Emotion Recognition from Speech Signals Using Deep Learning Methods

Sandeep Pathak\textsuperscript{1} & Mrs. Vaishali Kolhe\textsuperscript{2}
\begin{itemize}
\item \textsuperscript{1} PG Student, D. Y. Patil College of Engineering, Akurdi, Pune
\item \textsuperscript{2} Assistant Professor, D. Y. Patil College of Engineering, Akurdi, Pune.
\end{itemize}

Abstract: Human emotion recognition plays an important role in enhancing the Human Computer Interaction (HCI) in great length. Out of various ways Human Computer Interaction can be made speech turns out to be the most natural and fastest mode. Speech Emotion recognition is fundamentally pattern recognition, where a decided feature of speech signals is subjected to Machine learning algorithm for classifying pervasive emotion. In the paper Mel-Frequency Cepstral Co-efficient along with Energy, Pitch and Zero-crossing rate is used as features for classification of emotions and Deep Belief Network for classification algorithm. Experiments were aimed towards Developing a speaker independent emotion recognition system and obtained above average accuracy for six emotional states.

1. Introduction

Out of various modalities that are used in Human Computer Interaction (HCI) speech turns out to be the most natural and effective way. Speech is complex signal and contains information regarding message, speaker, language, emotions etc. Usually conversation has two parts verbal and non-verbal, verbal part conveys the message or what the speaker want to say but non-verbal part conveys the intent of the message [8]. Using speech in emotion detection helps in not only utilizing the message but also how the message in conveyed e.g. The word “OKAY” in English can be used to express admiration, disbelief, content, disinterest or an assertion etc.

Speech emotion recognition (SER) is pattern recognition where in which a decided feature or features are extracted, which is then classified using machine learning algorithm. Extraction of emotional state of the speaker from his/her speech is known as Speech Emotion Recognition [8]. The basic goals of a speech emotion recognition system are a) Understanding emotions present in speech. b) Synthesizing desired emotion in speech according to the intended message [8].

“Palette Theory” suggests that an emotion can be decomposed into primary emotions similar to a color which is a combination of two or more than two basic colors. The primary emotions are anger, fear, sadness, joy, disgust and surprise [1].

Definition of word ‘emotion’ is ambiguous and subjective in nature. To define ‘emotion’ is challenging task as it is uncertain in interpretation, which leads to first challenge in developing a SER system.

Building an SER system is a pattern recognition problem which contains five main modules speech input, feature extraction, feature selection, classification, and recognized emotional output [2] as shown in Figure 1.

![Figure 1. Speech Emotion Recognition System](image)

Constructing a suitable database or selecting an existing database is important task for SER system. The motivation behind developing a SER decides the method of collecting or choosing a database [8]. Feature extraction step extracts the parameters contained by the speech and change in these parameters results in corresponding change in emotions. Hence extracting these features helps in narrowing down the classification problem and help classifiers to label the emotions more accurately.

There are various application of Speech emotion recognition in our day to day life. It can improvise the naturalness in speech based human computer interaction. Accidents due to stressed mental state of driver can be avoided by alerting him/her while driving [2]. Analysis of call center conversation can be performed to study the behavior of call attendant with the customer to increase the productivity [4]. It
can be used in speech to speech translation of languages, source speech emotions are and synthesized into target speech [10]. Few challenges that lie in speech emotion recognition are constructing a suitable database or selecting an existing database is important task for SER system.

The motivation behind developing a SER decides the method of collecting or choosing a database [8]. Feature extraction step extracts the parameters contained by the speech and change in these parameters results in corresponding change in emotions. Hence extracting these features helps in narrowing down the classification problem and help classifiers to label the emotions more accurately.

2. Related Works

Different combination of features and classification methods are used in order to classify emotion the most used features are spectral and prosody. The spectral feature consists of MFCC (Mel Frequency cepstral coefficient) and prosody consists of pitch and energy. Most of the literature revolve around these two features Different classifier has been used for classifying the emotions by various researchers like Hidden Markov Model (HMM), Gaussian Mixture Model and Support Vector Machine (SVM) [10].

Different Neural Network model have also been like Convolution Neural Network (CNN) [15], Multilayer Perceptron (MLP), and Restricted Boltzmann Machine (RBM) [6] have been used recently for emotion recognition. Although a conclusion cannot be made which classifier is best, each classifier has its own advantage and disadvantage. Artificial neural network is used for emotion recognition because of its property of finding nonlinear boundaries separating the linear states. ANN reached the accuracy of 51.19% in speaker dependent recognition, and 52.87% for speaker independent [5][10].

MLP which is a class of neural network provided the accuracy rate of 68.10% for Leave One Text out (LOTO) scheme for testing and 51.65% for Leave One Speaker out (LOSO) scheme, both tests were conducted for speaker independent emotion recognition. MLP is a popular classifier in emotion recognition because it is easy to implement and well defined training algorithm once the architecture of ANN is defined. Similar result were obtained for deep learning method DBN and RBN with accuracy of 69.14% when tested for LOTO and 64.32% for LOSO scheme for speaker independent emotion recognition using the MFCC and prosody features [5].

CNN provided 79% accuracy for speaker independent emotion recognition system [15]. One of the most popular classifier in speech related research is Hidden Markov Model (HMM) because speech signal production mechanism is physically related to it. HMM provides good result in modeling temporal information in speech spectrum. The process of HMM is doubly stochastic and consist of first order markov chain which is hidden from the observer.

A random process is associated to each state which generates the observation sequence capturing the temporal structure of the data. HMM is trained for each emotion and an unknown sample is classified according to the model which illustrates derived feature sequence the best. HMM when used as classifier for emotion recognition provides accuracy of 76.12% for speaker dependent system using spectral features for speaker independent emotion Recognition HMM provided the accuracy of 64.77%.

The drawback of using HMM classifier is the feature selected should not only contain information about emotion but also fit the HMM structure as well. HMM classifier have lower recall rate when prosody and formant features are used than that of classifiers using spectral features [10] [1].

When global features are considered Gaussian Mixture Model tends to be more suitable for emotion recognition from speech signals. GMM is a probabilistic model for density estimation using convex combination of multi-variate normal densities. GMM can be considered as special case of continuous HMM with only one state. Due to low requirement for their training and testing and efficiency in modeling multimodal distributions, GMM are very efficient in emotion recognition when global features are extracted from training utterances. GMM provided maximum accuracy of 78.77% using the best features. Accuracy of 75% was achieved for speaker independent recognition and 89.12% for speaker dependent recognition using GMM [10].

3. Feature Extraction

Prominent features should be extracted for pattern recognition, intended information are represented by the selected features. In building SER system the speech signal is divided into small intervals called frames [3] and from each frames the features are extracted, such features are termed as local features.
3.1 Mel Frequency Cepstral Coefficients

MFCC is widely used speech feature in speech signal research. According to psychological studies, the perception of frequency content of speech signal does not follow linear scale in case of humans. Therefore, for each signal with frequency in Hertz a subjective frequency is measured on Mel scale [10] by the Equation 1.

\[ f_{mel} = 2595 \times \log\left(1 + \frac{f}{700}\right) \]  

The procedure for calculating MFCC is shown in Figure 2.

3.1.1 Pre-emphasis. Pre-emphasis is done in order to spectrally flatten the signal, this is done by taking Filter’s z transformation of the signal as shown in Equation 3.

\[ H(z) = 1 - \mu z^{-1} \]  

3.1.2 Frame Blocking. The signal after pre-emphasis is blocked into frames of N sample points with separation of M between each other, the value of M is lesser than M. The first sample point has N points and the second frame starts after Mth sample point.

3.1.3 Windowing. Each frame is multiplied by hamming window so the spectral distortion can be minimized it also minimizes the edge effect

3.1.4 Fast Fourier Transform (FFT). FFT is applied to the frames after blocking as it converts signals from frequency spectrum to data spectrum. Mathematically FFT is given by Equation 4.

\[ X_k = \sum_{n=0}^{N-1} x(n)e^{-j2\pi nk/N} \]  

3.1.5 Mel Scale FilterBank. After the signal is converted into frequency spectrum the frames are then filtered and power is calculated for each filter band. This step filters high and low frequency which is not processed by human ear.

3.1.6 Logarithm. Applying logarithm on output of MEL Scale Filter bank gives us MFCC.

3.1.7 Discrete Cosine Transformation. DCT converts the log power spectrum to time domain. After this MFCC is obtained as feature vector.

2.2 Fundamental Frequency

Fundamental Frequency (Fo) is related to pitch feature, it represents how high or low frequency of speech signal. It changes varying emotions anger is generally expressed with high fundamental frequency as compared to other using this relationship the emotions content in the signal can be classified. Fundamental frequency of a periodical signal is inverse of the pitch period length. The relationship between pitch variance in Hz and the emotion type is shown in Figure 3. The relationship between the glottal pulse and fundamental frequency is inversely proportional.

2.3 Energy

During expression of emotion there is variation in amplitude, it can be captured for classification by using Energy. It can be used to capture other features such as intensity, pitch and excitation mode.

While expressing “surprise” emotion the energy tends to be higher than other emotions and for sadness it is at low. Figure 4 shows relationship between energy and emotion.
Energy is required in order to classify emotion as it gives the intensity of the signal was spoken which is highly related to emotion.

![Figure 4. Relationship between Energy and Emotion](image)

### 3. Classification

The classifications in the experiments were performed using Multilayer Perceptron (MLP) and Deep Belief Network (DBN).

#### 3.1 Multilayer Perceptron (MLP)

MLP is used widely in field of emotion recognition. It contains set of process units which are stacked in layer. Units in the same layer are not connected but there is complete connection between nodes in different layers. Output from one layer is fed as the input in subsequent layers and the last layer is represented by Equation 5.

$$y = F\left( \sum_{i=1}^{n} w_i x_i + \theta \right) \quad [5]$$

In Equation 5 weight is represented by \( w \) and \( x \) represent the input feature vector, \( \theta \) is the bias. The output generated is weighted sum of input plus Bias term.

![Figure 5. Multilayer Perceptron](image)

#### 3.2 Deep Belief Network (DBN)

DBN consist of stacked RBM. The architecture is shown in Figure 6. Layers are trained in bottoms up approach with greedy layer wise learning. To perform classification the DBN can be trained and the classification algorithm can be used for classification. Secondly one more layer of output can be added and back propagation can be applied.

![Figure 6. Deep Belief Network](image)

### 4. Database

Developing or selecting an existing database is prerequisite for building emotion recognition or synthesis system. Selecting and developing a database is highly dependent on the purpose of the emotion recognition system one wants to develop. The speech corpora can be divided into three categories a) Actor (Simulated) b) Elicited (Induced) c) Natural emotional speech databases [8].

Database of Polish Emotional Speech comprises 240 recordings from 8 actors (4 females and 4 males). Recordings for every speaker were made during a single session. Each speaker utter five different sentences with six types of emotional load: joy, boredom, fear, anger, sadness and neutral with sampling rate 16 kHz a single channel audio tape recorder and 16 bit quantization.

### 5. Implementation Details

The experiments were performed on Polish Emotional database. Essentia speech library [18] was used in order to extract emotions. Fundamental frequency, Energy and MFCC was extracted and their first derivative was used for the classification purpose. 13 MFCC was extracted from the signal leaving the first and last 3 seconds and only of the middle part signal.

Python Theano framework for deep learning was used. The experiments were aimed to test speaker independent emotion recognition and all the files
were used for training the model leaving one speaker for testing, selection for speaker for testing was used randomly till all the speakers were once tested for recognition.

One hidden layer for MLP was used with four neurons applied using theano. The DBN was trained with Batch size = 50, Hidden units = 120 and 5 Layers. The classification was performed by adding one layer of RBM and back propagating the output. 500 epochs were used in DBN. The model was tested once the training attained maximum accuracy.

6. Result

Overall accuracy of the system for DBN was 58% in speaker independent model. Model using MLP gave 53.6% accuracy for the same database i.e. Polish Database. The final result was calculated by taking the mean of accuracy obtained for each speaker. Speaker dependent emotion recognition was also performed where a single text was used for testing the model and training the system on other text the accuracy for DBN reached 61.5% and for MLP accuracy was 59.16%. Table 1 summarizes the results of the experiments. Significant difference can be seen in Speaker dependent model which suggests DBN outperforms MLP in SER.

<table>
<thead>
<tr>
<th>Classifier</th>
<th>Speaker Independent</th>
<th>Speaker Dependent</th>
</tr>
</thead>
<tbody>
<tr>
<td>Multilayer Perceptron</td>
<td>59.16</td>
<td>53.6</td>
</tr>
<tr>
<td>Deep-Belief Network</td>
<td>61.5</td>
<td>58.6</td>
</tr>
</tbody>
</table>

7. Conclusion And Future Scope

In experiment the result suggests that the Deep Classifiers perform better than MLP when compared over same database and similar feature set. Different classifiers can be used once the DBN ends the learning phase. Different classifiers can be used in order to increase the accuracy of the system.

In future the work can be expanded to multiple speaker environments where conversation among different speakers is going on. The accuracy can also be increased in future by increasing the feature vector size and studying different features of speech signal that can capture emotional aspects more efficiently.

8. References


