Emotion Recognition from Audio Signal

Rajni, Dr. Nripendra Narayan Das

Abstract—Emotion recognition from Audio signal Recognition is a recent research topic in the Human Computer Interaction. The demand has risen for increasing communication interface between humans and digital media. Many researchers are working in order to improve their accuracy. But still there is lack of complete system which can recognize emotions from speech. In order to make the human and digital machine interaction more natural, the computer should be able to recognize emotional states in the same way as human. The efficiency of emotion recognition system depends on type of features extracted and classifier used for detection of emotions. In this project emotion from Hindi speech is developed. The database used was collected from various speakers belonging to different genders and age group. This work basically focused on eight emotions which comprises of a combination of fundamental emotions with some advance emotions and are listed as: Happy, Angry, Sad, Depressed, Bored, Anxiety, Fear and Nervous. These signals were preprocessed and analyzed using various techniques like: cepstral, linear prediction coefficient etc. In feature extraction various parameters used to form a feature vector are: fundamental frequency, pitch contour, formants, duration (pause length ratio) etc. These features are classified by using K Nearest Neighbor (KNN) classifier and Neural Network based classifiers. The performance based on both classifiers is measured in terms of their accuracy.

Index Terms—Audio Signal, Emotions, Cepstral, KNN, Neural Network Classifier.

I. INTRODUCTION

Emotion recognition from speech is a challenging problem in audio signal processing. Lot of information like: age, gender, emotion, person, and action can be estimated from a speech signal, emotion recognition is one of them. Emotion depends on voice generated from different parts of human vocal system. These systems can be helpful in detecting customers’ emotion, medical entertainment, crime detection, robotics voice and may other cases. Speech communication contains paralinguistic information of the speaker. Although enormous efforts are invested in recognising the emotions from speech but still much research is needed. A brief literature survey in this field is detailed in this section. Emotion recognition from Hindi speech has been done by Shashidhar et al [2], in this emotions like anger, disgust, fear, happy, neutral, sad, sarcastic and surprise are used to classify the emotions. Prosodic (energy, pitch, and duration) and spectral (MFCC) features are used to classify these emotions. A text independent emotion recognition has been proposed by Chauhan et al [3]. They used Mel Frequency Cepstral Coefficient (MFCC) and Gaussian mixture model for detecting emotions. Further speech emotion detection systems are classified using SVM and LIBSVM by Wankhade et al [4] and features are selected by using MFCC and MEDC. Further pitch contour based algorithm has been proposed by Ahmed [5]. The transformation of emotions using pitch parameter for Hindi speech is analyzed. An intonation pattern based Hindi speech analysis is proposed by Agrawal et al [6]. In this the neutral sentences are transformed into emotion rich sentences, or phrases. A study were made over transformation of emotion based on intonation patterns for hindi speech given by Agrawal et al [6], in which they worked to transform neutral sentences into emotion rich sentences or words with an changing an intonation pattern. Features are computed by using fundamental frequency($f_0$), energy contour as parameters to convert intonation emotion . In Bahugama and Raiwani [7] proposed MFCC and vector quantisation approach to identify speakers emotions. Emotions considered are: Happy, sad, anger, neutral in Hindi. In Kumar and RangaBabu [8], proposed a person’s emotion recognition from audio. In this six emotions are considered. The features are classified by using support vector machine. The system is composed of emotion recognition as well as gender recognition. In Albornoz et al [9], bio inspired features are selected for emotion recognition. In this spectral and prosodic features are used for emotion recognition in noisy environment. These features are classified by using neural network classifiers. Based on these literature survey it is clear that various emotions can be computed by using acoustic features. But still there is lack of dataset in this domain. Also feature and feature combination scheme may improve the accuracy of the system. In this thesis a comparative study on neural network and KNN classification scheme is proposed based on these features.

II. METHODOLOGY

The detailed steps are described in this section.

i. Input Audio Signal: The input signal used here is audio data collected in .wav format. The dataset is collected from different groups of people. The speech is collected from all these persons in different eight emotions. By using this dataset, 120 emotional files are collected for training and testing of these algorithms. Since these signals contain noise so, the signal is
preprocessed before processing. The eight emotions used in
data acquisition are shown in Figure 1.

### Figure 1 Emotions used for recognition.

After data acquisition, since it contains some noises which
have been acquired during signal acquisition are
pre-processed.

**ii. Preprocessing:**
The collected signal contains different types of noises. These
noises are filtered by using low pass filter. For this purpose
Butterworth low pass filter is applied here. The signal is sent
to further steps.

**iii. Feature extraction**

In order to collect emotions from audio signal the features are
extracted by using duration, pitch, energy, formant and ZCR.

**Duration:**
Durati0n specifies the time taken by speaker.
Since the duration also specify emotions. Considering this
into account duration of speech is used. The duration is
computed by equation as shown in equation (i)

\[ T = \left( N - \sum_{P=0}^{P=n} (P) \right) / dt \]  

Where, 
- \( T \) = duration of sample (in second)
- \( N \) = length of sample
- \( P \) = length of pause
- \( dt \) = time rate

**ZCR:** The zero crossing per unit time is computed by using
equation (ii)

\[ Z = n_c . (f / n) \]  

Where,
- \( n_c \) = number of zero crossing per frame
- \( f \) = sampling frequency (44100 hz)
- \( n \) = length of frame (30ms)

**Energy:**

Energy used in speaking into different emotions may vary.
Considering this into account energy of audio signal is
computed by using equation (iii).

\[ E(x) = \frac{\sqrt{\sum_{i=0}^{n} s_i^2}}{n} \]  

Where,
- \( E \) = energy of sample
- \( s \) = sample value of \( i \)th frame
- \( n \) = frame length, here 30ms.

**Pitch detection:**

Pitch is one of the essential components of emotion
recognition from audio signal. It defines rate of vibration of
speaker’s vocal cord. Although different sub features like
fundamental frequency, pitch, harmony etc. are used. In this
work the features selected are: cepstral fundamental
frequency, harmony and pitch contour.

**Formant frequency:**

The formant feature specifies phonetic content of speech
signals. As we know, that Hindi is more phonetic compared
to English. So considering this point formant frequency is
selected.

**Feature combination:**

The features obtained from the audio signal are of different
sizes. In order to make them uniform the feature vector is
resized. The duration and ZCR is resized into one pixel,
whereas energy, fundamental frequency, pitch, and formant
frequency are resized to 20 pixels. In this way total feature
length of 102 pixels is generated for each audio signal. These
features are further classified by using statistical (KNN) and
Neural Network based classifiers.

**A. Classification:**

1) *K nearest Neighbour classifier*

KNN is a statistical based classifier. The classification is
based on Euclidean distance. The KNN is an instance-based
learning or lazy learning classifier. In this the function is
only approximated locally and all computation is deferred till
the classification.

2) *Neural Network based classifiers*

A Neural network system is trained with the goal that
specific input information can be classified to a particular
target. In case of neural network system is adjusted based on
a comparison of the output and the target until it matches a
specific target. In this work a cascade forward neural
network system has been chosen. Since Cascade forward
network systems are like feed forward systems, however they
incorporate an association from the input and each previous
layer to the next layers. Likewise with feed forward network
systems, a two-or more layer cascade network system can
take in any finite input-output relationship arbitrarily well
given enough hidden neurons. Keeping in mind the end goal
to procure all acoustic features from the speech ,the extracted
acoustic data features are further resized and consolidated
together to frame a feature length of 102. The extracted
A comparative analysis between both classification schemes are shown in Table 1. From this table it is clear that KNN classifiers performed better compared to neural network classifier because KNN [10] classifier preserves samples feature more accurately.

<table>
<thead>
<tr>
<th>Feature length</th>
<th>KNN</th>
<th>Neural Network</th>
</tr>
</thead>
<tbody>
<tr>
<td>102</td>
<td>85.48</td>
<td>80.0</td>
</tr>
</tbody>
</table>

**REFERENCES**


