Audio based Emotion Detection and Recognizing Tool Using Mel Frequency based Cepstral Coefficient

Naveenkumar M1, Vishnu Kumar Kaliappan2

1,2Department of Computer Science and Engineering, KPR Institute of Engineering and Technology, Coimbatore, India
Naveenkumar2may94@gmail.com1, vishnudms@gmail.com2

Abstract—Human computer interaction (HCI) been more predominantly grown Speech Emotion appreciation in recent decade. Any interaction of mutual interest and benefits like service sectors including business and classrooms in turn could utilize this technology for ground analysis i.e., track the trace of emotional state of the receiver and so to adopt the appropriate strategy that helps the purpose of interaction. Interface with Robots upon analysis can provide details on the emotional state of the receiver of the service at the other end and paves way for application platforms such as Computer games, Customer relationship maintenance and management, e-learning, Banking and Classroom orchestration can be taken to the next level of sophisticated technology. The Work comprises of Emotional attribute extraction and Comparative Machine Classification. The comparative machine classification been done with Support vector machine (SVM). For final comparison purposes Mel-frequency Cepstrum Coefficients (MFCC) and with modulation spectral features (MSFs) are been explored and used whereas the final combinational comparisons are done between different databases for the specific features. The general experimental consequences display that the most amplitude of a voice sign varies from great sounds. For the sign it is taken on this paper the maximum amplitude is 5.3V.

Keywords—Voice Recognition, Emotion Recognition and Artificial Neural Networks.

1. INTRODUCTION

In human machine interaction techniques, Speech of the human had been the rapid and efficient characteristic to be taken for research. The Design of appropriate intelligence system for human voice recognition is the challenge over here. Out of plenty of researches over this discipline, the hindrance for natural interaction between human and machines is they aren't integrated with speech emotion analysis, thus a wide avenue for research platform. Initiated with technology for emotional extraction from the speaker's voice, it goes with recognizing linguistics of the speech to improve the performance of SER system. Services like PC tutorials and net movies which has to respond upon detected emotions of speaker would be enhanced with performance. For the subsequent reasons such recognition is a typical difficult one. The dominant emotion in complex feelings and how to distinguish between them is the challenging task. Pitch and energy contours construction from acoustic variability of speaker, design and rate of speaking could been done. Happiness, sadness, anger, fear, surprise, disgust are the six emotions the classifier focuses and analyze upon. Various aspects such as skin elasticity, heart rate and tone of voice, muscular tension and other objectives of the speaker are been noted and analyzed. Physical reflection of emotions is accessible externally but understanding the reflection of emotions is partially culture specific. Upon the results of art survey, speech emotion detection usage been considered as an appropriate technology in the applications intended.

Discourse is the critical form of communication among human beings. In arrange for the communication to require put, a speaker must produce a discourse flag within the shape of a sound weight wave, which voyages from the mouth of the speaker to the ears of the audience. The pathway of communication from speaker to listener begins by a thought that's made within the intellect of the
speaker. This thought is changed into words and sentences of a dialect. The brain of the audience at that point performs discourse acknowledgment and understanding. This movement between the speaker and the audience can be thought of as the transmitter, separately, within the discourse communication pathway. But there exist other functionalities other than fundamental communication. Discourse feeling acknowledgment is one of the most recent challenges in discourse preparing. Aside from human facial terminologies discourse has validated as one of the main principal promising modalities for the programmed acknowledgment of human state of mind. In particular within the area of security frameworks a developing intrigued may be watched all through the final 12 months.

This paper gives an overview about the disclose of a comprehensive review of voice emotion recognition tool targeting the voice feeling recognition scientists who do not originally have a deep experience in voice analysis. Segment 1 is devoted to the specificity of emotion revealing: the clinical historical past, of a brief assessment. The country of the art effects within the field of literature survey on segment 2. The numerous classification methods utilized in our proposed version are delivered in segment 3 and making preference is a section 4. Results and discussion of the paper is defined in segment 5 and the realization follows on segment6.

2. LITERATURE SURVEY

Md. Taufeeq Uddin, et.al[1] proposal comprises of Wrapper Random Forest-based feature selection method development. A Gabor filter is been used on speech emotion recognition performance study. These features extracts voice parameters and apply feature selection on wide feature set. Random forest algorithm is applied and importance score for the voice sequence features are generated. Using a multi-class SVM in CK database, chosen features are trained. The result reveals the expressions like sadness, neutral and anger which can be increased by method of feature selection.

Multimodal emotion popularity is an affective computing that is discussed by Cristian Torres, et.al [2] which simultaneously uses distinct physiological indicators and appears for comparing an emotional kingdom. Physiological indicators together with electroencephalogram, heat and electrocardiogram, had been used to evaluate feelings like happiness, disappointment or anger, or to assess degrees of valence. The main goal is to learning the discriminate strength of different capabilities related to numerous physiological indicators used for multimodal emotion popularity. Making use of recursive characteristic elimination and margin-maximizing function removal over databases, specifically, MAHNOB-HCI and DEAP. The outcomes display that electroencephalogram related functions illustrates the uppermost discrimination capacity. For the valence catalog, electroencephalogram functions are followed by way of galvanic pores and skin response functions profits the best discrimination electricity.

Face based and audio-primarily based emotion identification and popularity modalities were mentioned by Juan M. Mayor Torres, et.al [3] obtaining a success category price for valence tiers and a couple of emotion training settings. Face-primarily based sentiment recognition structures use a single picture channel illustrations including primary- additives-evaluation bleaching, isotropic smooth out. Speech emotion reputation tool use a uniform set of audio descriptors, together with first-class averaged Mel Occurrence Cepstrum factors. The strategies mean the inclusion of choice- fusion modalities to compensate the restricted function reparability and obtain excessive type prices. The face- primarily based and audio-based emotion popularity pipelines are evaluated the usage of a 5-Fold go-validation. For more efficiency in phrases of the category, the K-Nearest acquaintances shows an accelerated performance no matter the sorts of enter representation. In phrases of vocal sound expressions, the inclusion of a couple of standardized datasets yields an extra correct pipeline than the use of lined classifiers and tiny length datasets for figuring out valence stages.

Mingli song et.al[4] briefly worked on Audio/Video(AV) based recognition of emotion using Tripled Hidden Markov Model (THMM) that used both audio and video clips from the input video clips. THMM in turn preserves the natural correlation of audio and video over time and allows the
temporal relation. Since audio lags from visuals, it is inaccurate on assuming static synchronization. Upon the reliability of each and every modality THMM combines them. Basic emotional training is needed to be done with THMM initially. From the relative reliability of different levels of the acoustic noise, the influence of audio and video been weighted at recognition state. Here unlike HMM, THMM permits asynchronous by allowing the lag between audio and video and could be applied to various platforms of Human machine system.

Surabhi vaishnav, et al[5] novelly used Cauchy Bayers Classifier to recognize speech emotion and displayed them as video sequences. Their work system is a five compartmentalized model that includes input speech receiving database compartment, feature extraction block, selection compartment, classifier block and emotion recognition compartment. The problem is calculation of chi square for the extracted signals. Whether Gaussian or Cauchy assumption is preferable for the computation of recognition is a hindrance. After taking some of the extracted features of the input, appropriate recognition pattern model been produced hence that gives the state of emotion of the speaker. The speech acoustic front end through spectro temporal analysis and generate Raie material as short speech power spectrum in intervals is the first stage among the three stages of feature extraction. Static and dynamic feature vector composition is the second stage. Third stage reorganizes produced vectors as compact and robust, readily can be supplied into recognizer. From the selection results relevance emotions can be expressed. Aim of this work is to compare Cauchy and Gaussian assumptions of pattern recognition out of which Cauchy are better.

Ashish B, et.al[6] discussed recognition of emotion from text by implementing knowledge based ANN. Most traditional keyword-based method is used to detect and analyse the state of emotion of the input. KBANN uses KBANN algorithm to detect the emotion through the keywords in the text and during the absence of such keywords it is provided with implicit knowledge. Lack of training with the linguistic knowledge makes the empirical keyword system face problem with classification. It is with low accuracy in contrast with other methods.

| Type   | Rate          | Pitch Range | Voice Quality | Intensity |
|--------|---------------|-------------|---------------|-----------|
| Anger  | Slightly Fast | Wide        | Breathy       | Tense     |
| Happy  | Fast or Slow  | Wide        | Breathy       | Normal    |
| Sad    | Slightly Slow | Slightly Narrow | Resonant       | Slurring  |
| Fear   | More Fast     | Wide        | Irregular Voice | Precise   |
| Disgust| Very much Fast| Slightly Wide | Grumble chest tone | Normal   |

**Table 1: Various Parameters of Voice Signal**

![Plutchik's wheel simplified.]

The contribution refer to testing with popularity of feelings in German speech sign based totally mostly
on the identical principle as Gaussian Mixture Models of speakers by Martin Vondra, et.al [7]. The maximum strong set of rules for Voice reputation is based totally on Gaussian aggregate models. The authors observe three parameter sets: the first incorporates suprasegmental capabilities, within the second are segmental functions and the final is a mixture of the 2 previous parameter sets. The aim of this contribution is number of Gaussian Mixture Model additives and the best suitable choice of speech constraints for emotion reputation in German speech.

Table 2: Various Algorithms with the Voice Accuracy and Error Detection Rate

| S. No | MODEL                                 | VOICE ACCURACY (%) | ERROR DETECTION RATE (%) |
|-------|---------------------------------------|--------------------|--------------------------|
| 1     | Naive Bayes[5]                         | 0.56               | 45.1                     |
|       |                                       | 88                 | 5                        |
| 2     | Random Forest[1]                       | 0.59               | 47.5                     |
|       |                                       | 82                 | 6                        |
| 3     | Hidden Markov Model[4]                 | 0.60               | 48.8                     |
|       |                                       | 13                 | 9                        |
| 4     | Artificial Neural Network[6]           | 0.62               | 49.7                     |
|       |                                       | 47                 | 8                        |
| 5     | Multilayer Perceptron                  | 0.62               | 51.5                     |
|       |                                       | 98                 | 1                        |
| 6     | Restricted Boltzmann Machine           | 0.63               | 55.4                     |
|       |                                       | 76                 | 9                        |
| 7     | Gaussian Mixture Model[7]              | 0.64               | 59.8                     |
|       |                                       | 50                 | 6                        |
| 8     | Support Vector Machine[2]              | 0.64               | 63.0                     |
|       |                                       | 67                 | 1                        |

3. PROPOSED SERT METHOD

Recurrent Neural networks consists of enter devices, output devices and hidden gadgets, the most critical work is done with the aid of hidden unit. The model basically has a one way glide of facts from the preceding temporal concealment unit to the cutting edge timing hiding unit is shown in figure 2.
A. Signal Pre-Processing

The speech signal is extensively utilized in emotion detection tests. The step worried in SERT is proven in Figure 3. In recent works of literature, all are used signal dataset for processing and detecting the emotions from human’s speech. It also makes the wide variety of information affordable with in the schooling set and trying out set in the sort of manner that the classifier may classify more records. In this paper for speech signal data set the English language database are created and the voices are collected. The INTER1SP English emotional database includes utterances from professional actors (one woman and one man speakers). Paper has centered on as it were six fundamental feelings from the English dataset in arrange to attain the next and greater genuine charge of acknowledgment.

B. Tactic’s

The Speech Emotion Reputation Tool (SERT) framework is primarily based on two functionalities. They are noise reduction in the voice signal and emotion recognition in speech. To enhance the
accuracy of the voice signal the noise is eliminated through the Modulation Spectral approach. The speech signal is sampled using FFT. Based on the Mel Coefficients which are extracted from the MFCC, the speech emotions are recognized and classified and the noise from the signal and improving the quality of the speech signal predicting the accuracy.

Fig. 4: SERT Framework

The RNN organizes the entries in 6 instructions similar to the subsequent feelings: happiness, fear, disappointment, disgust, angeriness and marvel. Totally different mixtures of feeling options provide different emotion detection rate and the accuracy of the voice using the English database. The proposed architecture of the SERT is represented in figure 4.

1. Function Extraction the usage of MFCC

Auditory distinctive of the speech sign is characteristic. A tiny a part of statistics from the speech sign is mined to research the sign without disconcerting its acoustic houses. This extracted signal is used for schooling and trying out. It consists of the subsequent steps:

   i. Frame Stage Blockading

   Processing of speaking signals is achieved in lesser time periods are referred to as frames with length typically between 20 and 40 milliseconds [7]. Intersecting of mount is ready to smoothen the conversions among mounts by way of a predefined length. If the trial price is 12 kHz and the mount size is 258 trial factors, then the body duration is 258/12000 =21ms. The next mount initiates at “M” = 100 trials subsequently the first mount, and intersections it through N=M trials and so forth.

   ii. Windowing

   The speech sign is handled through a windowing feature to lessen breaks and spectral distortion at the limits of every mount. Each mount must be improved with a Hamming window to hold the continuity of the first and the closing points in the mount. The mount body is denoted through the use of
\[ Y(N) = [X(N) \ast W(N), \quad 0 \leq N \leq N - 1] \quad (I) \]

wherein \( W(N) \) is the window characteristic, \( X(N) \) as input mount and \( N \) is the wide variety of samples in each body of the frame.

iii. **Rapid Fourier Rework**

Rapid Fourier Rework is achieved on the windowing sign to transform it into the frequency region. The Unconnected Fourier Rework over a discrete sign \( x(N) \) of \( N \) trials, translates each mount body of \( N \) samples into the frequency region from its time area.

iv. **Changing the Mel Occurrence**

The Mel occurrence measure is linearly spaced for frequencies underneath one kHz and logarithmically space out more than one kHz. A pitch of one kHz tendency, which is comparable to one thousand Mels whilst it's far above the perceptual listening to aspect level by forty dB.

v. **Mel Occurrence Cepstrum factors**

Coefficients that represent audio primarily based on notion with their frequency bands logarithmically located and simulators the human talking reaction.

**Algorithm:** MFCC Feature Extraction Algorithm

**Input:** signal (Phonocardiogram signal)

**Output:** MFCC (Phonocardiogram signal)

**Function MFCC**

a. Initialize parameters;

b. Split into frames Phonocardiogram signal;

c. Apply Hamming window to frames;

d. Get spectrum by applying FFT to all frames;

e. Determine matrix for a mel spaced filterbank;

f. Transform spectrum to mel spectrum;

g. Obtain MFCC vector for each frame by Applying DCT

End Function

4. MAKING CHOICE

A. **Imply of MFCC**

Suggest for a statistics set is named as arithmetic imply. On this paper imply MFCC is taken to reduce the massive values which can be obtained from MFCC. When \( X = \{X_1, X_2, X_3, ..., X_N\} \) denotes MFCC values and general number of MFCC is “\( N \)” then the math suggest is taken as,

\[ X = \frac{X_1 + X_2 + X_3 + \cdots + X_N}{n} \quad (2) \]
B. Widespread Deviation

The usual deviation (σ) is a degree of the difference of a set of facts from their suggest. A small preferred deviation suggests that the facts to be very close to their imply and a high preferred deviation suggests that the records are prolonged out from their imply and among themselves. On this paper well known deviation of fifty audio system for unique feelings namely sadness, happiness, angriness, disgust, boredom and fear these are experimentally computed and range of trendy deviations for those emotions are optimized as follows: Sadness: 55.37 to 45.74; Happiness: 66.39 to 427.58; Angrienss: 55.97 to 434.36; Boredom: 43.29 to 328.51; Disgust: 64.97 to 359.07; Fear: 70.02 to 431.08.

5. RESULTS AND DISCUSSIONS

The experiments were designed to examine the overall performance and accuracy of the SERT framework to obtain the emotions clearly by using classifier. The analysis is performed against various feature extraction scheme. MATLAB tool is used as a testing platform for analysis. The algorithm is compared with other machine learning methods like, MLR, SVM and others and it is showing better results. The obtained results are given below.

![Fig. 5: GUI of SERTS Framework.](image)

Fig. 5: GUI of SERTS Framework.

The feature of SERT is to fetch the speech signal from the real time data saved inside the database. This speech information saved in a (.wav) record and given to the SERT model. The signal is divided into n number of small speech files and based on the Mel file the speech signal differs. The grid presentations, the mean and well known deviation of the speech signal a minimum and maximum value of the sign is displayed in figure 5.

![Fig. 6: Mel Filter bank window](image)

Fig. 6: Mel Filter bank window
In Figure 8 the graph shows that the amplitude is maximum at 170 KHz. The value of the amplitude is 5.5V in normalized one and 15dB in figure 9.
Fig. 9: Waveform of Magnitude in Frequency domain.

Fig. 10: Power and Logarithmic Spectrum

Original sign filtered with the aid of the hamming filter out and converted with the FFT. M3 elements are complex numbers and symmetrical due to the fact the FFT because used to transform the records. We take the absolute values (M=100, N=256) of the matrix elements to plot the power spectrum. Since the spectrum is symmetrical, we only plot half of it. The areas with the highest energy level are shown red in figure 9 and the red area is between 0.3 and 0.7 seconds. The output also shows that the majority of the power is intense in the lower frequencies (between 50 Hz and 1 kHz). The Play button is used to play the voice and the mean; SD values are displayed with the respective min, max values. When the classify button is pressed, the emotion is classified and the type of emotion is recognized and displayed in the SERT GUI in figure 10.
6. CONCLUSION

Within the present day state of affairs, increasingly more people are getting addicted to clever gadgets and lots of already have the tendency to engage with clever devices. The identical manner they do with exceptional humans. From the above study and comparison tables, it is clear that the accuracy and error detection rate increases with the combined modalities. By the mean of FFT and the mean values of MFCC in SERT framework, the implementation of standard deviation values for different emotions have been achieved. The output of the SERT was refined by means of comparing the improved values of standard deviation with the individual emotions. The SERT framework has produced better performance compared to existing methods. The Emotions are classified clearly and its accuracies are improved in the SERT framework.

7. REFERENCE

[1]. Md. Taufeeq Uddin, Md. Azher Uddin, "Random Forest based Feature Selection Approach for Recognition", DOI:10.1109/ICEEICT.2015.7307376

[2]. Cristian Torres-Valencia, Mauricio Alvarez-Lopez, Alvaro Orozco-Gutierrez, “SVM-based feature selection methods for emotion recognition from multimodal data”, Springer JMUI, 2017.

[3]. Juan M. Mayor Torres, Evgeny A. Stepanov, "Enhanced Face/Audio Emotion Recognition: Video and Instance Level Classification using ConvNets and Restricted Boltzmann Machines", IEEE 978-1-4503-4951-2, 2017.

[4]. Mingli Song, Chun Chen, Mingyu You, “Audio-visual based emotion recognition using tripled hidden markov model”, p-ISSN: 7803-8484-9, 2004.

[5]. NicuSebe, Michael S. Lew, Ira Cohen, AshutoshGarg, Thomas S. Huang, "Emotion Recognition Using a Cauchy Naive Bayes Classifier", p-ISSN: 1051-465U0, 2002

[6]. Yong-SooSeol, Dong-Joo Kim, Han-Woo Kim, “Emotion Recognition from text Using Knowledge-based ANN”, ITC-CSCC 2008.

[7]. Martin Vondra, Robert Vich, “Recognition of Emotions in German Speech Using Gaussian Mixture Models”, Springer LNCS, 2003.

[8]. Z ZAIN, “High Speed And Lowpower Gdi Based Full Adder”, Journal of VLSI Circuits And
[9]. Pb Agus Ristono*, "Design Of Reliable And Efficient Manchester Carry Chain Adder Based 8-Bit Alu For High Speed Applications", Journal Of VLSI Circuits And Systems, 1 (01), 1-4, 2019