International Journal of Advanced Trends in Engineering and Management (IJATEM)
Vol. 2(01), January 2023, pp. 17-24

Speech Emotion Recognition Using Machine Learning

J.A Smitha[1*], A. Tamizharasi[2]

[1]Professor, Department of Computer Science and Engineering, Sri Sairam college of Engineering, Bangalore, smithaja1234@gmail.com.
[2]Assistant Professor, Department of Computer Science and Engineering, R.M.D. Engineering College, kavarapettai, Thiruvallur, tamizh05834@gmail.com.

ABSTRACT

The aim of the paper is to analyse the reliability of predicting human emotion from speech using Long Short-Term Memory (LSTM) approach. Human-computer interaction is extensively utilised in Speech Emotion Recognition (SER) to examine various methods and datasets for determining the practical results and discover more regarding this unsolved problem. While it is complicated to identify audio and difficult to estimate a person's emotion because emotions are personal, SER makes this possible. Tone, pitch, expression, behaviour, and other states are used to determine emotion. A specific few of them are assumed to be capable of detecting emotion in speech. Here, different datasets are employed to recognize the emotions, such as Interactive Emotional Dyadic Motion Capture (IEMOCAP), Ryerson Audio-Visual Database of Emotional Speech and Berlin (EMO-DB). The final outcomes attained in trails, evidently demonstrate that the presented approach is used to achieve the task of voice emotion recognition. The proposed method implemented in python software for verifying its performance.

Keywords: SER, LSTM, RAVDESS, Berlin EMO-DB, IEMOCAP.

1 Introduction

In previous times, the fast advancement of elevated has made digital devices more desirable in our day to day lives. Chatbots, sales advertisements, intelligent healthcare, and entertainment all take into account the humanization of human-computer contact in addition to the completion of services. [1]. Empathy, which is closely linked to emotion in contrast to customer involvement, it was applied to enhance user experience in the layout of a human-computer interface dialogue system [2-4]. In point of fact, research on SER concentrates primarily on recognition modelling and emotion feature extraction. The ability to encode emotion components collected from speech signals is essential for efficient emotion recognition and is a key component of spoken emotion identification. Numerous studies sought acceptable audio elements or groups of features that would be ideal for representing emotions. Various feature units were used in other research to represent audio [5-7]. Numerous academics have attempted to extract features from the unprocessed audio waveform for SER, in recent times due to the advancement of Neural Networks (NN). According to some research, the majority of the feature extraction algorithm's settings are experientially updated, which results in the present acoustic feature set occasionally deficient in subjective emotional data [8]. Numerous investigations explored appropriate audio features for representing emotions. The main issue

of this paper is people are expressing mismatched manners of emotions in different dialogue turns. When describing ordinary daily life events, people exhibit high arousal emotions and consider neutral emotions. To overcome this issues, LSTM neural networks in this process do not alter their weights due to the feedback connection, which produces prospective activations for modern actions [9]. The time feature is constrictive for applications in which a real-time component is essential, such as music composition, speech processing and video description. This functionality can be useful in applications where timing is crucial, i.e. speech processing. [10]. Therefore, a unique datasets are deployed to recognize emotions, and the proposed technique effectively increasing the perfection of SER. The main contribution is to normalise to features to have a unit variance and remove the mean from the features. The research is conducted in a speaker-neutral environment.

2 Recent Works

Ying Zhou et al [2022] suggested a SER technology that is important in many employment such as health care and call centres social robots [11]. As a consequence, both industry and academia have focused on speech emotion recognition. As a result of the incorrect representation, their algorithms produce low accuracies. Authors address ambiguous speech emotions by suggesting a novel Multi-Classifier Interactive Learning (MCIL) technique inspired by optimally interacting theory.

Hengshun Zhou et al (2021) clarified that Multimodal Emotion Recognition (MER) task in emotions computation is challenging because it is problematic to extract discriminative characteristics to distinguish minute changes in person emotions with abstractions and various expressions. Regarding that, a self-attention based video stream and the suggested audio stream are combined to present a global FBP (G-FBP) approach to Audio-Visual Information Fusion (AVIF) [12].

Reem Hamed Aljuhani et al (2021) demonstrated that Due to the requirement for additional applications that include human interaction, machine learning (ML) algorithms have subsequently emerged as an active research area. The dataset was produced using publicly accessible YouTube videos and labelled with four different emotions: neutral, neutral, happiness, and happy. The classification techniques Multi-Layer Perceptron (MLP), Support Vector Machine (SVM) and K-Nearest Neighbour (KNN) were utilised after several spectral properties, including the Mel-Frequency Cepstral Coefficient (MFCC) and Mel Spectrogram (MS) were obtained [13]. The results of the three models were discussed, analysed, and compared using various feature extraction methods.

Chenghao Zhang et al (2021) have presented SER, It is a crucial step in the human-computer interaction process and has gained more attention in recent years. Although many different ideas have been put out in SER, performance has not yet been improved. The way to effectively extract emotion-oriented features is a major challenge in the SER system poor implementation. Unlike many previous works, our model incorporates instance normalisation, a general technique in the style transfer field, rather than batch normalisation [14].

Jia-Hao Hsu et al (2021) explained that nonverbal vocalisations inside an utterance, such as laughing, cries or other emotion interjections, play a significant part in emotion display in real-life communication. Few nonverbal vocalisations, which typically occur during normal speech, were taken into consideration by emotion recognition systems in earlier studies [15]. First, a verbal and nonverbal sound detector depends on support SVM is created. Using a prosodic phrase auto-tagger, the nonverbal/verbal sound segments are extracted. The emotion and sound feature embeddings for each segment are extracted using deep residual networks (ResNets).

3 Proposed Work Explanation

The input speech signal is converted into data form through data preparation module, then the data augmentation process increase the amount of data by adding modified existing data. Here, the acoustic analysis is utilized for extracting features, it is separate the signal and reduce the unwanted information (i.e.). Noise. Finally the LSTM model used to process the data and making the accurate predictions in SER.

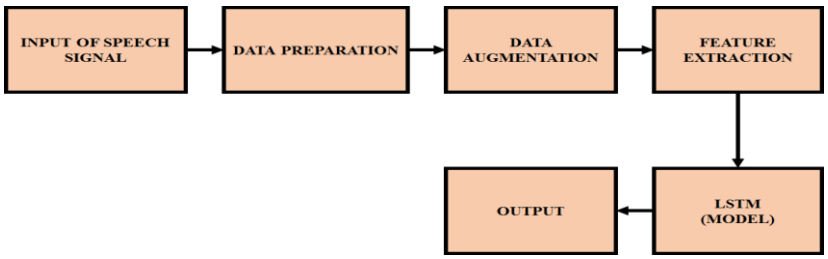

Figure 1: *Block diagram of proposed system*

The Figure 1 shows the block diagram for proposed methodology.

3.1 Data Augmentation

- When the initial data size is constrained, data augment is required to tackle the problem of data imbalance.
- The generation of large training data samples via a series of deformations on the data in the training dataset is known as data augment.
- There are several techniques to add data to an image, such as rotating it in various directions, both horizontally and vertically, introducing noise, and altering the colour of the image.

Algorithm 1 Data Augmentation Algorithm

Input: *sp*: Spectrogram
Sr: Sample Rate
bg: Background Noise
tm: Time Shifting,
bgnr: Background Noise Range
tsh: Time shifting rate
Algorithm:|

```
1: Initialize and assign input parameters (sp, bg, tm, bgnr, tsh)
2: sp = read_folder ('folder name/'+Filename');
3:       for i: = 1 to length (sp) do
4:           data = Read Spectrogram File (sp, Name);
5:           bg = add.random.bgnoise (data, bgnr);
6:           tm = Time. Shifting (data, sr * tsh);
7:           Write spectrogram (bg_noise.jpg.bg);
8:           Write spectrogram (time_shifting.jpg. tm);
9            end for
```

The data augmentation algorithm's pseudo-code is given in Algorithm 1.

3.2 Feature Extraction

- The objective of the acoustic analysis is to categorise the speech signal into its individual components and provide parametric measurements for each one.
- Physical characteristics in terms of loudness, frequency and amplitude are known as acoustic features.
- Before acoustic features are extracted, speech signals are further pre-processed.
- The Butterworth filter is employed in this research to eliminate background distortion from speech signals.

3.3 LSTM Networks

- The feedback connection of recurrent neural networks allows them to learn and react to recent events without changing their slowly shaped weights. This creates short-term activations.
- This feature can be advantageous in applications in which time is critical, such as speech processing, music composition, and video description.
- Nevertheless, because they have been trained using Back Propagation through Time, error signals flowing backward in time can grow or disappear depending on the size of the weights.
- As a result, the network will either produce oscillating weights or train and converge slowly.

4 Results and Discussion

In this research, 3 separate voice datasets are employed, which are widely used by authors in emotion recognition. This collection of video and audio recordings comprises 12 female and 12 male individuals reading English lines while displaying 8 different facial expressions. For this analysis, only speech samples are utilized. Additionally, there are 1440 records in the collection as a whole. It is commonly employed by scholars studying SER identification and

enables to conduct more thorough comparisons with earlier research. German dataset consists of 535 audio outputs subdivided into 7 different emotion categories. The dataset comprises the following emotion classes: disgust, rage, sadness, anger, fear/anxiety, neutral, and boredom**.** This dataset includes audio, video, and facial movement samples taken from 5 actor pairings, two of whom were male and one of whom was female. Ten emotion categories: angry, joyful, sad, neutral, annoyed, eager, afraid, astonished, disgusted are represented in the data series' audio recordings. The suggested framework for unreliable data. In this paper, we only take into account the datasets for the four emotion classes (angry, happy, neutral, and sad). There are 889 audio files total throughout four classes.

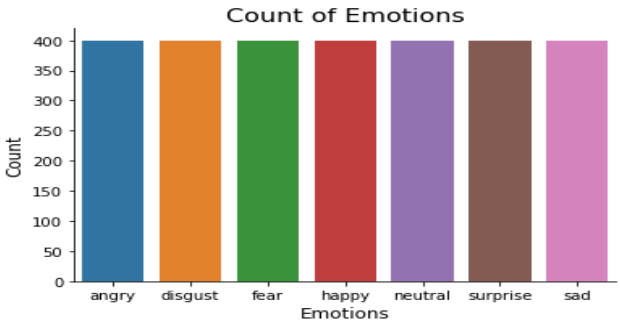

Figure 2: *Data Set*

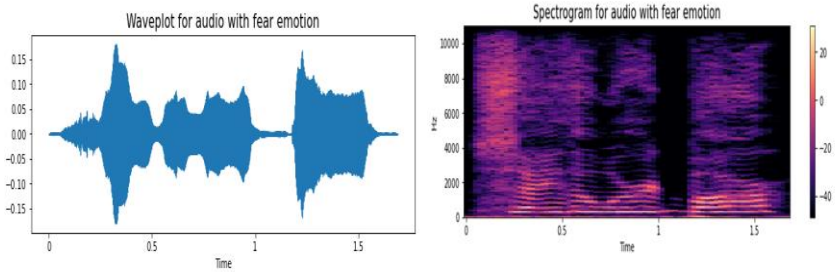

Figure 3: *Wave plot and Spectrogram for Audio with Fear*

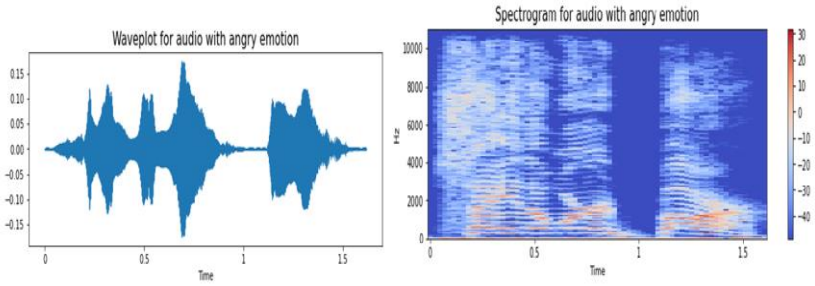

Figure 4: *Wave plot and Spectrogram for Audio with Angry Emotion*

Article Title: Speech Emotion Recognition Using Machine Learning

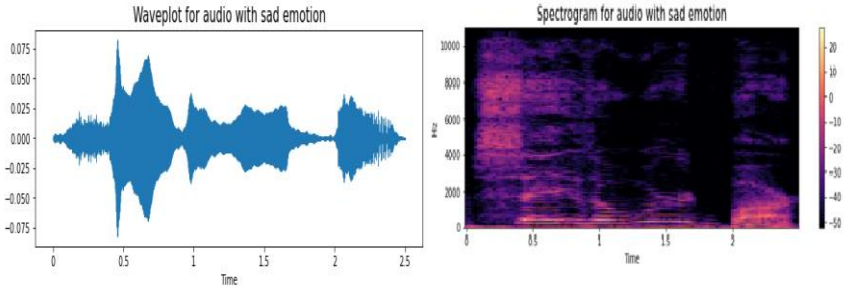

Figure 5: *Wave plot and Spectrogram for Audio with Sad Emotion*

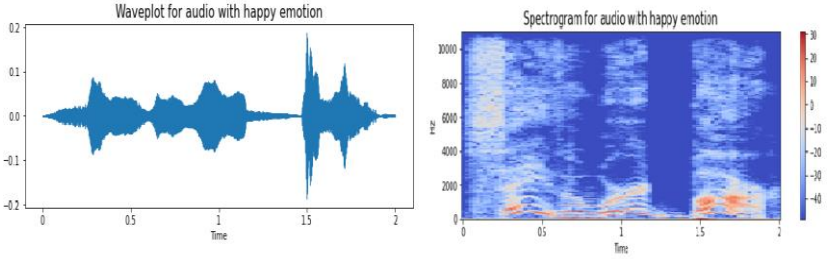

Figure 6: *Wave plot and Spectrogram for Audio with Happy Emotion*

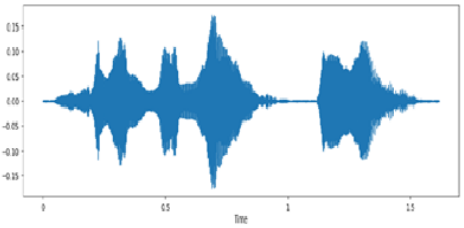

Figure 7: *Simple Audio*

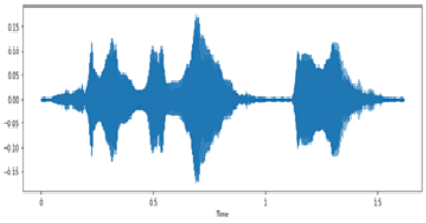

Figure 8: *Noise Injection*

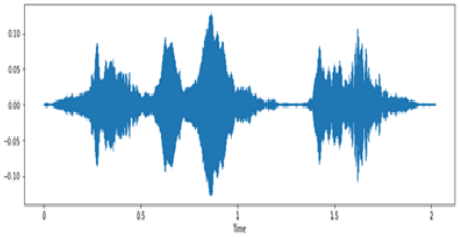

Figure 9: *Stretching*

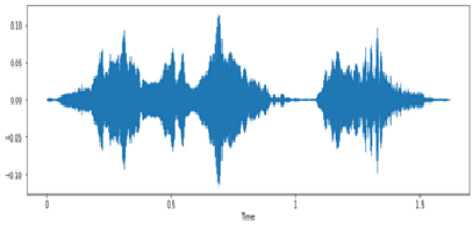

Figure 10: *Shifting*

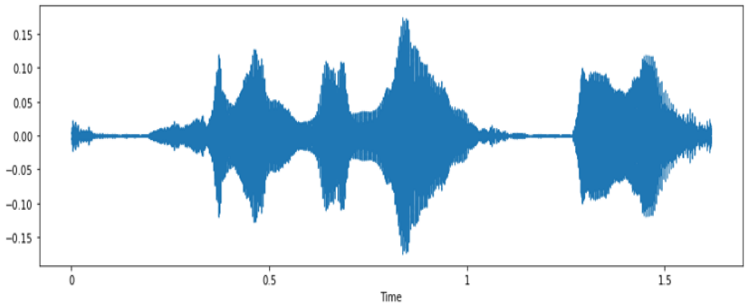

Figure 11: *Pitch*

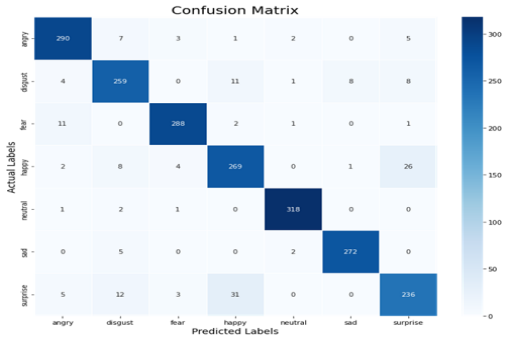

Figure 12: *Model Loss and Accuracy*

5 Conclusion

In this paper, a multiple emotional SER techniques and the related speech databases, comparing them from various angles. Including IEMOCAP, a semi-natural database that is frequently cited, and VAM, a natural database that is frequently utilised, in addition to the simulated databases. The Deep convolutional LSTM structures and LSTM networks have been deployed to elevate the problem to a new level and give the network long-term memory so it can recognise enduring paralinguistic patterns. Additionally, they have demonstrated improved speaker-independent emotion identification abilities. Last but not least, the addition of the attention approaches has given the classifiers a new level of nonlinearity, which can assist build a more effective system with fewer components. In effort to get models closer to production in actual circumstances, future research might examine more reliable and dataset-independent solutions.

References

1. Zhao Jianfeng; Xia Mao; Lijiang Chen, Year: 2019, "Speech emotion recognition using deep 1D & 2D CNN LSTM networks," Biomedical signal processing and control, Vol: 47, pp. 312–323.

2. Wang Jianyou et al, Year: 2020, "Speech emotion recognition with dual-sequence LSTM architecture," ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP).

3. Lim Wootaek; Daeyoung Jang; Taejin Lee, Year: 2016, "Speech emotion recognition using convolutional and recurrent neural networks," 2016 Asia-Pacific signal and information processing association annual summit and conference (APSIPA).

4. Khalil Ruhul Amin et al, Year: 2019, "Speech emotion recognition using deep learning techniques: A review," IEEE Access, Vol: 7, pp. 117327–117345.

5. Srinivas Parthasarathy; Carlos Busso, Year: 2020, "Semi-Supervised Speech Emotion Recognition with Ladder Networks", IEEE/ACM Transactions, Vol: 28, pp. 2697–2709.

6. Yi Lei; Shan Yang; Xinsheng Wang, Year: 2022, "MsEmoTTS: Multi-Scale Emotion Transfer, Prediction, and Control for Emotional Speech Synthesis", IEEE/ACM Transactions, Vol: 30, pp. 853–864.

7. Na Liu; Baofeng Zhang; Bin Liu; Jingang Shi; Lei Yang; Zhiwei Li; Junchao Zhu, Year: 2021, "Transfer Subspace Learning for Unsupervised Cross-Corpus Speech Emotion Recognition", IEEE Access, Vol: 09, pp. 95925–95937.

8. Kerkeni Leila; Youssef Serrestou; Mohamed Mbarki; Kosai Raoof; Mohamed Ali Mahjoub; Catherine Cleder, Year: 2019, "Automatic speech emotion recognition using machine learning."

9. Abdelhamid Abdelaziz A et al, Year: 2022, "Robust speech emotion recognition using CNN+ LSTM based on stochastic fractal search optimization algorithm," IEEE Access, Vol: 10, pp. 49265–49284.

10. Reza Lotfian; Carlos Busso, Year: 2019, "Curriculum Learning for Speech Emotion Recognition from Crowd sourced Labels", IEEE/ACM Transactions, Vol: 27, no: 04, pp. 815–826.

11. Ying Zhou; Xuefeng Liang; Yu Gu; Yifei Yin; Longshan Yao, Year: 2022, "Multi-Classifier Interactive Learning for Ambiguous Speech Emotion Recognition", IEEE/ACM Transactions, Vol: 30, pp. 695–705.

12. Hengshun Zhou; Jun Du; Yuanyuan Zhang; Qing Wang; Qing-Feng Liu; Chin-Hui Lee, Year: 2021, "Information Fusion in Attention Networks Using Adaptive and Multi-Level Factorized Bilinear Pooling for Audio-Visual Emotion Recognition", IEEE/ACM Transaction, Vol: 29, pp. 2617–1629.

13 Reem Hamed Aljuhani; Areej Alshutayri; Shahd Alahdal, Year: 2021, "Arabic Speech Emotion Recognition from Saudi Dialect Corpus", IEEE Access, Vol: 09, pp. 127081–127085.

14. Chenghao Zhang; Lei Xue, Year: 2021, "Auto encoder With Emotion Embedding for Speech Emotion Recognition", IEEE Access, Vol: 09, pp. 51231–51241.

15. Jia-Hao Hsu; Ming-Hsiang Su; Chung-Hsien Wu; Yi-Hsuan Chen, Year: 2021, "Speech Emotion Recognition Considering Nonverbal Vocalization in Affective Conversations", IEEE/ACM Transaction, Vol: 29, pp. 1675–1686.