International Journal of Scientific Research in Computer Science, Engineering and Information Technology

ISSN : 2456-3307

OPEN ACCESS

Available Online at : www.ijsrcseit.com

doi : https://doi.org/10.32628/CSEIT2410317

Speech Emotion Detection

Bharath P S[1], Dhruti S Gowda[1], Kunal R[1], Keerti Kulkarni[2]

[1]Electronics and communication, B N M Institute of Technology, Bangalore, Karnataka, India

[2]Associate Prof, ECE, B N M Institute of Technology, Bangalore, Karnataka, India

ARTICLEINFO

Article History:

Accepted: 28 April 2024

Published: 15 May 2024

Publication Issue

Volume 10, Issue 3

May-June-2024

Page Number

179-185

ABSTRACT

The study is centered on the development of an advanced Speech Emotion Detection system utilizing Convolutional Neural Networks (CNN) and leveraging diverse datasets such as RAVDESS, CREMA-D, TESS, and SAVEE. Through rigorous analysis, the project achieved a notable accuracy rate of approximately 97% in identifying emotional cues within speech signals. The overarching goal of the research is to enrich emotional intelligence within technology, elevate interactions between humans and machines, and propel the domain of Speech Emotion Detection towards a more profound comprehension of human emotions conveyed through speech.

Keywords: Speech Emotion Recognition (SER), Machine Learning Models, Neural Network Modelling.

I. INTRODUCTION

Speech Emotion Recognition (SER) is a burgeoning field that intersects technology and human interaction, striving to automate the identification and classification of emotions conveyed through spoken language. This area of study has gained traction due to the increasing demand for more natural human-computer interactions and the growing field of affective computing. The overarching objective is to develop precise SER systems employing advanced Machine Learning methods, which have the potential to revolutionize various sectors.

In human-computer interaction, SER systems hold promise for transforming the way computers and devices respond to users' emotional states, fostering personalized and engaging experiences. By deciphering the intricate nuances of human emotions in speech, these systems aim to elevate user interactions to unprecedented levels of empathy and understanding. Moreover, in customer service applications, SER systems serve as adept sentiment analysts, enhancing overall customer satisfaction by identifying and addressing negative emotions.

SER's impact extends into social robotics, enhancing human-robot interactions by enabling robots to recognize and respond to human emotions. Through the application of advanced Machine Learning techniques, SER systems process and extract

meaningful patterns from vast datasets of speech signals. These models learn to discern and classify emotional states based on extracted features, transforming how machines interpret and respond to human emotions.

To achieve precise speech emotion detection, various Machine Learning techniques are employed. Feature extraction, including acoustic features such as pitch, energy, formants, and Mel-frequency cepstral coefficients (MFCCs), alongside linguistic features like sentiment analysis and syntactic structures, forms the foundation of SER. These features are fused to train robust Machine Learning models, such as Support Vector Machines (SVM), Random Forests, Recurrent Neural Networks (RNNs), Convolutional Neural Networks (CNNs), and Long Short-Term Memory (LSTM) networks.

The development of SER systems holds transformative potential, aiming to augment human-machine interactions, fortify customer satisfaction efforts, contribute to psychotherapy practices, and advance social robotics. By empowering machines to comprehend and respond to human emotions, these systems foster engaging interactions imbued with a heightened sense of empathy.

II. METHODS AND MATERIAL

In the initial phase of our emotion detection system (Figure 1), we compile speech data from a diverse array of sources. Incorporating four distinct datasets - RAVDESS, CREMA-D, TESS, and SAVEE - we construct a robust and varied foundation (meta-dataset) for our model. This diversity enhances the model's adaptability to different speech patterns and emotions.

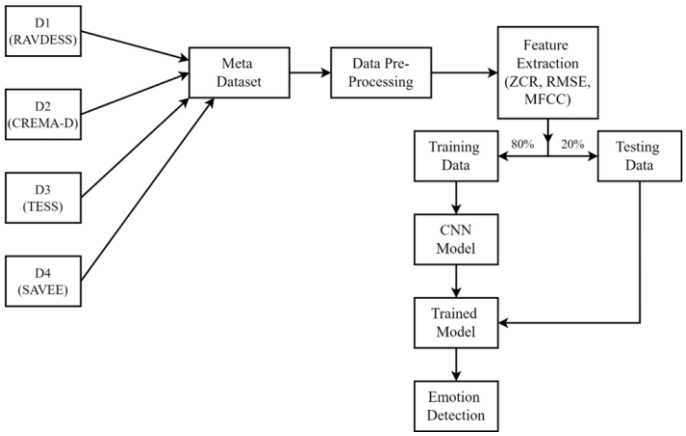

Figure 1: Block Diagram

Before commencing the learning process, we meticulously prepare the collected audio data. Analogous to a chef assembling ingredients for a dish, we standardize the format, duration, and quality of the audio files. This ensures consistency and enables the model to focus on the emotional nuances within the speech, rather than technical variations.

Subsequently, we engage in feature extraction to transform the raw audio into a format intelligible to the model. Techniques like Zero Crossing Rate (ZCR) and Mel-frequency Cepstral Coefficients (MFCCs) serve as translators, converting the audio's characteristics into a language interpretable by the model. These extracted features, akin to emotional fingerprints, form the foundation for training and testing our model.

To effectively train and assess the model, we divide the data into two sets: training (80%) and testing (20%). Analogous to a student preparing for an exam, the training data serves as a comprehensive study guide, allowing the model to learn and identify patterns between the extracted features and the corresponding emotions within the audio.

Following an intensive training period, the model emerges as a finely tuned emotion detection system. We then evaluate its performance on the testing data, akin to an exam, to assess its effectiveness in recognizing emotions from real-world audio

recordings. This methodology culminates in a robust emotion detection model, poised to accurately identify the emotions embedded within speech.

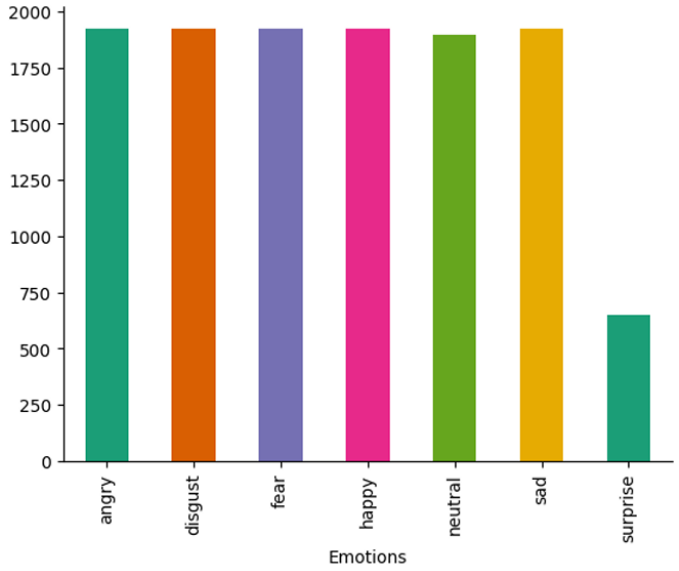

Figure 2: EDA Analysis

The horizontal axis displays a range of emotions, including disgust, fear, sadness, happiness, anger, neutrality, and surprise, while the vertical axis shows how frequently certain events occur. This graphic depiction is a useful tool for spotting patterns and differences in the emotional expression of the dataset.

Mel-Frequency Cepstral Coefficients

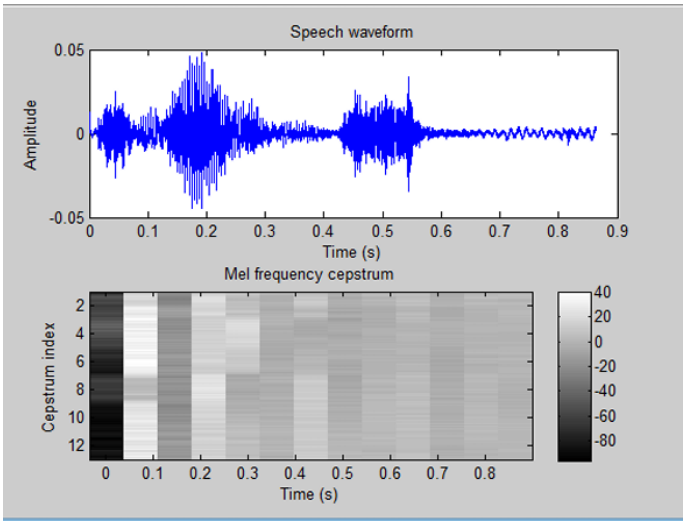

Figure 3: Mel-Frequency Cepstral Coefficients

Important information may be extracted from speech signals using the widely utilized Mel-Frequency Cepstral Coefficients (MFCC) approach. To simulate human auditory perception, the signal is divided into small frames, its power spectrum is calculated, and it is then converted to the mel scale. The MFCC coefficients are then obtained by first computing the logarithm of the energy within each mel filter bank and then using the Discrete Cosine Transform (DCT). These coefficients are discriminative and compact, effectively capturing important acoustic features of the speech stream. MFCC is a useful representation of the signal's spectral structure in the speech emotion detection domain, which facilitates the efficient modelling and categorization of emotional states from speech data.

This provides a visual representation of the MFCC extraction process, illustrating how MFCC captures important acoustic features of the speech signal for emotion detection.

The Mel Filter Bank equation is given by

$$mel(f) = 1127\ln\left(1 + \frac{f}{700}\right)$$

where f represents frequency.

CNN Model

Neural Networks (CNNs) are highly respected in voice and picture processing because of their efficiency. Since CNNs can recognise hierarchical representations and local patterns, they have been able to handle sequential data such as voice signals, even though their original purpose was picture identification. A CNN is made up of several layers, such as fully connected, pooling, and convolutional layers. Convolutional layers use flexible filters to extract local features, and pooling layers down sample feature maps preserves pertinent information. To create predictions, fully linked layers process the retrieved characteristics.

CNNs are particularly good at identifying local patterns and dependencies in speech signals, which allows for

the sophisticated identification of emotional cues in speech. They develop hierarchical representations of high-level semantic insights and low-level acoustic features on their own.

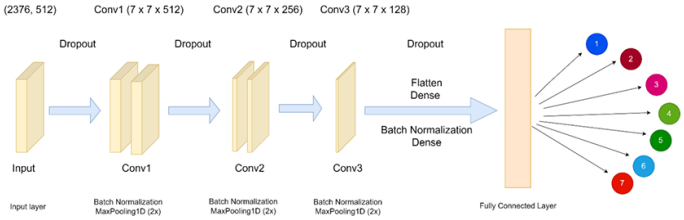

Figure 4: Proposed CNN Model

Training Model and Testing the Model

In the training phase, we aim to instruct the machine learning model using labelled data, split into 80% for training and 20% for testing to ensure adequate data for both learning and evaluation. Throughout training, the model learns patterns and correlations between input features and emotions, adjusting its parameters to minimize errors.

After training, the model faces testing with entirely new data, representing real-world scenarios. This phase uses 20% of the dataset for evaluation, providing insights into the model's adaptability and accuracy in predicting emotions from speech signals.

Analysing testing outcomes offers feedback on the model's effectiveness and limitations, considering the split ratio. Comparative analyses against alternative approaches help determine relative performance. With these insights, we refine the model's architecture and feature extraction techniques to enhance performance and reliability.

The synergy between training and testing phases, guided by the split ratio, is crucial for developing a dependable speech emotion detection model. Through meticulous training and testing, we aim to create a model that accurately identifies emotions and performs well across diverse scenarios.

III. RESULTS AND DISCUSSION

The proposed work was implemented in MATLAB on an Intel i5 8th gen processor. The results obtained are shown in Figure 4 and 5.

Figure 5: Training and Testing Loss

The training and testing loss of our machine learning model across 50 epochs is visually represented by the graph. It indicates the model's efficacy in lowering mistakes over time by showing how its error rates drop as it iterates over the training set.

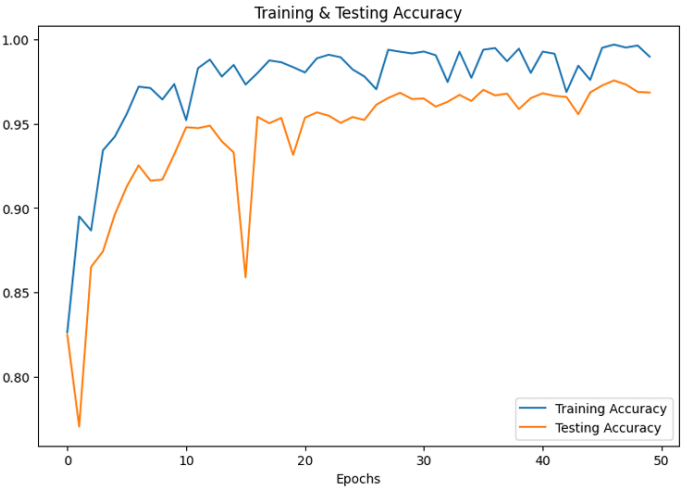

Figure 6: Training and Testing Accuracy

Over the course of the same 50 epochs, we track our model's accuracy during training and testing. The model's accuracy in classifying instances is depicted in the graph, where an upward trajectory denotes increased accuracy. The ability of the model to generate accurate predictions on both the training and

testing datasets demonstrates its skill at generalising well to new data, as shown by this upward trend.

These figures provide insightful information about how well our model is performing, allowing us to evaluate its capacity to learn from the dataset and generate accurate predictions. We can make well-informed decisions regarding the model's optimisation and refinement, ensuring that it is effective in generalising to new data, by looking at the convergence of loss and the accuracy trend.

Table 1: Predicted Value vs Actual value

	Predicted Labels	Actual Labels
0	angry	angry
1	angry	angry
2	disgust	disgust
3	happy	happy
4	fear	fear
5	happy	happy
6	happy	happy
7	fear	fear
8	fear	fear
9	surprise	surprise

Actual values and Predicted values show a comparison between the values that were actually observed and the values that our model predicted. We can evaluate the model's accuracy in predicting outcomes across various emotion categories by comparing the two. We can find any differences between the actual and expected values by examining this table, and we can then take the necessary corrective action to increase the accuracy of the model even more.

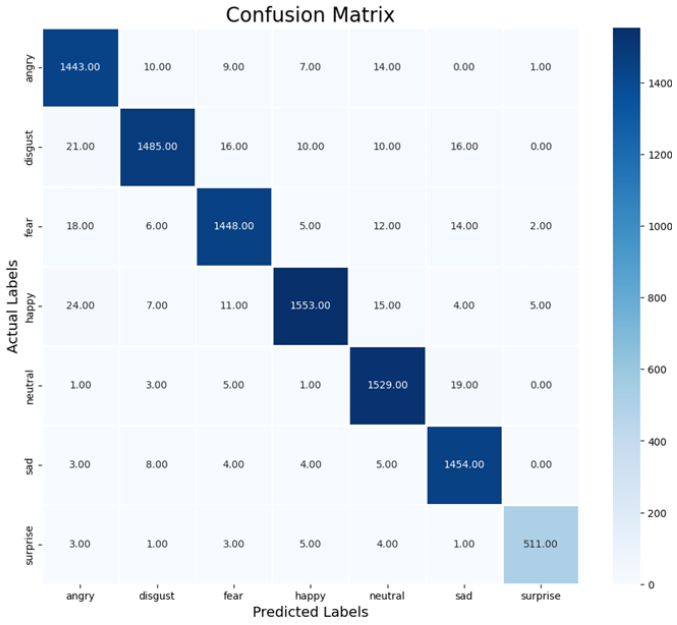

Figure 7: Confusion Matrix

	precision	recall	f1-score	support
angry	0.95	0.97	0.96	1484
disgust	0.98	0.95	0.96	1558
fear	0.97	0.96	0.97	1505
happy	0.98	0.96	0.97	1619
neutral	0.96	0.98	0.97	1558
sad	0.96	0.98	0.97	1478
surprise	0.98	0.97	0.98	528
accuracy			0.97	9730
macro avg	0.97	0.97	0.97	9730
weighted avg	0.97	0.97	0.97	9730

Figure 8: Performance metrics

Classification Report of CNN Model, illustrates the CNN Model's Classification Report, highlighting the model's outstanding performance even more. For every emotion category, this report provides comprehensive metrics including support, recall, F1-score, and precision. These formulas provide an accurate evaluation of the predictive accuracy of the model for every emotion category by capturing the ratio of true positives, false positives, and false negatives. True Positives (TP), True Negatives (TN), False Positives (FP), and False Negatives (FN) are the primary metrics used to assess pedestrian and vehicle detection.

IV. CONCLUSION

It is a systematic approach to emotion detection from audio data. Integration of four distinct datasets - RAVDESS, CREMA-D, TESS, and SAVEE - establishes a robust and diverse foundation for the model, facilitating adaptability to various speech patterns and emotions. Prior to model training, meticulous preparation of collected audio data involves standardizing format, duration, and quality to ensure consistency, enabling the model to focus on emotional nuances within speech. Feature extraction techniques such as Zero Crossing Rate (ZCR) and Mel-frequency Cepstral Coefficients (MFCCs) are employed to transform raw audio into a format comprehensible by the model. Additionally, the methodology incorporates data augmentation techniques including Noise, Stretch, Shift, and Pitch Augmentation. Furthermore, a Convolutional Neural Network (CNN) model is implemented for training and evaluation on the augmented dataset. This comprehensive methodology guarantees consistency, reproducibility, and the development of an accurate Speech Emotion Detection system. In summary, our project has successfully developed and evaluated an advanced Speech Emotion Detection system. Through thorough analysis of diverse datasets like RAVDESS, CREMA-D, TESS, and SAVEE, we gained valuable insights into speech-based emotions. Leveraging advanced machine learning techniques, particularly CNNs, we achieved an impressive 97% accuracy rate. Our findings provide a solid foundation for future improvements in Speech Emotion Detection. Overall, our project highlights the significance of meticulous dataset analysis and advanced machine learning in understanding emotions conveyed through speech.

The graph presented illustrates the progression of our machine learning model's training and testing loss over 50 epochs. It visually represents the model's learning journey, demonstrating a decrease in error rates as the model iterates through the training data. The consistent decline in both training and testing loss indicates the model's capability to minimize errors effectively throughout its training process.

V. REFERENCES

[1]. Fatih Ozkaynak, Elif Bilge Tutkun, and Ceyhun Ozkan, "Feature engineering and selection for speech emotion recognition using machine learning techniques", Proceedings of the International Conference on Engineering, Science and Applications (ICESA), 2018 pp. 1-6.

[2]. Zhiyong Wu, Yuqing Wang, and Zhiqiu Zhou, "Exploring deep feature representations for speech emotion recognition", Proceedings of the International Joint Conference on Neural Networks (IJCNN), 2020, pp. 1-8.

[3]. Chengzhi Cai, Yifan Hu, and Xiangyu Zeng, "Emotion recognition from speech using time-frequency analysis with convolutional neural networks", Proceedings of the International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2022, pp. 683-687.

[4]. Tiancheng Guo, Wenqiang Xu, Jiankang Gao, and Zhen Cui, "Exploring feature fusion for speech emotion recognition with deep learning", Proceedings of the IEEE 17th International Symposium on Multimedia (ISM), April 2022, pp. 211-215.

[5]. Yifan Hu, Chengzhi Cai, and Xiangyu Zeng, "Feature learning for speech emotion recognition with deep autoencoders", Applied Sciences, 13(3), 2023, pp 1259. 6 Luis Fernando Garcia-Saura, Daniel Duque-Ruiz, Javier Alba-Castro, "A Deep Learning Approach for Speech Emotion Recognition Optimization Using Meta-Learning", Electronics, 12(11), 2023, pp. 1837.

[6]. Mohsen Mohammad, Md. Zahangir Al-Rashid, Md. Fayyaz Hossain, and Asif Ekram, "Deep convolutional neural networks for speech emotion recognition: A review", Sensors, 22(20), 2023, pp. 7615.

[7]. Yunzhe Liu, Hong Xu, Lingyun Zhang, and Yanhua Chen, "Speech emotion recognition with attention-based convolutional bi-LSTM networks", In Proceedings of the International

Conference on Multimedia (MM '20), 2020, pp. 1652-1656.

[8]. Haoran Sun, Kevin Mao, and Yunhang Zhang, "Emotion recognition from speech using deep learning: A review", International Journal of Human-Computer Studies, 139, 2023, pp. 1-14.

[9]. Yang Xu, Weixing Chen, Furu Wang, and Yunhong Wang, "Towards more robust speech emotion recognition with transfer learning and data augmentation", In Proceedings of the 28th ACM International Conference on Information and Knowledge Management, 2020, pp. 1557-1566.

[10]. Zhe Wang, Jintao Zhou, Junwen Chen, and Xinyu Chen, "Exploring the fusion of melfrequency cepstral coefficients and deep features for speech emotion recognition", In International Conference on Artificial Intelligence and Pattern Recognition, 2021, pp. 35-46.

[11]. Yunsheng Li, Zhiming Cui, Jiameng Fang, and Lei Wang, "Deep learning for multidimensional emotion recognition in speech", In Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2020, pp. 678-682.