciation of vowels and consonants, stress, and prosody; although grammar, semantics, vocabulary, and other language characteristics often vary concurrently with accent.

4. **Loudness :**

   Loudness is the quality of a sound that is the primary psychological correlate of physical strength (amplitude). More formally, it is defined as "that attribute of auditory sensation in terms of which sounds can be ordered on a scale extending from quiet to loud.

Loudness is also affected by parameters other than sound pressure, including frequency, bandwidth and duration. In acoustics *volume* is related to amplitude, sound pressure, and dynamics.

5. **Formant frequency :**

   Formants are defined by as the spectral peaks of the sound spectrum |P(f)|' of the voice. **Formant** is also used to mean an acoustic resonance and, in speech science and phonetics, a resonance of the human vocal tract. It is often measured as an amplitude peak in the of the sound, using a spectrogram. In acoustics, it refers to a peak in the sound envelope and/or to a resonance in sound sources, notably musical, as well as that of sound chamber

6. **Pitch:**

   Pitch represents the perceived fundamental frequency of a sound. It is one of the major auditory attributes of speech signal. Pitch allows the construction of melodies; pitches are compared as "higher" and "lower", and are quantified as frequencies Pitch is related to frequency, but they are not equivalent. Frequency is the scientific measure of pitch.

7. **Timbre:**

   Timbre is the quality of a musical note or sound or tone that distinguishes different types of sound production, such as voices or musical instruments. The physical characteristics of sound that mediate the perception of timbre include spectrum and envelope. Timbre is also known in psychoacoustics as *tone quality* or *tone color*. For example, timbre is what, with a little practice, people use to distinguish the saxophone from the trumpet in a jazz group, even if both instruments are playing notes at the same pitch and loudness. Timbre has been called a "wastebasket" attribute.

4. EMOTION

4.1 Emotion Definition:

The word emotion includes a wide range of observable behaviors, expressed feelings, and changes in the body state. This diversity in intended meanings of the word emotion make it hard to study. For many of us emotions are very personal states, difficult to define or to identify except in the most obvious instances. Moreover, many aspects of emotion seem unconscious to us. Even simple emotional states appear to be much more complicated than states as hunger and thirst.

Emotion is the complex psycho physiological experience of an individual's state of mind as interacting with biochemical (internal) and environmental (external) influences. In humans, emotion fundamentally involves "physiological arousal, expressive behaviors, and conscious experience".

Emotion is associated with mood, temperament, personality and disposition, and motivation. Motivations direct and energize behavior, while emotions provide the affective component to motivation, positive or negative.

A related distinction is between the emotion and the results of the emotion, principally behaviors and emotional expressions. People often behave in certain ways as a direct result of their emotional state, such as crying, fighting or fleeing. If one can have the emotion without the corresponding behavior, then we may consider the behavior not to be essential to the emotion.

4.2 Types of emotions:

Though there is no rule as to classify various emotions they can be separated into two categories.

Primary emotions and secondary emotions:
Emotions such as happiness, sadness, anger, etc. Are the most basic emotions that a person displays during communication .they easy to identify and simple to classify.
Secondary emotions like disgust, pride, sarcasms, etc. Are complex in nature and made up from primary emotions .to classify and identify is different.

Conscious and Unconscious Emotions :
Conscious emotions: These emotions are visible in person's behavior .example: when angry the person may raise his hand .they include emotions such as happiness, sadness, angered.
Unconscious emotions: Emotions such as sarcasms may sometimes be concealed under the smile of the speaker and may not be directly visible to the observer.

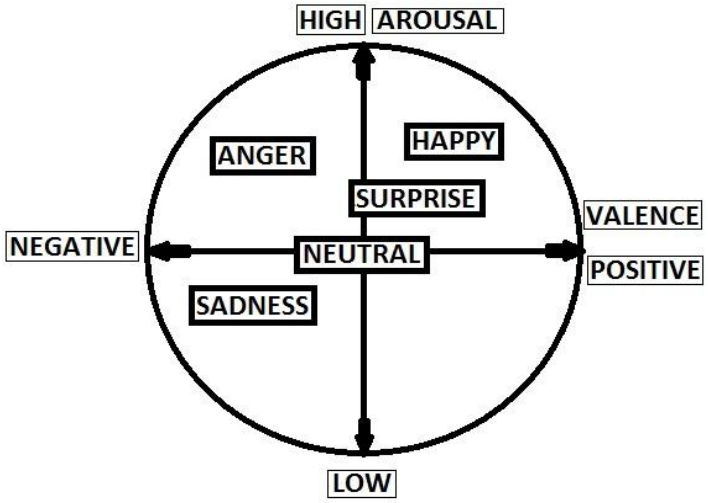

Figure 4.1 Arousal-Valence Matrix for Emotion Recognition. [2]

4.3 Parameter variation in emotion:

The table below shows the variation of various physical parameters under the influence of different emotions.

Table 4.1 Acoustic Parameters of Tone Sequences

	Anger	Happiness	Sadness	Fear	Disgust
Speech Rate	Slightly faster	Faster or slower	Slightly slower	Much faster	Very much faster
Average Pitch	Very much higher	Much higher	Slightly slower	Very much higher	Very much slower
Pitch Range	Much wider	Much wider	Slightly narrower	Much wider	Slightly wider
Intensity	Higher	Higher	Lower	Normal	Lower
Voice Quality	Breathy chest	Breathy blaring tone	Resonant	Irregular voicing	Grumble chest tone
Articulation	Tense	Normal	Slurring	Precise	Normal
Pitch changes	Abrupt	Smooth and upward inflection	Downward inflection	Normal	Wide downward terminal inflects

5. EXISTING SYSTEM

5.1 System Prototype

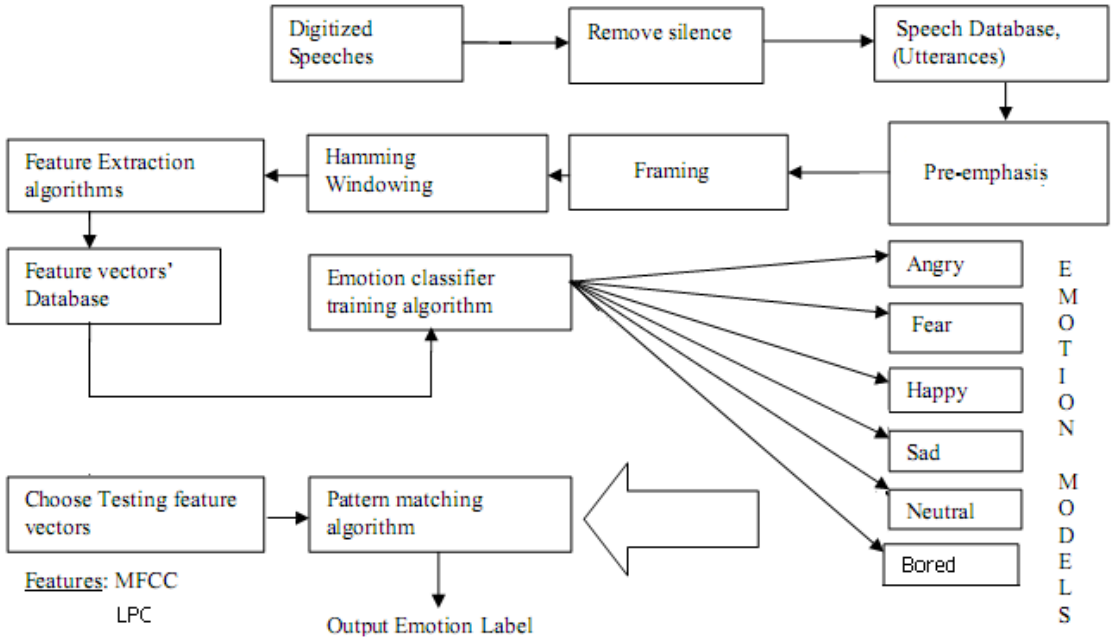

Figure 5.1 Basic block diagram of Emotion Recognition

1. Pre processing:

Firstly, the voiced analog signal from the speaker is converted into digital signal using microphone which acts as an A to D converter. in the preprocessing stage first each signal is having a cd component which has to be removed and since the silence parts of the signal do not carry any useful information, those parts including the leading and trailing edges are eliminated by thresh holding the energy of the signal. The signals are dividing into frames using a hamming window of length 23msec.

2. Feature extraction:

In these study prosodic features such as pitch, energy, formants which were generally used in the both speech recognition and emotional speech recognition a set of novel acoustic features in this experiment. We will be using features such as energy, pitch, formant, MFCC and LPC. All the features are extracted from each

frame and then the mean and the standard deviation for each feature is considered to constitute the feature vector.

5.2 Feature extraction:

5.2.1 MFCC:

In sound processing, the 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.

Mel-frequency cepstral coefficients (MFCCs) are coefficients that collectively make up an MFC. They are derived from a type of cepstral representation of the audio clip (a nonlinear "spectrum-of-a-spectrum"). The difference between the cepstrum and the mel-frequency cepstrum is that in the MFC, the frequency bands are equally spaced on the mel scale, which approximates the human auditory system's response more closely than the linearly-spaced frequency bands used in the normal cepstrum. This frequency warping can allow for better representation of sound, for example, in audio compression.

For speech/speaker recognition the most commonly used acoustic features are Mel–scale frequency cepstral coefficient (MFCC). MFCC takes human perception sensitivity with respect to frequencies into consideration and therefore are best for speech/speaker recognition.

Mel scale

The mel scale, proposed by Stevens, Volkman and Newman in 1937 is a perceptual scale of pitches judged by listeners to be equal in distance from one another. The reference point between this scale and normal frequency measurement is defined by equating a 1000 Hz tone, 40 dB above the listener's threshold, with a pitch of 1000 mels. Above about 500 Hz, larger and larger intervals are judged by listeners to produce equal pitch increments. As a result, four octaves on the hertz scale above 500 Hz are judged to comprise about two octaves on the mel scale. The name mel comes from the word melody to indicate that the scale is based on pitch comparisons.

A popular formula to convert f hertz into mel is:

$$m = 2595 \log_{10}\left(\frac{f}{700} + 1\right) = 1127 \log_{e}\left(\frac{f}{700} + 1\right)$$

Steps to be followed for the calculation of MFCC

1. **Pre-emphasis:**

   The speech signals s (n) is sent to the high pass filter

   s_2 (n) = s (n) – a * s (n-1)

   Where s_2 (n) is the output signal and the value of a is between 0.9 and 1.

   The z -transform of the filter is

   H (z) = 1 – a * Z^{-1}

   The goal of pre-emphasis is to compensate the high frequency part that was suppressed during the sound production mechanism of humans. Moreover it can also amplify the importance of high frequency formants.

2. **Frame blocking:**

   The input speech signal is segmented into frames of 20-30 ms with optional overlap of 1/3- ½ of the frame size .usually the frame size (in terms f sample points) is equal power of 2 in order to facilitate the use of FFT. If this is not the case we need to do zero padding to the nearest length of power of 2. If sample rate is 16 kHz and frame size is 320 the sample points, then the frame duration is 320/16000= 0.02 secs = 20ms additional, if the overlap is 160 points then the frame rate is 16000/ (320-160) =100 frames per second.

3. **Hamming window:**

   Each frame has to be multiplied with the hamming window in order to keep the continuity of the first and the last points in the frame. if the signal in a frame is denoted by s(n),where n is equal to 0 to N-1, then the signal after hamming windowing is s(n)*w(n);where w(n) is the hamming window defined by :

   $W(n, \alpha) = (1-\alpha)-\alpha * \cos(2\pi n/(N-1)), \quad 0 \leq n \leq N-1$

   In practice the value of α is set to 0.46.

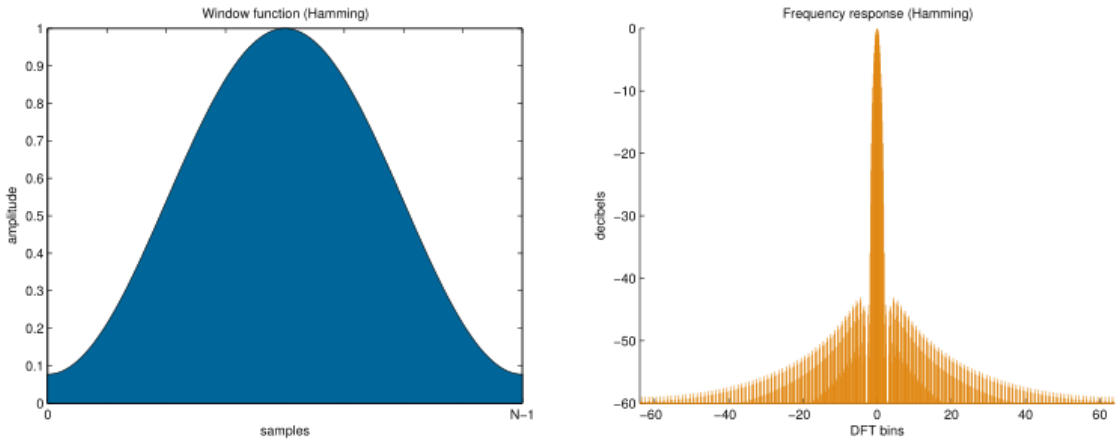

Figure 5.2 Hamming Window [7]

4. **Fast Fourier transform(FFT):**

Spectral analysis show that different timbre in speech signals corresponds to different energy distribution over frequency. Therefore we usually perform FFT to obtain the magnitude frequency response of each frame.

When we perform FFT on a frame we assume that the signal within the frame is periodic and continuous when wrapping around. If this is not the case we still perform FFT but the in continuity at the frame's first and last points is likely to introduce undesirable effects in the frequency response. To deal with this problem we have two strategies

- Multiply each frame by the hamming window to increase is continuity at the first and last points.
- Take a frame of variable size it always contains an integer multiple number of the fundamental periods of the speech signal.

5. **Triangular band pass filters:**

We multiply the magnitude frequency response by a set of 20 triangular band pass filters to get the log energy of each triangular band pass filter. The position of these filters are equally spaced along the Mel frequency which is related to the common linear frequency f by the following equation

$$\text{Mel}(f) = 1125 * \ln(1 + f/700)$$

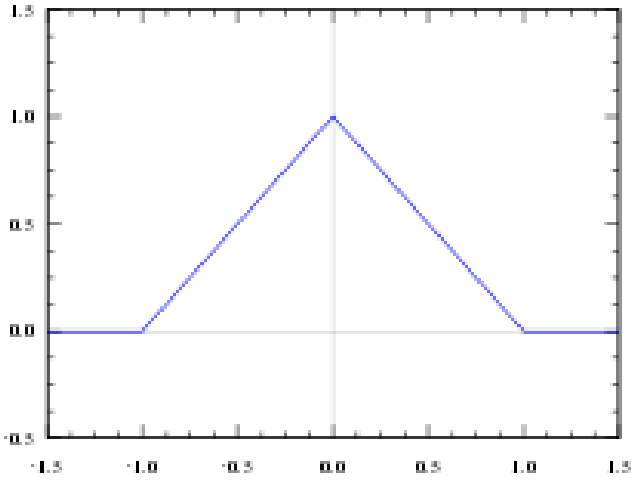

Figure 5.3 Triangular Filter [7]

6. Take the logs of the powers at each of the Mel frequencies.
7. Take the discrete cosine transform of the list of Mel log powers, as if it were a signal.
8. The MFCCs are the amplitudes of the resulting spectrum.

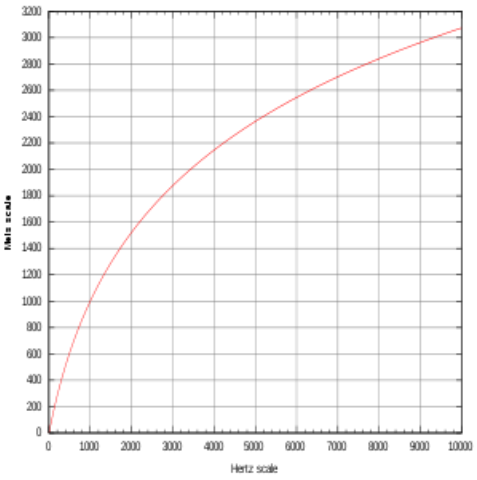

Figure 5.4 Mel Scale[7]

5.2.2 Linear Predictive Coding:

Linear predictive coding (LPC) is a tool used mostly in audio signal processing and speech processing for representing the spectral envelope of a digital signal of speech in compressed form, using the information of a linear predictive model. It is one of the most powerful speech analysis techniques, and one

of the most useful methods for encoding good quality speech at a low bit rate and provides extremely accurate estimates of speech parameters.

Source Filter Model:

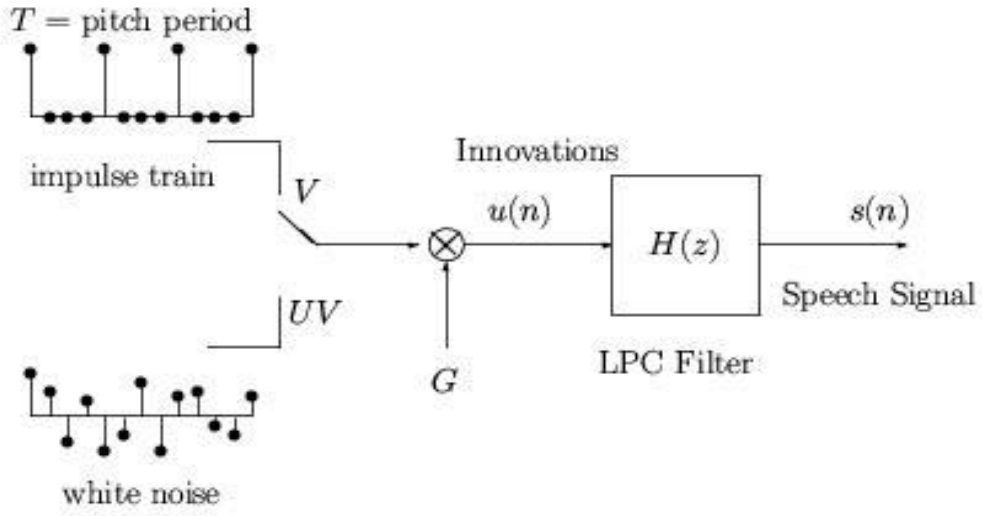

Figure 5.5 Source filter[19]

Where 'v' stands for voiced signals;

'uv' stands for non-voiced signals;

' G'is the air pressure in the wind pipe and

U(n) represents the transfer function of the trachea

The **source–filter model of speech production** models speech as a combination of a sound source, such as the vocal cords, and a linear acoustic filter, the vocal tract (and radiation characteristic). An important assumption that is often made in the use of the source-filter model is the independence of source and filter. In such cases, the model should more accurately be referred to as the "independent source-filter model".

While only an approximation, the model is widely used in a number of applications because of its relative simplicity. To varying degrees, different phonemes can be distinguished by the properties of their source(s) and their spectral shape. Voiced sounds (e.g., vowels) have (at least) a source due to (mostly) periodic glottal excitation, which

can be approximated by an impulse train in the time domain and by harmonics in the frequency domain, and a filter that depends on, e.g., tongue position and lip protrusion. On the other hand, fricatives have (at least) a source due to turbulent noise produced at a constriction in the oral cavity (e.g., the sounds represented by orthographically by "s" and "f"). So called *voiced fricatives* (such as "z" and "v") have two sources - one at the glottis and one at the supra-glottal constriction.

The source-filter model is used in both speech synthesis and speech analysis, and is related to linear prediction. The development of the model is due, in large part, to the early work of Gunnar Fant, although others, notably Ken Stevens, have also contributed substantially to the models underlying acoustic analysis of speech and speech synthesis.

In implementation of the source-filter model of speech production, the sound source, or excitation signal, is often modeled as a periodic impulse train, for voiced speech, or white noise for unvoiced speech. The vocal tract filter is, in the simplest case, approximated by an all-pole filter, where the coefficients are obtained by performing linear prediction to minimize the mean-squared error in the speech signal to be reproduced. Convolution of the excitation signal with the filter response then produces the synthesized speech.

Basic Principles

LPC starts with the assumption that the speech signal is produced by a buzzer at the end of a tube. The glottis (the space between the vocal cords) produces the buzz, which is characterized by its intensity (loudness) and frequency (pitch). The vocal tract (the throat and mouth) forms the tube, which is characterized by its resonances, which are called *formants*.

LPC analyzes the speech signal by estimating the formants, removing their effects from the speech signal, and estimating the intensity and frequency of the remaining buzz. The process of removing the formants is called *inverse filtering*, and the remaining signal is called the *residue.*

The numbers which describe the formants and the residue can be stored or transmitted somewhere else. LPC synthesizes the speech signal by reversing the process: use the

residue to create a source signal, use the formants to create a filter (which represents the tube), and run the source through the filter, resulting in speech.

Because speech signals vary with time, this process is done on short chunks of the speech signal, which are called frames. Usually 30 to 50 frames per second give intelligible speech with good compression.

Applications:

LPC is generally used for speech analysis and resynthesis. It is used as a form of voice compression by phone companies, for example in the GSM standard. It is also used for secure wireless, where voice must be digitized, encrypted and sent over a narrow voice channel; an early example of this is the US government's Navajo I. LPC synthesis can be used to construct vocoders where musical instruments are used as excitation signal to the time-varying filter estimated from a singer's speech.

6. PROPOSED SYSTEM

System Design

In our approach we assume that there is only one emotion in each utterance (sentence). If, for some reasons, several emotions are recognized in one sentence which is the common case, one predominant emotion is selected as the emotion of this sentence. Two different measures are used: Either the emotion with the maximum duration (time measure) or the emotion with the maximum number of occurrence (frequency measure) is selected as the emotion of the sentence. As opposed to the previous scenarios where speech and emotions are recognized at the same time, in the two-step recognition approach we use the output of a speech recognizer to construct an optimized language model for the speech-emotion recognizer. By that, emotion recognition is performed with the aid of the knowledge of what the user has said.

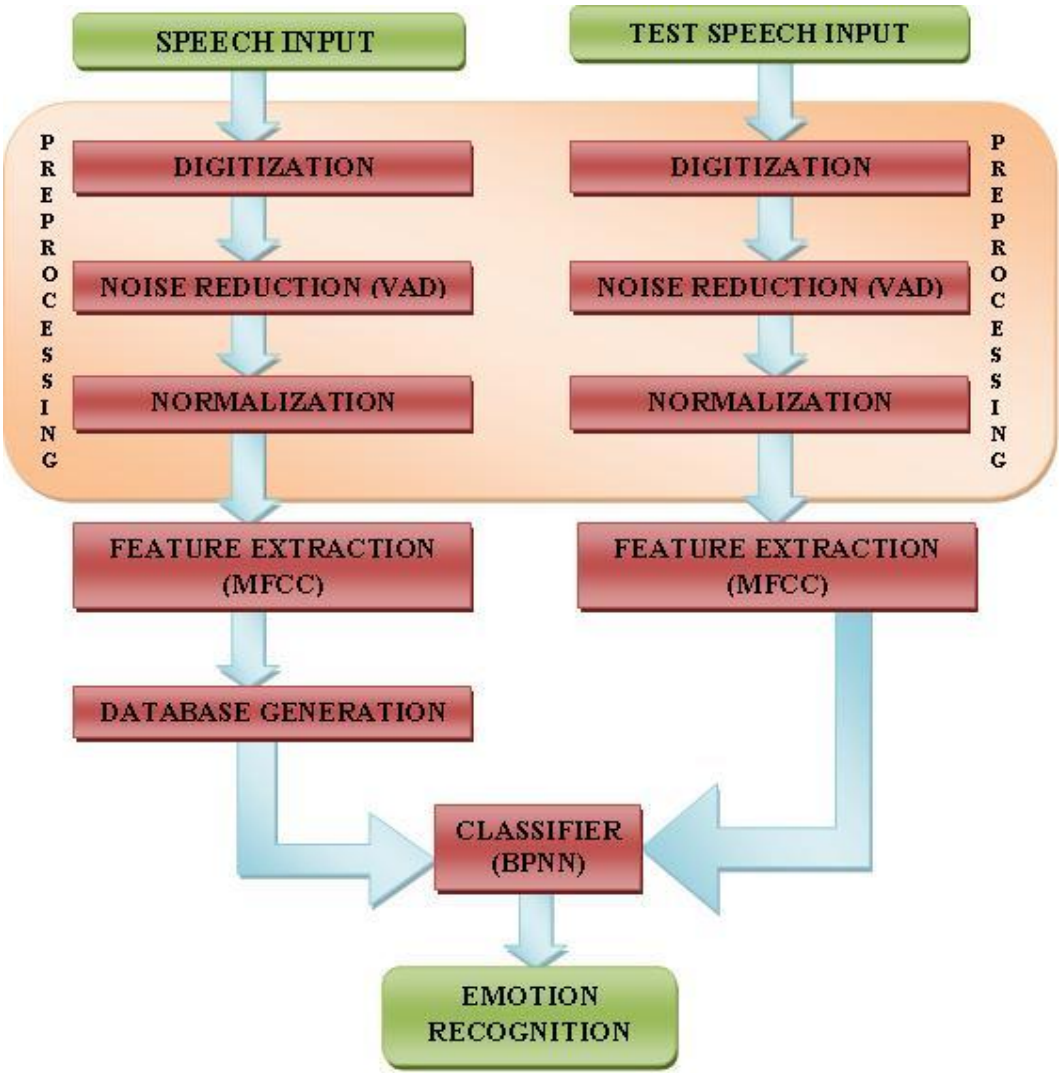

Figure 6.1 Emotion Recognition Flow-graph

6.1 Data Acquisition

A headphone-mic, a computer and GOLDWAVE software were used for single channel recording of emotionally biased utterances of fixed lengths in each emotion from 3 male and 3 female speakers. Each speaker was asked to utter 5 times a fixed set of 3 short sentences, each of a different emotion. The necessary emotional acting was obtained by narrating to the speaker a short emotional story so as to sufficiently arouse the same emotion in the dialogues. Utterances corresponding to sad, surprise, happy, anger and neutral utterances are recorded. MATLAB 7 Software was used for all computations.

6.2 Preprocessing

After data acquisition, silence periods within the samples were removed. The usage of Voice Activity Detection (VAD) technique was used to delete such silence frames. Then the speech samples were passed through the LPF ($1 - 0.97z^{-1}$) which gives a spectral tilt to the speech samples.

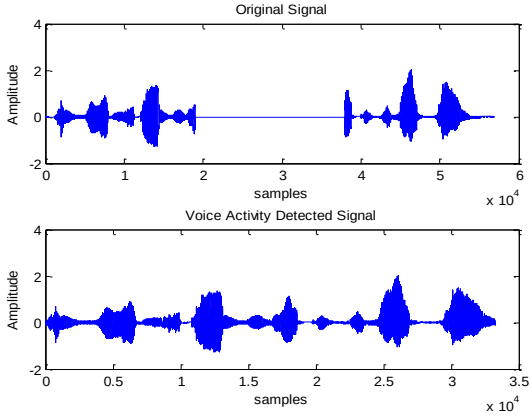

Figure 6.2 Voice Activity Detection (VAD)

6.3 Windowing

The filtered voice samples were segmented into 23.22 ms frames with each frame having 50% overlap with the adjacent frames. Each frame is then multiplied by a Hamming window of the same length i.e. 246 samples per frame length [8].

6.4 Feature Extraction

20 MFCC and 1 total Log-Energy features were calculated from each frame using 24 triangular Mel-frequency filter banks. Then each feature vector was normalized by mean

and standard deviation, which were computed from the feature vectors extracted from the trained utterances [8].

6.5 Pattern Matching

The trained Artificial Neural Network (ANN) was required to be tested with features extracted from the test-utterances. The ANN was trained with multiple voice samples taken at different instances uttering the same phrase at all times. The total log-likelihood of these test-vectors of one test utterance with respect to the trained matrix corresponding to each emotion-class was computed. The test utterance is considered to belong to that emotion-class with respect to which the total log-likelihood becomes the largest [8].
An ambiguity may arise when surprise may be expressed along with any other emotion such as anger-surprise, happy-surprise, etc. Also some of the emotions like surprise-anger, surprise-happy, anger-happy and sad-neutral appear to have similar acoustic characteristics. So a Confusion-Matrix was prepared which would take care of these uncertainties up to a certain level.

6.6 Display of Output

We have to take the input from the user. Then the system processes it and extracts its features using the MFCC coefficients. This vector is given alongside as input to the trained matrix and the evaluated output is compared with the available feature model. The GUI for this system is then prepared using MATLAB.

6.7 Results and Discussions

The use of three different language databases for emotion recognition has resulted in the following observations. The recognition accuracies were obtained using the same feature extraction method i.e. MFCC and BPNN classification techniques. Correspondingly, the confusion-matrix shows that the confusion between anger and surprise is high in comparison with any other pair of emotion. This is due to the approximation of near formant and pitch acoustic features of these two emotions. The back-propagation algorithm also proves to be an efficient method for emotion recognition with reference to the graphical result.

7. IMPLEMENTATION

7.1 Feature Extraction

The purpose of this module is to convert the speech waveform, using digital signal processing (DSP) tools, to a set of features (at a considerably lower information rate) for further analysis. This is often referred as the*signal-processing front end*.

The speech signal is a slowly timed varying signal (it is called *quasi-stationary*). An example of speech signal is shown in Figure 2. When examined over a sufficiently short period of time (between 5 and 100 msec), its characteristics are fairly stationary. However, over long periods of time (on the order of 1/5 seconds or more) the signal characteristic change to reflect the different speech sounds being spoken. Therefore, *short-time spectral analysis* is the most common way to characterize the speech signal.

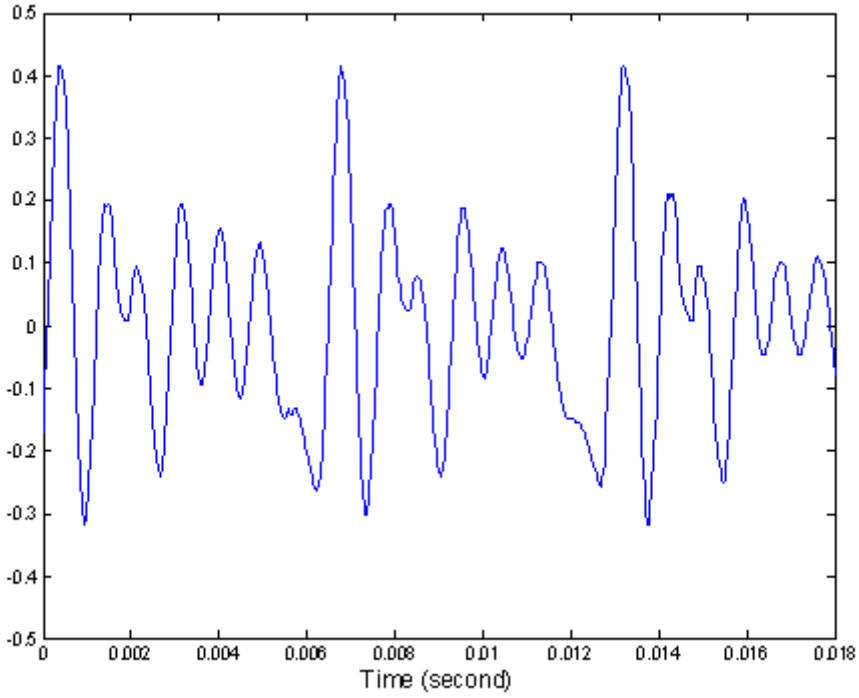

Figure 7.1 An example of speech signal

A wide range of possibilities exist for parametrically representing the speech signal for the speaker recognition task, such as Linear Prediction Coding (LPC), Mel-Frequency Cepstrum Coefficients (MFCC), and others. MFCC is perhaps the best known and most popular, and this is used in this project.

MFCCs are based on the known variation of the human ear's critical bandwidths with frequency, filters spaced linearly at low frequencies and logarithmically at high frequencies have been used to capture the phonetically important characteristics of speech. This is expressed in the *mel-frequency* scale, which is linear frequency spacing below 1000 Hz and a logarithmic spacing above 1000 Hz. Here the mel scale is being used which translates regular frequencies to a scale that is more appropriate for speech, since the human ear perceives sound in a nonlinear manner. This is useful since our whole understanding of speech is through our ears, and so the computer should know about this, too. Feature Extraction is done using MFCC processor

7.2 Mel-Frequency Cepstrum Coefficients Processor

A block diagram of the structure of an MFCC processor is given in Figure 7.1. The speech input is typically recorded at a sampling rate above 10000 Hz. This sampling frequency was chosen to minimize the effects of *aliasing* in the analog-to-digital conversion. These sampled signals can capture all frequencies up to 5 kHz, which cover most energy of sounds that are generated by humans. As been discussed previously, the main purpose of the MFCC processor is to mimic the behavior of the human ears. In addition, rather than the speech waveforms themselves, MFFC's are shown to be less susceptible to mentioned variations.

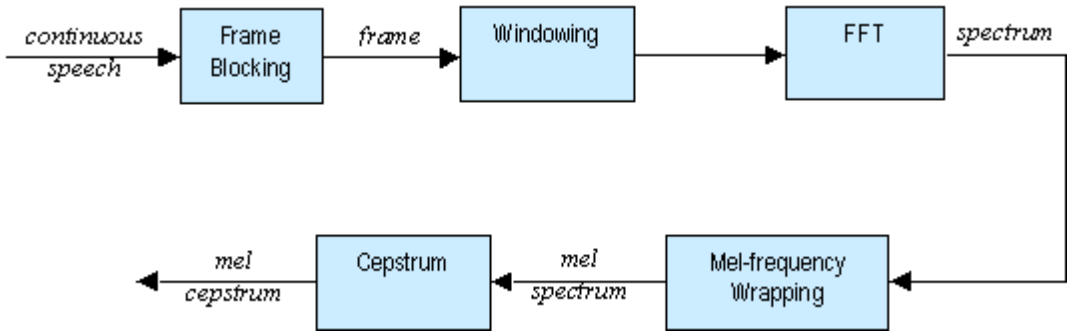

Figure 7.2 Block Diagram of the MFCC Processor

7.2.1 Frame Blocking

In this step, the continuous speech signal is blocked into frames of N samples, with adjacent frames being separated by M ($M < N$). The first frame consists of the first N samples. The second frame begins M samples after the first frame, and overlaps it by $N - M$ samples. Similarly, the third frame begins $2M$ samples after the first frame (or M samples after the second frame) and overlaps it by $N - 2M$ samples. This process continues until all the speech is accounted for within one or more frames [18].

The values for *N* and *M* are taken as *N* = 256 (which is equivalent to ~ 30 msec windowing and facilitate the fast radix-2 FFT) and *M* = 100. Frame blocking of the speech signal is done because when examined over a sufficiently short period of time (between 5 and 100 msec), its characteristics are fairly stationary. However, over long periods of time (on the order of 1/5 seconds or more) the signal characteristic change to reflect the different speech sounds being spoken. Overlapping frames are taken not to have much information loss and to maintain correlation between the adjacent frames.
N value 256 is taken as a compromise between the time resolution and frequency resolution. One can observe these time and frequency resolutions by viewing the corresponding power spectrum of speech files which was shown in the figure 3.2. In each case, frame increment M is taken as N/3.

For N = 128 we have a high resolution of time. Furthermore each frame lasts for a very short period of time. This result shows that the signal for a frame doesn't change its nature. On the other hand, there are only 65 distinct frequencies samples. This means that we have a poor frequency resolution.

For N = 512 we have an excellent frequency resolution (256 different values) but there are lesser frames, meaning that the resolution in time is strongly reduced.

It seems that a value of 256 for N is an acceptable compromise. Furthermore the number of frames is relatively small, which will reduce computing time.
So, finally for N = 256 we have a compromise between the resolution in time and the resolution in frequency.

7.2.2 Windowing

The next step in the processing is to window each individual frame so as to minimize the signal discontinuities at the beginning and end of each frame. The concept here is to minimize the spectral distortion by using the window to taper the signal to zero at the beginning and end of each frame. If we define the window as $w(n), 0 \le n \le N-1$, where *N* is the number of samples in each frame, then the result of windowing is the signal.

$$y_l(n) = x_l(n)w(n), \quad 0 \le n \le N-1$$

Typically the *Hamming* window is used, which has the form:

$$w(n) = 0.54 - 0.46\cos\left(\frac{2\pi n}{N-1}\right), \quad 0 \le n \le N-1$$

7.2.3 Fast Fourier Transform

The next processing step is the Fast Fourier Transform, which converts each frame of N samples from the time domain into the frequency domain. The FFT is a fast algorithm to implement the Discrete Fourier Transform (DFT), which is defined on the set of N samples $\{x_n\}$, as follow:

$$X_k = \sum_{n=0}^{N-1} x_n e^{-j2\pi kn/N}, \qquad k = 0,1,2,\ldots,N-1$$

In general X_k's are complex numbers and we only consider their absolute values (frequency magnitudes). The resulting sequence $\{X_k\}$ is interpreted as follow: positive frequencies $0 \le f < F_s/2$ correspond to values $0 \le n \le N/2-1$, while negative frequencies $-F_s/2 < f < 0$ correspond to $N/2+1 \le n \le N-1$. Here, F_s denotes the sampling frequency.

The result after this step is often referred to as *spectrum* or *periodogram*.

7.2.4 Mel-frequency Warping

As mentioned above, psychophysical studies have shown that human perception of the frequency contents of sounds for speech signals does not follow a linear scale. Thus for each tone with an actual frequency, *f*, measured in Hz, a subjective pitch is measured on a scale called the 'mel' scale. The *mel-frequency* scale is a linear frequency spacing below 1000 Hz and a logarithmic spacing above 1000 Hz.

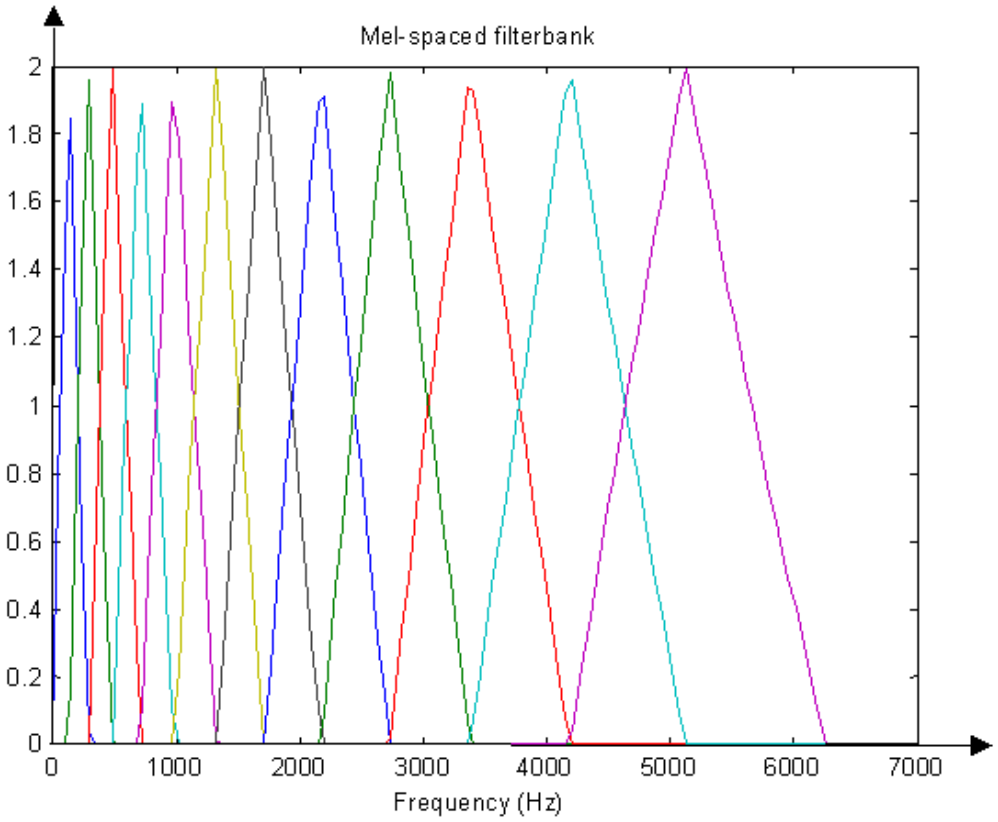

Figure 7.3 An example of mel-spaced filter bank

One approach to simulating the subjective spectrum is to use a filter bank, spaced uniformly on the mel-scale (see Figure 4). That filter bank has a triangular bandpass frequency response, and the spacing as well as the bandwidth is determined by a constant mel frequency interval. The number of mel spectrum coefficients, *K*, is typically chosen as 20. Note that this filter bank is applied in the frequency domain, thus it simply amounts to applying the triangle-shape windows as in the Figure 4 to the spectrum. A useful way of thinking about this mel-wrapping filter bank is to view each filter as a histogram bin (where bins have overlap) in the frequency domain.

7.2.5 Cepstrum

In this final step, we convert the log mel spectrum back to time. The result is called the mel frequency cepstrum coefficients (MFCC). The cepstral representation of the speech spectrum provides a good representation of the local spectral properties of the signal for the given frame analysis. Because the mel spectrum coefficients (and so their logarithm) are real numbers, we can convert them to the time domain using the Discrete Cosine Transform (DCT). Therefore if we denote those mel power spectrum coefficients

that are the result of the last step are $\tilde{S}_0, k = 0,2,\ldots, K-1$, we can calculate the MFCC's, $\tilde{c}_n$, as

$$\tilde{c}_n = \sum_{k=1}^{K} (\log \tilde{S}_k) \cos\left[n\left(k - \frac{1}{2}\right)\frac{\pi}{K}\right], \qquad n = 0,1,\ldots, K\text{-}1$$

Note that we exclude the first component, $\tilde{c}_0$, from the DCT since it represents the mean value of the input signal, which carried little speaker specific information.

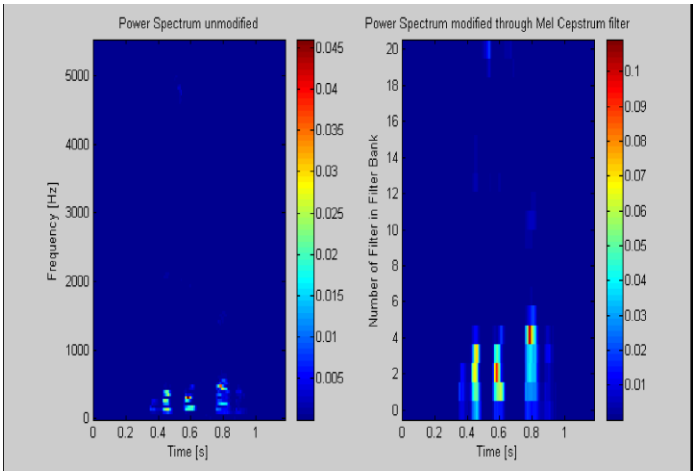

Figure 7.4 Power spectrum modified through mel spaced filter bank

8. VECTOR QUANTIZATION

8.1 Introduction

The problem of speaker recognition belongs to a much broader topic in scientific and engineering so called *pattern recognition*. The goal of pattern recognition is to classify objects of interest into one of a number of categories or classes. The objects of interest are generically called *patterns* and in our case are sequences of acoustic vectors that are extracted from an input speech using the techniques described in the previous section. The classes here refer to individual speakers. Since the classification procedure in our case is applied on extracted features, it can be also referred to as *feature matching*.

Furthermore, if there exist some set of patterns whose individual classes are already known, then one has a problem in *supervised pattern recognition*. This is exactly our case, since during the training session, we label each input speech with the ID of the speaker (S1 to S8). These patterns comprise the *training set* and are used to derive a classification algorithm. The remaining patterns are then used to test the classification algorithm; these patterns are collectively referred to as the *test set*. If the correct classes of the individual patterns in the test set are also known, then one can evaluate the performance of the algorithm.

The state-of-the-art in feature matching techniques used in speaker recognition includes Dynamic Time Warping (DTW), Hidden Markov Modeling (HMM), and Vector Quantization (VQ). In this project, the VQ approach will be used, due to ease of implementation and high accuracy. VQ is a process of mapping vectors from a large vector space to a finite number of regions in that space. Each region is called a *cluster* and can be represented by its center called a *codeword*. The collection of all code words is called a *codebook*.

8.2 Speaker Modeling

Using Cepstral analysis as described in the previous section, an utterance may be represented as a sequence of feature vectors. Utterances spoken by the same person but at different times result in similar yet a different sequence of feature vectors. The purpose of voice modeling is to build a model that captures these variations in the extracted set of features.

There are two types of models that have been used extensively in speaker recognition systems**:** stochastic models and template models .The stochastic model treats the speech production process as a parametric random process and assumes that the parameters of the underlying stochastic process can be estimated in a precise, well defined manner. The template model attempts to model the speech production process in a non-parametric manner by retaining a number of sequences of feature vectors derived from multiple utterances of the same word by the same person. Template models dominated early work in speaker recognition because the template model is intuitively more reasonable.

However, recent work in stochastic models has demonstrated that these models are more flexible and hence allow for better modeling of the speech production process. The state-of-the-art in feature matching techniques used in speaker recognition includes Dynamic Time Warping (DTW), Hidden Markov Modeling (HMM), and Vector Quantization (VQ).

In a speaker recognition system, each speaker must be uniquely represented in an efficient manner. This process is known as vector quantization. Vector quantization is the process of mapping vectors from a large vector space to a finite number of regions in that space. Each region is called a *cluster* and can be represented by its center called a codeword.

The collection of all codewords is called a codebook. The data is thus significantly compressed, yet still accurately represented. Without quantizing the feature vectors, the system would be too large and computationally complex. In a speaker recognition system, the vector space contains a speaker's characteristic vectors, which are obtained from the feature extraction described above. After the completion of vector quantization, only a few representative vectors remain, and these are collectively known as the speaker's *codebook.* The codebook then serves as delineation for the speaker, and is used when training a speaker in the system.

8.3 Vector Quantization

Vector quantization *(VQ)* is the process of taking a large set of feature vectors and producing a smaller set of feature vectors that represent the centroids of the distribution, i.e. points spaced so as to minimize the average distance to every other point. We use vector quantization since it would be impractical to store every single feature vector that we generate from the training utterance [8][11]. While the VQ algorithm does take a while to compute, it saves time during the testing phase, and therefore is a compromise that we can live with.

A vector quantizer maps *k-dimensional* vectors in the vector space *Rk* into a finite set of vectors $Y = \{yi: i = 1, 2, ..., N\}$. Each vector *yi is* called a code vector or a *codeword* and the set of all the codewords is called a *codebook.* Associated with each codeword, *yi*, is a nearest neighbor region called Voronoi region, and it is defined by:

$$V_i = \left\{ x \in R^k : \|x - y_i\| \leq \|x - y_j\|, \, for\, all\, j \neq i \right\}$$

The set of Voronoi regions partition the entire space *Rk* such that:

$$\bigcup_{i=1}^{N} V_i = R^k$$

$$\bigcap_{i=1}^{N} V_i = \phi \qquad \text{for all } i \neq j$$

As an example, take vectors in the two dimensional case without loss of generality. Figure 7.4 shows some vectors in space. Associated with each cluster of vectors is a representative codeword. Each codeword resides in its own Voronoi region. These regions are separated with imaginary lines in figure 3.6 for illustration. Given an input vector, the codeword that is chosen to represent it is the one in the same Voronoi region. The representative codeword is determined to be the closest in Euclidean distance from the input vector. The Euclidean distance is defined by:

$$d(x, y_i) = \sqrt{\sum_{j=1}^{k} (x_j - y_{ij})^2}$$

Where x_j is the jth component of the input vector, and y_{ij} is the jth is component of the codeword y_i.

8.4 Clustering Mechanism

In the figure, only two speakers and two dimensions of the acoustic space are shown. The circles refer to the acoustic vectors from the speaker 1 while the triangles are from the speaker 2. In the training phase, using the clustering algorithm described in Section 4.2, a *speaker-specific* VQ codebook is generated for each known speaker by clustering his/her training acoustic vectors. The result codewords (centroids) are shown in Figure 5 by black circles and black triangles for speaker 1 and 2, respectively. The distance from a vector to the closest codeword of a codebook is called a VQ-distortion. In the recognition phase, an input utterance of an unknown voice is "vector-quantized" using each trained codebook and the *total VQ distortion* is computed. The speaker corresponding to the VQ codebook with smallest total distortion is identified as the speaker of the input utterance.

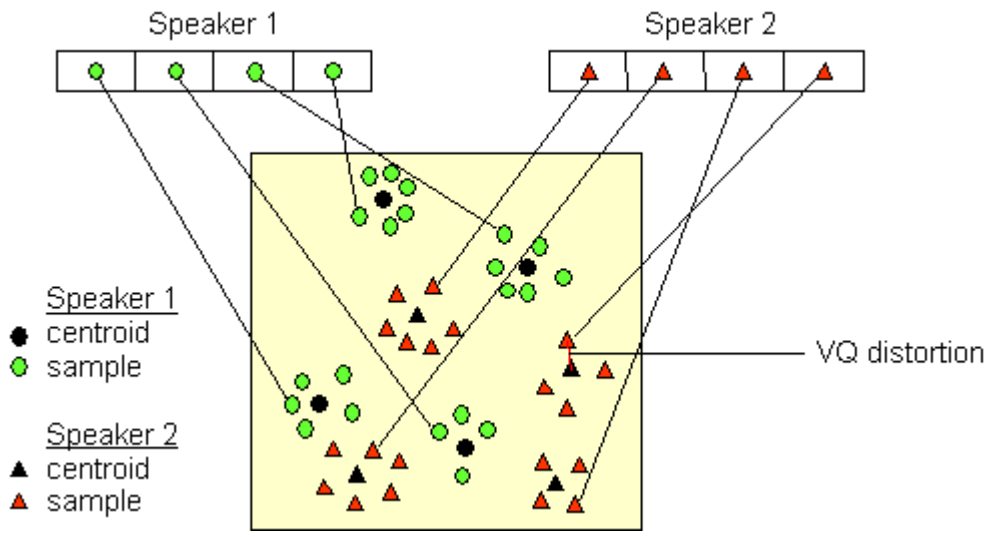

Figure 8.1 Conceptual diagram illustrating vector quantization codebook formation. One speaker can be discriminated from another based of the location of centroids.

After the enrolment session, the acoustic vectors extracted from input speech of each speaker provide a set of training vectors for that speaker. As described above, the next important step is to build a speaker-specific VQ codebook for each speaker using those training vectors. There is a well-know algorithm, namely LBG algorithm [Linde, Buzo and Gray, 1980], for clustering a set of L training vectors into a set of Mcodebook vectors. The algorithm is formally implemented by the following recursive procedure:

1. Design a 1-vector codebook; this is the centroid of the entire set of training vectors (hence, no iteration is required here).
2. Double the size of the codebook by splitting each current codebook $\mathbf{y}_n$ according to the rule

$$\mathbf{y}_n^+ = \mathbf{y}_n(1+\varepsilon)$$
$$\mathbf{y}_n^- = \mathbf{y}_n(1-\varepsilon)$$

where n varies from 1 to the current size of the codebook, and ε is a splitting parameter (we choose ε=0.01).

3. Nearest-Neighbor Search: for each training vector, find the codeword in the current codebook that is closest (in terms of similarity measurement), and assign that vector to the corresponding cell (associated with the closest codeword).

4. Centroid Update: update the codeword in each cell using the centroid of the training vectors assigned to that cell.

5. Iteration 1: repeat steps 3 and 4 until the average distance falls below a preset threshold

6. Iteration 2: repeat steps 2, 3 and 4 until a codebook size of *M* is designed.

Intuitively, the LBG algorithm designs an *M*-vector codebook in stages. It starts first by designing a 1-vector codebook, then uses a splitting technique on the codewords to initialize the search for a 2-vector codebook, and continues the splitting process until the desired *M*-vector codebook is obtained.

Figure 8.2 shows, in a flow diagram, the detailed steps of the LBG algorithm. "*Cluster vectors*" is the nearest-neighbor search procedure which assigns each training vector to a cluster associated with the closest codeword. "*Find centroids*" is the centroid update procedure. "*Compute D (distortion)*" sums the distances of all training vectors in the nearest-neighbor search so as to determine whether the procedure has converged.

Concluding this section, we have elaborated the introduction to Vector Quantization and the Linde, Buzo and Gray algorithm for VQ is discussed, and formation of a speaker specific codebook is formed using LBG VQ algorithm on the MFCC's obtained. This is explained in the figure 8.3.

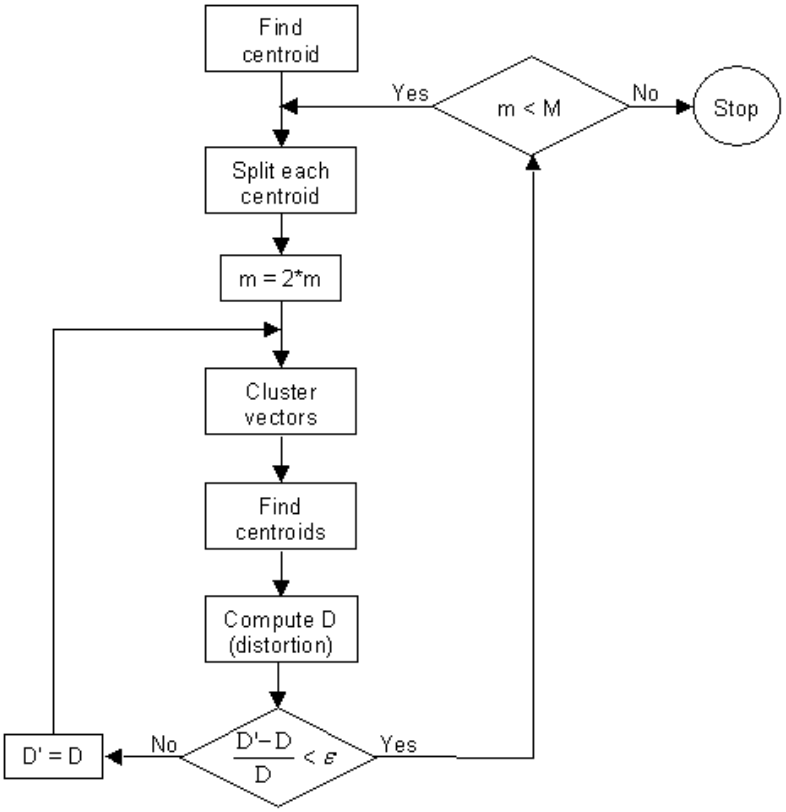

Figure 8.2 Flow diagram of the LBG algorithm
(Adapted from Rabiner and Juang, 1993)

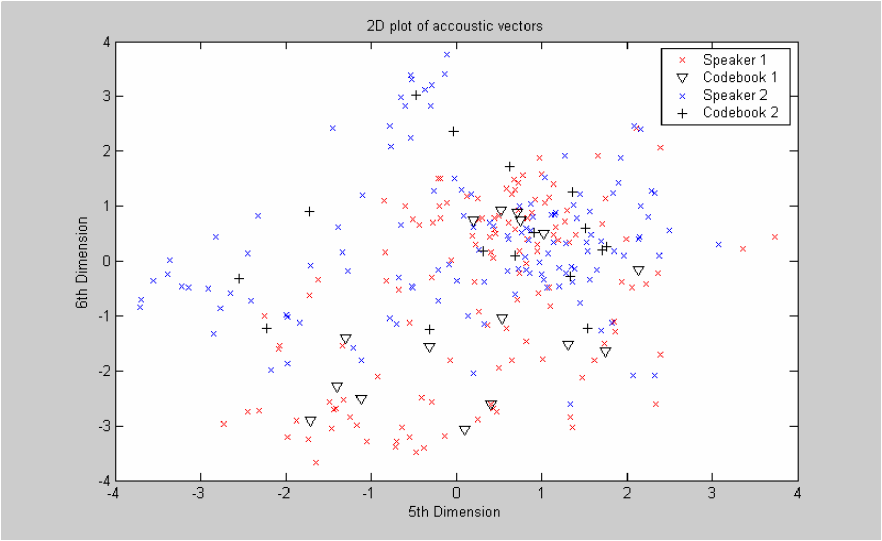

Figure 8.3 Codebooks and MFCCs corresponding to speaker 1 and 2.

9. CLASSIFIER

9.1 Neural Network

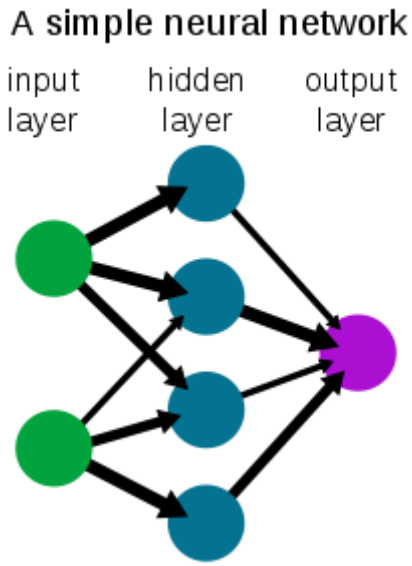

Figure 9.1 A simple neural network

The term neural network was traditionally used to refer to a network or circuit of biological neurons. The modern usage of the term often refers to artificial neural networks, which are composed of artificial neurons or nodes. Thus the term has two distinct usages:

1. Biological neural networks are made up of real biological neurons that are connected or functionally related in the peripheral nervous system or the central nervous system. In the field of neuroscience, they are often identified as groups of neurons that perform a specific physiological function in laboratory analysis.
2. Artificial neural networks are composed of interconnecting artificial neurons (programming constructs that mimic the properties of biological neurons). Artificial neural networks may either be used to gain an understanding of biological neural networks, or for solving artificial intelligence problems without necessarily creating a model of a real biological system. The real, biological nervous system is highly complex: artificial neural network algorithms attempt to

abstract this complexity and focus on what may hypothetically matter most from an information processing point of view. Good performance (e.g. as measured by good predictive ability, low generalization error), or performance mimicking animal or human error patterns, can then be used as one source of evidence towards supporting the hypothesis that the abstraction really captured something important from the point of view of information processing in the brain. Another incentive for these abstractions is to reduce the amount of computation required to simulate artificial neural networks, so as to allow one to experiment with larger networks and train them on larger data sets.

This article focuses on the relationship between the two concepts; for detailed coverage of the two different concepts refer to the separate articles: Biological neural network and Artificial Neural Network.

A biological neural network is composed of a group or groups of chemically connected or functionally associated neurons. A single neuron may be connected to many other neurons and the total number of neurons and connections in a network may be extensive. Connections, called synapses, are usually formed from axons to dendrites, though dendrodendritic microcircuits and other connections are possible. Apart from the electrical signaling, there are other forms of signaling that arise from neurotransmitter diffusion, which have an effect on electrical signaling. As such, neural networks are extremely complex.

Artificial intelligence and cognitive modeling try to simulate some properties of biological neural networks. While similar in their techniques, the former has the aim of solving particular tasks, while the latter aims to build mathematical models of biological neural systems.

In the artificial intelligence field, artificial neural networks have been applied successfully to speech recognition, image analysis and adaptive control, in order to construct software agents(in computer and video games) or autonomous robots. Most of the currently employed artificial neural networks for artificial intelligence are based on statistical estimation, optimization and control theory.

The cognitive modeling field involves the physical or mathematical modeling of the behaviour of neural systems; ranging from the individual neural level (e.g. modeling the spike response curves of neurons to a stimulus), through the neural cluster level (e.g. modeling the release and effects of dopamine in the basal ganglia) to the complete organism (e.g. behavioural modeling of the organism's response to stimuli). Artificial intelligence, cognitive modeling, and neural networks are information processing paradigms inspired by the way biological neural systems process data.

9.2 The brain, computer and neural networks:

Neural networks, as used in artificial intelligence, have traditionally been viewed as simplified models of neural processing in the brain, even though the relation between this model and brain biological architecture is debated, as little is known about how the brain actually works.

A subject of current research in theoretical neuroscience is the question surrounding the degree of complexity and the properties that individual neural elements should have to reproduce something resembling animal intelligence.

Historically, computers evolved from the von Neumann architecture, which is based on sequential processing and execution of explicit instructions. On the other hand, the origins of neural networks are based on efforts to model information processing in biological systems, which may rely largely on parallel processing as well as implicit instructions based on recognition of patterns of 'sensory' input from external sources. In other words, at its very heart a neural network is a complex statistical processor (as opposed to being tasked to sequentially process and execute).

Neural coding is concerned with how sensory and other information is represented in the brain by neurons. The main goal of studying neural coding is to characterize the relationship between the stimulus and the individual or ensemble neuronal responses and the relationship among electrical activity of the neurons in the ensemble. It is thought that neurons can encode both digital and analog information.

9.3 Neural networks and artificial intelligence

Artificial neural network

A neural network (NN), in the case of artificial neurons called artificial neural network (ANN) or simulated neural network (SNN), is an interconnected group of natural or artificial neurons that uses a mathematical or computational model for information processing based on a connectionist approach to computation. In most cases an ANN is an adaptive system that changes its structure based on external or internal information that flows through the network.

In more practical terms neural networks are non-linear statistical data modeling or decision making tools. They can be used to model complex relationships between inputs and outputs or to find patterns in data.

However, the paradigm of neural networks - i.e., implicit, not explicit , learning is stressed - seems more to correspond to some kind of natural intelligence than to the traditional symbol-based Artificial Intelligence, which would stress, instead, rule-based learning.

KNN

In pattern recognition, the k –nearest neighbor algorithm (KNN) is a method fr classifying objects based on closest training examples in the feature space. KNN is a type of instance based learning or learning where the function is only approximated locally and all computation is deferred until classification. The KNN algorithm is amongst the simplest of all, machine learning algorithms.

10. BACK PROPAGATION

10.1 Multi Layer Back Propagation Algorithm

A single-layer network has severe restrictions: the class of tasks that can be accomplished is very limited. Minsky and Papert (Minsky & Papert, 1969) showed in 1969 that a two layer feed-forward network can overcome many restrictions, but did not present a solution to the problem of how to adjust the weights from input to hidden units. The central idea behind this solution is that the errors for the units of the hidden layer are determined by back-propagating the errors of the units of the output layer. For this reason the method is often called the back-propagation learning rule. Back-propagation can also be considered as a generalization of the delta rule for non-linear activation functions and multilayer networks.

10.2 Multi Layer Feed Forward Networks

A feed-forward network has a layered structure. Each layer consists of units which receive their input from units from a layer directly below and send their output to units in a layer directly above the unit. There are no connections within a layer. The Ni inputs are fed into the first layer of Nh;1 hidden units. The input units are merely 'fan-out' units; no processing takes place in these units. The activation of a hidden unit is a function Fi of the weighted inputs plus a bias, as given in equation

$$y_k(t+1) = \mathcal{F}_k(s_k(t)) = \mathcal{F}_k\left(\sum_j w_{jk}(t)\, y_j(t) + \theta_k(t)\right),$$

The output of the hidden units is distributed over the next layer of Nh;2 hidden units, until the last layer of hidden units, of which the outputs are fed into a layer of No output units .

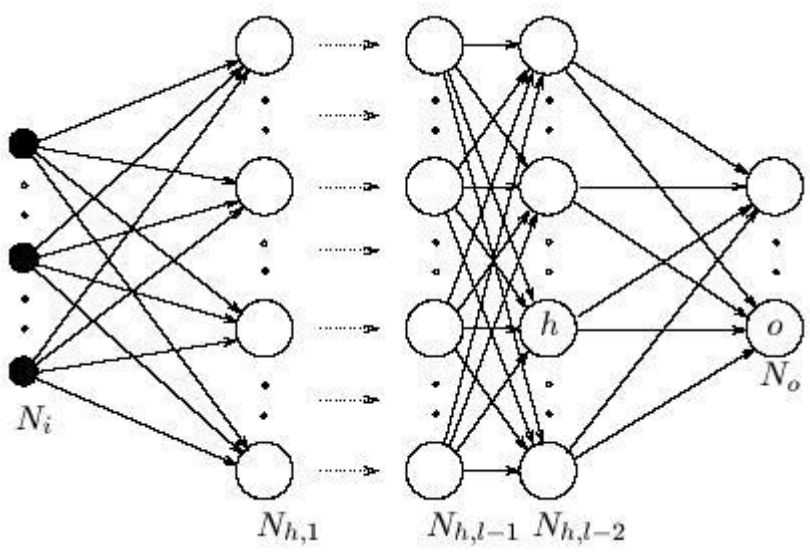

Figure 10.1 Back propagation network [16]

Although Back Propagation can be applied to networks with any number of layers, just as for networks with binary units it has been shown (Hornik, Stinchcombe, & White, 1989; Funahashi, 1989; Cybenko, 1989; Hartman, Keeler, & Kowalski, 1990) that only one layer of hidden units success to approximate any function with finitely many discontinuities to arbitrary precision, provided the activation functions of the hidden units are non-linear (the universal approximation theorem). In most applications a feed-forward network with a single layer of hidden units is used with a sigmoid activation function for the units.

10.3 Delta Rule

Since we are now using units with nonlinear activation functions, we have to generalize the delta rule:

The activation is a differentiable function of the total input, given by

$y_k^p = \mathcal{F}(s_k^p)$, in which $s_k^p = \sum_j w_{jk} y_j^p + \theta_k$. To get the correct generalisation of the delta rule as presented in the previous chapter, we must set $\Delta_p w_{jk} = -\gamma \frac{\partial E^p}{\partial w_{jk}}$.

The error measure Ep is defined as the total quadratic error for pattern p at the output units:

$E^p = \frac{1}{2} \sum_{o=1}^{N_o} (d_o^p - y_o^p)^2$, where dpo is the desired output for unit o when pattern p is clamped. We further set $E = \sum_p E^p$ as the summed squared error. We can write $\frac{\partial E^p}{\partial w_{jk}} = \frac{\partial E^p}{\partial s_k^p} \frac{\partial s_k^p}{\partial w_{jk}}$.

By equation $s_k^p = \sum_j w_{jk} y_j^p + \theta_k$ we see that the second factor is $\frac{\partial s_k^p}{\partial w_{jk}} = y_j^p$. When we define $\delta_k^p = -\frac{\partial E^p}{\partial s_k^p}$, we will get an update rule which is equivalent to the delta rule as described in the previous chapter, resulting in a gradient descent on the error surface if we make the weight changes according to: $\Delta_p w_{jk} = \gamma \delta_k^p y_j^p$. The trick is to figure out what δp k should be for each unit k in the network. The interesting result, which we now derive, is that there is a simple recursive computation of these δ's which can be implemented by propagating error signals backward through the network.

To compute δp k we apply the chain rule to write this partial derivative as the product of two factors, one factor reflecting the change in error as a function of the output of the unit

and one re ecting the change in the output as a function of changes in the input. Thus, we have

$$\delta_k^p = -\frac{\partial E^p}{\partial s_k^p} = -\frac{\partial E^p}{\partial y_k^p}\frac{\partial y_k^p}{\partial s_k^p}.$$

Let us compute the second factor. By equation $y_k^p = \mathcal{F}(s_k^p),$ we see that $\frac{\partial y_k^p}{\partial s_k^p} = \mathcal{F}'(s_k^p),$ which is the same result as we obtained with the standard delta rule. Substituting this and equation $\delta_k^p = -\frac{\partial E^p}{\partial s_k^p} = -\frac{\partial E^p}{\partial y_k^p}\frac{\partial y_k^p}{\partial s_k^p}.$ in equation $\frac{\partial y_k^p}{\partial s_k^p} = \mathcal{F}'(s_k^p),$, we get $\delta_o^p = (d_o^p - y_o^p)\,\mathcal{F}_o{}'(s_o^p)$

for any output unit o. Secondly, if k is not an output unit but a hidden unit k = h, we do not readily know the contribution of the unit to the output error of the network. However, the error measure can be written as a function of the net inputs from hidden to output layer; Ep = Ep(sp 1,sp 2,..... sp j.....) and we use the chain rule to write

$$\frac{\partial E^p}{\partial y_h^p} = \sum_{o=1}^{N_o}\frac{\partial E^p}{\partial s_o^p}\frac{\partial s_o^p}{\partial y_h^p} = \sum_{o=1}^{N_o}\frac{\partial E^p}{\partial s_o^p}\frac{\partial}{\partial y_h^p}\sum_{j=1}^{N_h} w_{ko}y_j^p = \sum_{o=1}^{N_o}\frac{\partial E^p}{\partial s_o^p}w_{ho} = -\sum_{o=1}^{N_o}\delta_o^p w_{ho}.$$

Substituting this in equation $\delta_k^p = -\frac{\partial E^p}{\partial s_k^p} = -\frac{\partial E^p}{\partial y_k^p}\frac{\partial y_k^p}{\partial s_k^p}.$ yields $\delta_h^p = \mathcal{F}'(s_h^p)\sum_{o=1}^{N_o}\delta_o^p w_{ho}.$

Equations $\delta_o^p = (d_o^p - y_o^p)\,\mathcal{F}_o{}'(s_o^p)$ and $\delta_h^p = \mathcal{F}'(s_h^p)\sum_{o=1}^{N_o}\delta_o^p w_{ho}.$ give a recursive procedure for computing the δ's for all units in

the network, which are then used to compute the weight changes according to equation. This procedure constitutes the generalised delta rule for a feed-forward network of non-linear
units.

10.4 Understanding Back Propagation

The equations derived in the previous section may be mathematically correct, but what do they actually mean? Is there a way of understanding back-propagation other than reciting the necessary equations? The answer is, of course, yes. In fact, the whole back-propagation process is intuitively very clear. What happens in the above equations is the following. When a learning pattern is clamped, the activation values are propagated to the output units, and the actual network output is compared with the desired output values, we usually end up with an error in each of the output units. Let's call this error eo for a particular output unit o. We have to bring eo to zero The simplest method to do this is the greedy method: we strive to change the connections in the neural network in such a way that, next time around, the error eo will be zero for this particular pattern. We know from the delta rule that, in order to reduce an error, we have to adapt its incoming weights according to-

$$\Delta w_{ho} = (d_o - y_o) y_h.$$

That's step one. But it alone is not enough: when we only apply this rule, the weights from input to hidden units are never changed, and we do not have the full representational power of the feed-forward network as promised by the universal approximation theorem. In order to adapt the weights from input to hidden units, we again want to apply the delta rule. In this case, however, we do not have a value for δ for the hidden units. This is solved by the chain rule which does the following: distribute the error of an output unit o to all the hidden units that is it connected to, weighted by this connection. Differently put, a hidden unit h receives a delta from each output unit o equal to the delta of that output unit weighted with (= multiplied by) the weight of the connection between those units.

10.5 Working with Back-Propagation

The application of the generalised delta rule thus involves two phases: During the first phase the input x is presented and propagated forward through the network to compute the output values yp o for each output unit. This output is compared with its desired value do, resulting in an error signal δp o for each output unit. The second phase involves a backward pass through the network during which the error signal is passed to each unit in the network and appropriate weight changes are calculated.

Weight adjustments with sigmoid activation function.

- The weight of a connection is adjusted by an amount proportional to the product of an error signal δ, on the unit k receiving the input and the output of the unit j sending this signal along the connection: $\Delta_p w_{jk} = \gamma \delta_k^p y_j^p.$
- If the unit is an output unit, the error signal is given by $\delta_o^p = (d_o^p - y_o^p)\, \mathcal{F}'(s_o^p).$ Take as the activation function F the 'sigmoid' function as defined $y^p = \mathcal{F}(s^p) = \frac{1}{1+e^{-s^p}}.$ In this case the derivative is equal to

$$\begin{aligned} \mathcal{F}'(s^p) &= \frac{\partial}{\partial s^p}\frac{1}{1+e^{-s^p}} \\ &= \frac{1}{(1+e^{-s^p})^2}(-e^{-s^p}) \\ &= \frac{1}{(1+e^{-s^p})}\frac{e^{-s^p}}{(1+e^{-s^p})} \\ &= y^p(1-y^p). \end{aligned}$$

such that the error signal for an output unit can be written as: $\delta_o^p = (d_o^p - y_o^p)\, y_o^p(1-y_o^p).$

- The error signal for a hidden unit is determined recursively in terms of error signals of the units to which it directly connects and the weights of those connections.

For the sigmoid activation function:

$$\delta_h^p = \mathcal{F}'(s_h^p) \sum_{o=1}^{N_o} \delta_o^p w_{ho} = y_h^p (1 - y_h^p) \sum_{o=1}^{N_o} \delta_o^p w_{ho}.$$

10.6 Learning rate and momentum

The learning procedure requires that the change in weight is proportional to $\partial E^p / \partial w$ True gradient descent requires that large number of steps are taken. The constant of proportionality is the learning rate . For practical purposes we choose a learning rate that is as large as possible without leading to oscillation. One way to avoid oscillation at large , is to make the change in weight dependent of the past weight change by adding a momentum term:

$\Delta w_{jk}(t+1) = \gamma \delta_k^p y_j^p + \alpha \Delta w_{jk}(t)$, where t indexes the presentation number and F is a constant which determines the effect of the previous weight change.

11. DATABASE

11.1 German Database

Data based Used:

We are making use of Berlin database. Speech samples collection contains about 500 utterances spoken by actors in a happy, angry, anxious, fearful, bored and disgusted way as well as in a neutral version.

Every utterance is named according to the same scheme:

- Positions 1-2: number of speaker
- Positions 3-5: code for text
- Position 6: emotion (sorry, letter stands for German emotion word)
- Position 7: if there are more than two versions these are numbered a, b, c ....

Example: 03a01Fa.wav is the audio file from Speaker 03 speaking text a01 with the emotion "Freude" (Happiness).

Information about the speakers

- 03 - male, 31 years old
- 08 - female, 34 years
- 09 - female, 21 years
- 10 - male, 32 years
- 11 - male, 26 years
- 12 - male, 30 years
- 13 - female, 32 years
- 14 - female, 35 years
- 15 - male, 25 years
- 16 - female, 31 years

Table 11.1 The translation text of Berlin Database

	Text (german)	Closest English Translation
a01	Der Lappen liegt auf dem Eisschrank.	The tablecloth is lying on the frigde.
a02	Das will sie am Mittwoch abgeben.	She will hand it in on Wednesday.
a04	Heute abend könnte ich es ihm sagen.	Tonight I could tell him.
a05	Das schwarze Stück Papier befindet sich da oben neben dem Holzstück.	The black sheet of paper is located up there besides the piece of timber.
a07	In sieben Stunden wird es soweit sein.	In seven hours it will be.
b01	Was sind denn das für Tüten, die da unter dem Tisch stehen?	What about the bags standing there under the table?
b02	Sie haben es gerade hochgetragen und jetzt gehen sie wieder runter.	They just carried it upstairs and now they are going down again.
b03	An den Wochenenden bin ich jetzt immer nach Hause gefahren und habe Agnes besucht.	Currently at the weekends I always went home and saw Agnes.
b09	Ich will das eben wegbringen und dann mit Karl was trinken gehen.	I will just discard this and then go for a drink with Karl.
b10	Die wird auf dem Platz sein, wo wir sie immer hinlegen.	It will be in the place where we always store it.

Table 11.2 Code of Emotion from Berlin Database

Letter	emotion (English)	letter	emotion (German)
A	Anger	W	Ärger (Wut)
B	Boredom	L	Langeweile
D	Disgust	E	Ekel
F	anxiety/fear	A	Angst
H	Happiness	F	Freude
S	Sadness	T	Trauer

12. PROGRAMMING

12.1 Software Used

PRAAT:

Praat is a program for doing phonetic analyses and sound manipulations (Boersma and Weenink (1992 2001)). It is available for many different platforms (Windows, Macintosh, Unix, Linux). Using praat a variety of files may be read. Such as follows:

1. Sound files (.wav, .aiff files etc.)
2. TextGrid files (label files)
3. Formant files
4. Spectrogram files

It also allows us to perform the following functions.

1. Play: Plays the Sound object
2. Edit: Allows you to view the waveform, spectrogram, formants, and do various types of
3. queries.
4. Draw or Paint: Makes a picture of the selected object. See below.
5. Periodicity: Calculates (called "Pitch" in PRATT).
6. Spectrum: Calculates spectra and spectrograms.
7. Formants & LPC: Calculates formants.
8. To Manipulation: Resynthesizes.

One of the very many advantages of PRAAT is that it includes a scripting language that allows you to automate or semi-automate labeling, phonetic analyses, and sound manipulations.

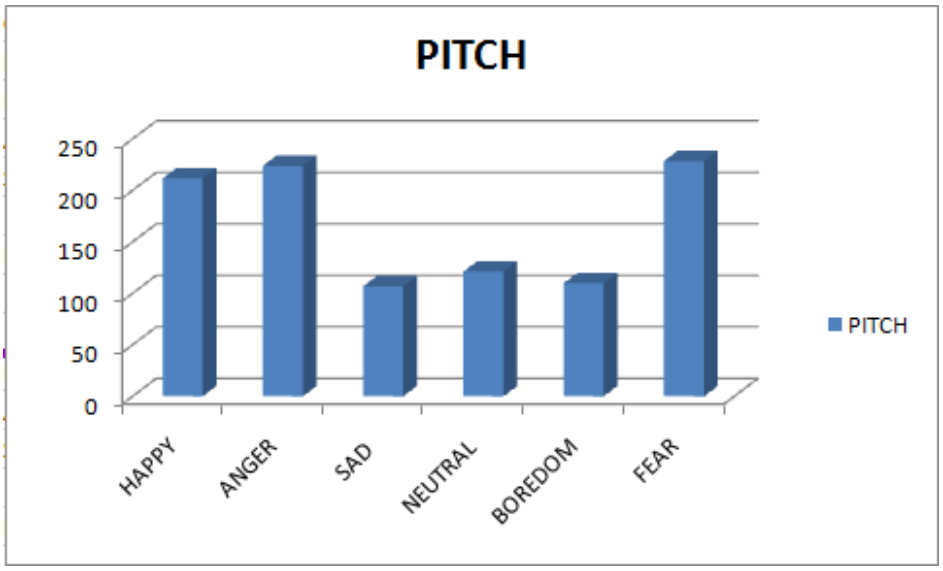

Figure 12.1 Variation of Pitch with emotions

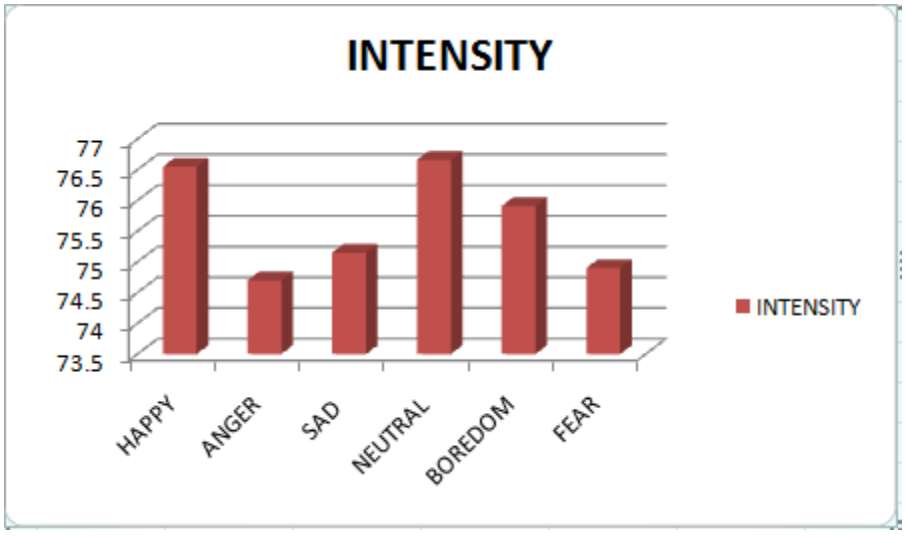

Figure 12.2 Variation of Intensity with emotions

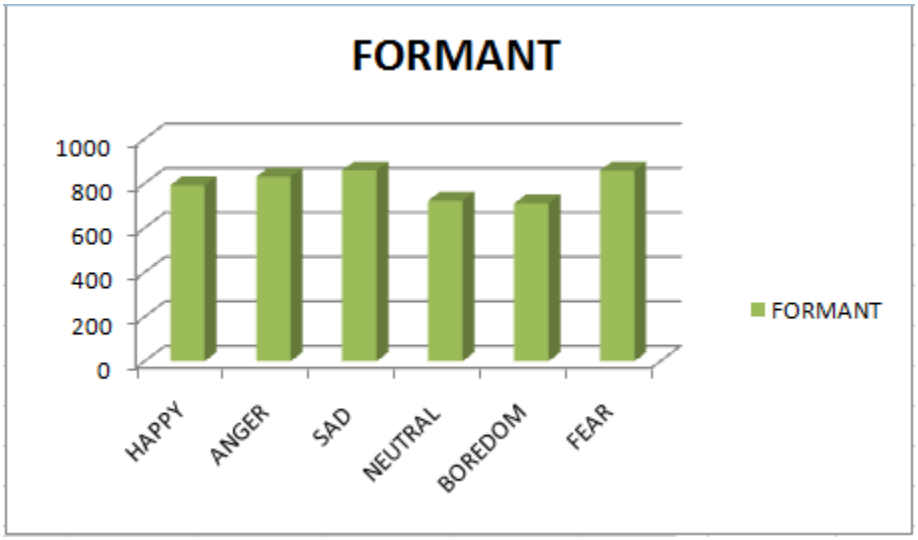

Figure 12.3 Variation of Formant Frequency with emotions

Conclusion:

It was observed that the pitch was greater in case of active emotions like happiness and anger and lower comparatively for other emotions. The intensity of happy and neutral states was higher compared to others. Passive emotions like sad and fear dominated the formant frequency.

MATLAB:

12.2 MATLAB introduction and advantages

MATLAB is the name used to refer to the class of matrix calculator environments derived from first, called MATLAB. MATLAB was developed by Cleve Moler in late 1970s at the University of New Mexico and other locations with support from National Science Foundation. Since then, many work –a-likes have been developed or are in development commercially, in government and university labs, and so forth.

MATLAB is a high performance language for technical computing. It integrates computation, visualization, and programming in an easy-to-use and interactive environment where problems and solutions are expressed in familiar mathematical notation.

MATLAB is an interactive system whose basic data element is an array that doesn't require dimensioning. This allows you to solve many technical computing problems, especially those with matrix and vector formulations, in a fraction of time it would take a program in scalar non-interactive languages as C or Fortran.

The name of product available from MathWorks is "MATLAB". MathWorks holds a registered trademark on the uppercase version of MATLAB. For our project we mainly rely on the image processing toolbox of MATLAB.

Advantages of MATLAB:

- High-level language for technical computing
- Development environment for managing code, files, and data
- Interactive tools for iterative exploration, design, and problem solving
- Mathematical functions for linear algebra, statistics, Fourier analysis, filtering, optimization, and numerical integration
- 2-D and 3-D graphics functions for visualizing data

- Tools for building custom graphical user interfaces

Result analysis:

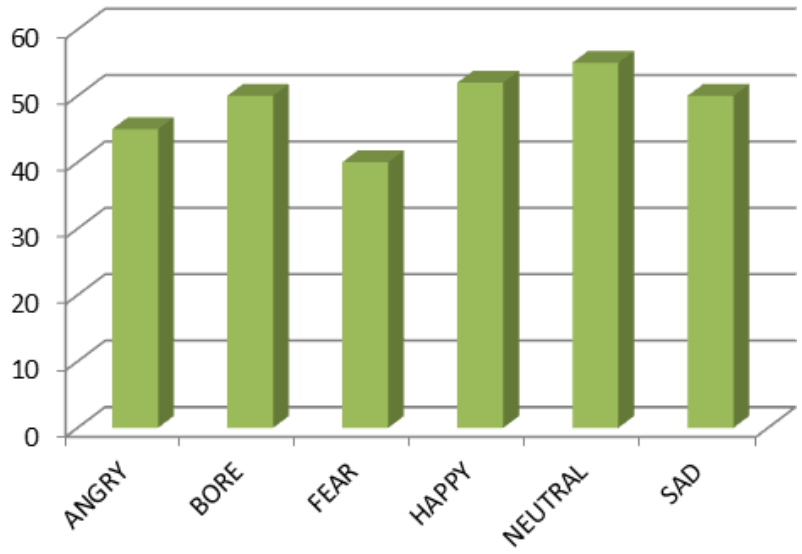

Figure 12.4 Using LPC

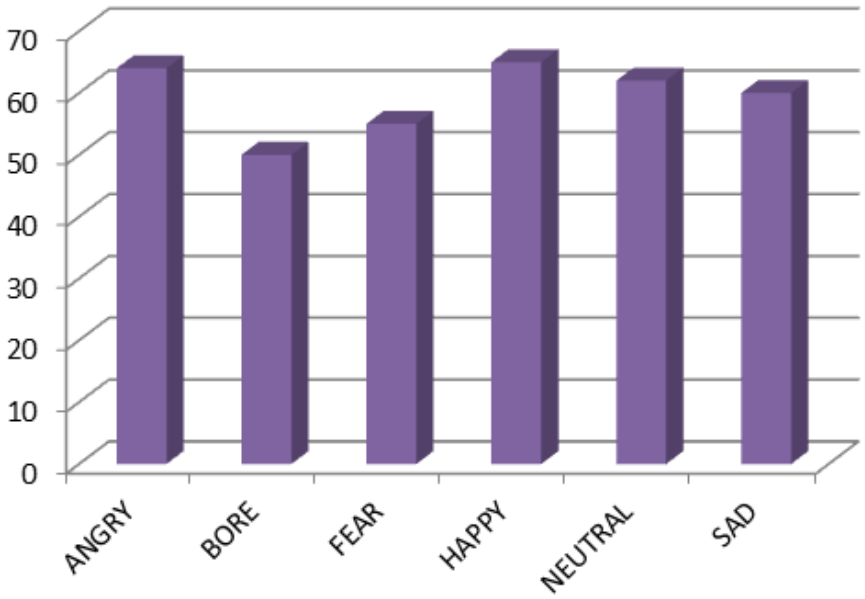

Figure 12.5 Using MFCC

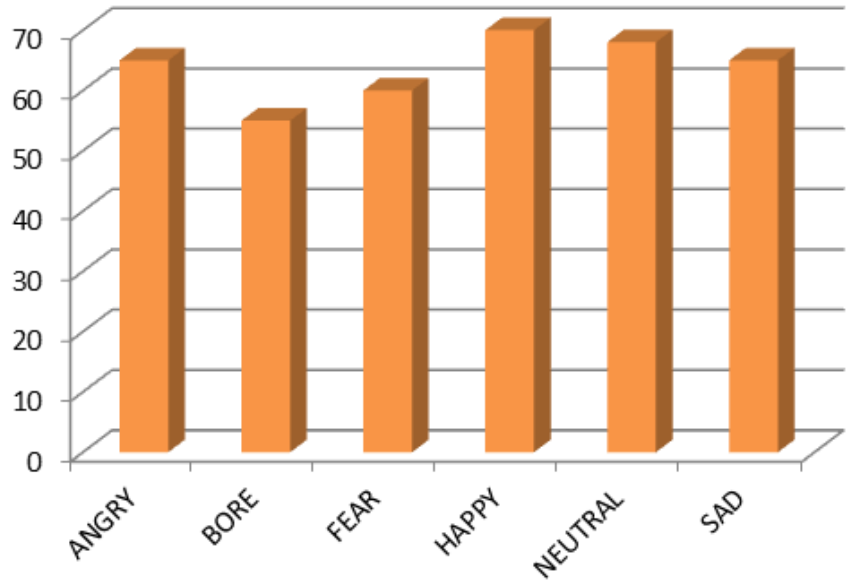

Figure 12.6 Using both LPC and MFCC

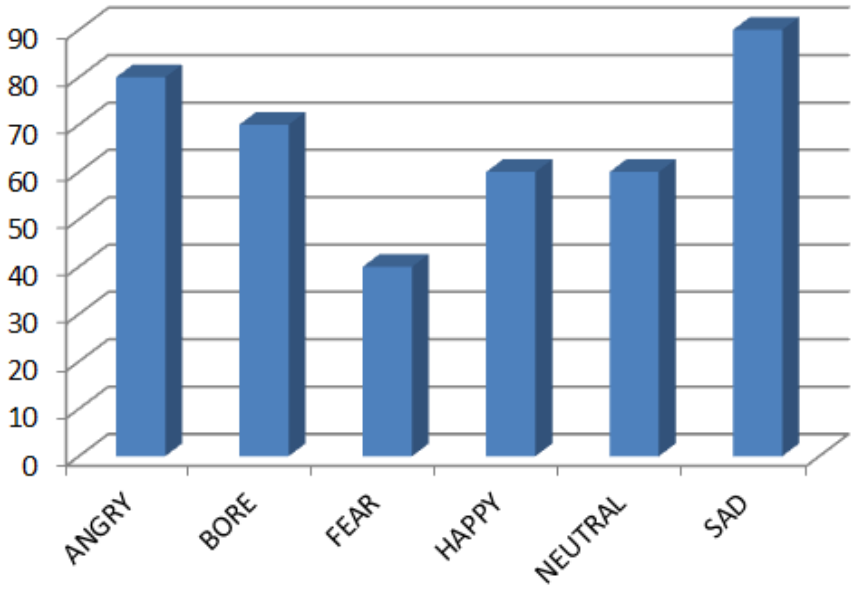

Figure 12.7 Neural Network

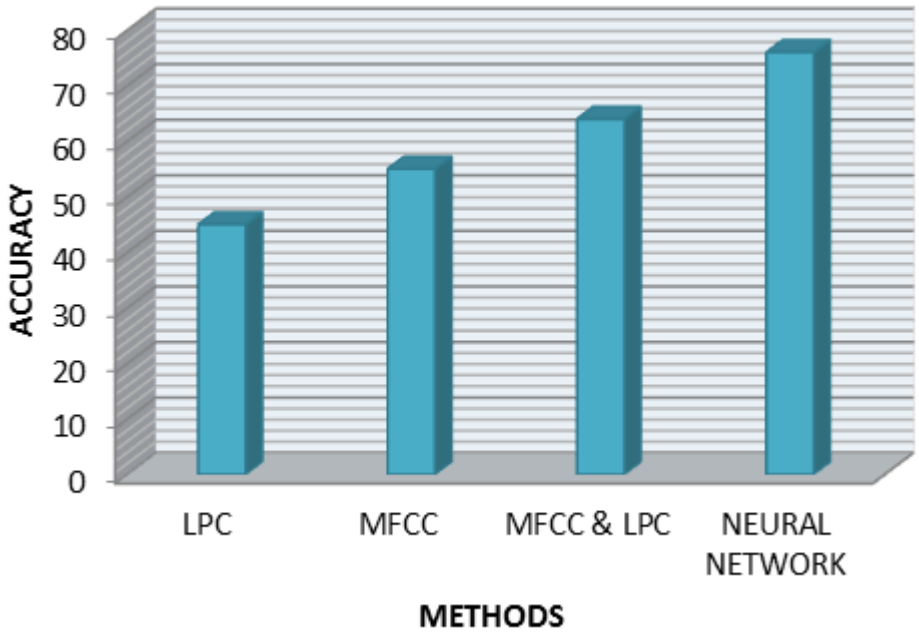

Figure 12.8 Comparison with different methods

Table 12.1 Accuracy Estimation

	LPC	MFCC	LPC & MFCC	Neural Networks
Accuracy	45	55	63	76

13. GRAPHICAL USER INTERFACE

13.1 Definition

A graphical user interface (GUI) is a human-computer interface (i.e., a way for humans to interact with computers) that uses **windows**, **icons** and menus and which can be manipulated by a mouse (and often to a limited extent by a keyboard as well).
A window is a (usually) rectangular portion of the monitor screen that can display its contents (e.g., a program, icons, a text file or an image) seemingly independently of the rest of the display screen. A major feature is the ability for multiple windows to be open simultaneously. Each window can display a different application, or each can display different files (e.g., text, image or spreadsheet files) that have been opened or created with a single application.

An icon is a small picture or symbol in a GUI that represents a program (or command), a file, a directory or a device (such as a hard disk or floppy). Icons are used both on the desktop and within application programs. Examples include small rectangles (to represent files), file folders (to represent directories), a trash can (to indicate a place to dispose of unwanted files and directories) and buttons on web browsers (for navigating to previous pages, for reloading the current page, etc.).

Commands are issued in the GUI by using a mouse, trackball or touchpad to first move a pointer on the screen to, or on top of, the icon, menu item or window of interest in order to select that object. Then, for example, icons and windows can be moved by dragging (moving the mouse with the held down) and objects or programs can be opened by clicking on their icons.

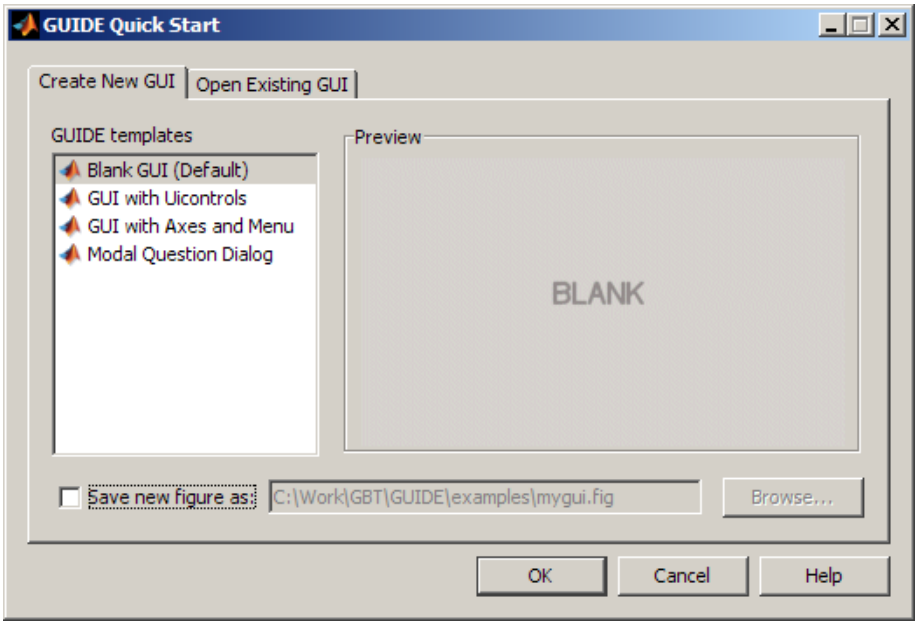

Figure 13.1 The GUI Start Dialogue Box in MATLAB

13.2 Advantages

A major advantage of GUIs is that they make computer operation more intuitive, and thus easier to learn and use. For example, it is much easier for a new user to move a file from one directory to another by dragging its icon with the mouse than by having to remember and type seemingly arcane commands to accomplish the same task.

Adding to this intuitiveness of operation is the fact that GUIs generally provide users with immediate, visual feedback about the effect of each action. For example, when a user deletes an icon representing a file, the icon immediately disappears, confirming that the file has been deleted (or at least sent to the trash can). This contrasts with the situation for a Command Line Interface, in which the user types a delete command (inclusive of the name of the file to be deleted) but receives no automatic feedback indicating that the file has actually been removed.

In addition, GUIs allow users to take full advantage of the powerful multitasking (the ability for multiple programs and/or multiple instances of single programs to run simultaneously) capabilities of modern operating systems by allowing such multiple programs and/or instances to be displayed simultaneously. The result is a large increase in the flexibility of computer use and a consequent rise in user productivity.

But the GUI has become much more than a mere convenience. It has also become the standard in human-computer interaction, and it has influenced the work of a generation of computer users. Moreover, it has led to the development of new types of applications and entire new industries. An example is desktop publishing, which has revolutionized (and partly wiped out) the traditional printing and typesetting industry.

13.3 Implementation

The GUIDE Toolbox provided by MATLAB allows advanced MATLAB programmers to provide Graphical User Interfaces to their programs. GUIs are useful because they remove end users from the command line interface of MATLAB and provide an easy way to share code across nonprogrammers. In addition by using special compilers the mathematical ability of MATLAB seamlessly blends in with the GUI functionality provided.

13.3.1 Layout

Guide simplifies the creation and manipulation of Handle Graphics objects. This means that:

- It simplifies access to the properties of Handle Graphics objects.
- It lets you place and arrange GUI elements on a figure by clicking and dragging with the mouse.

Specifically, Guide consists of five MATLAB tools that offer a streamlined approach to working with MATLAB figure windows.

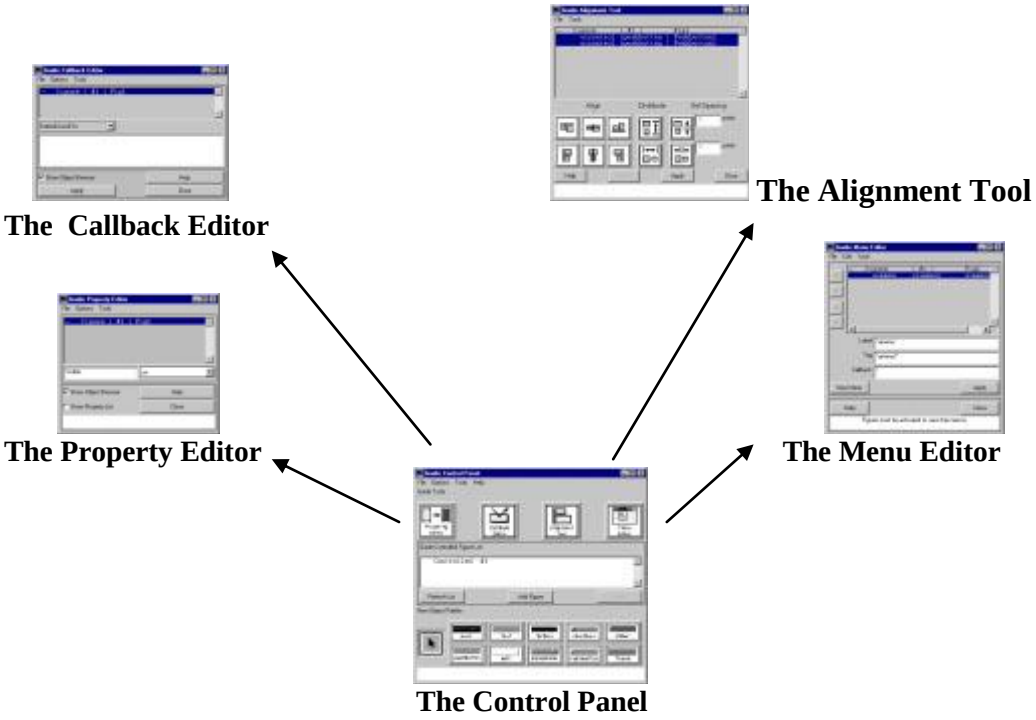

Figure 13.2 The five tools of GUIDE Command in MATLAB

The Guide tools are the Control Panel, the Property Editor, the Callback Editor, the Alignment Tool, and the Menu Editor. Each tool performs a distinct task and also is aware of and interacts with other tools. You can open each Guide tool by entering one of these commands at the MATLAB prompt.

- guide (for the Control Panel)
- propedit (for the Property Editor)
- cbedit (for the Callback Editor)
- align (for the Alignment Tool)
- menuedit (for the Menu Editor)

13.3.2 Callback Editor

The Callback Editor lets you modify the callbacks of selected objects. You can change multiple lines of code at once using the Edit Box. The Edit Box also makes it easier to enter Handle Graphics code because it allows you to omit nested quotes and enter code on multiple lines.

You can initialize the Callback Editor by entering cbedit from the MATLAB prompt, or by clicking on the appropriate button in the Control Panel. You can use the Callback Editor to make changes to any figure. For example, you might want to change the ButtonDownFcn for a figure display the current figure handle.

Begin by selecting the Show Object Browser check box to display a list of objects. Select the property you want to modify, in this case the ButtonDownFcn.

Select the Apply push button when you have finished making changes. If you click in the figure now, it displays the current figure in the MATLAB workspace.

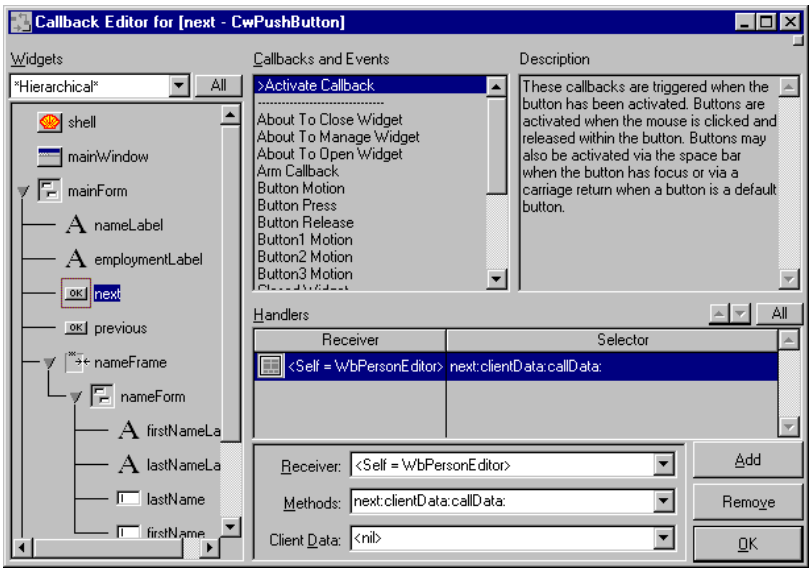

Figure 13.3 Callback Editor

13.3.3 Flowchart Algorithm

At the start step, the program begins its execution. It may perform such initialization tasks as instantiating classes and connecting to a database. It then displays a menu. The program will need to wait until the user makes a selection. After the user makes a selection, the program will process the user's input to determine which menu command was selected. If the user selected quit, the program goes to the end step. Before the program terminates, it may perform such "clean up" tasks as de-allocating memory and disconnecting from the database. If the user selected a command other than quit, the command's task is performed. For example, if the selection was the show branch command, the program will query the database and then display the results. The program then redisplays the menu and waits for the user to make another selection.

The flow chart in the following page shows the control flow of a of the user-defined handling using controlled-commands for the emotion recognition:

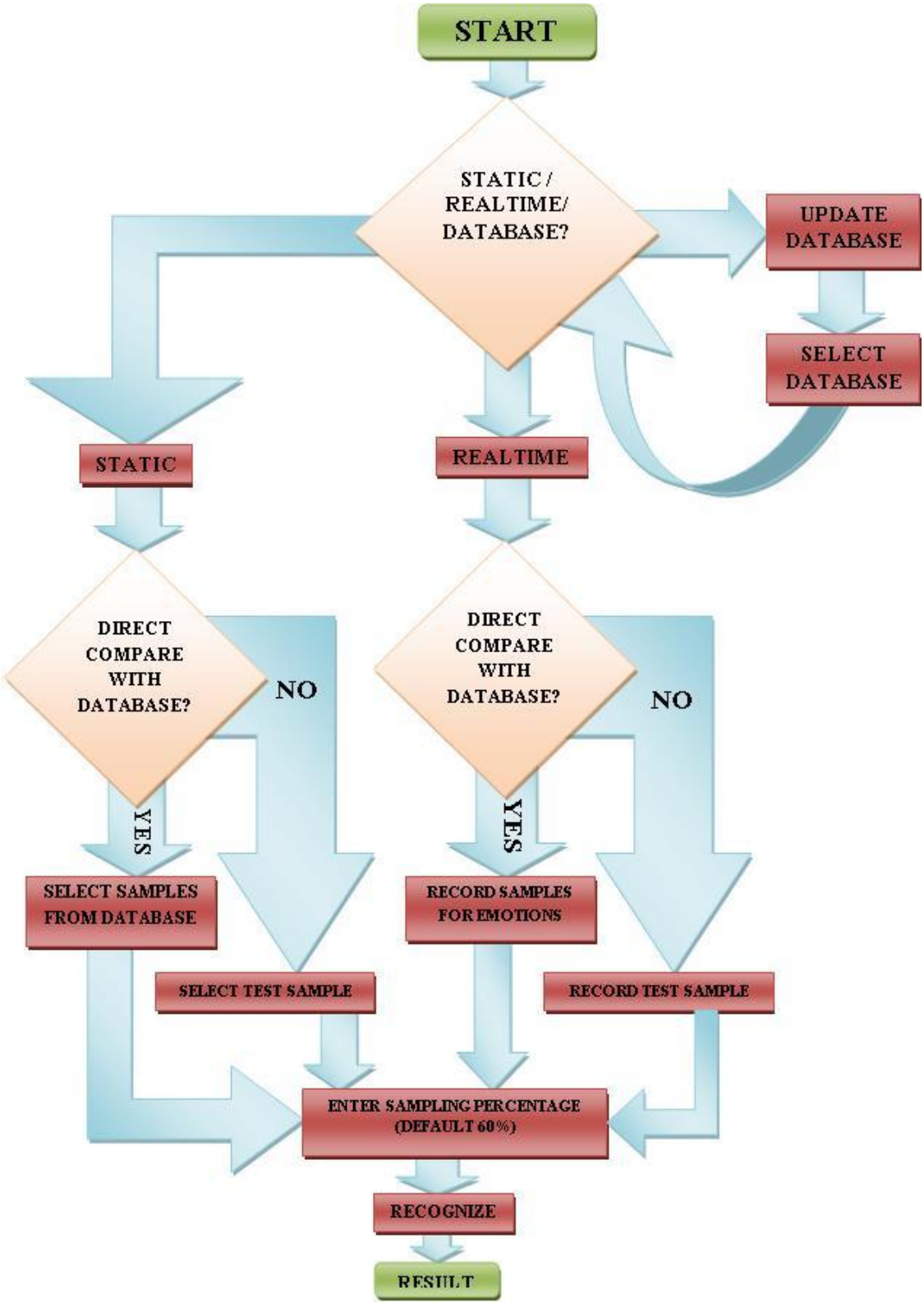

Figure 13.4 GUI Flowchart

13.34 Front End Display

The ultimate stage of a recognition model is its capacity to display the computed results in an efficient and intelligent mode. The graphical user interface (GUI) aids in this aspect to describe the response of the system using figures, text-box and plots. One such example of this technique is shown below. The two-dimensional acoustic plot is generated by the system once the user pushes the 'display plot' button. Consequently, a number of features can be obtained in the built-in GUI Toolbox of MATLAB that provides a suitable technique for human computer interaction.

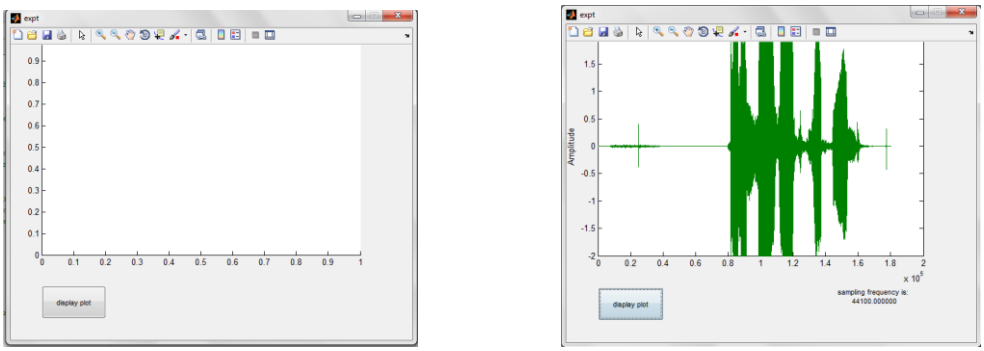

Figure 13.5 Graphical Display of Speech Signal

14. SYSTEM RESTRICTIONS

DIFFICULTIES IN AUTOMATIC EMOTION RECOGNITION (AER)

Human comprehension of speech compared to AER

Humans use more than their ears when listening; they use the knowledge they have about the speaker and the subject. Words are not arbitrarily sequenced together, there is a grammatical structure and redundancy that humans use to predict words not yet spoken. Furthermore, idioms and how we 'usually' say things makes prediction even easier.

In AER we only have the speech signal. We can of course construct a model for the grammatical structure and use some kind of statistical model to improve prediction, but there are still the problem of how to model world knowledge, the knowledge of the speaker and encyclopedic knowledge. We can, of course, not model world knowledge exhaustively, but an interesting question is how much we actually need in the AER to measure up to human comprehension.

- **Body language**

A human speaker does not only communicate with speech, but also with body signals - hand waving, eye movements, postures etc. This information is completely missed by AER.

This problem is addressed within the research area multimodality, where studies are conducted how to incorporate body language to improve the human-computer communication.

- **Noise**

Speech is uttered in an environment of sounds, a clock ticking, a computer humming, a radio playing somewhere down the corridor, another human speaker in the background etc. This is usually called noise, i.e., unwanted information in the speech signal. In AER we have to identify and filter out these noises from the speech signal.

Another kind of noise is the echo effect, which is the speech signal bounced on some surrounding object, and that arrives in the microphone a few milliseconds later. If the place in which the speech signal has been produced is strongly echoing, then this may give raise to a phenomenon called reverberation, which may last even as long as seconds.

- **Spoken language ≠ Written language**

Spoken language has for many years been viewed just as a less complicated version of written language, with the main difference that spoken language is grammatically less complex and that humans make more performance errors while speaking. However, it has become clear in the last few years that spoken language is essentially different from written language. In AER, we have to identify and address these differences.

Written communication is usually a one-way communication, but speech is dialogue-oriented. In a dialogue, we give feed-back to signal that we understand, we negotiate about the meaning of words, we adapt to the receiver etc.

Another important issue is disfluences in speech, e.g. normal speech is filled with hesitations, repetitions, changes of subject in the middle of an utterance, slips of the tongue etc. A human listener does usually not even notice the disfluences, and this kind of behavior has to be modeled by the AER system.

Another issue that has to be identified is that the grammaticality of spoken language is quite different to written language at many different levels. Some differences are pointed out:

•In spoken language, there is often a radical reduction of morphemes and words in pronunciation.

•The frequencies of words, collocations and grammatical constructions are highly different between spoken and written language.

•The grammar and semantics of spoken language is also significantly different from that of written language; 30-40% of all utterances consist of short utterances of 1-2-3 words with no predicative verb.

- **Channel variability**

One aspect of variability is the contexts were the acoustic wave is uttered. Here we have the problem with noise that changes over time, and different kinds of microphones and everything else that effects the content of the acoustic wave from the

speaker to the discrete representation in a computer. This phenomenon is called *channel variability*.

- **Speaker variability**

All speakers have their special voices, due to their unique physical body and personality. The voice is not only different between speakers; there are also wide variations within one specific speaker.

We will in the subsections below list some of these variations.

- **Realization**

If the same words were pronounced over and over again, the resulting speech signal would never look exactly the same. Even if the speaker tries to sound exactly the same, there will always be some small differences in the acoustic wave you produce. The *realization* of speech changes over time.

- **Speaking style**

All humans speak differently; it is a way of expressing their personality. Not only do they use a personal vocabulary, they have an unique way to pronounce and emphasize. The speaking style also varies in different situations; we do not speak in the same way in the bank, as with our parents, or with our friends.

Humans also communicate their emotions via speech. We speak differently when we are happy, sad, frustrated, stressed, disappointed, defensive etc. If we are sad, we may drop our voice and speak more slowly, and if we are frustrated we may speak with a more strained voice.

- **The sex of the speaker**

Men and women have different voices, and the main reason to this is that women have in general shorter vocal tract than men. The fundamental tone of women's voices is roughly two times higher than men's because of this difference.

- **Anatomy of vocal tract**

Every speaker has his/hers unique physical attributes, and this affects his/her speech. The shape and length of the vocal cords, the formation of the cavities, the size of the lungs

etc. These attributes change over time, e.g. depending on the health or the age of the speaker.

- **Speed of speech**

We speak in different modes of speed, at different times. If we are stressed, we tend to speak faster, and if we are tired, the speed tends to decrease. We also speak in different speeds if we talk about something known or something unknown.

- **Regional and social dialects**

Dialects are group related variation within a language. Janet Holmes de-fines regional and social dialects as follows:

- **Regional dialect**

Regional dialects involve features of pronunciation, vocabulary and grammar which differ according to the geographical area the speaker come from.

- **Social dialect**

Social dialects are distinguished by features of pronunciation, vocabulary and grammar according to the social group of the speaker.

In many cases, we may be forced to consider dialects as 'another language' in AER, due to the large differences between two dialects.

- **Same Sentence - Different Emotion**

The concept of *SSDE* is that two sentences that sound the same but have different emotions. In the table below, we give some examples of SSDE:

Questioning	Angry
What are you doing?	What are you Doing!
What is wrong with you?	What is wrong with you!

15. CONCLUSION AND APPLICATION

Conclusion:

Thus we found neural network to have highest efficiency in the detection of emotion. This is worth noting as we have considered a speaker independent and text independent database. The use of MFCC and BPNN as its classifier also produces good results; which is improved further where we make use of MFCC and LPC with BPNN as the classifier. The overall accuracy seems to vary from emotion to emotion. The emotions detected in the order of accuracy using Neural Networks are as follows:

Table 15.1 Confusion-Matrix obtained as an average of the two language databases

Emotional Class	Neutral	Happy	Sad	Anger	Surprise
Neutral	79%	10.2%	15.4%	0%	4%
Happy	13.2%	75.2%	12.7%	11.76%	35.2%
Sad	13.59%	14.5%	71.35%	0%	0%
Anger	0%	0%	0%	87.43%	54.3%
Surprise	3.1%	36.8%	5.5%	49.28%	85.9%

Table 15.2 The Recognition Accuracies for the two languages corresponding to the five emotions

Emotional Class	Neutral	Happy	Sad	Anger	Surprise
English	65%	68.7%	73.5%	77.5%	78.6%
German	73.2%	77.2%	75.7%	85 %	76.2%

Dealing with the vocal emotions is one of the challenges for speech processing technologies. Whereas the research on facial emotions recognition has been quite extensive, and which focuses on speech modality, both for automated production and recognition by machines The selection of a feature set is a critical issue for all recognition systems. In the conventional approach to emotion classification of speech signals, the features typically employed are the fundamental frequency, energy contour, duration of silence and voice quality.

The effect of any emotion on speech will depend on the attention and cognitive conflict between the speaker's emotional response and the focus of speech; physiological changes such as different breathing styles. for instant, when one is in a state of anger fear or joy the nervous system is aroused, the heart rate and blood pressure increase the mouth becomes dry and there are occasional muscle tremors speech is then loud fast enunciated with strong high frequency energy. when one is bored or sad the parasympathetic nervous system is aroused, the heart rate and blood pressure decrease and salivation increases which results in slow pitched with little high frequency energy, In vocal transmission characteristics visual contact between the speaker and the listener is not necessary, sound can be used to attract attention as listeners always have their listening "turn on" and open

to sound coming from any source. Vocal channel should be more suitable for the signaling of certain emotions than of others. Fear and alarm are the most clearly vocally expressed emotions in long distance communication where as disgust is the bad one for the long distance.

According to the language very short term changes have been observed, in fundamental frequency and expressed emotion in tone language than in indo American languages.

Difficulties in vocal Emotion recognition is also due to human comprehension of speech compared to emotion recognition human can predict the words with the idioms, in computer based emotion recognition used some kind of statistical model to improve the prediction it is difficult to model the world knowledge, knowledge of the speaker and encyclopedic knowledge. body language, noise ,spoken language is dialog oriented ,and with influences like hesitations repetitions, change of subject in the middle of an utterance , slip of tongue, continuous speech , channel variability and speaker variability.

Characteristics of speech samples changes with respect to speaking style, gender anatomy of vocal tract, speed of speech and also changes as per social and regional dialect, natural language has inherent ambiguities like homophones word boundary ambiguities [7]. Further improvement and expansion may be achieved according to the following suggestions: The set of the most efficient features for emotion recognition is still vague.

A possible approach to extracting non-textual information to identify emotional states in speech is to apply all known feature extraction methods. Thus, we may try to incorporate the information of different features into our system to improve the accuracy of emotion recognition. Recognizing emotion translation in real human communication is also a challenge. Thus, it will be worthwhile to determine the points where emotion transitions occur.

Applications:

- Psychiatry

  It can be applied for better understanding of human emotions. Also helps in distinguishing between the various types of emotions displayed. E.g. anger-hot (anger which is clearly visible in voice) and anger-cold (concealed anger in voice).

- Lie detector

  A person may on speaking lies show a sudden change in his emotional state .e.g. on speaking a lie his state may change from anger to fear.

- Virtual teachers

  In order to better replicate the class room environment in a better way in distance learning courses this application may be use full . e.g. the virtual teacher will be angry when a student gives a wrong answer and happy when he gives a right answer.

- Communications for disabled

  Depending on the emotion displayed monitoring of the disabled patient, may be carried out efficiently. E.g. emotions such as fear and sadness may require the immediate attention being provided to the patient.

- Brain mapping

  The human brain is a very complex structure responsible for our actions. Researchers are yet to decode the way in which it functions. By observing which part of the brain remains active for various emotions we can understand better the working of the brain. Also the parts of the brain responsible for displaying a particular emotion are identified.

- Training actors

  By training actors through this application they can better their ‘dramatic ‘performance. Also their ability to produce such emotions on stage artificially becomes enhanced.

- Call centers

  In call centers it is often required to convince people regarding the purchase of different products. As such by knowing the emotional state of the customer the seller may accordingly plan his marketing strategy.

Future Development:

We have successfully implemented vocal emotion recognition using artificial neural network. The data base used in this case was Berlin Data Base. However a real time system may be developed where the user enters the speech signal to the system. This signal is then processed and classified according to the emotion. Thus emphasis may be laid to make the system not only speaker independent and text independent, but also language independent. This means that the same data base and classifier be used for classifying emotions regardless of language. Therefore such a system would be truly of universal application. Example: the system may take inputs as Mandarin, French, Spanish, Hindi, Mexican, etc. and still provide accurate results. The system in this case would have to choose parameters for classification such that they are universally displayed in the emotions.

REFERENCES

1. Carlos Busso, Sungbok Lee, Member, and Shrikanth Narayanan, Fellow, IEEE "*Analysis of Emotionally Salient Aspects of Fundamental Frequency for Emotion Detection*" IEEE TRANSACTIONS on audio, speech and language processing, VOL. 17, NO. 4, MAY 2009.

2. Tsang-long Pao, Yu-Te Chen, Junheng Yeh, "*Emotion Recognition and Evaluation from Mandarin Speech Signals*", ICIC International, July 2008, Volume 4, (Tatung University, Taiwan), pp. 1695-1709.

3. Noam Amir, "Classifying Emotions in Speech: a comparison of methods", Academic College of Tel Aviv Yaffo.

4. Valery A. Petrushin, "*Emotion Recognition in Speech Signal: Experimental Study, Development and Application*", International Conference in Spoken Language Processing, 2000.

5. Albino Nogueiras, Antonio Bonafonte and Jose B Marino (Spain) , "*Speech Emotion Recognition using Hidden Markov Models*", EUROSPEECH, 2001.

6. Manish Gaurav, IIT Kanpur," Performance analysis of spectral and prosodic features in speech".

7. Ian McLoughlin, "*Applied Speech And Audio Processing With MATLAB Examples*", Cambridge University Press 2009.

8. Rabiner L. R. and Schafer R.W., "*Digital Processing Of Speech Signals*", Prentice Hall, 1978, Pearson Education (LPE Indian Edition).

9. David Kreisel, "*A Brief Introduction To Neural Networks*", A German Manuscript translated into English, *http://www.dkriesel.com/*

10. S. Rajasekaran, G. A. Vijayalakshmi Pai, "*Neural Networks, Fuzzy Logic and Genetic Algorithms*", PHI Learning.

11. A.B. Kandali, A.B. Routray, Basu T.K., "*Emotion Recognition From Assamese Speeches Using MFCC And GMM Classifier*", IEEE Region 10 Conference TENCON 2008, NOV. 19-21 Hyderabad, India, 2008, pp. 1-5.

12. Chul Min Lee, *Student Member, IEEE,* and Shrikanth S. Narayanan, *Senior Member, IEEE* "*Toward Detecting Emotions in Spoken Dialogs*" IEEE TRANSACTIONS on speech and audio processing, VOL. 13, MARCH 2005.

13. F. Burkhardt, A. Paeschke, M. Rolfes, W. Sendlmeier, B. Weiss, "*A Database of German Emotional Speech*", T-Systems, TU Berlin, Department of Communication Science, LKA Berlin, HU Berlin.

14. Prasad Reddy P.V.G.D., Prasad A., Srinivas Y., Brahmaiah P., "*Gender Based Emotion Recognition System for Telugu Rural Dialects Using Hidden Markov Models*" Journal of Computing, Volume 2, Issue 6, June 2010, ISSN 2151-9617.

15. Yi-hao Kao and Lin-shan Lee, "*Feature Analysis for Emotion Recognition from Mandarin Speech Considering the Special Characteristics of Chinese Language",* INTERSPEECH 2006 - ICSLP September 17-21, Pittsburgh, Pennsylvania.

16. Johannes Pittermann, Angela Pittermann, Wolfgang Minker "*Emotion recognition and adaptation in spoken dialogue systems*". International Journal in Speech Technology, 2010, 13: 49–60.

17. Ben J. Shannon, Kuldip K. Paliwal, "*A comparative study of filter bank spacing for speech recognition*", Microelectronics Engineering Research Conference 2003.

18. R V Pawar, P P Kajave and S N Mali, "*Speaker identification using neural networks*", World Academy Of Science, Engineering And Technology 2005.

19. K R Aida-Zade , C Ardil and S S Rustamov , "*Investigation of combined use of MFCC and LPC features in speech recognition system*", World Academy of Science, Engineering And Technology 2006.

20. Mahdi Shaneh,Azizollah & Taheri, "*Voice command recognition system based on MFCC and VQ algorithm*", World Academy of Science, Engineering And Technology 2009.

21. Madhavi S. Pednekar, Kavita Tiware and Sachin Bhagwat, "*Continuous Speech Recognition for Marathi Language Using Statistical Method*", IEEE International Conference on 'Computer Vision and Information Technology, Advances and Applications', ACVIT-09, December 2009, pp. 810-816.

22. Firoz Shah. A, Raji Sukumar. A, and Babu Anto. P, "*Discreet Wavelet Transforms and Artificial Neural Networks for Speech Emotion Recognition*", International Journal of Computer Theory and Engineering, Vol. 2, No. 3, 1793-8201, June 2010, pp.319-322.

23. S.N. Sivanandam, S.N. Deepa, "*Principles of Soft Computing*", WILEY India, 2009.

24. A.B. Kandali, A.B. Routray, Basu T.K., "*Emotion Recognition From Assamese Speeches Using MFCC And GMM Classifier*", IEEE Region 10 Conference TENCON 2008, NOV. 19-21 Hyderabad, India, 2008, pp. 1-5.

INDEX