EMOTION RECOGNITION THROUGH SPEECH USING DEEP LEARNING

K Ananya[1]
Department of Networking and Communications,
SRMIST, Kattankulathur – 603203,
Chennai, India
kv6187@srmist.edu.in

Priyanshi Sharma[2]
Department of Networking and Communications,
SRMIST, Kattankulathur – 603203,
Chennai, India
ps2058@srmist.edu.in

Chenna Vanshika[3]
Department of Networking and Communications,
SRMIST, Kattankulathur – 603203,
Chennai, India
vc8427@srmist.edu.in

C. Fancy[4*]
Department of Networking and Communications,
SRMIST, Kattankulathur – 603203,
Chennai, India
fancyc@srmist.edu.in
(* - Corresponding Author)

***Abstract*—The research aims to develop a voice emotion identification system that utilises empirical feature optimisation techniques and all accessible acoustic data. The methodology used in this work involves gathering data from several emotive speech datasets and utilising audio visualisation techniques to interpret the data. Next, apply augmentation methods to the dataset, such as pitch shifting, temporal stretching, and noise injection. Key audio properties such as Mel Frequency Cepstral Coefficients, Zero Crossing Rate, RMS value, and MelSpectrogram are involved in speech emotion identification. Pitch, tone, and energy shifts corresponding to various emotions are all captured in exquisite detail. A mixture of feature selection and applied deep learning model hyperparameter tuning—specifically, Convolutional Neural Networks—is used to achieve optimisation. The model's accuracy in identifying emotions from speech is evaluated using pre-processed training data and unseen testing data. In addition to improving customer service and mental health monitoring, the results of this project will help to make human-computer interaction more responsive and intuitive. The innovation in the research is ensured by combining the use of Convolutional Neural Networks as the hyperparameter tuning tool for deep learning models with an optimisation process feature filter. It will be ensured by this study that the model learns to recognise emotions from speech data in an efficient manner. The proposed methodology exhibits higher accuracy in identifying emotions from speech when compared to the latest related study. Assessing the model's performance against pre-processed training and unseen testing data, the results are extremely noteworthy.**

Keywords*—Speech Emotion Recognition, Acoustic Features, Feature Optimization, Data Augmentation, Deep Learning, Convolutional Neural Networks, Emotion Detection, Human-Computer Interaction, Customer Service, Mental Health Monitoring

I. INTRODUCTION

Technological innovation has been changing many fields including artificial intelligence (AI) and machine learning. Among the emerging fields is emotion recognition, which aims at making human-computer interactions more intuitive and efficient. Speech Emotion Recognition (SER) utilizes natural modes of communication to identify emotional states with the potential to enhance user experiences across multiple applications. This project aims to develop a strong deep-learning model for accurately automating emotion recognition from speech. The problem statement addresses the challenge of striking a balance between a comprehensive representation of acoustic features and empirical feature tuning to inculcate emotional intelligence into technology that can better understand human emotions.

II. LITERATURE SURVEY

A novel method for speech emotion recognition (SER) based on the mechanics of emotion perception in the human brain was reported by Liu G., Cai S., and Wang C. [1]. In contrast to the current SER models, which mostly rely on computer vision and natural language processing (NLP), this approach takes emotional brain regions into account based on neuroscience research. It is a multi-task learning-based methodology for improving SER through neuroscience to enhance precision and stability. The study carefully delves into errors associated with traditional methods, showing that this neuroscience approach has superior performance potential. However, there are challenges in combining brain science with SER such as modelling complex brain functions accurately as well as addressing individual differences in emotional responses. In case these challenges are not well tackled, the overall accuracy and effectiveness of the SER system may be seriously compromised

Jones E., Jan T., Babar M. I., Khalil R. A., Zafar M. H., and Alhussain T. [2] thoroughly assessed the most recent deep learning-based SER techniques. This paper gives an overview of the sources of emotional data by discussing several acting, induced, and natural emotion databases. Furthermore, the significance of prosodic features and Mel-frequency cepstral coefficients (MFCC), which are employed in feature computation, and other fundamental acoustic characteristics for SER will be covered in relation to emotion recognition. Additionally, by illustrating how these advanced techniques have increased the accuracy and efficiency of SER systems, the work highlights the recent shift towards deep learning methodologies for automated feature extraction technologies. However, it is crucial to recognise problems such as variation in the included face expressions or even employ suitable vast diverse datasets for DL model training. In general, SER systems' overall performance may suffer if these problems are not resolved.

Ben Ayed Y. and Aouani H. [3] offer the SER system, which has a two-stage architecture made up of feature extraction and

979-8-3503-0692-7/23/$31.00 ©2023 IEEE

categorisation. After extracting a 42-dimensional feature vector for the feature extraction stage, which includes 39 MFCCs, ZCR, HNR, and TEO, lower-level features are extracted for dimension reduction using autoencoders. Moreover, support vector machines (SVM) with several different kernel functions are employed for the classification, including radial basis function, polynomial, and linear kernels to assess their effectiveness. The efficiency of combining strong classifiers with rich feature extraction has been demonstrated by this work. Nevertheless, choosing the right parameters is a big difficulty, as using the wrong parameters might lead to information loss during the dimension reduction process.

Shaila S. G., Sindhu A., Monish L., Shivamma D. and Vaishali B. [4] used the Ravdess database in this study, which provides high-quality vocal emotional data which includes 7356 files (24 actors). In this work, we have selected Chroma features, MFCC, spectrogram(Melspectrogram), Spectral contrast and Tonnetz features as they are capable of capturing the evidence hidden within the speech domain. The impact of these variables on SER was investigated in this work. The results have shown that if used properly inside an SER system, there will be accuracy improvement in speech-emotion systems. But till now this aspect has not been considered well enough and a lack still exists in information expression among speech emotions and the design of effective feature extractors for these languages.

For feature extraction and selection, Jakubec M., Chmulík M., Lieskovská E., and Jarina R. [5] used a variety of deep learning models, including CNN, RNN, LSTM, and GRU. To increase the accuracy of SER, they used an attention mechanism that highlights important information while disregarding unnecessary ones. The outcomes showed that the performance of SER might be improved by utilising end-to-end deep learning models with attention mechanisms. The performance and operating time of the entire SER system will be greatly impacted if we are unable to resolve several difficult problems, such as the computational expense of these models and the need for a sizable training database.

Deep learning and feature extraction techniques are crucial to the methodology used in other Speech Emotion Recognition (SER) studies [6] – [16]. To train models for emotion categorisation, most research includes a variety of auditory data, including pitch, energy levels, Mel Frequency Cepstral Coefficients (MFCC), and extra spectral information. Many works take a two-step approach: first, the features are extracted, then algorithms like Support Vector Machines (SVM), Recurrent Neural Networks (RNN), and Convolutional Neural Networks (CNN) are used to classify the features. To improve model performance, some research also used cutting-edge methods including multi-task learning and attention mechanisms. Preprocessing and data augmentation techniques are frequently used to increase the models' resilience.

III. Datasets Overview

Four datasets were used in this project:

- 7,442 audio samples from 91 actors—48 men and 43 women—of different racial and ethnic backgrounds, ages ranging from 20 to 74—were captured for the CREMA-D dataset [17]. Six sorts of emotions can be expressed through the recordings: disgust, fear, anger, neutral, happy, and sad. Four intensity levels for each emotion were noted: Low, Medium, High, and Unspecified. The richness of emotional expression in various circumstances is ensured by the diversity of actors and the intensity levels of the emotional states that should be evoked.
- 1,012 audio-only recordings performed by 24 professional actors—12 of whom are men and the remaining six are women—are included in the RAVDESS dataset [18]. A neutral North American accent was used for all of the recordings. From serenity and happiness to fear, rage, and grief, these are the gamut of emotions. There are two distinct expressions for each of these emotions: strong and normal, along with an additional neutral expression for each. Consistency across samples is ensured by standardising the recordings using two utterances that are matched in vocabulary.
- Four male native English speakers, ages 27 to 31, are featured in the SAVEE dataset [19] through audio recordings. Fear, happiness, sadness, rage, disgust, surprise, and neutrality are among its seven emotion classes. It varies throughout sentences to encompass all emotional expressions and provide a broad overview of the various ways that emotions are expressed in speech.
- Two female actors, one 26 years old and the other 64, narrate 2,800 utterances on the TESS dataset [20]. 200 target phrases were recorded by each actress using the same carrier phrase, "Say the word." This dataset contains the following seven unique emotional capture classes: disgust, anger, sadness, fear, neutrality, happiness, and pleasant surprise. The dataset is reliable for analysing speech emotion because of its enormous quantity of recordings and standardisation of recording circumstances.

Together, these datasets include a wide range of audio samples, including different speaker demographics and emotional expressions, which are critical for creating and testing reliable speech emotion detection systems.

IV. Data Preparation and Data Visualization

The audio files from the Crema-D, Ravdess, Savee, and Tess datasets are organised into structured data frames with categorised emotion labels. Every audio file path associated with an emotion label is placed in a single format. Processing is made simple by extracting paths and related audio files from metadata, label mapping, and integrating them into a single data frame. Bar graphs showing the distribution of all emotions in the datasets, including totals for each emotion, were drawn to ensure a fair representation of the emotions represented. Ensuring that an imbalance in the portrayal of a certain emotion was countered at its source and that a stable and equitable foundation for the model's training was formed was imperative.

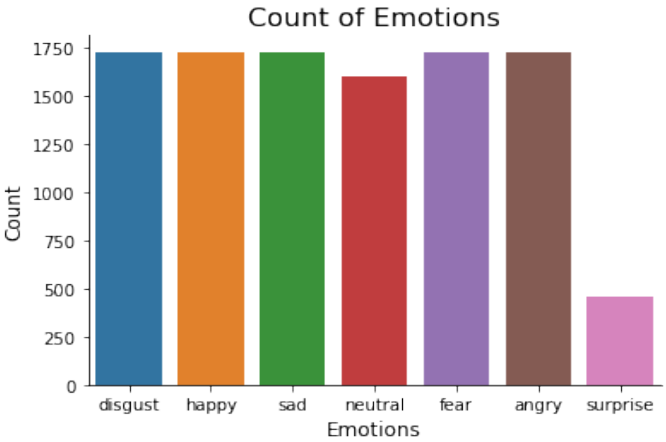

Fig. 1. Count of emotions

V. Data Augmentation

To improve the generalisation and robustness of the model, data augmentation techniques were used on the audio data. This is crucial for introducing variances into the training set, enabling a model to be less sensitive to inputs of diversely characterised audio samples in the real world. Some of the techniques used for augmentation are as follows:

- Noise Injection: Random noise is added to the digital signal, to make the model capable of tolerating environmental noises found in actual sound recordings.
- Pitch Shifting: The pitch of the sound signal is changed up or down without affecting its pace, enabling the system to identify emotions irrespective of pitch changes across various speakers.
- Time Stretching: Slowing down or speeding up an audio signal without modifying its pitch, to teach the model how to handle differences that exist in speech duration and therefore speech speed.

- Shifting: Slightly moving forward or backwards through time concerning an audio signal, ensuring that even slight timing mismatches present in audio information are dealt with by the model

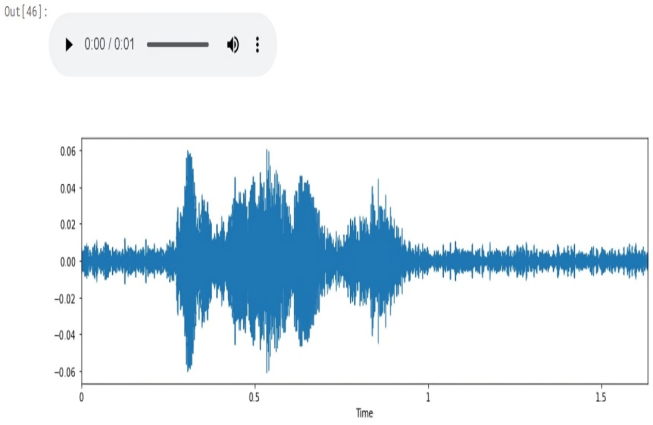

Fig. 2. Data Augmentation Wave Plot

VI. Feature Extraction

The feature extraction process is vital in converting raw audio data into formats that machine learning models can comprehend. In this project, the following key features of audio signals were extracted:

- Zero Crossing Rate (ZCR): The ZCR is the rate at which the sign of the audio signal changes. It ascertains whether or not there are high frequencies in the area. This will distinguish between different sounds and emotions since the rate at which signs cross might vary significantly depending on the emotional intonation.
- Mel Frequency Cepstral Coefficients (MFCC): Symbolising the audio short-wavelength power spectrum. They are obtained by applying a Fourier transform to a signal, then obtaining the logarithms of the powers, mapping the signal's powers onto a mel scale, and then applying an inverse Fourier transform. Certain timbral elements of speech, which are important for picking up on minor emotional cues, can be eliminated by MFCC. They mimic the way that human ears distinguish between tones based on variations in specific frequencies.
- Chroma short-time Fourier transform (STFT): Displays the distribution of energy of an audio signal among 12 pitch classes (or chroma bins), corresponding to the 12 semitones in a musical octave. It captures the harmonics and melodies present in the sound, which are indicative of the speaker's emotional state.
- Root Mean Square (RMS) Value: Measures the magnitude of an audio signal, serving as a rough indicator for its loudness. RMS is useful for capturing speech energy and intensity that can greatly fluctuate with different emotions.
- MelSpectrogram: A spectrogram that uses mel scale to represent the frequency axis, providing detailed time-frequency representation of audio signals. MelSpectrograms give a comprehensive picture of the frequency content of the sound over time that helps to capture both temporal and spectral features of speech.

The first step towards having raw audio data interpreted by machine learning models will be the extraction of features, such as ZCR, MFCC, Chroma STFT, RMS Value, and MelSpectrogram, from the data. These features help capture other important aspects of the audio signal and can enhance the model's ability to distinguish between different emotional states in speech.

Fig. 3. Feature Extraction

VII. Model Architecture and Training

A single Convolutional Neural Network (CNN) was employed to build the classifier for voice emotion recognition. The architecture of the CNN was designed to enable it to accurately identify emotions by efficiently capturing and learning intricate information from audio data.

A. Convolutional Layers

A number of convolutional layers precede each max pooling layer in the CNN architecture. From the input audio signals, local features are extracted using convolutional layers.

- Layer 1: The ReLU activation function comes after the first convolutional layer, which is made up of 376 filters with a kernel size of 3.
- Layer 2: ReLU activation is used in the second convolutional layer, which likewise contains 255 filters and a kernel size of 3.
- Layer 3: One more of these is included, consisting of 128 filters with a size of 3 for the kernel and a hyperbolic tangent activation function.
- Layer 4: This additional layer comprises 64 filters, a hyperbolic tangent activation function, and a size of 3 for the kernel.

Whenever the data passes through these convolutions, each captures different aspects of the audio signals thus abstracting higher-level features progressively as they go through the number of data layers at once.

B. Pooling Layers

After each convolutional layer, a max-pooling layer is added with a pool size of 2 for downsampling the feature maps. This layer lowers the feature maps' spatial dimensions, which lowers the model's computational complexity. Max-pooling improves generalisation ability by reducing the number of features to only those that are truly significant, hence avoiding fitting the model too closely to the training set. Some translation invariance is also included inside the pooling function, which makes the model even more robust to tiny shifts and picture distortions in the input data.

C. Dropout Layers

Dropout layers have been included after several layers to prevent overfitting. During the training phase, a random fraction rate of the input units is set to 0.

- Dropout Layer 1: Applied with a dropout rate of 0.5 immediately following the third max pooling layer.
- Dropout Layer 2: Applied immediately following the dense layer, with a 0.5 dropout rate

D. Fully Connected Layers

The output will be flattened and fed into fully connected (dense) layers for classification after the convolutional and pooling layers.

- Dense Layer 1: This will have 32 units with ReLU activation.
- Dense Layer 2: The last dense layer will be of size 7 with a softmax activation, corresponding to the 7 emotion classes.

E. Training

The procedures listed below demonstrate how to create a graph of accuracy and loss during the testing and training stages. This will demonstrate the model's performance in detail.

- Model Training: The Adam optimiser was used for 30 epochs with a learning rate of 0.001. The loss function in this instance of multi-class classification was categorical cross-entropy. To minimise loss and maximise potential classification accuracy, the model is trained by iteratively going through the dataset and modifying the model's weights.
- Performance tracking: Throughout the training process, the performance parameters for validation and training, such as accuracy and loss, have been regularly examined. This enables the hyperparameters to be adjusted for optimal performance if necessary.
- Plotting: Accuracy and loss values were noted at each training and validation epoch, and a graph was created using these numbers. The y-axis displayed the accuracy and loss numbers, while the x-axis showed the total number of epochs. To compare the model's performance over time, the plots for the training and validation sets were given back as separate lines.

```
8042/8042 [==============================] - 1s 120us/step
Accuracy of our model on test data :  61.626458168029785 %
```

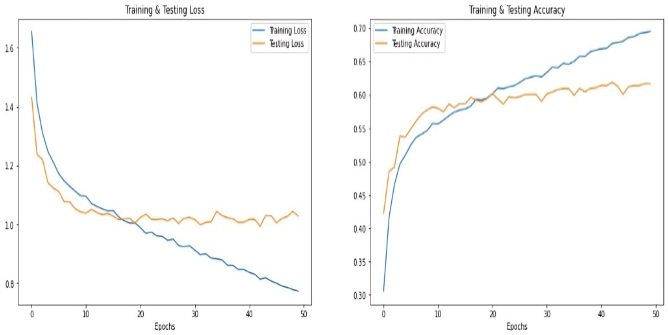

Fig. 4. Model Training

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

Fig. 5. Model Prediction

VIII. Evaluation

On testing, the speech emotion recognition model reached an accuracy of about 0.62. With this result, performance indicates that the model rightly identifies the emotion in most cases, although much is yet to be improved.

The model's performance across the many emotion classes is clearly broken out in the confusion matrix. The matrix makes it easier to analyse where the model performs well or poorly by displaying the proportion of accurate and inaccurate predictions made in each class.

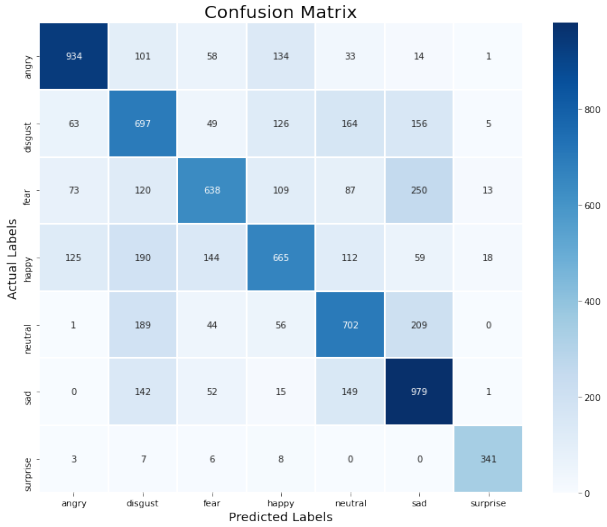

Fig. 6. Confusion Matrix

This confusion matrix depicts that the model does an outstanding job in classifying emotions such as 'surprise' and 'sad'. The high accuracy in the prediction of 'surprise' might be explained by the fact that this emotion generally has more unique audio features, which make the model more capable of identifying it. Similarly, pitch and tone for 'sad' emotions are usually unique and hence easily separable.

The classification report includes extensive metrics for each emotion class, including precision, recall, and F1-score. These measurements provide a thorough analysis of the model's performance, pointing out both its advantages and disadvantages.

	precision	recall	f1-score	support
angry	0.78	0.73	0.76	1275
disgust	0.48	0.55	0.52	1260
fear	0.64	0.49	0.56	1290
happy	0.60	0.51	0.55	1313
neutral	0.56	0.58	0.57	1201
sad	0.59	0.73	0.65	1338
surprise	0.90	0.93	0.92	365
accuracy			0.62	8042
macro avg	0.65	0.65	0.65	8042
weighted avg	0.62	0.62	0.62	8042

Fig. 7. Classification Report

Recall gauges the model's capacity to locate every pertinent example in a dataset, whereas precision assesses the accuracy of the positive predictions. The F1-score is a statistic that offers a balance between recall and precision.

- Surprise: The model produced a high F1-score (0.92) based on its high recall (0.93) and precision (0.90). This suggests that the model is quite good at identifying feelings of surprise.
- Angry: With a recall of 0.73 and a precision of 0.78, the model likewise fared well in recognising angry emotions.
- Disgust and Fear: In terms of recognising disgust and fear, the model does a passable job, with F1-scores of 0.52 and 0.56, respectively.
- Happy, Neutral, and Sad: With F1 values ranging from 0.55 to 0.65, the model demonstrated respectable performance for

happy, neutral, and sad emotions.

IX. Inferences

The analysis of the confusion matrix and classification report reveals several key insights:

- High Distinguishability: Emotions like surprise and anger, very different in audio signature, are easier for the model to recognize.
- Moderate Performance: The model will have more difficulties with emotions like disgust and fear, whose audio features may overlap.
- Balanced Accuracy: The model maintains balanced performance across most emotions, although visible variance in the accuracy can be noted.

To enhance the speech emotion recognition model's capacity for generalisation, the dataset's size and diversity can be expanded. Secondly, to improve the robustness of the model, sophisticated augmentation methods like time stretching and pitch shifting can be used on the dataset. Deeper CNN or RNN neural network topologies can be tested, as well as more complex feature extraction techniques that can be refined. Contextual features also can be used in order to maintain class correctness and relevance, and the model can be updated on a regular basis with fresh data.

X. Results

It attested to an overall test set accuracy of 0.61. Performance varied greatly between the various emotion classes. With a 0.92 accuracy rate for "surprise" and a 0.78 accuracy rate for "angry," the model demonstrated excellent performance. These feelings have more distinctive auditory characteristics, which the model may have successfully identified. On the other hand, with 0.52 and 0.56 F1-scores for "disgust" and "fear," respectively, the performance was subpar. The reason for this low accuracy is that these emotions are difficult to discern from one another due to overlapping acoustic features. In the range of 0.55 to 0.65 F1-scores, the class model demonstrated medium accuracy for "happy," "neutral," and "sad" responses.

XI. Conclusion

In summary, this work sheds information on the possible effects of machine learning and artificial intelligence on speech emotion recognition (SER), a type of emotion recognition technology. We put out a deep learning model for speech-based emotion identification that is incredibly accurate. It addresses the challenge between a complete representation of acoustic features and extensive feature-tuning work based on experimental results. Technology with an emotional intelligence nature will definitely facilitate better understanding and interaction between human and machine. The progress obtained in this research can be used to develop better applications for enhancing user experience in real-life scenarios as SER becomes more common to understand and reproduce human emotion. The results and techniques presented in this paper would contribute towards the continuous efforts in enrooting emotional intelligence into technology; entering an era where technologies are aware of users' emotions.

XII. Future Work

1) Model Optimization and Enhancement: Improvising and fine-tuning the classical learning model to attain more accurate and generalized sound digital files. Thus, the usage of dissimilar kinds of NN architectures, for example, the performance of RASINGE-Transformer that got the fifth place at Scale, the use of More-Transformer and the successful trial results with RASINGE-5 that got the second place at Scale, has facilitated capturing the necessary emotional features in the speech.
2) Feature Engineering and Selection: Increasing the significant acoustic feature representation through the combination of new types of hardware and the use of the method of feature selection. It should be thoroughly investigated over the model that is more appropriate with feature extraction since the subtle emotional cues are enigma that requires the usage of techniques like prosody, tone, and speech rate.
3) Cross-Linguistic and Cross-Cultural Adaptability: The objective is to create models reserved and trained for a particular linguistic and cultural environment and that the test occurs in many different linguistic and cultural environments to guarantee that these models work in all these different linguistic and cultural environments. Thereby gaining insights into the emotional expressions of how people differ in various countries and also strengthening the model in its adaptability aspect.
4) Real-Time Processing and Deployment: The model's primary objective is to be used in real-time, and this can be achieved by lowering the computational load and delay time. To make SER useful in practical applications, hardware acceleration methods and effective algorithms should be utilised.
5) Integration with Other Modalities: In order to fully comprehend emotional recognition, research is being done on the integration of speech emotion recognition with other modalities like physiological signals or facial expression analysis. When multimodal approaches were used, emotion recognition performance and reliability significantly increased.
6) User Experience and Feedback: Performing UX research to get inputs on the working of the system in real-life scenarios. Improve user experience by the obtained feedback that will be the base of the continuous improvements posed to fit the system to user needs and characteristics.
7) Ethical Considerations and Privacy: Addressing privacy issues related to emotion detection technology, including consent and potential threat to the user's private data. The procedure of setting the standards and protocols, that will lead to feasible and ethical usage of the technology as well as clear and transparent disclosure, should be invented.
8) Scalability and Generalization: Being in the cloud allows people to manipulate it to expand or contract when handling a larger amount of data. The steady model performance across various physical environments and exact scenarios is achieved by the model

References

[1] Liu, G., Cai, S., Wang, C. "Speech emotion recognition based on emotion perception." EURASIP Journal on Audio, Speech, and Music Processing 2023, 22.

[2] Khalil, R. A., Jones, E., Babar, M. I., Jan, T., Zafar, M. H., Alhussain, T. "Speech Emotion Recognition Using Deep Learning Techniques: A Review." IEEE Access 7 (2019): 115749–115767.

[3] Aouani, H., Ben Ayed, Y. "Speech Emotion Recognition with Deep Learning." In: 24th International Conference on Knowledge-Based and Intelligent Information and Engineering Systems, 2020. MIRACL, University of Sfax, Tunisia.

[4] Shaila, S. G., Sindhu, A., Monish, L., Shivamma, D., Vaishali, B. "Speech Emotion Recognition Using Machine Learning Approach." Department of CSE (Data Science), Dayananda Sagar University, Bangalore, Karnataka, India.

[5] Lieskovská, E., Jakubec, M., Jarina, R., Chmulík, M. "A Review on Speech Emotion Recognition Using Deep Learning and Attention Mechanism." Electronics 10 (10) (2021): 1163.

[6] Byun, S.-W., Lee, S.-P. "A Study on a Speech Emotion Recognition System with Effective Acoustic Features Using Deep Learning Algorithms." Applied Sciences 11 (4) (2021): 1890.

[7] Singh, J., Saheer, L. B., Faust, O. "Speech Emotion Recognition Using Attention Model." International Journal of Environmental Research and Public Health 20 (6) (2023): 5140.

[8] Aswani, R., Gawale, A., Dhawale, B., Shivade, A., Donde, N., Tambe, U. "Speech Emotion Recognition." International Journal of Creative Research Thoughts (IJCRT) 9 (5) (2021): e123.

[9] Ingale, A. B., Chaudhari, D. S. "Speech Emotion Recognition." International Journal of Soft Computing and Engineering (IJSCE) 2 (1) (2012): 61-65.

[10] Utane, A. S., Nalbalwar, S. L. "Emotion Recognition through Speech." International Journal of Applied Information Systems 5 (1) (2013): 1–7.

[11] Koolagudi, S. G., Rao, K. S. "Emotion Recognition from Speech: A Review." International Journal of Speech Technology 15 (2012): 99–117.

[12] Rawat, A., Mishra, P. K. “Emotion Recognition through Speech Using Neural Network.” International Journal of Advanced Research in Computer Science and Software Engineering 5 (5) (2015): 1–7.
[13] Wanare, A. P., Dandare, S. N. “Human Emotion Recognition From Speech.” International Journal of Engineering Research and Applications 4 (7) (2014): 74–78.
[14] Kwon, O.-W., Chan, K., Hao, J., Lee, T.-W. “Emotion Recognition by Speech Signals.” Institute for Neural Computation, University of California, San Diego, USA.
[15] Wani, T. M., Qadri, S. A. A., Gunawan, T. S., Mirakartwi, M., Ambikairajah, E. “A Comprehensive Review of Speech Emotion Recognition Systems.” IEEE Access 9 (2021): 70232–70251.
[16] Khan, M., Gueaieb, W., El Saddik, A. “MSER: Multimodal Speech Emotion Recognition Using Cross-Attention with Deep Fusion.” Expert Systems with Applications 219 (2023): 122946.
[17] Crowd-sourced Emotional Mutimodal Actors Dataset (Crema-D) (Accessed on 20 Jun 2024)
[18] Ryerson Audio-Visual Database of Emotional Speech and Song (Ravdess) (Accessed on 20 Jun 2024)
[19] Surrey Audio-Visual Expressed Emotion (Savee) (Accessed on 20 Jun 2024)
[20] Toronto Emotional Speech Set (Tess) (Accessed on 20 Jun 2024)