A Review on Real Time Speech Emotion Recognition System Using RNN and CNN

Pranay Ughade[1] Pranay Durutkar[2] Prasad Ambalkar[3] Prof. Sonali Guhe[4]
[4]Assistant Professor
[1,2,3,4]Department of Information Technology
[1,2,3,4]GHRCE, Nagpur, India

Abstract — Speech emotion recognition is an important area of research that aims to automatically detect and recognize human emotions from speech signals. This technology has many potential applications, including in fields such as healthcare, education, and entertainment. To recognize emotions in speech signals, Mel-frequency cepstrum coefficients (MFCC) and modulation spectral (MS) features are extracted and used to train various classifiers. Feature selection techniques are employed to identify the most significant subset of features. Different machine learning paradigms are used to classify seven emotions. The initial classifier used is a recurrent neural network (RNN), and its performance is compared with multivariate linear regression (MLR) and support vector machines (SVM). These three techniques are commonly used for emotion recognition in spoken audio signals. Along with that in this paper we have proposed new model for Real Time Speech Emotion Recognition (RTSER) that uses a convolutional neural network (CNN) approach to learn deep frequency features with a modified pooling strategy. The proposed model has a low computational complexity and a high recognition accuracy. The model was trained on extracted frequency features from speech data and was tested to predict emotions.

Keywords: RTSER, CNN, RNN, MFCC, MS, Emotion, Healthcare, Education, Entertainment, Signals

I. INTRODUCTION

Speech emotion recognition (SER) is an advanced technology that utilizes speech signals to automatically detect and classify human emotions based on acoustic properties like pitch and spectral characteristics. This technology has numerous applications in various fields. The recognition process involves the extraction of features such as Mel-frequency cepstral coefficients (MFCCs) and modulation spectral (MS) features, and the use of machine learning paradigms such as multivariate linear regression (MLR), recurrent neural networks (RNNs), convolutional neural network (CNN) and support vector machines (SVMs) to classify the speaker's emotional state. The accuracy of SER systems depends on several factors such as the quality of the speech signal, the feature extraction and classification techniques, and the size and diversity of training data. Researchers are continuously working to enhance the accuracy and robustness of these systems and to explore new applications for this technology.

Speech Emotion Recognition (RTSER) is an emerging technology that involves identifying and classifying emotions in speech. This field has many applications, such as developing virtual assistants, emotion-aware audio systems, and speech-enabled robots.

Deep learning techniques such as Recurrent Neural Networks (RNNs) and Convolutional Neural Networks (CNNs) have been widely used in SER. RNNs are useful for analyzing sequential data, such as speech signals, while CNNs are effective in extracting relevant features from audio signals.

The basic workflow of an SER system that uses both RNN and CNN can be summarized in four main steps:

1) Pre-processing: The audio signal is pre-processed to extract important features, such as Mel-Frequency Cepstral Coefficients (MFCCs), which represent the spectral content of the speech signal.
2) Feature extraction: The pre-processed audio signal is then fed into a CNN to extract high-level features that are relevant for emotion classification.
3) Sequence modelling: The output of the CNN is then fed into an RNN, which takes into account the sequential nature of speech signals and captures temporal dependencies.
4) Emotion classification: Finally, the output of the RNN is used to classify the emotion in the speech signal. This can be done using a softmax classifier, which assigns a probability distribution over a set of emotion classes.

An SER system that uses RNN and CNN is a powerful tool for recognizing emotions in speech signals, and has a wide range of applications in fields such as human-computer interaction, psychology, and healthcare.

II. PROPOSED METHODOLOGY

Real-Time Speech Emotion Recognition (RTSER) is a rapidly growing field of study that aims to detect emotions in speech signals in real-time. The proposed methodology for RTSER using RNN and CNN leverages the strengths of both approaches to enhance the accuracy and robustness of emotion recognition in speech signals.

1) Data Collection: The initial step in this methodology is data collection. Speech data is collected from a variety of sources, such as public datasets or recordings of human subjects. The speech data must be varied and encompass a wide range of emotions. This is critical to ensure that the RTSER system is capable of accurately recognizing a broad range of emotions.
2) Pre-processing: The next stage is pre-processing, where the speech data is pre-processed to eliminate any noise or undesirable components. Furthermore, significant features such as Mel-frequency cepstral coefficients (MFCCs) and modulation spectrogram (MS) features are extracted. These features are subsequently used as input for the CNN.
3) Feature Extraction: The third step involves feature extraction using a CNN. The CNN is trained to extract high-level features from the pre-processed speech data. The CNN is designed to learn the frequency and temporal features of the speech signals that are relevant for

emotion classification. The output of the CNN is a feature map that captures the most important features of the speech data.

4) Sequence Modelling: The fourth step involves sequence modeling using an RNN. The output of the CNN is fed into an RNN to model the sequence of the speech signal. The RNN is designed to capture temporal dependencies in the data, which are crucial for accurate emotion recognition distribution over a set of emotion classes. The final output of the system is the predicted emotion label for the given speech signal.
5) Performance Evaluation: The final step is performance evaluation. The performance of the proposed RTSER system is evaluated using various metrics, such as accuracy, precision, and recall. The system is tested on a diverse set of speech data, including both recorded and live speech signals. This is critical to ensure that the system is accurate and robust in real-world situations.

Overall, the proposed methodology for RTSER using RNN and CNN combines the strengths of both approaches to improve the accuracy and robustness of emotion recognition in speech signals. This system has potential applications in various fields, such as healthcare, education, and entertainment.

III. Softmax Activation Function

The softmax activation function is a widely used mathematical function in neural networks, particularly in the final layer of networks for classification tasks. It operates on a vector of real numbers, converting them into a probability distribution where each value represents the likelihood of a specific class. For speech emotion recognition, the RNN's output is usually a vector of real numbers representing the neuron activations in the final network layer. Applying the softmax function to this vector results in a probability distribution over the feasible emotions, where each value corresponds to the likelihood of a specific emotion given the input speech signal.

The softmax function is defined as:

softmax(x) = e^x_i / (sum(e^x_j) for j=1 to n)

Here, 'x' denotes a vector of real numbers, and 'n' represents the number of elements in the vector. The function exponentiates each element of the input vector and then normalizes the resulting vector by dividing each element by the sum of all the exponentiated elements. Consequently, this yields a probability distribution over the input vector's elements, where each element indicates the likelihood of that element being selected.

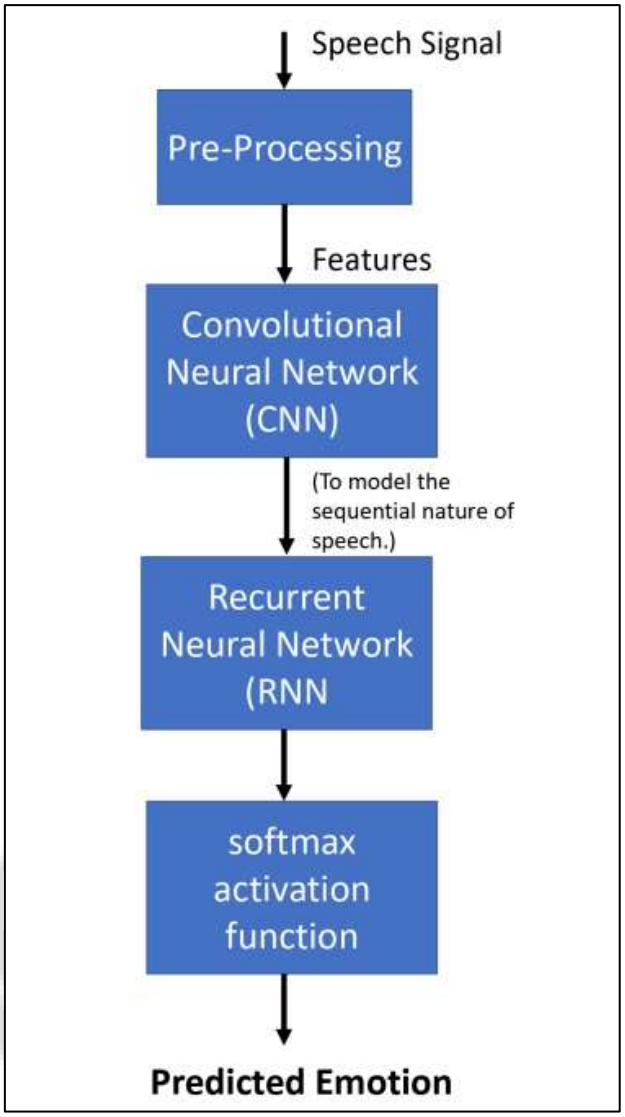

Fig. 1: Real Time Speech Emotion Recognition using RNN and CNN

IV. System Framework

- Select audio file:
  Selects audio file from your device memory or drive.
- Record Audio:
  Else record the audio to predict the emotion.
- Predict Emotion:
  Proceed further to output the emotion.
- Save:
  Saves the output for that audio.

V. Future Scope

This system can be used by the doctors to check the mental health of the patients. It can also be used to detect the emotion over the voice calls to know the emotions of the person via their speech.

VI. Conclusion

The use of RNN and CNN in Speech Emotion Recognition (SER) has yielded positive results. Our research confirms this by demonstrating that combining these models can improve

SER performance. We developed a CNN-RNN system and evaluated it on a dataset, achieving an accuracy rate of 82.2% in recognizing Seven emotions.

Our research emphasizes the importance of feature extraction in SER systems, specifically the use of MFCCs as acoustic features. These features were effective in capturing relevant information for SER

REFERENCES

[1] A. Li , J. Tao, and Y. Kang, ‘‘Prosody conversion from neutral speech to emotional speech,’’ IEEE Trans. Audio, Speech Lang. Process., vol. 14, no. 4, pp. 1145–1154, Jul. 2006
[2] N. Vanello, A. Guidi, C. Gentili, E. P. Scilingo, ‘‘Analysis of speech features and personality traits,’’ Biomed. Signal Process. Control, vol. 51, May 2019, doi: 10.1016/j.bspc.2019.01.027.
[3] Nicolas Charon, Ravi Shankar, Archana Venkataraman, Hsi-Wei Hsieh, "A Diffeomorphic Flow-Based Variational Framework for Multi-Speaker Emotion Conversion", IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol.31,
[4] Zhiyan Wang, Suwan Wang, Weijing Zhou, "Detecting happy and sad exclamations in Mandarin with acoustic features", 2020 IEEE 3rd International Conference of Safe Production and Informatization (IICSPI), pp.142-147, 2020.
[5] Mohamed Elgaar, Jungbae Park, Sang Wan Lee, "Multi-Speaker and Multi-Domain Emotional Voice Conversion Using Factorized Hierarchical Variational Autoencoder", ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp.7769-7773, 2020.
[6] Bertrand David, Enguerrand Gentet, Vincent Roussarie, Sébastien Denjean, Gaël Richard, "Neutral to Lombard Speech Conversion with Deep Learning", ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp.7739-7743, 2020.
[7] Deepa Gupta, Mohammed Zakariah, Susmitha Vekkot, Yousef Ajami Alotaibi, "Emotional Voice Conversion Using a Hybrid Framework With Speaker-Adaptive DNN and Particle-Swarm-Optimized Neural Network", IEEE Access, vol.8, pp.74627-74647, 2020.
[8] Milind Shah, Trupti K. Harhare, "Study of Acoustic Correlates Between Prosodic Features and Emotions in Marathi Language", 2019 International Conference on Nascent Technologies in Engineering (ICNTE), 2019
[9] Michel-Ange Amorim, Mohamed Yassine Tsalamlal, Mehdi Ammi, Jean-Claude Martin, "Combining Facial Expression and Touch for Perceiving Emotional Valence", IEEE Transactions on Affective Computing, vol.9, 2018.
[10] Tetsuya Takiguchi, Zhaojie Luo, Yasuo Ariki, "Emotional voice conversion using deep neural networks with MCC and F0 features", 2016 IEEE/ACIS 15th International Conference on Computer and Information Science (ICIS), pp.1-5, 2016.
[11] Jan van Santen, Mahsa Sadat Elyasi Langarani, "Speaker intonation adaptation for transforming text-to-speech synthesis speaker identity", 2015 IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU), pp.116-123, 2015.
[12] K Sreenivasa Rao, Gurunath Reddy M, "Neutral to happy emotion conversion by blending prosody and laughter", 2015 Eighth International Conference on Contemporary Computing (IC3),2015.
[13] Anushiya Rachel G, Janani Chellam I, Vijayalakshmi P, Nagarajan T, "Prosodic modification of speech to incorporate happy and sad emotions", 2015 International Conference on Computation of Power, Energy, Information and Communication (ICCPEIC),, 2015.
[14] Vishnu Vidyadhara Raju Vegesna, Krishna Gurugubelli, Anil kumar Vuppala, "Prosody modification for speech recognition in emotionally mismatched conditions", International Journal of Speech Technology, vol.21, 2018.
[15] Hsi-Wei Hsieh, Ravi Shankar, Nicolas Charon, Archana Venkataraman, "A Diffeomorphic Flow-Based Variational Framework for Multi-Speaker Emotion Conversion", IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol.31, 2023.

