JETIR.ORG ISSN: 2349-5162 | ESTD Year : 2014 | Monthly Issue
JOURNAL OF EMERGING TECHNOLOGIES AND INNOVATIVE RESEARCH (JETIR)
An International Scholarly Open Access, Peer-reviewed, Refereed Journal

Speech Emotion Recognition: Investigation survey

Mohamad Emad Bitar

Ph.D. Scholar, Department of Computer Science, CMS College of Science & Commerce
Coimbatore, India - 641049
t22h.12345@gmail.com

Dr. V. Sujatha

Vice Principal, CMS College of Science and Commerce,
Coimbatore, India – 641049
Sujathapadmakumar4@gmail.com

-

Abstract - This paper's main goal is to provide a speech emotion recognition framework. Many aspects of the speech can be extracted to identify emotions. The removal of sentiments from waves in SERs has been accomplished using a variety of tried-and-true discourse evaluation and classification techniques. In this study, a number of challenges connected to recent work on speech-based emotion identification are described. The selection of a database, identifying numerous speech-related variables, and making an appropriate classification model choice are the main problems in emotion recognition. We have conducted literature research on the many traits that are utilized to identify emotions in human speech. The importance of several classification models has been highlighted in addition to some recent study work reviews.

***Key Words*: Speech Emotion Recognition; SVM; Classification; CNN; Mel frequency cepstral coefficients**

I. INTRODUCTION

The human vocal system has several characteristics that can convey context and information, including speech, tone, and pitch [1]. SER is frequently utilized to meet natural human-computer communication requirements. Speech emotion detection is known to take the emotional tone out of a speaker's speech. Speech recognition systems are designed to be utilized with this type of recognition to derive valuable semantics.

Discrete speech emotion models and continuous speech emotion models are both included in the SER model [2]. The second model indicates that the emotion is in the emotion space, and every emotion has a unique strength on each percentage. The first model conveys numerous individualistic emotions, indicating that a particular voice has a specific individualistic emotion. Any model for recognizing feelings can use the distinctive job of determining a person's emotional state as a level.

It makes use of a range of emotions, including neutral, disgust, wrath, fear, surprise, joy, happiness, and sadness.

The three processes of pre-processing, feature extraction, and feature classification make up the bulk of the SER technique [3].

• Pre-processing: Pre-processing refers to all the transformations that raw data goes through before being put into the machine. It involves the removal of pre-emphasis, normalisation, and windowing, making it a necessary step to obtain the pure signal utilized in the following stage, namely feature extraction. In order to expedite training, it is also crucial.

• Feature Extraction: A little amount of information from the speech signal is removed for analysis without altering the speech's characteristics [4]. The feature parameters known as Mel frequency cepstral coefficients are commonly utilized in voice recognition [5]. MFCC employs the Mel scale [6] based on the hearing organ's response. In this study, various features are retrieved from the speech signals, and analysis are then done on those features [7]. Max-pooling, an activation function, and many convolutional layers are required for feature extraction. The accuracy of a device's ability to recognize speech emotions is improved by using the various feature extraction methods. The focus of the study is on the pre-processing of the acquired audio samples, where noise is removed using filters from speech samples.

• Feature Classifier: There are two primary categories of classifiers for Speech Emotion Recognition, namely non-linear classifiers and linear classifiers [8].
To build the best classifier to represent emotional states, a variety of classification techniques are applied such as K-Nearest Neighbor, Gaussian Mixture Models (GMM), Support Vector Machine (SVM), Hidden Markov Models (HMM), and Neural Network.

II. LITERATURE SURVEY

In [1], Girija Deshmukh et al. proposed a technique for measuring the Short-Term Energy (STE), Pitch, and MFCC coefficients of the emotions of annoyance, happiness, and melancholy using audio samples. Natural speech was recorded using open source North American English as expressiveness and feedback. As a result, only anger, happiness, and sorrow were identified as emotions. In-depth characteristics of the speaker, such as sound, energy, and pitch, were also recognized. Train and test sets are manually created from the entire Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS) dataset. Using feature vectors as input, the multi-class Support vector machine (SVM) generates a model for each emotion.

In [2], Peng Shi developed the discrete model and continuous model of speech emotion recognition; several traits are examined to produce better emotional descriptions. The Deep Belief Network (DBN) is superior than Artificial Neural Networks (ANNs) and Support Vector Machines (SVMs).

Networks (DBNs) have an accuracy rate that is around 5% greater than the conventional techniques. The results demonstrate how much better the features recovered using Deep Belief Networks are than the original feature. Because SVM classifies at small sizes better than DBN-DNN, DBN-SVM had slightly better results than DBN-DNN. Better categorization results from DBN's conversion of superficial qualities into deep abstract characteristics.

Gray Level Co-occurrence Matrix (GLCM) and Mel Frequency Cepstral Coefficient were used to describe the feature extraction in J. Umamaheswari et al[3] .'s presentation of pre-processing utilizing K-Nearest Neighbour (KNN) and Pattern Recognition Neural Network (PRNN) algorithms (MFCC). Standard algorithms like Hidden Markov Models (HMMs) and Gaussian mixture models (GMMs) were identified as a superior production than the benchmark algorithms when the results were evaluated for their precision rate, accuracy, and f-Measure. The pattern that is created by the emotional waves is afterwards detected by PRNN. K-NN method is used to find the signal's plausible nearest pattern. The Speech database has the following basic classes, for example:

- Neutral
- Angry
- Sad
- Happy

According to M.S. Likitha et al. in [4], observed recognition necessitates a verbal communication wave assessment in order to categorise the necessary feeling based on training of its properties, such as sound, format, and phoneme. A large variety of speech signal-based algorithms were created on the side of functionality withdrawal and examination. The communication kinesics' acoustic accuracy is a feature. The practise of eliminating a little amount of information from the audio signal used to reflect each speaker later on is known as feature withdrawal. Most extraction techniques are readily available, but coefficient extraction is the most popular (MFCC). The speech signal's audio feature is the sound.

Feature extraction is the process of extracting a little amount of information from vocal expression that can then be utilized to act for each speaker.

According to Zhang Lin et al. in [5], SER technology is utilized to track the driver's erratic emotions and employs specific phrase recognition technology to select parking instructions in emergency scenarios. Prosodic, spectral-based, and quality features are the three categories of speech features that are extracted. The Linear Predictive Cepstral Model and the Mel-cepstral coefficient (MFCC) are often utilized characteristic characteristics in voice recognition (LPCC). The characteristics are extracted using SVM. This device can identify an emergency situation based on the driver's voice's anxiety or terror. On that premise, the parking instructions are listed.

Since the speech emotion database and voice parking guidance database utilized in this study were compiled in a variety of contexts, the recognition efficiency is dramatically reduced as soon as the parking guidance voice's emotion is acknowledged.

According to Asaf Varol et al. in [6, sound is described as a pressure wave that results from a substance's vibration within a molecule. The inquiry has looked into the properties of sound energy. Experiments combining the voice signal spectrogram and Artificial Neural Networks produce more useful findings (ANNs). Role extraction methods including acoustic analysis and spectrogram analysis are used in SER on the EMO-DB dataset. The expanding application of SERs in areas including psychiatry, pattern recognition, and signal processing has also been covered in these current trends. Additionally, the author claims that multiple machine learning techniques should be used on various kinds of datasets with varied kinds of tests in order to produce higher success rates.

In [7], Abhijit Mohanta et al. analysed emotional speech signals to identify emotions including anger, fear, happiness, and neutrality using parameters like loudness, voice detection, and excitation energy. These traits are referred to as sub-segmental features. The study of these emotions has been done utilizing features such instantaneous fundamental frequency (F0) using Zero Frequency Filtering (ZFF), signal energy, formant frequencies, and dominating frequencies. Instead than classifying the actor's emotional states, the study examines the features of four separate emotion states' generation. Some signal processing techniques, such as ZFF and STE, were utilized to find instantaneous F0 and the zero-crossing rate (ZCR).

Speech emotion recognition was viewed as an interesting component of human computer interaction by Edward Jones et al. in [8]. (HCI). Feature extraction and feature classification must be the primary SER strategies. For feature classification, both linear and non-linear classifiers can be utilized. Support Vector Machines (SVMs), Bayesian Networks, and other classifiers are widely employed in linear classifiers (BN). These kinds of classifiers are useful for SER since speech signals are thought to vary. Deep learning approaches have more benefits for SER than conventional approaches. Deep learning approaches have the capacity to recognize complicated internal structure and do not require manual feature extraction or tweaking.

In [9], Michael Neumann et al. presented their findings as an example of how speech emotion recognition might benefit from learning about unlabelled voice entities (SER). To analyse visuals of various representations, they used t-distributed neighbour embeddings (t-SNE). However, the 2D projections do not reveal any divisible clusters. Due to the bandwidth these plots require, they are excluded. A sizable dataset is used to train the autoencoder. They have added representations produced by autoencoders, which has resulted in consistent improvements in the SER model's identification accuracy. The study of procreative adversarial networks for representation learning and numerous alternatives to autoencoders are also made possible by the research.

The German Corpus (Berlin Database of Emotional Speech) data, which included more than 250 recordings, was used by Radim Burget et al. in [10]. To prevent overlapping, each recording was split into 20 millisecond chunks. The data was then divided into training, validation, and testing sets by removing 3098 quiet parts. During pre-processing, Google WebRTC vocal workout detector was utilized to eliminate the silent audio segments. Following that, all of the files are normalised to have a mean and module variance of 0. Deep Neural Networks (DNN) were given input data in batches with each iteration lasting 21 iterations. Every batch has a comparable amount of divisions, and the pattern is followed by patterns for various divisions, including neutral, furious, sad, and so on. DNN had no understanding of the nature or the genuine experience that the performer is trying to express.

Table 2.1 Literature survey

S.No	Title	Summary
1	Speech based Emotion Recognition Using Machine Learning[1]	Was only limited to classify three emotions i.e., Angry, Sad and Happy
2	Speech Emotion Recognition Based on Deep Belief Network[2]	Emotions such as shame and surprise cannot be identified, affecting the overall rate of recognition and average rate of recognition
3	An Enhanced Human Speech Emotion Recognition Using Hybrid of PRNN and KNN[3]	4 universal feelings are classified like neutral, sadness, happiness, angry, over the given input.
4	Speech based human emotion recognition using MFCC[4]	Only one feature extraction is used i.e., MFCC. Due to which, only three emotions are confirmed irrespective of the gender
5	Research on Emergency Parking Instruction Recognition Based on Speech Recognition and Speech Emotion Recognition[5]	If the system recognizes the feeling of the parking education expression, the efficiency is degraded due to that fact the voice stopping guidance database utilized in this paper are recorded totally extraordinary surroundings.
6	New Trends in Speech Emotion Recognition[6]	It obtained less accuracy. Changing dataset can prove as a solution to this problem.
7	Emotion recognition from speech signal[7]	It just analyses characteristics of various four emotion states, and does not include the classification of emotion states.
8	Speech Emotion Recognition Using Deep Learning Techniques: A Review[8]	The research work is done for different DNN techniques but no such implementation are done using this technique. It was just a theoretical concept.
9	Improving Speech Emotion Recognition with Unsupervised Representation Learning on Unlabeled Speech[9]	Representation became similar to factors. Henceforth, in 2D projections, separate clusters were not found which were bound to the space limitations
10	Speech emotion recognition with deep learning[10]	DNN had no understanding of the real sense of what the actor is try to saying it. Neither it had any understanding of the speech, vibrations etc.

III. CONCLUSIONS

Speech emotion recognition (SER) is a field that analyses speech to identify human emotions. An accurate and precise database with audible actors and minimal background noise is ideal. In order to extract audio features from speech samples using SER techniques, several classifier algorithms are employed to identify the emotion.

Numerous features are utilized to identify emotions, however feature extraction with MFCC appeared to be crucial in identifying emotions in speech. While selecting the appropriate and best classifier is a crucial stage in SER. The Database, the characteristics retrieved from the Databases, and the classification model (algorithm) used to classify the Emotions all have a role in how accurate the Speech Emotion Recognition System is.

REFERENCES

[1] Sukanya, Girija Deshmukh, Apurva Gaonkar, and Using machine learning to recognize speech-based emotions, Kulkarni, Institute of Electrical and Electronics Engineers, Mar. 2019.
[2] Peng Shi, Institute of Electrical and Electronics Engineers, "Speech Emotion Recognition Based on Deep Belief Network," March 2018.
[3] A. Akila and J. Umamaheswari, "An Enhanced Human Speech Emotion Recognition Using Hybrid of PRNN and KNN," Institute of Electrical and Electronics Engineers, February 2019.
[4] "Speech Based Human Emotion Recognition Using MFCC", M.S. Likitha, A. Upendra Raju, and K. Hasitha 2017 March,

"MFCC", Institute of Electrical and Electronics Engineers.

[5] "Research on Emergency Parking Instruction Recognition Based on Speech Recognition and Speech Emotion Recognition", Institute of Electrical and Electronics Engineers, November 2019. Tian Kexin, Huang Yongming, Zhang Guobao, and Zhang Lin

[6] Institute of Electrical and Electronics Engineers, "New Trends in Speech Emotion Recognition," Ye Sim Ülgen Sonmez and Asaf Varol, June 2019.

[7] Esther Ramdinmawii, Abhijit Mohanta, and Vinay K. Mittal, "Emotion recognition from speech signal," Institute of Electrical and Electronics Engineers, November 2017.

[8] "Speech Emotion Recognition Using Deep Learning Techniques: A Review", Institute of Electrical and Electronics Engineers, August 2019 Ruhul Amin Khalil, Edward Jones, Mohammad Inayatullah Babar, Tariqullah Jan, Mohammad Haseeb Zafar, and Thamer Alhussain

[9] Ngoc Thang Vu and Michael Neumann published "Improving Speech Emotion Recognition with Unsupervised Representation Learning on Unlabeled Speech" in the May 2019 issue of the Institute of Electrical and Electronics Engineers.

[10] "Speech emotion recognition with deep learning," PavolHarár, RadimBurget, and Malay Kishore Dutta, Institute of Electrical and Electronics Engineers, February 2017.

AUTHORS PROFILE

Mohamad Emad Bitar, Ph.D. Scholar at CMS College of Science and Commerce, graduated with a master's degree of Compute Science and bachelor's degree in Information Technology. As a researcher, the areas of interest are speech detection, sign language detection.

Dr. V.Sujatha, has 19 years of teaching experience and 2 years of IT Industrial experience. Her area of specialization is web mining, Big Data Analysis. She has Published 24 research articles in National and International Journals and also presented papers in several National Conferences, Seminars and Workshops. She is currently guiding M.Phil. and Ph.D. Scholars. She also sets question papers for universities in Native Nadu.