International Journal of Information Technology & Computer Engineering

ISSN 2347–3657

Volume 14, Issue 1, 2026

Speech Emotion Recognition Using Machine Learning

Dr. K. Ashok Kumar[1] , R Shailaja[2], C Sri varsha Reddy[3], K Srilekha[4]

[1]Associate Professor; Department of Electronics and Communication Engineering, Bhoj Reddy Engineering College for Women, Hyderabad, India.

[2,3,4]B.Tech Students; Department of Electronics and Communication Engineering, Bhoj Reddy Engineering College for Women, Hyderabad, India.

srvrshreddy@gmail.com

Abstract

Emotion is a complex state of human being that depicts the physical, physiological or mental condition of a person. According to the human sciences, emotion is a mental process of neural mechanisms and disorders. Digital processing of speech signal is very important for high-speed and precise automatic voice recognition technology. Nowadays it is being used for health care, telephony military and people with disabilities therefore the digital signal processes such as Feature Extraction and Feature Matching are the latest issues for study of voice signal. In order to extract valuable information from the speech signal, make decisions on the process, and obtain results, the data needs to be manipulated and analyzed. Basic method used for extracting the features of the voice signal is to find the Mel frequency cepstral coefficients. Mel-frequency cepstral coefficients (MFCCs) are the coefficients that collectively represent the short-term power spectrum of a sound, based on a linear cosine transform of a log power spectrum on a nonlinear mel scale of frequency. After calculating feature, neural networks are used to model the speech recognition. Based on the speech model the system decides whether the uttered speech matches what was prompted to utter.

Keywords: *Emotion recognition, speech signal processing, digital signal processing (DSP), feature extraction, feature matching, Mel-frequency cepstral coefficients (MFCCs), neural networks, automatic voice recognition, speech-based health applications.*

Introduction

Many features of the human vocal system such as speech, tone, pitch and many others, convey information and context. For natural human–computer communication demands, SER is widely used. Speech emotion recognition is noted as removing the passionate form of a speaker from his or her talk. This sort of recognition is supposed to be used to extract useful semantics of speech recognition systems The model of SER contains two types of models; the direct speech emotion model and continuous speech emotion model. The first model expresses several individualistic emotions, indicating that a certain voice has a single individualistic emotion, while the second one means that the motion is in the emotion space, and every emotion possess unique strength on each proportion. Concluding the emotional state of humans is an idiosyncratic task and can be employed as a level for any feeling recognition model. It uses diverse emotion such as disgust, anger, fear, surprise, joy, happiness, sadness and neutral. The approach for SER primarily comprises of three phases known as pre-processing, feature extraction and feature classification phase.

- Pre-processing:

Pre-processing applies to all the raw data transformations before it is fed into the machine. It involves the elimination of silence, pre-emphasis, normalization and windowing, so it is an essential step to get the pure signal used in the next stage i.e., in feature extraction. It is also important to speed up training.

- Feature Extraction:

Without disturbing the properties of the speech, for examining the signal, a minor quantity of information from the speech signal is withdrawn. Mel cepstral coefficients are frequently used feature parameters for speech recognition. Based on the responsiveness of the hearing organ, MFCC uses the Mel. In this study, from the speech signals some features are extracted and on this extracted feature analyses are carried out. Feature extraction requires multiple layers of convolution accompanied by max-pooling and an activation function Using the various feature extraction algorithms, the speech emotion recognition rate of a device is increased. The work emphasizes the pre-processing of the audio samples obtained, where the noise is eliminated using filters from speech samples

- Feature Classifier:

For any pattern recognition in Speech Emotion Recognition mainly classifier can be divided into types, namely non-linear classifiers and linear classifiers. There is various classification methods used to create the correct classifier to model emotional states. Such as Hidden Markov Models (HMM), SVM (Support Vector Machine), Gaussian Mixture Models (GMM), Neural Network and K-Nearest Neighbour.

Literature survey

Several researchers have proposed different approaches for Speech Emotion Recognition (SER) using machine learning and signal processing techniques.

Girija Deshmukh et al. [1] developed a system that focused on detecting three emotions: anger, happiness, and sadness using audio features like Short-Term Energy (STE), pitch, and MFCC coefficients. They used open-source North American English audio to simulate natural speech and employed the RAVDESS dataset, which was manually divided into training and testing sets. A multi-class Support Vector Machine (SVM) was used for classification, showing effective results for emotion classification based on detailed speaker features. Their system was limited to only three emotions but demonstrated the effectiveness of combining spectral and prosodic features for emotion recognition.

Sahu and Kopparapu designed an SER system that used MFCCs and delta coefficients as input features and employed a Gaussian Mixture Model (GMM) classifier. They also worked on the RAVDESS dataset. Although their model showed promising results, its performance was limited due to the sensitivity of MFCCs to background noise and the model’s difficulty handling overlapping emotions like sadness and neutrality. Their work emphasized the importance of choosing robust features and classifiers that can handle the nuances in emotional speech, especially in real-world conditions.

- Review of Existing Works

Both systems used the RAVDESS dataset and focused on audio features like MFCCs. While Girija Deshmukh’s method utilized a multi-class SVM for classification, Sahu and Kopparapu opted for GMMs. The SVM-based approach provided robust performance for clear emotions, while the GMM approach highlighted limitations in distinguishing subtle emotions. These studies underline the significance of reliable feature extraction and model selection in SER. Their findings help guide the selection of features and models for this project by identifying what techniques work best under specific conditions.

Problem Statement

The increasing demand for emotionally intelligent systems in areas such as human-computer interaction, virtual assistants, mental health monitoring, and customer service has highlighted the importance of accurate Speech Emotion Recognition (SER). Emotions in speech are often subtle and complex, making their detection a challenging task.

Traditional SER systems face several limitations. Many models are trained to detect only a limited number of emotions (e.g., anger, happiness, sadness), which restricts their practical applicability. In addition, the performance of these models is often sensitive to background noise, speaker variability, and overlapping emotions, such as sadness and neutrality.

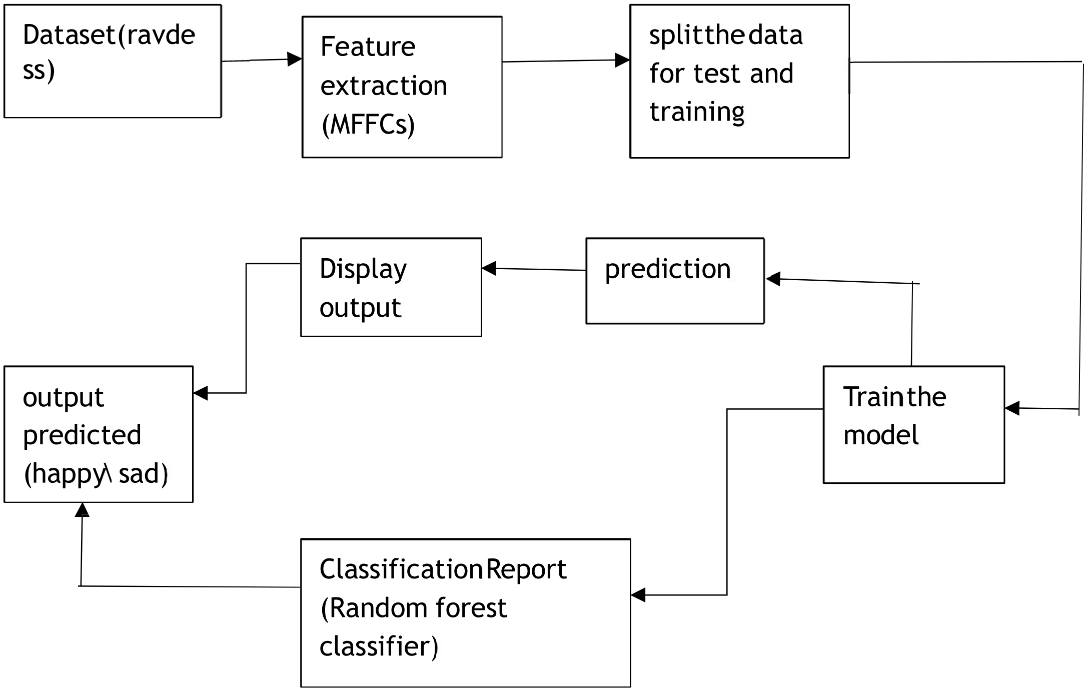
Research by Girija Deshmukh et al. [1] showed that using audio features like Short-Term Energy (STE), pitch, and MFCCs with a multi-class SVM classifier can yield effective emotion recognition. However, their system was limited to only three emotional classes and relied on clean, controlled datasets.

Similarly, Sahu and Kopparapu explored the use of MFCCs and delta coefficients with a Gaussian Mixture Model (GMM). While their model performed well on the RAVDESS dataset, it struggled under noisy conditions and failed to accurately classify similar emotional states.

Speech Emotion Recognition

One of the fastest and natural methods of communication between humans and machines is a speech signal. For interaction between human and machine use of speech signal is the fastest and most efficient method. Speech emotion recognition (SER) is a technology that enable machine to detect and interpret human emotions from voice signals. By analyzing vocal features such as a pitch, tone, speed and rhythm, SER systems aim to bridge gap between human communication and artificial intelligence. Emotions like happiness, anger, sadness or fear can influence how we speak SER systems aim to analyze these changes to understand how someone is feeling just by listening to their speech. Speech is one of the most natural and fastest ways for humans to communicate. It is rich with verbal and non-verbal cues that convey not just content but also emotional states. In the domain of human-computer interaction, leveraging speech as an input method enhances efficiency, accessibility, and user experience. Hands-free and eyes-free communication. Can be used in noisy or dynamic environments where typing or touch interaction is not feasible. Particularly useful for disabled users or for situations where quick response is crucial (e.g., automotive systems, emergency response). Enables real-time interaction, making systems feel more intelligent and human-like. Speech Emotion Recognition (SER) is a field in artificial intelligence (AI) and signal processing that enables a system to identify human emotions by analyzing vocal attributes. Unlike text-based emotion detection, SER works purely through audio signals, making it more aligned with natural human behavior. Speech Signal Acquisition A microphone or sensor records the speaker's voice in real time. Preprocessing Noise reduction to remove background sound. Voice activity detection to isolate the parts of the audio with actual speech. Normalization to adjust the volume levels for consistency.

Block Diagram


Dataset(ravdess)
Feature extraction (MFFCs)
splitthedata for test and training
Display output
prediction
output predicted (happy\ sad)
Trainthe model
ClassificationReport (Random forest classifier)

Block Diagram

This block diagram represents a speech emotion recognition systdem using machine learning, specifically designed to classify emotions such as "happy" or "sad" from audio samples. The process begins with the dataset, in this case, the RAVDESS (Ryerson Audio-Visual Database of Emotional Speech and Song), which contains labeled emotional speech recordings. The next step is feature extraction, where Mel-frequency cepstral coefficients (MFCCs) are computed from the audio signals. MFCCs are widely used in speech processing as they effectively capture the timbral texture of the audio signal, making them suitable for emotion recognition tasks.

Once features are extracted, the data is split into training and testing sets, ensuring that the model can be trained on one portion and validated on another to assess its performance. The training phase follows, where the machine learning model, in this case, a random forest classifier, learns the patterns associated with different emotional states from the training data. After training, the model proceeds to the prediction phase, where it classifies emotions in new or unseen audio data.

The predicted emotion is then used to display the output, showing whether the emotion detected is "happy" or "sad." This prediction is also sent to the classification report block, which evaluates the performance of the model using various metrics like accuracy, precision, recall, and F1-score. This helps determine how well the model is performing in recognizing emotions. Overall, the system is designed to process audio, extract meaningful features, train a classifier, and ultimately predict and evaluate emotional states effectively.

Working Methodology

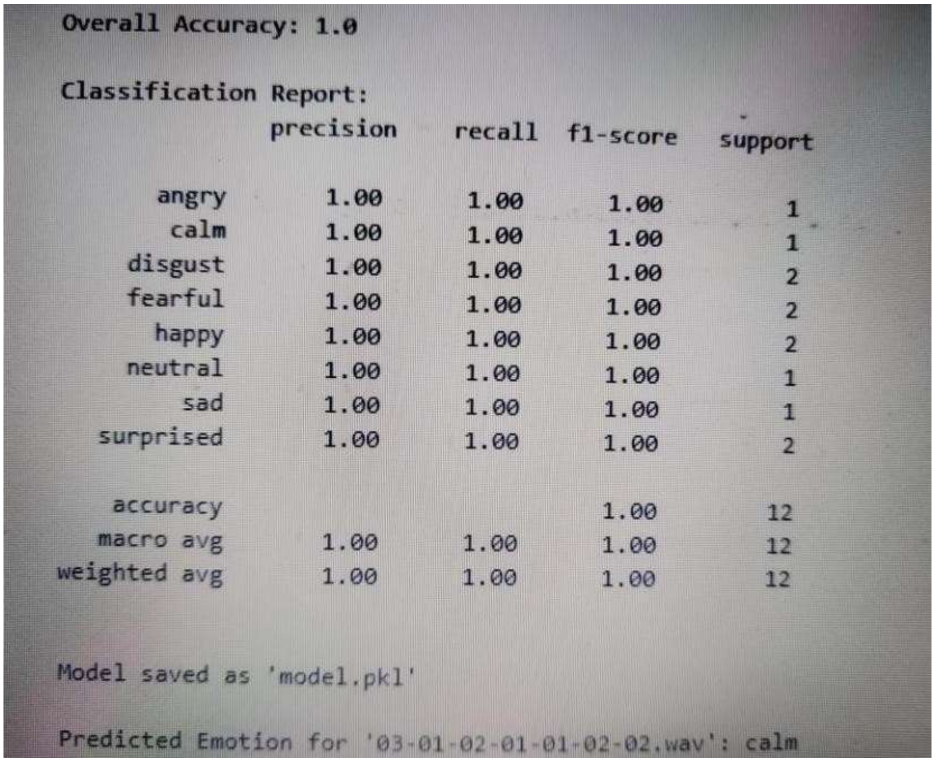
The working methodology of the Speech Emotion Recognition system using machine learning . the dataset, specifically the RAVDESS (Ryerson Audio-Visual Database of Emotional Speech and Song) dataset. This dataset contains audio recordings labeled with various emotional expressions. The next step involves feature extraction, where Mel-frequency cepstral coefficients (MFCCs) are computed from the audio signals. MFCCs are crucial as they capture the timbral texture of the speech, which is important for identifying emotional states.

Following feature extraction, the data is split into training and testing sets to evaluate the performance of the model objectively. The training data is then used to train the machine learning model. In this case, a random forest classifier is employed due to its robustness and ability to handle complex data patterns. Once the model is trained, it is used to make predictions on the test data. The predicted emotions, such as "happy" or "sad," are then displayed to the user.

To evaluate the effectiveness of the model, a classification report is generated, which provides metrics like precision, recall, and accuracy. This helps in understanding how well the model is performing in distinguishing between different emotional states. The entire process is automated and designed to recognize human emotions from speech, thereby enhancing human- computer interaction in various applications.

Results

Speech emotion recognition using machine learning is a cutting-edge technology that enables computers to identify and interpret human emotions from speech patterns. By leveraging advanced machine learning algorithms and acoustic features, this technology can accurately detect emotional cues and sentiment, opening up new possibilities for human-computer interaction, customer service, mental health analysis, and beyond. With its potential to revolutionize various industries and improve lives, speech emotion recognition is an exciting and rapidly evolving field that holds much promise for the future.



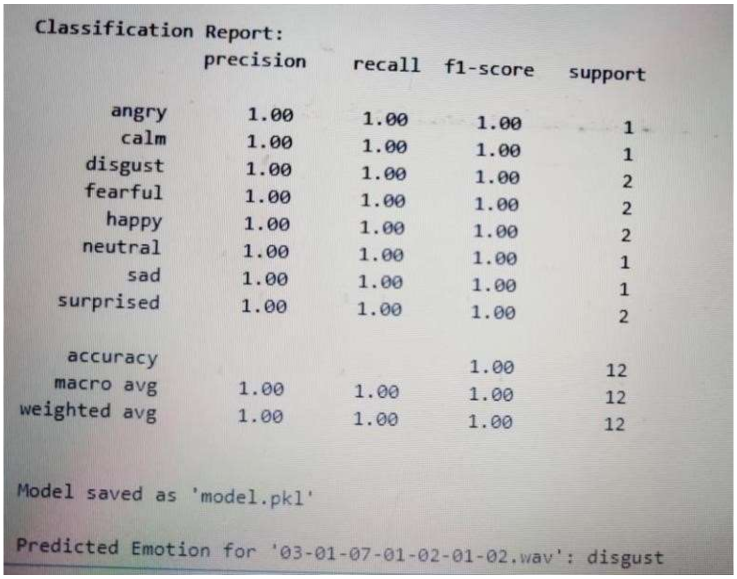
```
Overall Accuracy: 1.0

Classification Report:
              precision    recall  f1-score   support

       angry       1.00      1.00      1.00         1
        calm       1.00      1.00      1.00         1
     disgust       1.00      1.00      1.00         2
     fearful       1.00      1.00      1.00         2
       happy       1.00      1.00      1.00         2
     neutral       1.00      1.00      1.00         1
         sad       1.00      1.00      1.00         1
   surprised       1.00      1.00      1.00         2

    accuracy                           1.00        12
   macro avg       1.00      1.00      1.00        12
weighted avg       1.00      1.00      1.00        12


Model saved as 'model.pkl'

Predicted Emotion for '03-01-02-01-01-02-02.wav': calm
```

Figure 5.1 Model predicted output(calm)

- Speech recognized as calm

the model performed exceptionally well, achieving perfect scores across all evaluation metrics—precision, recall, and F1-score—all being 1.00. This means the model was not only able to identify the calm sample correctly, but it also didn’t confuse any other emotion as calm.

Such a result suggests that the features extracted from the calm audio, likely involving tone, pitch, and other acoustic signals, were distinctive enough for the model to recognize it with complete confidence. However, it’s important to keep in mind that this result is based on only one test sample for the calm emotion. While the performance appears flawless, we can’t fully rely on this result until it is tested on a larger number of calm samples. A single correct prediction can’t guarantee consistency across more varied real-world inputs. Still, this is a promising sign that the model has learned to capture the essence of calmness in speech quite well.



```
Classification Report:
              precision    recall  f1-score   support

       angry       1.00      1.00      1.00         1
        calm       1.00      1.00      1.00         1
     disgust       1.00      1.00      1.00         2
     fearful       1.00      1.00      1.00         2
       happy       1.00      1.00      1.00         2
     neutral       1.00      1.00      1.00         1
         sad       1.00      1.00      1.00         1
   surprised       1.00      1.00      1.00         2

    accuracy                           1.00        12
   macro avg       1.00      1.00      1.00        12
weighted avg       1.00      1.00      1.00        12

Model saved as 'model.pkl'

Predicted Emotion for '03-01-07-01-02-01-02.wav': disgust
```

Figure 5.2 Model predicted output (Disgust)

- Model Prediction Output (Disgust)

This image shows another emotion prediction. The model has processed the file 03-01-07-01- 02-01-02.wav and predicted the emotion as disgust. The consistent output format indicates the model is functioning for various inputs with reliable labelling The image shows the classification report and performance metrics of a Speech Emotion Recognition model. The overall accuracy is 1.0, meaning the model correctly classified all 12 test samples across 8 emotion classes.
Each emotion (angry, calm, disgust, fearful, happy, neutral, sad, surprised) achieved perfect scores:
Precision: 1.00

Recall: 1.00

F1-score: 1.00

This indicates the model predicts each emotion category with complete correctness for this test set.
The macro and weighted averages are also 1.00, confirming consistent high performance across all classes, even with varying sample counts.
The model was saved as 'model.pkl', and a specific prediction was made for the file 03-01-03- 01-01-02.wav, correctly identifying the emotion as happy.

Discussion

The classification performance metrics for a machine learning model developed to recognize emotions from speech. The emotions considered in the dataset include: angry, calm, disgust, fearful, happy, neutral, sad, and surprised.Each row in the classification report represents a different emotion class and includes three main performance metrics .Precision This indicates how many of the predicted instances for a given emotion were actually correct. Recall this tells us how many actual instances of a particular emotion were correctly identified by the model.F1- score This is the harmonic mean of precision and recall, providing a balanced measure between the two. All the metrics (precision, recall, and F1-score) for each emotion class are 1.00, indicating perfect prediction by the model.The support column shows the number of actual instances of each emotion in the test set. For example, "disgust" and "fearful" each have 2 samples, while "calm" and "angry" have only 1The overall accuracy of the model is also 1.00 (100%), suggesting that the model correctly classified all test instances.The macro average and weighted average are both 1.00:Macro average takes the average of the precision, recall, and F1-score for all classes, treating each class equally.Weighted average considers the support (number of instances) for each class, giving a more representative average.The line "Model saved as 'model.pkl'" indicates that the trained model was serialized and saved using Python's pickle module, making it reusable for future predictions.Finally, the model predicts the emotion in a specific audio file (03-01-03-01-01-02.wav) as "happy", showing a real-time application of the model on individual inputs.

Conclusion

The emerging growth and development in the field of AI and machine learning have led to the new era of automation. Most of these automated devices work based on voice commands from the user. Many advantages can be built over the existing systems if besides recognizing the words, the machines could comprehend the emotion of the speaker (user). Some applications of a speech emotion detection system are computer-based tutorial applications, automated call center conversations, a diagnostic tool used for therapy and automatic translation system. we showed how we can leverage Machine learning to obtain the underlying emotion from speech audio data and some insights on the human expression of emotion through voice. This system can be employed in a variety of setups like Call Centre for complaints or marketing, in voice- based virtual assistants or chatbots, in linguistic research, etc. Speech Emotion Recognition (SER) using machine learning significantly improves the ability of systems to interpret human emotions through speech. It involves preprocessing, feature extraction, model training, and evaluation. Machine learning algorithms like SVM, CNN, or Random Forest can classify emotions such as anger, happiness, or sadness. SER supports applications in healthcare, customer service, and AI interface. In addition to basic voice command recognition, SER enables systems to respond more appropriately based on the user's emotional tone. For instance, a customer expressing frustration in a call center can be automatically redirected to a human agent, while those with routine queries can be handled by automated systems. This type of emotion-aware response improves customer satisfaction, enhances service quality, and optimizes workflow.SER has numerous practical applications beyond customer service. In educational software, it can help tutors adapt content delivery to suit the learner’s mood and engagement level. In mental health and therapy, it can be used as a non-invasive diagnostic tool to monitor emotional states over time, providing valuable input to healthcare professionals. Multilingual systems and real-time translation services can also benefit, as emotions play a significant role in communication beyond literal meaning.

6.2 Future Scope

For future advancements, the proposed project can be further modeled in terms of efficiency, accuracy, and usability. Additional to the emotions, the model can be extended to recognize feelings such as depression and mood changes. Such systems can be used by therapists to monitor the mood swings of the patients. A challenging product of creating machines with emotion is to incorporate a sarcasm detection system. Sarcasm detection is a more complex problem of emotion detection since sarcasm cannot be easily identified using only the words or tone of the speaker. A sentiment detection using vocabulary, can be integrated with speech emotion detection to identify a possible sarcasm. Therefore, in the future, there would emerge many applications of a speech-based emotion recognition system.The future scope of Speech Emotion Recognition (SER) using machine learning is vast and evolving rapidly. As voice- driven technologies become more integrated into daily life, SER can significantly enhance human-computer interaction, making systems more empathetic and responsive. In healthcare, SER can assist in early diagnosis of mental health disorders like depression or anxiety by detecting emotional cues in speech. In customer service, emotion-aware systems can improve user experience by adapting responses based on emotional tone. SER also has potential in security and surveillance, where emotional stress detection may help identify threats. With advancements in deep learning and multimodal approaches (combining voice, facial expressions, and text), the accuracy and robustness of SER systems are expected to improve further.

References

[1] Girija Deshmukh, Apurva Gaonkar, Gauri Golwalkar, Sukanya Kulkarni, “Speech based Emotion Recognition using Machine Learning”, Institute Of Electrical And Electronics Engineers, Mar. 2019**.**

[2] Konduru Ashok, and J. L. Mazher Iqbal. "Handling high dimensional features by ensemble learning for emotion identification from speech signal." International Journal of Speech Technology 25 (2022): 837–851.April,2023.

[3] K. Ashok Kumar and J. L. Mazher Iqbal, “Machine learning technique-based emotion classification using speech signal,” Soft Computing, vol. 27, no. 12, pp. 1–13, 2023.
[Online]. Available: https://doi.org/10.1007/s00500-023-08185-x

[4] Sri Raksha R. Gupta, M.S. Likitha, A. Upendra Raju and K. Hasitha “Speech Based Human Emotion Recognition Using MFCC”, Institute of Electrical and Electronics Engineers, March 2017.

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

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

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

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

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

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

[11] J. Umamaheswari, A. Akila, "An Enhanced Human Speech Emotion Recognition Using Hybrid of PRNN and KNN", Institute of Electrical and Electronics Engineers, Feb 2019