ENTROPY BASED
AUDIO EMOTION ANALYSIS

A Major Project Report submitted in the partial fulfillment
of the requirements for the award of the degree of

MASTER OF TECHNOLOGY

(INFORMATION SYSTEMS)

Submitted By:

AMANDEEP KAUR

(Roll No. 2K13/ISY/03)

Under the esteemed guidance of

Dr. SEBA SUSAN

Assistant Professor

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING
DELHI TECHNOLOGICAL UNIVERSITY
BAWANA ROAD, DELHI-110042
SESSION: 2013-2015

CERTIFICATE

This is to certify that work entitled "**Entropy based Audio Emotion Analysis**" submitted by **Amandeep kaur (2k13/ISY/03)**, to Delhi Technological University, Delhi for the award of the degree of Master of Technology is a bonafide record of research work carried out by her under my supervision.

The content of this thesis, in full or parts have not been submitted to any other institute or university for the award of any degree or diploma.

Dr. Seba Susan
Project Guide
Assistant Professor
Department of Computer Science and Engineering
Delhi Technological University
Shahbad Daultpur, Bawana Road, Delhi-110042

Date:_______________

ACKNOWLEDGEMENT

I would like to thank my project guide, **Dr. Seba Susan** for her valuable guidance and wisdom in coming up with this project. I humbly extend my words of gratitude to **Dr. O. P**. **Verma**, Head of Department, and other faculty members of IT department for providing their valuable help and time whenever it was required. I thank all my friends at DTU who were constantly supporting me throughout the execution of this thesis.

Special thanks to the Almighty Lord for giving me life and the strength to persevere through this work. Last but not least, I thank my family for believing in me and supporting me.

Amandeep kaur
Roll No. 2k13/ISY/03
M.Tech (Information Systems)
E-mail: **kaur.aman123@gmail.com**
Department of Computer Science and Engineering
Delhi Technological University

ABSTRACT

Humans commonly interact with each other using speech. Extracting information from the speech helps in effective interaction between humans and computers. Thus analyzing and recognizing emotions in humans has attracted a lot of researchers in past two decades. The major challenge in this area is to recognize the features to be extracted from the speech that can effectively and efficiently classify the emotions in humans. Audio Emotion Analysis includes the detection of a scream, an extreme emotion of fear and classification of emotions. Scream detection is done using 2 methods. One involving extraction of log energy, auto-correlation and MFCC features from the input speech and other involving the use of non-extensive entropy on MFCC features. For audio emotion analysis, two methods are proposed. First method makes use of weighted mean and entropy of MFCC features and second method uses different entropy values of different audio features like pitch and energy. Speech database used are freely available for research purpose. SAVEE database is used for emotion recognition. For scream detection, many sounds like applause, laugh, different type of screams etc are taken from the web directly.

LIST OF FIGURES

TABLE OF CONTENTS

CHAPTER 1

INTRODUCTION

1. INTRODUCTION

Speech is the most common medium used by humans to interact with each other. This fact motivated the researchers to explore speech as a medium for human and machine interaction. This requires understanding the speech and emotional state of humans by the machines. Speech emotion recognition involves extracting useful semantics from speech to study emotional content of speech signals.

Humans can express emotions in various forms like face, speech, gait etc. Emotions in speech mainly consists of “what is said” and “how is said”. The first component deals with the linguistic components of the speech. Emotion analysis can be performed by analyzing either one or both of these components.

The design of audio emotion recognition systems involves three main aspects. First is the correct choice of features to be extracted. Second is the use of appropriate classification method. Third is the audio database to be used.

1.1 Challenges involved in emotion analysis

Recognizing emotions from human speech involves lot of challenges. It is difficult to decide the type of features to be extracted from the speeches which are powerful enough to distinguish between different emotions. Also, each speaker has a different speaking style, accent and speed. This directly impacts pitch and energy levels which are commonly extracted from the speech signals. The problem is aggravated by the fact that emotions expressed by a speaker depend largely on his culture and environment. Apart from this, there are no clear boundaries between different emotional states and it differs from person to person. Sometimes even humans cannot distinguish between natural emotions.

1.2 Features for audio emotion recognition

It is important to choose suitable features that can effectively classify emotions in a speech. But, there are many issues involved in feature extraction. According to Moataz [1] there are 4 main issues which are discussed below.

1.2.1 Local vs. Global features

This issue refers to the region of analysis in the speech signal which is used for the extraction of features. The region of analysis can be complete speech. Global features are extracted from it.

Also, Speech can be divided into small intervals, called frames and they can be used as a region of analysis. Thus local features such as pitch and energy can be extracted from each frame. The signal is divided into frames because it is believed that in a frame, signal is approximately stationary.

Features can also be extracted by segmenting speech signal on basis of phonemes and analyzing the variation in spectral shape for each phoneme for various emotions.

1.2.2 Categories of speech features

It is important to choose the best features for efficient classification of emotion. It is also necessary that these features are speaker independent and do not depend on the lexical content. Speech features can fall into one of 4 categories: continuous features, qualitative features, spectral features, and TEO (Teager energy operator)-based features

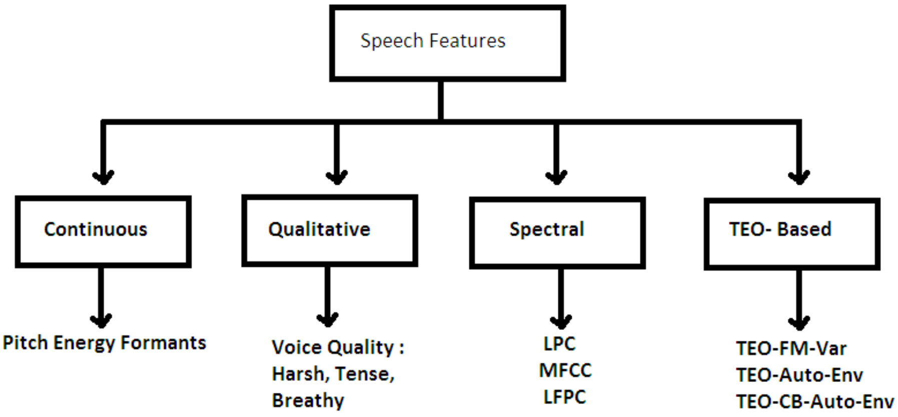

Figure 1.1 Categories of speech features

1.2.3 Processing of Speech

Preprocessing operations are performed on the speech signal before extracting features. Some of the preprocessing operations are energy normalization, multiplying each frame with a hamming window. Sometimes the silence intervals carry important information and these intervals are not altered.

Some post processing steps can also be involved after feature extraction. Some of them are feature normalization, dimensionality reduction etc.

1.2.4 Combining acoustic with other information sources

There can be situations where non acoustic information sources such as facial expression, gesture or pose can help in emotion recognition. Thus different information sources can be combined to improve the results.

1.3 Classification Scheme

After the necessary feature extraction, it is important to appropriately classify the emotions on the basis of these features. Various types of classifiers can be used for this purpose. Some of them are

- Hidden Markov Model (HMM)
- Gaussian Mixture Model (GMM)
- Support Vector Machine (SVM)
- Artificial Neural Networks (ANN)
- K-NN

Each classifier has its own advantages and disadvantages

1.4 Speech Database

The performance of emotion recognition program depends largely on the choice of database. A low-quality database may lead to incorrect conclusions and results. It is also important to choose the database according the requirement. Example we may need a noisy database with specific sounds, specific emotion database of male/female/children, stress related database. There are some standard databases available on the internet for research purposes. Also, own database can be created.

1.5 Applications of audio emotion recognition

Speech recognition systems can be used where man and machine interactions are needed like web movies, where, the response of the systems depend on the emotion detected. It can be used in cars where depending on the emotion of the driver, safety procedures are initiated. The most common medical application is for diagnosis of emotions to provide therapies. It can be used in translation mechanisms where the emotional state of the person plays important role during conversation. It can be applied to call center systems to detect the mood or frustration of the speaker.

CHAPTER 2

LITERATURE REVIEW

2. Literature Review

2.1 Scream Detection

Scream can be considered as an anomaly in the audio sample. Scream detection is helpful in many everyday applications such as old age homes, home care or security applications etc. Some of the work done in this area is described below.

2.1.1 Scream detection for home applications

For scream detection in an audio, both analytic and statistical features are used for classification. According to Huang [3], 3 features are used for scream identification. First feature used is log energy. Since scream lasts longer than many other sounds, and has high energy, this feature represents the energy continuity of an audio segment. Since a scream represents not only high energy but high pitch also, the second feature used is auto-correlation. It helps to find out the high pitch in an audio segment. Finally, to validate scream like sound, SVM classifier is used on the MFCC features extracted from an audio segment.

2.1.1.1 Log Energy

Scream generally represents an audio segment with continuous and relatively high energy. Log energy helps us to classify a speech segment as scream or non-scream on the basis of energy continuity.

For a given sampling rate F_s, we divide the speech into number of frames, size of each frame being 256 samples. The log energy of the frame is computed as

$$E = 10\log(\sum_{i=1,2..256} x_i^2) \qquad (2.1)$$

A segment of an audio signal S(t) consisting of overlapping frames, with 128 samples overlap between any 2 consecutive frames, can be classified as scream or non-scream segment by analyzing the energy of all frames within this segment.

If, all the frames in the segment have energy greater than a threshold T, that segment is classified as a scream segment.

2.1.1.2 Auto-Correlation

A scream not only represents a high energy, it also represents a high pitch. Auto-correlation is used to extract the high pitch.

The speech signal is divided into number of segments, each of size N=1024 samples. Thus, Auto-correlation R of a segment of signal x_i, i=0…N-1 is

$$R(t) = \sum_i (x_i x_{i+t}) \quad (2.2)$$

Here t=1,2…256 to minimize the noise. This means that for each segment, 256 values of auto-correlation are found. A smoothing filter is applied before computing the peaks.

For each segment, the auto-correlation values are normalized as R = R / R(0).

The peaks and valleys are determined as

$$\begin{aligned} R_1(t) &= R(t) - R(t+1) \\ R_2(t) &= R(t-1) - R(t) \\ Peaks &= \{t \mid R_1(t) > 0 \,\&\, R_2(t) > 0\} \\ Valleys &= \{t \mid R_1(t) < 0 \,\&\, R_2(t) < 0\} \end{aligned} \quad (2.3)$$

Then the threshold, thre is applied to these Peaks and Valleys

$$\begin{aligned} P_1 &= \{Peaks \mid R_1(Peaks) > thre\} \\ V_1 &= \{Valleys \mid R_1(Valleys) < -thre\} \\ V &= \min(R(V_1)) \end{aligned} \quad (2.4)$$

For each segment, Pitch_high is set as the first peak P_1 after lowest valley V

$$Pitch\_high = P_1(i), i = \min(k), so\ P_1(k) > V \quad (2.5)$$

After obtaining the Pitch_high for each segment, we calculate the sound sample value at these points. If the sound sample values exceed a threshold value th, the segment is classified as a scream segment.

2.1.1.3 MFCC Features

Calculating high pitch and high and continuous energy for a segment doesn't always ensure that it's a scream segment. Thus another feature, Mel frequency cepstral coefficients (MFCCs) are calculated for the input audio sample.

MFCC features are extracted for the sound sample using frame of size 256 samples and an overlap of 128 samples in 2 consecutive frames. Thus, for each frame we get 36 MFCCs (12 MFCC without energy, plus 1st and 2nd derivative of the 12 MFCCs).

20 frames are considered at one time and the mean, minimum, maximum and standard deviation is found for each coefficient. Thus we obtain a 36 X 4 size matrix from a 36 X 20 matrix of MFCC features.

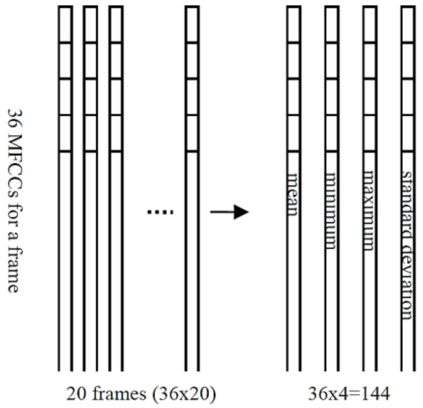

Figure 2.1 Features Extraction from MFCC features

This 36 X 4 matrix is converted into row vector of size 144 and fed into SVM classifier for classification as a scream or non-scream segment. SVM classifier is initially trained with same amount of MFCC features extracted for scream and non-scream segments like applause, laugh, footsteps etc.

2.1.1.4 System Diagram

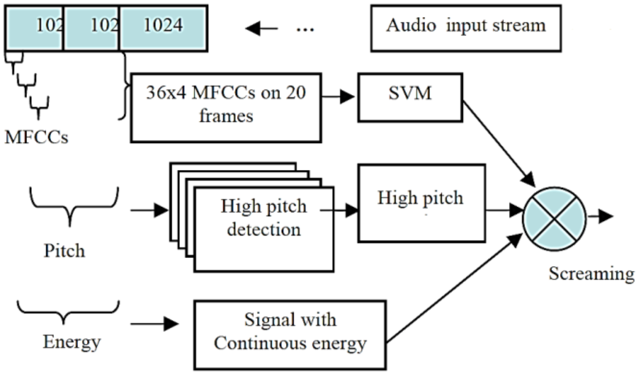

Figure 2.2 System Diagram for Scream detection

2.1.2 Scream and gunshot detection for Audio-Surveillance Systems

For audio analysis, considerable number of features is extracted from the speech. Those features are chosen which are not too much sensitive to SNR conditions.

2.2.2.1 Audio Features

Valenzise [4] finalized the feature set which is depicted in table below. The features are extracted at analysis frame of 23ms and with sampling frequency of 22050Hz and 1/3 overlap.

Table 2.1 Audio features used for classification

#	Feature Type	Features	Ref
1	Temporal	ZCR	[7]
2-6	Spectral	4 spectral moments + SFM	[8]
7-36	Perceptual	30 MFCC	[9]
37-39	Spectral distribution	spectral slope, spectral decrease, spectral roll-off	[8]
40-49	Correlation-based	(filtered) periodicity, (filtered) correlation slope, decrease and roll-off, modified correlation centroid, correlation kurtosis	[7][8]

2.2.2.2 Feature Selection

From the 49 feature set obtained, certain number of features is selected such that 1<=l<=49. Generally l is kept small. Two main feature selection approaches are adopted. First approach is filter method. A performance evaluation metric is calculated from the data directly and according to it, features which do not provide much help in classification are removed. The second approach is wrapper approach. The performance evaluation metric is a feedback from the classifier which helps to decide the features to be removed.

2.2.2.3 Classification

Two Gaussian Mixture Model (GMM) classifiers are used to differentiate between scream and background noise and between gunshots and noise. Each classifier is trained separately with the samples of their respective classes. For testing, the speech is divided into number of frames and each frame is classified independently by 2 classifiers. The final answer is the logical OR of the two answers.

2.2 Audio Emotion Recognition

Recognizing emotions from audio samples help in numerous practical applications. It can also be utilized for therapies.

2.3.1 Emotion recognition using Eigen Values of Autocorrelation Matrix (EVAM)

A feature set of 5 most significant Eigen Values of Autocorrelation Matrix (EVAM) of each frame of speech signal is used. They represent the powers of 5 most prominent frequency components in the speech. Gaussian Mixture Model (GMM) is used as a classifier. 7 emotions are considered for this experiment. They are Anger, Disgust, Fear, Happy, Neutral, Sad and Surprise.

2.3.1.1 Feature Extraction

The extraction of features from the speech usually involves various steps.

i. A pre-processing task of silence removal is performed for each speech

ii. The input speech signal is divided into number of fixed size frames.

iii. For each frame, Autocorrelation matrix is computed with lag p=32.

iv. 5 most significant Eigen values are computed for each frame.

v. These features are normalized by subtracting mean and standard deviation of the features

2.3.1.2 Classification

The feature vectors are extracted for both training and testing samples. For each emotion, one GMM is trained for all the feature vectors of that emotion using Expectation-Maximization Algorithm. Thus 7 GMMs are trained corresponding to the 7 emotions. After training, testing of all the test speech samples for all the emotions is carried out taking one at a time. The mean log likelihood of EVAM feature vectors of each test sample w.r.t. trained GMM of each emotion class is computed. The one with largest mean log likelihood is the emotion-class for that particular test sample.

2.3.1.3 System Diagram

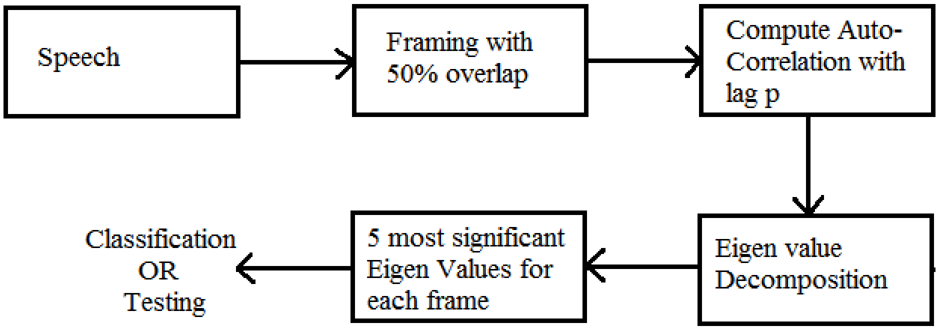
Speech
Framing with 50% overlap
Compute Auto-Correlation with lag p
Eigen value Decomposition
5 most significant Eigen Values for each frame
Classification OR Testing

Figure 2.3 System Diagram for Emotion recognition using EVAM

2.3.2 Emotion Recognition using Spectro-Temporal features

Recognizing emotions from the speech depict the emotional state of a human being. One of the challenges in emotion recognition is designing the effective measures. The features which are extracted over short frame duration (around 30 ms) with longer temporal info like derivatives (e.g. MFCC) depict signal’s short term spectral properties only. As per the proposed approach, the emotions are recognized from audio sample by extracting modular spectral features (MSFs) from the spectro-temporal representation of an audio signal.

Modulation spectral features capture both spectral and temporal properties of the speech signal. This is because they are based on the frequency analysis of the temporal envelopes (i.e. amplitude modulations) of multiple acoustic frequency bins. The steps to represent a speech signal in spectro-temporal representation and further extract MSFs from these are detailed below.

2.3.2.1 Spectro-temporal representation of speech

The audio samples first need to be converted to spectro-temporal representation so that MSFs can be calculated from them. Auditory filter bank and modulation filer bank helps to capture both acoustic frequency and temporal modulation frequency components, thereby conveying information that is important for human speech perception but missing from conventional short-term spectral features.

The various steps involved are

2.3.2.1.1 PRE-PROCESSING AND WINDOWING

The preprocessing of a signal involves resampling the speech to 8 kHz and normalizing its active speech level. This sampling rate has been considered adequate for emotion recognition. Speech is then divided into number of non-overlapping frames. G.729 voice activity detection (VAD) algorithm is used to label the frames as active or inactive. Frames labeled as active are retained. This is because non active frames are considered to hold noise and other non-speech activities which are unnecessary for emotion recognition.

Windowing involves converting a preprocessed signal s(n) into long term segments $s_k(n)$. This is done by multiplying a 256 ms Hamming Window with 64 ms frame shift. Hamming window is used since it reduces the spectral leakage

2.3.2.1.2 AUDITORY FILTERBANK

Human auditory system can be represented as series of over-lapping band pass frequency channels, namely auditory filters with critical bandwidth that increase with filter center frequencies. A gammatone filter bank with 19 filters is used where the center frequency of filters is proportional to their bandwidth. These are depicted by equivalent rectangular bandwidth (ERB).

The output for each frame k, for ith critical band filter is given as

$$s_k(i,n) = s_k(n) * h(i,n) \quad (2.6)$$

The output is the convolution of the long term segment with the impulse response of the ith critical band filter

2.3.2.1.3 HILBERT TRANSFORM

The Hilbert envelop can be computed from $s_k(i,n)$ as a magnitude of a complex analytical signal. It can be depicted as

$$\begin{aligned} H_k(i,n) &= |\hat{s}_k(i,n)| \\ H_k(i,n) &= \sqrt{s_k^2(i,n) + H\{s_k(i,n)\}} \end{aligned} \quad (2.7)$$

Here, H{.} denote Hilbert transform.

2.3.2.1.4 MODULATION FILTERBANK

For the proper interpretation of human auditory system, an M-band modulation filterbank is used along with a gammatone filterbank. Modulation filterbank is applied to each $H_k(i,n)$ and M outputs $H_k(i,j,n)$ are generated, where j denotes the jth modulation filter. Here 5-band modulation filterbank is used. Thus j value varies from 1 to 5. The filter center frequencies of the filterbank are equally spaced on a log scale.

2.3.2.1.5 ENERGY SUMMATION

The energy of the $H_k(i,j,n)$ for each frame is measured as

$$E_k(i,j) = \sum_{n=1}^{L} |H_k(i,j,n)|^2 \quad (2.8)$$

K represents the number of frames and L is the number of samples per frame.

2.3.2.2 Flowchart for deriving ST representation

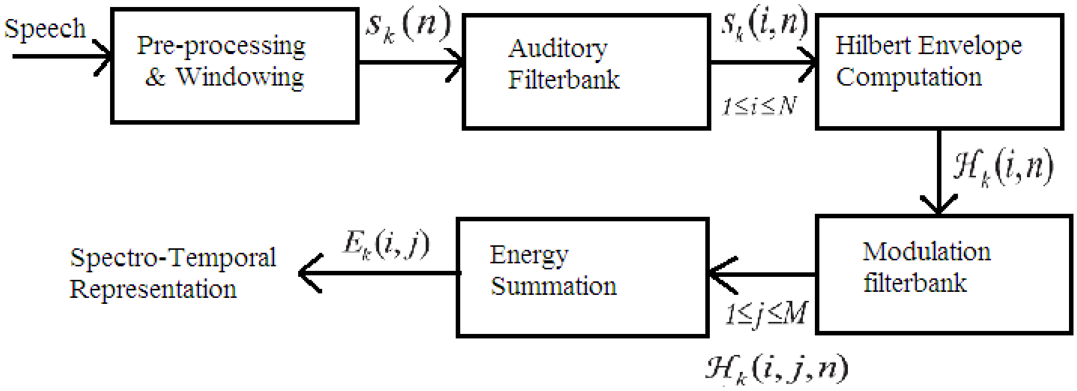

Figure 2.4 System Diagram for Emotion Recognition using ST Representation

2.3.2.3 Extraction of Modulation Spectral Features

Once the speech signal is represented in spectro-temporal format, Modulation Spectral Features (MSFs) are extracted from them. Two types of MSFs are calculated from the spectro temporal representation of the speech, one are the spectral measures and another are the linear predication parameters. The process of extracting features is given below.

Before calculating the spectral measures and LP parameters, for each frame k, the ST representation $E_k(i, j)$ is scaled to unit energy.

$$\sum_{i,j} E_k(i,j) = 1 \quad (2.9)$$

For each frame, 6 spectral measures can be calculated as follows

2.3.2.3.1 SPECTRAL MEASURE-1 : MEAN

For each frame k, mean is calculated for the energy samples belonging to the jth modulation channel (j varies from 1 to 5)

$$\phi_{1,k}(j) = \frac{\sum_{i=1}^{N} E_k(i,j)}{N} \quad (2.10)$$

This measure depicts the speech energy distribution along the 5 modulation channels.

2.3.2.3.2 SPECTRAL MEASURE-2 : SPECTRAL FLATNESS

Spectral flatness is given by the ratio of geometric mean by the arithmetic mean and is defined as

$$\phi_{2,k}(j) = \frac{\sqrt[N]{\prod_{i=1}^{N} E_k(i,j)}}{\phi_{1,k}(j)} \quad (2.11)$$

Spectral flatness close to 1 indicates that geometric mean and arithmetic mean are approximately same and thus spectrum is flat. Spectral flatness value close to 0 indicates that arithmetic mean value is far more than geometric mean and thus spectrum has widely different spectral amplitudes.

2.3.2.3.3 SPECTRAL MEASURE-3 : SPECTRAL CENTROID

This measure indicates center of mass of the spectrum for each channel.

$$\phi_{3,k}(j) = \frac{\sum_{i=1}^{N} f(i) E_k(i,j)}{\sum_{i=1}^{N} E_k(i,j)} \quad (2.12)$$

f(i) = i i.e. index of ith critical band filter.

It is observed that adjacent modulation channels show high correlation. Thus spectral flatness and centroid parameters also show high correlation. Thus spectral measure 2 and 3 are calculated only for j= 1,3, and 5 to reduce such information redundancy.

2.3.2.3.4 SPECTRAL MEASURE-4 : MODULATION SPECTRAL CENTROID

The 19 acoustic channels are grouped into 4 groups as 1–4, 5–10, 11– 15, and 16–19. The channels in the same group are added.

$$E_{k(l,j)} = \sum_{i \in D_l} E_k(i,j) \quad (2.13)$$

Here D_l denotes the groups formed and l varies from 1 to 4.

Thus, the spectral measure is calculated as

$$\phi_{4,k}(l) = \frac{\sum_{j=1}^{M} jE_k(l,j)}{\sum_{j=1}^{M} E_k(l,j)} \quad (2.14)$$

2.3.2.3.5 SPECTRAL MEASURE-5 and 6

The spectral measure 5 and 6 are the linear regression coefficient (slope) and the corresponding regression error which is the root mean squared error, RMSE obtained by fitting a first-degree polynomial to E_k(i,n) where j=1,2… M, in a least squares sense

Once the spectral measures are calculated, the linear predication measures are computed from the selected modulation channels i.e. j=1,3 and 5. This selection is also done to remove redundancy created high adjacent correlation of modulation channels. The LP coefficients obtained are further transformed to cepstral coefficients i.e. LPCC. This is done because LPCC are shown to be more robust and reliable for speech recognition than using direct LP coefficients.

2.3.2.4 Classification

We get 23 spectral features and 18 LPCC features. Thus a total of 41 MSFs are calculated for each frame. These features are extracted for each emotion and classified using and Support Vector Machine (SVM) classifier. After performing the training of SVM, test samples are applied for the classification.

2.3.3 Speech Emotion Classification using Machine Learning Algorithms

2.3.3.1 Feature Extraction

Classification of emotion is performed using the architecture of Distributed Speech Recognition System (DSR). Acoustic features are extracted from the input speech using speech recognition front-end algorithm of the ETSI ES 202 211 V1.1.1 standard. The input speech is sampled at 16kHz and divided into overlapping frame size of 400 samples with an overlap of 160 samples. For each frame, 15 coefficients are extracted. These features are: 1 log-energy coefficient and 12 MFCC features and their first and second order derivatives, the pitch period and jitter, and the voicing class. All the frames which are marked as silence i.e. unvoiced segments are removed. Also, various statistical features like mean, variance, maxima, minima, quartiles etc are also calculated. Over 3800 statistical components are calculated. CFSSubsetEval method provided by WEKA (Waikato Environment for Knowledge Analysis) is used for the feature selection. Finally, feature discretization and feature normalization is performed.

2.3.3.2 Classification

Various emotional databases are available for research purposes. 6 emotions are considered. They are anger, disgust, fear, happiness, boredom and sadness. Different algorithms which are provided by WEKA are considered for classification. SMO algorithm is found to give best classification results.

2.3.4 Emotion recognition using a hierarchical binary decision tree approach

It is important to track the emotions in human speech for study of human communications and behaviors. Many classification schemes have been worked on for emotion recognition systems. Here, a hierarchical computation structure is used for emotion classification.

2.3.4.1 Feature Extraction

The table below shows the features used. These features are extracted using OpenSmile toolbox. There are 16 low level descriptors like ZCR, pitch, HNR etc. 12 statistical features for each low level descriptor eg mean, standard deviation, minimum, maximum etc. In all, there are 384 acoustic features.

Table 2.2 Features used for Emotion Recognition

Raw Acoustic Features + Deltas	Statistical Functionals
Pitch (f0) Root mean square energy (rms) Zero crossing rate (zcr) Harmonic to noise ratio (hnr) Mel-frequency cepstral coefficients (1–12 mfcc)	Mean, standard deviation, kurtosis Skewness, minimum, maximum Relative position, range Two linear regression coefficients Mean square error of linear regression

The extracted features are then normalized wrt mean and variance of the neural speech of the database used. The feature selection is performed using binary logistic regression with step wise forward selection in a standard statistic software SPSS.

2.3.4.2 Classification

A hierarchical binary decision tree used for emotion classification takes the easiest task at first stage and most ambiguous at the last stage. Thus, a multiclass problem is split into numerous two-class problems where the easy classification is taken at top and difficult ones at last. The realization varies for each database. The order of classification is important. Different binary classifiers such as Bayesian Logistic Regression (BLR), Support Vector Machine (SVM) etc are considered.

2.3.4.3 System Diagram

Below figure explains the classification framework: a hierarchical binary decision tree with easiest task at the first stage and most ambiguous task at the last stage.

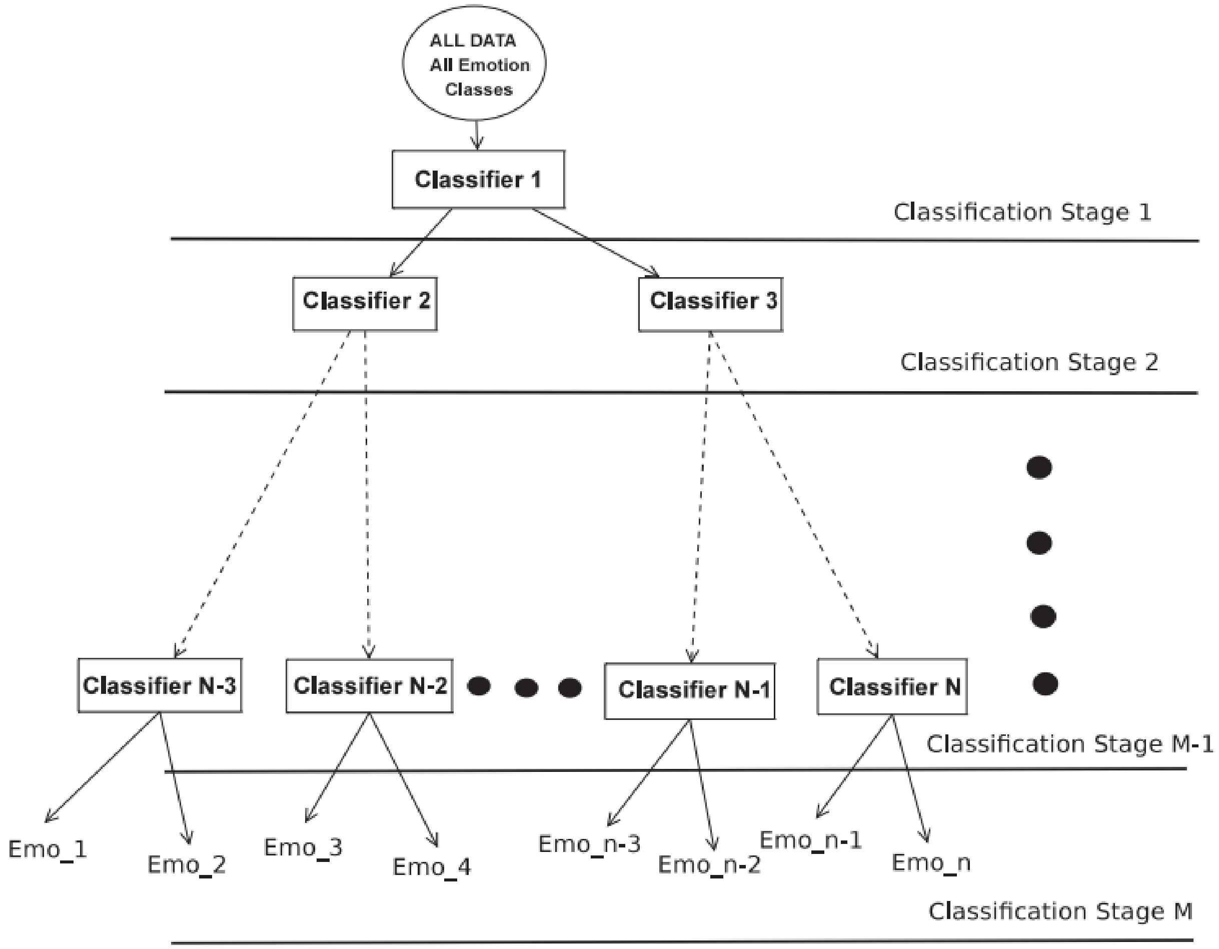

Figure 2.5 System Diagram for Emotion Recognition using hierarchical binary decision tree approach

CHAPTER 3

PROPOSED SCHEME – AUDIO EMOTION ANALYSIS

3. AUDIO EMOTION ANALYSIS

Humans can express emotions in various forms like face, speech, gait etc. Most common emotion analysis in humans is performed by analyzing face, speech etc. Here only audio emotion analysis is considered. There are numerous applications of emotion analysis. Thus it is important to effectively recognize different emotions. There are different kinds of emotions in human beings. Some of them are anger, disgust, fear, happy, neutral, sad, surprise etc. The emotions can be broadly classified as positive or negative emotions. Analyzing emotions involves extracting various features of a speech and classifying using an appropriate classifier.

The most commonly and widely used acoustic features for speech or speaker recognition is Mel-scale frequency cepstral coefficients. 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.

This feature extraction is a lossy and non-invertible transformation. This means that once the features are extracted from the sound sample, the transformation cannot lead to a perfect reconstruction. The primary reason for allowing the loss of information is computational complexity and robustness.

MFCC features can be extracted from a sound sample in various steps

1. Pre-emphasis
2. Framing
3. Hamming Window
4. Fast Fourier Transform
5. Triangular Band-pass Filter
6. Discrete Cosine Transform

For better operations, Log Energy and Delta Operations are also performed. Thus, each frame is converted to 12 MFCCs + 1 normalized energy parameter. The first and second derivatives (D’s and DD’s) of MFCCs and energy are estimated, resulting in 39 numbers representing each frame.

Increasing the number of parameters leads to an increase in complexity and a large amount of training sequence.

LOG ENERGY: We consider the energy within the frame as an important feature and thus 13th feature of MFCC is log energy

DELTA CEPSTRUM: The time derivatives of energy and MFCC are added as a new feature, which shows the velocity (differential coefficient) and acceleration coefficient of energy and MFCC. If we add the velocity, the feature dimension is 26. If we add both the velocity and the acceleration, the feature dimension is 39. Most of the speech recognition systems use 39-dimensional features for recognition.

The experiments performed in different dimensions of emotion analysis are

- Scream detection in a speech
- Detection of change of emotion
- Emotion classification

3.1 Scream detection

Scream in an audio represents an anomaly. Scream detection is useful in various applications like in old age homes and public places where a scream would mean some disturbances which need to be urgently catered to. Scream detection is carried out by 2 methods which are explained in detail.

3.1.1 Scream detection using entropy

Either Non-extensive entropy or Shannon entropy is used for scream detection. The Non-extensive entropy with Gaussian gain is used for the scream detection. Non-extensive entropy performs better than other entropies for the representation of correlated texture patterns containing non-additive information. The non-extensive entropy can be defined as follows.

Consider a random variable X = $\{x_1, x_2 \dots x_n\}$
With associated probabilities P = $\{p_1, p_2 \dots p_n\}$

Assuming probability distribution as complete
i.e. $p_i \in [0,1]$

and $\sum_{i=1}^{n} p_i = 1$ for i=1,2…n

The information gain on the ith event of X with an associated probability p$_i$ can be defined by Gaussian function as

$$I(p_i) = e^{-p_i^2} \quad (3.1)$$

The entropy of X is defined as

$$H(P) = E(I(p_i)) = \sum_{i=1}^{n} p_i I(p_i)$$

$$H(P) = \sum_{i=1}^{n} p_i e^{-p_i^2} \quad (3.2)$$

Non-extensive entropy was proposed for representation of textures containing non-additive information content and that repeat in regular manner over space. However, for random textures, entropies like Shannon being additive work better. Shannon entropy is defined as

$$H = -\sum p(x)\, logp(x) \quad (3.3)$$

The Gaussian information gain function of the non-extensive entropy ensures that the low probability event falls inside the bell of Gaussian. Thus the anomalies can be detected. Non-extensive entropy with Gaussian gain is used for the scream detection. Non-extensive entropy performs better than other entropies for the representation of correlated texture patterns containing non-additive information.

Non-extensive entropy has been used earlier for anomaly detection in videos and neural networks. Here, it is used for scream detection. Scream is an anomaly in an audio. Non-extensive entropy is used for textures containing non-additive information. But here, we are using sum of non-extensive entropy for anomaly detection. This is done because non-extensive entropy helps to determine regular patterns. The sum of entropies will behave abnormally for an abnormal behavior than for a normal behavior.

The steps involved in scream detection from audio using non-extensive entropy are as follows

i. For an input sound sample, MFCC features are found

ii. 39 MFCC features (12 MFCC features and 1 normalized energy parameter, appended by delta and acceleration coefficients) are found for each sample of frames considered.

iii. MFCC features are divided into number of overlapping slices. Thus, each slice contains 39 features for few frames. 50% overlap is considered.

iv. For each of 39 features, non-extensive entropy is found for all overlapping slices.

v. For each slice, the 39 entropies found are added up and graph is plotted

vi. The resulting graph shows anomalies where scream is detected. This means that the graph shows a spike where scream is found.

3.2 Detection of Change of Emotion

In day to day life, it is possible that emotion of a person may change while speaking. Thus, in a speech, it is possible that several emotions exist. Tracking the change of emotion helps to determine and understand the emotional state of a person.

SAVEE database which is freely available for research purpose is considered. This database consists of different emotions for different speakers. A mosaic of all emotions is created for each speaker and used for testing. The change of emotion is found using non-extensive entropy. The steps involved are as follows.

i. For an input sound sample, MFCC features are found

ii. 39 MFCC features (12 MFCC features and 1 normalized energy parameter, appended by delta and acceleration coefficients) are found for each sample of frames considered.

iii. MFCC features are divided into number of overlapping slices. Thus, each slice contains 39 features for few frames. 50% overlap is considered.

iv. For each of 39 features, non-extensive entropy is found for all overlapping slices.

v. The entropy for a particular array of features is found by calculating the probability values of unique values in the array and then applying the formula of non-extensive entropy.

vi. For each slice, the 39 entropies found are added up and graph is plotted.

vii. The resulting graph shows the change in emotion points.

3.3 Emotion Classification

It is important to extract features from the speech to understand the emotional state of a person. It is difficult to know the amount and type of features necessary for the feature extraction. Also, it is important to choose the correct classifier for emotion classification.

SAVEE database, which is created for research purpose is used. The database consists of 15 samples each of 7 emotions (Anger, Disgust, Fear, Happy, Neutral, Sad and Surprise). In all there are 4 speakers. Different emotion classification techniques involving extraction of different features from the input speech sample are as follows

3.3.1 Emotion Recognition using MFCC and entropy

Three features are calculated from each speech sample. These features are

- For each speech signal, 39 MFCC features are calculated by dividing the speech signal into different frames. For each 39 MFCC feature, we calculate the weighted mean across all frames. The weighted mean is calculated as follows

  Let the number of frames f = 1, 2, 3….N

  The histogram of ith MFCC feature, where i = 1, 2, 3...39 denoted by the probability value $p(z_{if})$

  The weighted mean of ith MFCC feature across all frames is calculated as

  $m_i = \sum_{f=1}^{N} z_{if} p(z_{if})$ (3.4)

  Thus weighted mean of each of 39 features is calculated across all frames and we get a 1X39 feature vector for each speech signal. This is the first feature.

- Also, for each 39 MFCC feature, we calculate non-extensive entropy across all frames. The entropy is calculated as follows

  For a random variable $X = \{x_1, x_2 \dots x_n\}$

  With associated probabilities $P = \{p_1, p_2 \dots p_n\}$

  Assuming probability distribution as complete, the entropy of X is defined as follows

$$H(P) = \sum_{i=1}^{n} p_i e^{-p_i^2} \quad (3.5)$$

So, for each 39 MFCC features, we get an entropy value. Thus we get a 1X39 entropy vector for each signal. This is second feature.

- The third feature is sum of sampled data of an input speech

These features are extracted for training samples and testing samples. Different classifiers are used for classification like SVM, GMM.

3.3.2 Emotion Recognition using different entropy values of pitch and energy

The main features used here are pitch and energy. Pitch represents the perceived frequency of the sound sample. Pitch and energy varies differently for different emotions. For example, pitch and energy will be low when the person is sad than when he is in anger. Entropies are calculated for different scenarios involving pitch and entropy values of the input speech. Following features are extracted

- Pitch entropy and energy entropy is calculated from the co-occurrence matrix of pitch and energy values of input speech. Co-occurrence matrix indicated the distribution of co-occurring values at a given offset.
- 39 non-extensive entropy values of 39 MFCC features across speech sample are calculated.
- Successive differences of input samples are calculated and then pitch entropy and energy entropy is found.
- Mean of pitch entropy and energy entropy of speech.
- Entropy of pitch entropy and energy entropy

So, in all we get a feature vector of size 56 for each speech sample.

These features are calculated for testing and training samples both and then classified using SVM classifier.

CHAPTER 4

EXPERIMENTION RESULTS AND DISCUSSION

4. RESULTS

Many speech emotion databases are available for research purposes. Scream detection involves working with different types of screams like male scream, female scream, short or long duration of scream. These different types of screams are downloaded from internet. The speech involved is taken from database IITKGP-SEHSC: Hindi speech corpus. This database is created by IIT Kharagpur. For emotion recognition, Surrey Audio-Visual Expressed Emotion (SAVEE) database which consists of phonetically-balanced TIMIT sentences is used. This database has been provided for research purposes in emotion recognition systems. The experimentation of audio emotion analysis is carried out on MATLAB 32-bit version 7.10.0.499 and 32-bit Windows 7 operating System.

Audio emotion analysis is categorized into detection of extreme emotion of scream and classification among different emotions. Scream is a negative emotion and thus, detection of scream can prove helpful in many practical applications. Detection of scream is done using 2 methods. First method involves detecting scream by analyzing 3 features: Log energy, Auto-correlation values and MFCC features. The second method involves using non-extensive entropy on the MFCC features extracted from speech. A table is constructed showing how many screams have been correctly classified using these methods. It can be seen that using entropy values of MFCC features give us better accuracy.

Recognizing emotions from human speech help us to analyze the emotional state of a person. Extracting features from speech and classifying them improves the computer human interaction so that emotions are recognized by computers. Here emotion classification is performed using 2 methods. First method involves extraction of entropy and weighted mean from MFCC features of speech. Second method involves extraction of different types of entropy values from pitch and energy of speech. A table showing percentage accuracy for each emotion of each speaker is indicated.

A confusion matrix for each speaker is drawn to understand which emotion gets classified correctly and incorrectly. There is a good indication that there can be a good classification between positive, negative and neutral emotions. Thus a percentage accuracy table and confusion matrix is created for the same. At the end, the result of different techniques of emotion classification, which are discussed in chapter 2 literature review section, are compared against each other.

4.1 Scream detection by analyzing various audio features

Speech samples are taken from database IITKGP-SEHSC, created by IIT Kharagpur. A small database is created by adding different types of screams and other sounds like applause, laugh, footsteps etc in various combinations in the background of the speech.

All the 3 features are extracted for all samples of this database and analyzed. Thus a comparison is made between all these 3 features and analyzed which all non-scream features are classified as scream. Following results are obtained.

Table 4.1 Classification of scream by different features [3]

Sound Sample	**Total number of samples**	**Sound classified as Scream by:-**		
		Log Energy	**Auto-correlation**	**36 MFCC**
Applause	**5**	5	3	4
Footsteps	**5**	0	0	5
Laugh	**4**	2	1	3
Telephone	**6**	3	1	5
Scream	**12**	12	12	12
Can	**1**	0	0	1
Door	**1**	0	0	0
Pour	**1**	0	0	1
Sneeze	**1**	1	0	1
TOTAL CORRECT CLASSIFICATION	**36**	55%	80%	13.3%

This matrix tells us that which all scream and non-scream sounds are classified as scream by these 3 features. For example, there are total of 5 Applause samples in the complete database, out of which all 5 are wrongly classified as scream segment by Log Energy, only 3 out of 5 are classified as scream by Auto-correlation and 4 out 5 are classified wrongly as scream by MFCC features.

Comparing the results of these 3 features, we can see that 55% of samples are correctly classified as scream by Log Energy feature, 80% by auto-correlation and approx. 14% by MFCC features. Thus Auto-correlation gives us the best result out of all.

4.2 Scream detection using entropy

For the detection of scream using entropy, different type of scream samples are taken from the web and added in the background of speech samples taken from IITKGP-SEHSC database. The screams are added in various combinations and the result is as follows.

In the below graph, the input speech consists of 2 screams, one in the beginning and another towards the end which are added in the background. Following figure is the input speech with background noise and scream detection using Non-extensive entropy

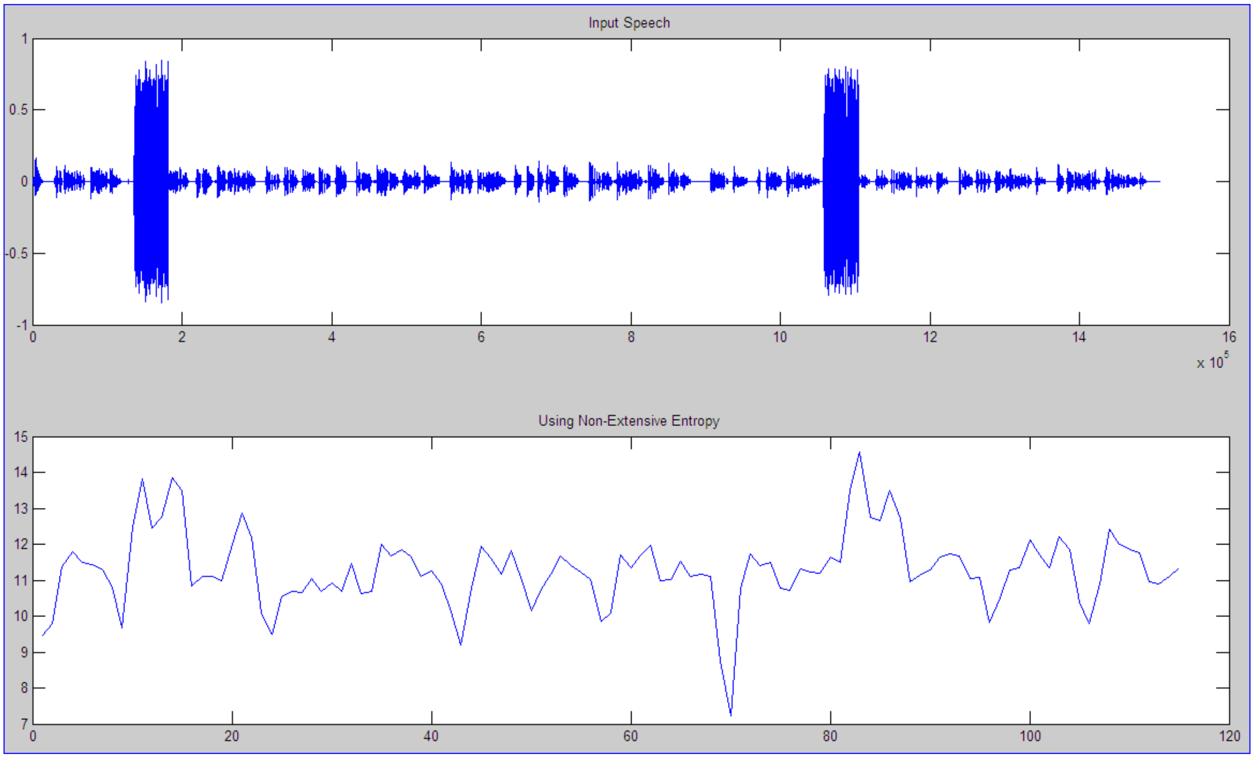

Figure 4.1 Scream Detection using Non-Extensive Entropy

If we use Shannon entropy instead of non-extensive entropy, we get graph almost same as the one for non-extensive. Below figure shows input speech with background scream and scream detection using Shannon entropy.

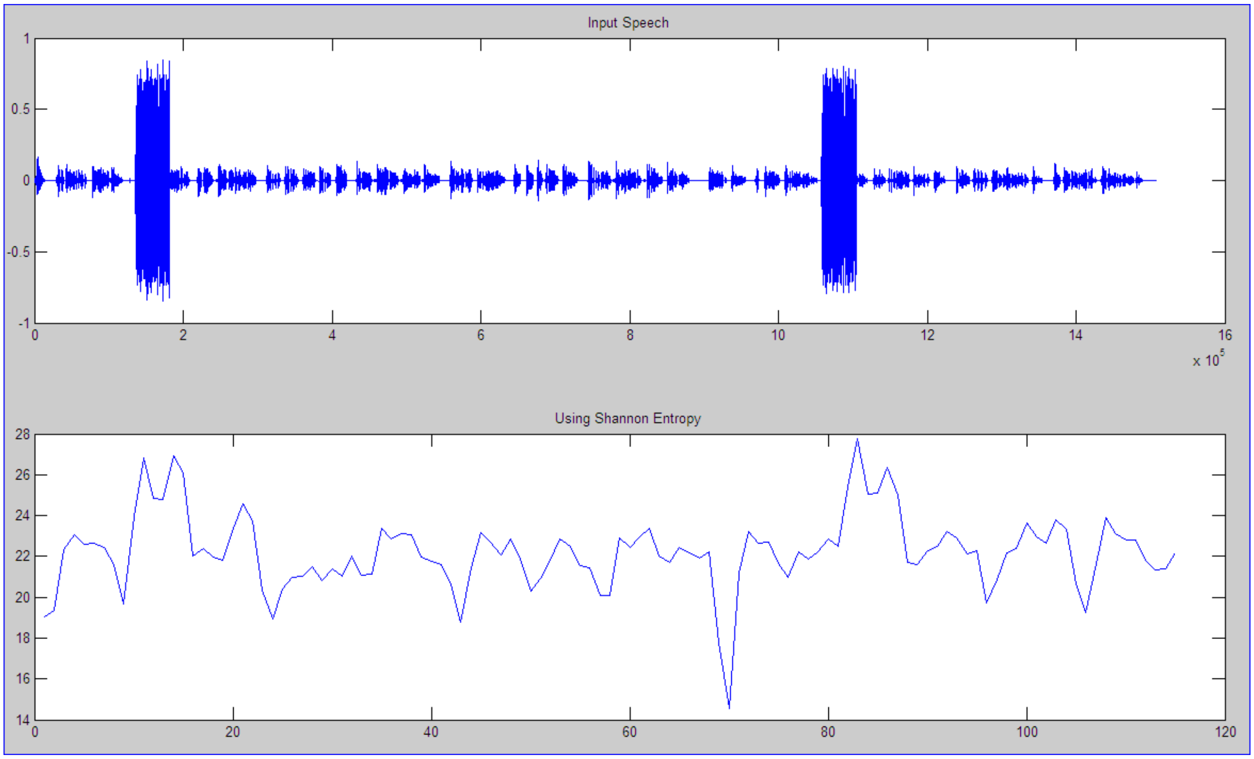

Figure 4.2 Scream Detection using Shannon Entropy

By analyzing the graphs of using both entropies, we find that the graph shows a peak where the scream is detected. The specific positions of scream are found and shown in graph below. The spikes show the anomalies i.e. scream in an audio. The first graph shows spikes of scream when non-extensive entropy is used. Second graph shows spikes of scream when Shannon entropy is used.

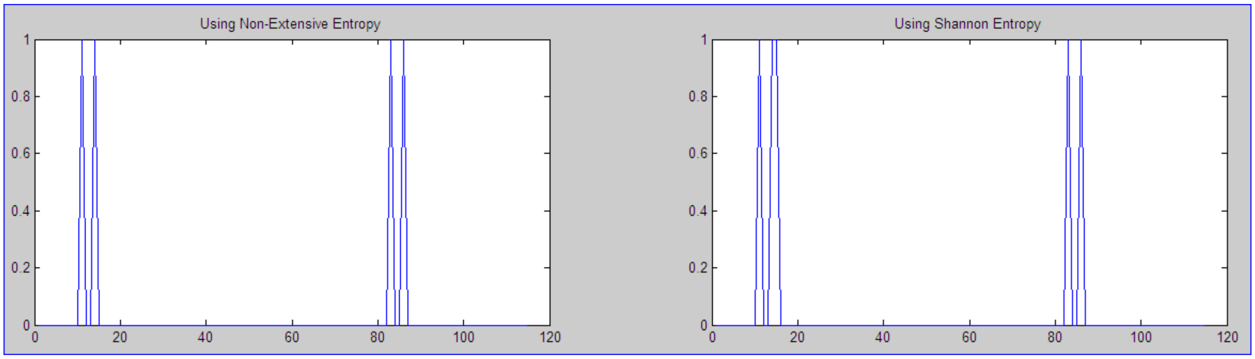

Figure 4.3 Detection of Scream points

4.3 Detection of change of emotion

SAVEE database consist of many samples for 7 different emotions viz. anger, disgust, fear, happy, neutral, sad and surprise. A mosaic of all these emotions for a particular speaker is created. This means that for a particular speaker, the emotions are added back to back. Now, entropy is applied to detect the change of emotion.

Following graph shows the mosaic of all the 7 emotions of a speaker.

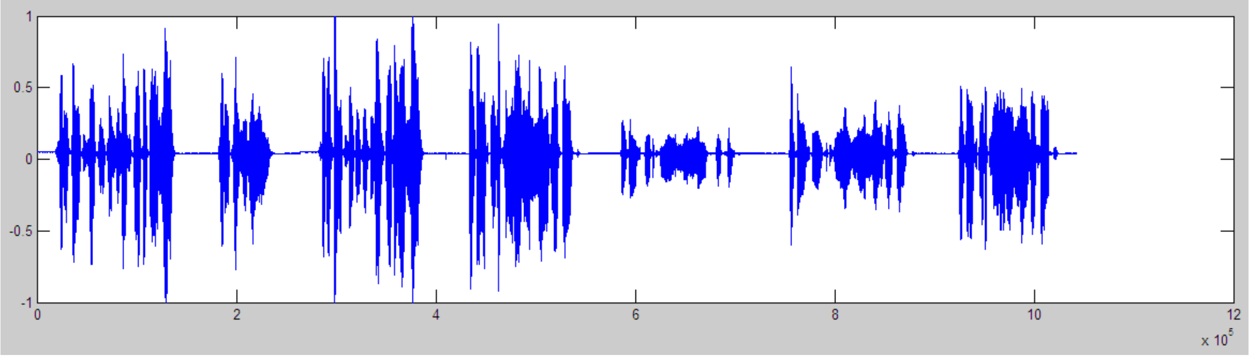

Figure 4.4 Mosaic of emotions

Following graph is obtained after applying entropy. The dips in the graph show the change in the emotion.

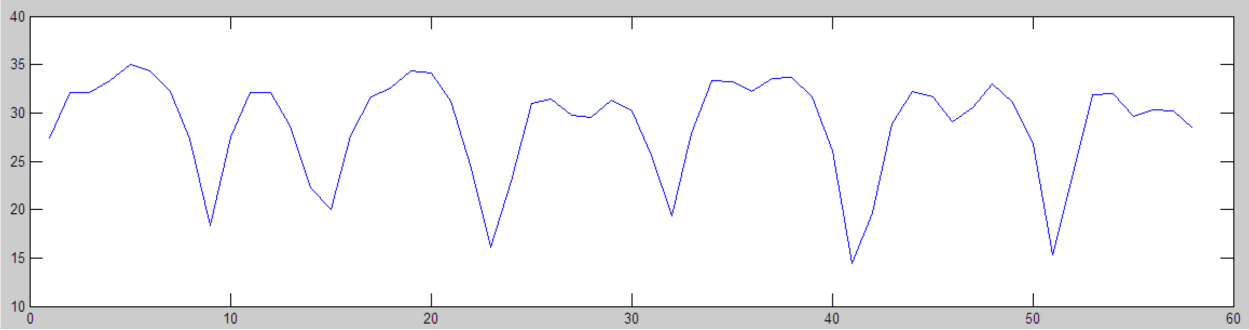

Figure 4.5 Detection of change of emotion

4.4 Emotion Classification

SAVEE database is considered for emotion classification experiments. There are 4 different speakers, with 15 samples for each of the 7 emotions (anger, disgust, fear, happy, sad, surprise and neutral). Out of these only 5 are considered viz anger, happy, sad, surprise and neutral. Emotion classification is performed by extracting weighted mean and entropy of MFCC features of input speech and the sum of input sound samples. The emotions are classified using SVM classifier. Below table shows the percentage accuracy of each emotion for each speaker.

Table 4.2 Percentage Accuracy for Emotion Classification for different speakers

SPEAKER	Percentage accuracy for each emotion					Total percentage Accuracy
	Anger	Happy	Neutral	Sad	Surprise	
1	100	40	100	100	100	88
2	40	60	100	80	60	68
3	40	80	80	20	100	60
4	20	60	40	20	60	40

The above table depicts the percentage accuracy for each emotion for each speaker. Not all the emotions give 100% accuracy result; this means that some of these emotions are misclassified. To know the complete details of classification, a confusion matrix is created for each speaker to understand which emotion is getting classified as which emotion. This further helps in proper distinguishing of emotions. The confusion matrix is created for each speaker separately for all 5 emotions.

Table 4.3 Confusion Matrix

SPEAKER 1		anger	happy	neutral	sad	surprise
		1	2	3	4	5
Anger	1	5	0	0	0	0
Happy	2	0	2	0	1	2
Neutral	3	0	0	5	0	0

Sad	4	0	0	0	5	0
Surprise	5	0	0	0	0	5

SPEAKER 2		anger	happy	neutral	sad	surprise
		1	2	3	4	5
Anger	1	2	1	0	0	2
Happy	2	1	3	0	0	1
Neutral	3	0	0	5	0	0
Sad	4	0	0	0	4	1
Surprise	5	0	1	0	0	5

SPEAKER 3		anger	happy	neutral	sad	surprise
		1	2	3	4	5
Anger	1	2	1	0	0	2
Happy	2	0	4	0	0	1
Neutral	3	0	0	4	1	0
Sad	4	0	0	5	0	0
Surprise	5	0	0	0	0	5

SPEAKER 4		anger	happy	neutral	sad	surprise
		1	2	3	4	5
Anger	1	1	1	0	0	3
Happy	2	1	3	0	0	1
Neutral	3	0	2	3	0	1
Sad	4	1	0	3	1	1
Surprise	5	2	0	0	0	3

The above confusion matrix gives an indication that there can be good classification between neutral, negative and positive emotions. Anger is taken as negative emotion and Happy is taken as positive emotion. Below table shows the percentage accuracy for emotion classification.

Table 4.4 Percentage Accuracy for classification of positive, negative and neutral emotions

SPEAKER	Percentage Accuracy
1	100
2	94
3	74
4	60

The confusion matrix for above classification is shown below

Table 4.5 Confusion Matrix for positive, negative and neutral emotions

SPEAKER 1		Happy	Neutral	Anger
		1	2	3
Happy	1	5	0	0
Neutral	2	0	5	0
Anger	3	0	0	5

SPEAKER 2		Happy	Neutral	Anger
		1	2	3
Happy	1	4	0	1
Neutral	2	0	5	0
Anger	3	0	0	5

SPEAKER 3		Happy	Neutral	Anger
		1	2	3
Happy	1	4	0	1
Neutral	2	0	5	0
Anger	3	3	0	2

SPEAKER 4		Happy	Neutral	Anger
		1	2	3
Happy	1	3	0	2
Neutral	2	2	3	0
Anger	3	2	0	3

The left out emotions are tested against this classification to see whether the emotions like fear, disgust get classified as positive, negative or neutral emotions

Table 4.6 Confusion matrix for all emotions

SPEAKER 1	Anger	Neutral	Happy
Anger	5		
Disgust	2	3	
Fear	3		2
Happy			5
Neutral		5	
Sad	1	4	
Surprise	5		

SPEAKER 2	Anger	Neutral	Happy
Anger	5		
Disgust	1	2	2
Fear	4		1
Happy	1		4
Neutral		5	
Sad		5	
Surprise	2		3

SPEAKER 3	Anger	Neutral	Happy
Anger	2		3
Disgust		5	
Fear	1		4
Happy	1		4
Neutral		5	
Sad		5	
Surprise			5

SPEAKER 4	Anger	Neutral	Happy

Anger	3		2
Disgust	2	2	1
Fear	4	1	
Happy	2		3
Neutral		3	2
Sad		5	
Surprise	4		1

The proposed scheme is compared with various other techniques of emotion recognition, some of which are discussed in chapter 2 literature review. The various techniques for comparisons are

- Extracting weighted mean and entropy of the MFCC features and using SVM classifier for classification.
- Extracting weighted mean and entropy of the MFCC features and using GMM classifier for classification.
- Eigen values are extracted from the auto-correlation matrix of the input speech
- 6 Spectral features are extracted from the spectro-temporal representation of the input speech.
- Extracting the below mentioned features for the input sample and classifying using SVM
    - Pitch and energy entropy of co-occurrence matrix
    - 39 entropy values of MFCC
    - Pitch and energy entropy of successive difference of input sample
    - Mean of pitch and energy entropy of input
    - Entropy of pitch entropy and energy entropy

The percentage accuracy for each technique for a particular speaker is shown in the table below.

Table 4.7 Percentage Accuracy for different emotion classification techniques

Emotion Classification Method	Percentage Accuracy
Using different entropy values of pitch and energy	90%
Using entropy of MFCC features and SVM classifier	88%
Using Spectro-Temporal features of speech [13]	68%
Using entropy of MFCC features and GMM classifier	52%
Using Eigen Values of Auto-Correlation Matrix [11]	37.1%

CHAPTER 5

CONCLUSION

AND FUTURE WORK

5. CONCLUSION

Emotion analysis in human speech is an important aspect of today’s world to help us understand the emotional state of a human being. It is difficult to list down the features to be extracted from the speech which can effectively and efficiently identify an emotion. Also, the method of classification plays an important role. Comparing 5 different methods of emotion classification, we can conclude that entropy plays vital role in increasing the accuracy of recognition.

Entropy gives us the idea of amount of information stored. Taking out information from the speech using entropy helps us to understand the areas of anomalies and differences which help us in classification.

However, there are many aspects that need to be improved. It is difficult to train the system for a particular speaker’s emotion and test using other speaker’s emotion. This can be improved upon by further refining the features to be extracted. Also, in scream detection the method can be more effective even in various background sounds like a child crying or glass breaking, which resemble a scream.

REFRENCES

[1] El Ayadi, Moataz, Mohamed S. Kamel, and Fakhri Karray. "Survey on speech emotion recognition: Features, classification schemes, and databases." *Pattern Recognition* 44.3 (2011): 572-587.

[2] Anagnostopoulos, Christos-Nikolaos, Theodoros Iliou, and Ioannis Giannoukos. "Features and classifiers for emotion recognition from speech: a survey from 2000 to 2011." *Artificial Intelligence Review* 43.2 (2015): 155-177.

[3] Huang, Weimin, et al. "Scream detection for home applications." Industrial Electronics and Applications (ICIEA), 2010 the 5th IEEE Conference on. IEEE, 2010.

[4] Valenzise, Giuseppe, et al. "Scream and gunshot detection and localization for audio-surveillance systems." Advanced Video and Signal Based Surveillance, 2007. AVSS 2007. IEEE Conference on. IEEE, 2007.

[5] Koolagudi, Shashidhar G., et al. "IITKGP-SEHSC: Hindi speech corpus for emotion analysis." *Devices and Communications (ICDeCom), 2011 International Conference on*. IEEE, 2011.

[6] Jackson, P., and S. Haq. "Surrey Audio-Visual Expressed Emotion(SAVEE) Database." (2014).

[7] L. Lu, H. Zhang, and H. Jiang, “Content analysis for audio classification and segmentation,” Speech and Audio Processing, IEEE Transactions on, vol. 10, no. 7, pp. 504–516, 2002.

[8] G. Peeters, “A large set of audio features for sound description (similarity and classification) in the CUIDADO project,” CUIDADO Project Report, 2004.

[9] S. Sigurdsson, K. B. Petersen, and T. Lehn-Schiøler, “Mel frequency cepstral coefficients: An evaluation of robustness of mp3 encoded music,” in Proceedings of the Seventh International Conference on Music Information Retrieval (ISMIR), 2006.

[10] Susan, Seba, and Madasu Hanmandlu. "Unsupervised detection of nonlinearity in motion using weighted average of non-extensive entropies." Signal, Image and Video Processing (2013): 1-15.

[11] Kandali, Aditya Bihar, Aurobinda Routray, and Tapan Kumar Basu. "Comparison of Features Based on MFCCs and Eigen Values of Autocorrelation Matrix for Cross-Lingual Vocal Emotion Recognition in Five Languages of Assam." India Conference (INDICON), 2009 Annual IEEE. IEEE, 2009.

[12] @misc{Ellis05-rastamat, Author = {Daniel P. W. Ellis}, Year = {2005}, Title = {{PLP} and {RASTA} (and {MFCC}, and inversion) in {M}atlab}, Url = http://www.ee.columbia.edu/~dpwe/resources/matlab/rastamat/}, Note = {online web resource}}

[13] Wu, Siqing, Tiago H. Falk, and Wai-Yip Chan. "Automatic recognition of speech emotion using long-term spectro-temporal features." *Digital Signal Processing, 2009 16th International Conference on*. IEEE, 2009.

[14] Casale, Salvatore, et al. "Speech emotion classification using machine learning algorithms." *Semantic Computing, 2008 IEEE International Conference on*. IEEE, 2008.

[15] Lee, Chi-Chun, et al. "Emotion recognition using a hierarchical binary decision tree approach." *Speech Communication* 53.9 (2011): 1162-1171.

[16] Susan, Seba, and Mayank Dwivedi. "Dynamic Growth of Hidden-Layer Neurons Using the Non-extensive Entropy." Communication Systems and Network Technologies (CSNT), 2014 Fourth International Conference on. IEEE, 2014.

[17] S. Haq and P.J.B. Jackson, "Multimodal Emotion Recognition", In W. Wang (ed), Machine Audition: Principles, Algorithms and Systems, IGI Global Press, ISBN 978-1615209194, DOI 10.4018/978-1-61520-919-4, chapter 17, pp. 398-423, July 2010.

[18] Wu, Siqing, Tiago H. Falk, and Wai-Yip Chan. "Automatic speech emotion recognition using modulation spectral features." Speech Communication 53.5 (2011): 768-785.

[19] F. Zheng, G. Zhang, Z. Song, "Comparision of Different Implementations of MFCC", Journal of Computer Science & Technology, vol. 16, no. 6, September 2001, pp. 582-589

[20] Susan, Seba, and Srishti Sharma. "A Fuzzy Nearest Neighbor Classifier for Speaker Identification." Computational Intelligence and Communication Networks (CICN), 2012 Fourth International Conference on. IEEE, 2012.

[21] Gonzalez, Rafael C., Richard E. Woods, and Steven L. Eddins. Digital image processing using MATLAB.Vol. 2. Knoxville: Gatesmark Publishing, 2009.

[22] Giannakopoulos, Theodoros. "A method for silence removal and segmentation of speech signals, implemented in Matlab." *Department of Informatics and Telecommunications, University of Athens,*

Greece, Computational Intelligence Laboratory (CIL), Insititute of Informatics and Telecommunications (IIT), NCSR DEMOKRITOS, Greece (2009).

[23] I. Tsochantaridis, T. Joachims, T. Hofmann, and Y. Altun, Large Margin Methods for Structured and Interdependent Output Variables, Journal of Machine Learning Research (JMLR), 6(Sep):1453-1484, 2005.

[24] http://www.cic.unb.br/~lamar/te073/Aulas/mfcc.pdf

[25] Ross, A. and Jain, A. K., "Multimodal biometrics:an overview," *Procc.EUSIPCO*, pp. 1221-1224, Sept.2004.

[26] http://labrosa.ee.columbia.edu/matlab/rastamat/

[27] http://www.scribd.com/doc/59159000/Speaker-Recognition-Using-MATLAB

[28] E. Douglas-Cowie, N. Campbell, R. Cowie, and P. Roach, "Emotional speech: towards a new generation of databases," Speech Communication, vol. 40, pp. 33–60, 2003.

[29] Lin, Yi-Lin, and Gang Wei. "Speech emotion recognition based on HMM and SVM." *Machine Learning and Cybernetics, 2005. Proceedings of 2005 International Conference on*. Vol. 8. IEEE, 2005.

[30] Shen, Peipei, Zhou Changjun, and Xiong Chen. "Automatic speech emotion recognition using support vector machine." *Electronic and Mechanical Engineering and Information Technology (EMEIT), 2011 International Conference on*. Vol. 2. IEEE, 2011.

[31] Sayedelahl, Aya, et al. "Audio-based emotion recognition from natural conversations based on co-occurrence matrix and frequency domain energy distribution features." *Affective computing and intelligent interaction*. Springer Berlin Heidelberg, 2011. 407-414.

[32] K. Oatley and P. N. Johnson-Laird, "Towards a Cognitive Theory of Emotions", *Cognition and Emotion*, pp 29-50, 1987.

[33] S. Furui, *Digital Speech Processing, Synthesis and Recognition*, New York: Marcel Dekker, 1989.

[34] L. R. Rabiner and B. H. Juang, B.H., *Fundamentals of Speech Recognition*, Englewood Cliffs, NJ: Prentice Hall, 1993.

[35] I. R. Murray and J. L. Arnott, “Toward the simulation of emotion in synthetic speech: A review of the literature on human vocal emotion”, *J. Acoust. Soc. Amer.*, 93 (2), pp 1097–1108, 1993.

[36] Jegou, Herve, Matthijs Douze, and Cordelia Schmid. "Product quantization for nearest neighbor search." *Pattern Analysis and Machine Intelligence, IEEE Transactions on* 33.1 (2011): 117-128.

[37] P.K. Atrey, N.C. Maddage, and M.S. Kankanhalli. Audio Based Event Detection for Multimedia Surveillance. Acoustics, Speech, and Signal Processing, 2006. ICASSP-97., 2006 IEEE International Conference on, 2006.

[38] C. Clavel, T. Ehrette, and G. Richard. Events Detection for an Audio-Based Surveillance System. Multimedia and Expo, 2005. ICME 2005. IEEE International Conference on, pages 1306–1309, 2005.

[39] W. Zajdel, J. D. Krijnders, T. Andringa, D. M. Gavrila, “CASSANDRA: Audio-video Sensor Fusion for Aggression Detection”, AVSS 2007, pp.200-205

[40] P.K. Atrey, N.C. Maddage, and M.S. Kankanhalli. Audio Based Event Detection for Multimedia Surveillance. ICASSP 2006

[41] A. F. Smeaton, M. McHugh, Towards event detection in an audio-based sensor network, Proceedings of the third ACM international workshop on Video surveillance & sensor networks, 2005, pp.87-94

[42] S. Ntalampiras, I. Potamitis, N. Fakotakis, "On acoustic surveillance of hazardous situations," icassp, pp.165-168, 2009 IEEE International Conference on Acoustics, Speech and Signal Processing, 2009

[43] R. Radhakrishnan, A. Divakaran, "Systematic Acquisition of Audio Classes for Elevator Surveillance", SPIE Image and Video Communications and Processing, Vol. 5685, March 2005, pp. 64-71

[44] L. Gerosa, G. Valenzise, F. Antonacci, M. Tagliasacchi, A. Sarti, "Scream And Gunshot Detection In Noisy Environments", EURASIP European Signal Processing Conference, September, Poznan, Poland, 2007

[45] A. Rabaoui, M. Davy, S. Rossignol, Z. Lachiri and N. Elouze, "Improved One-Class SVM Classifier for Sounds Classification", *IEEE AVSS*, London, Sept 2007

[46] Kwon, Oh-Wook, et al. "Emotion recognition by speech signals."*INTERSPEECH*. 2003.

[47] Wu, Siqing, Tiago H. Falk, and Wai-Yip Chan. "Automatic speech emotion recognition using modulation spectral features." *Speech communication* 53.5 (2011): 768-785.

[48] Pan, Yixiong, Peipei Shen, and Liping Shen. "Speech emotion recognition using support vector machine." *International Journal of Smart Home* 6.2 (2012): 101-107.

[49] D. Ververidis and C. Kotropoulos, “Emotional speech recognition: Resources, features, and methods,” Speech Communication, vol. 48, pp. 1162–1181, 2006.

[50] M. Shami and W. Verhelst, “An evaluation of the robustness of existing supervised machine learning approaches to the classification of emotions in speech,” Speech Communication, vol. 49, pp. 201–212, 2007.