SPEECH EMOTION RECOGNITION USING ADRNN CLSTM HYBRID ARCHITECTURE

1. Sudhakar S, MCA Student, PES Institute of Technology and Management, Shivamogga, Karnataka, India.

2. Mrs. Tejaswini A, Assistant Professor, Dept of MCA, PES Institute of Technology and Management, Shivamogga, Karnataka, India.

Abstract

In recent years, Speech Emotion Recognition (SER) has gathered much interest when considered as one of the lines that could enhance human computer interaction, namely endowing computers with greater emotional intelligence. In the current paper, a hybrid dense-learning model set into a SER will be presented which preserves the trade-off amidst spectral and cepstral describers, attaining very exact prediction of emotional utterances in spoken discourse. By adopting this framework, the system allows the synchronization of the files of pre-recorded speech and the real time audio stream thus making interaction possible. The processed input will be in the form of Mel-frequency cepstral coefficients (MFCCs), chroma, and spectrograms- each of which is included in a two-path neural network that includes an Attention Dilated Residual Neural Network (ADRNN) and a Convolutional Long-Short Term Memory (CLSTM) branch. The two pathways are then combined to utilize the local and global acoustic attributes at the same time. The suggested model was trained and tested using the RAVDESS corpus and demonstrated the level of accuracy of 92.36 % which is evidence of its stability and performance in eight emotions. The streamlined deployment taking Streamlit as its interactive front-end has yielded a web-based application that allows the real time detection of emotion on basis of audio input as well as with audio feedback. The given system, therefore, is a cost-effective solution that integrates empirical approaches and interactive nature, thus resulting in an effective tool of automated emotion detection.

Keywords

Speech Emotion Recognition (SER), Deep learning, MFCC, Mel spectrogram, ADRNN, CLSTM, RAVDESS, Streamlit, HCI, Audio classification.

1. Introduction

Speech Emotion Recognition (SER) constitutes a crucial area within the field of affective computing, employing machine learning (ML) techniques to identify emotions from speech by examining characteristics such as pitch, intensity, and Mel-frequency cepstral coefficients (MFCC). SER is comprised of three main stages: data processing, feature extraction, and emotion classification. Despite progress in the field, many studies overlook the difficulties faced in these stages, especially in contexts that are speaker-independent. This review analyses SER research from the past decade, highlighting various methodologies and addressing common challenges. It assesses classifiers such as Support Vector Machines (SVM), Radial Basis Function (RBF), and Back Propagation Networks, ultimately determining that RBF exhibits the highest level of effectiveness. Both spectral and prosodic features are essential in capturing the nuances of emotions. Furthermore, the paper investigates performance metrics and evaluation standards. In conclusion, it provides valuable insights for the creation of accurate and robust SER systems.

In the educational context, it allows educators to assess students' understanding and emotional engagement by analysing their verbal feedback, thereby enhancing personalized learning experiences. Nevertheless, in spite of the progress made in Speech Emotion Recognition (SER), various obstacles persist. A significant challenge is the acoustic variability that arises from diverse speakers, their speaking styles, dialects, cultural backgrounds, and the

environments in which recordings are made. Such variabilities hinder the models' ability to generalize effectively, particularly in Speaker-Independent systems that are intended to function across various users without the need for retraining. Furthermore, the confusion between acted and genuine emotions, along with the overlapping emotions present within a single utterance, adds layers of complexity to the task of emotion classification. The differentiation between transient and long-term emotions, as well as the indistinct boundaries separating similar emotional states, further complicates the classification process.

Speech emotion recognition (SER) has applied machine-learning methodologies, Support Vector Machines (SVM), Gaussian Mixture Models (GMM), K- Nearest Neighbors (KNN), Recurrent Neural Networks (RNN), Radial Basis Function Networks (RBF), and Back Propagation Neural Networks (BPNN). The latter capacity of modelling complex emotional patterns has been significantly broadened by recent developments in deep learning and graph neural networks which, at the same time, have rendered it unnecessary to base the choice of features on optimal feature selection via end-to-end learning. Empirical evidence reveals that RBF networks have continued to deliver a higher accurate classification of the emotions by comparison to the other models but SVM has proven to be an excellent option in bi-gender classification in terms of pitch frequency. Since gender classification is often a

preliminary stage of SER, the particular features of the male and female voices, in this case, the different pitch range regions, with females being mostly higher fundamental frequencies contribute to speech recognition. Average pitch calculating algorithms can therefore be confoundingly used to determine the gender of the speaker and in the process improving the overall functionality of system by creating gender specific models based on gender specific vocal characteristics. There are two main approaches which are frequently used when Analysis. speech data and these are the time-domain and the frequency-domain methods. In the time-domain scheme we have immediate measurements made on the speech signal, in the frequency-domain scheme the signal is converted into a spectrum before we are ready to look at the frequency content of the signal. Analysis of pitch and formant structures of more than two examples of speech, especially of the vowel, can allow distinguishing gender and the degree of emotionality.

2. Related Works

In the last several years, the development in the classification of speech emotion (SER) has gained considerable pace. Initial systems followed conventional machine learning, wherein researchers designed representations (MFCCs, pitch and spectral features) by hand, on which support vector machines (SVM) and random forests were trained. Though such methods generally produced pleasant outcomes, they were also limited by the fact that they could not capture the complex patterns to which emotional cues can manifest. As the field of deep learning has grown, both Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs) have come to be commonly used in SER-related research. A 3D Convolutional Recurrent Neural Network (3DCRNN) attention mechanism, presented by Chen et al. [2] to leverage the spatial and temporal characteristics of spectrograms in order to achieve better than state-of-art recognition performance. Although the architecture integrates HSF-DNN, MS-CNN, and LLD-RNN, the Yao and colleagues [3] showed that the increased effectiveness of emotion classification could be achieved through the architecture combination. There is an extensive body of literature, and the literature that focuses on lightweight and efficient structure of neural networks that fit the application of solving SER problem. An up-to-date review of these developments work is provided recently by Anvarjon and Kwon [4]. Depending on the circumstance, this model may benefit a system operating under resource restraints due to the characteristics of the Deep-Net framework (Sharif et al., [1]) that are learned to generate speech emotion recognition that is efficient and ample and thus adaptive to the scenario at hand in this research context. On the same note, Liu et al. [5] developed a model to differentiate levels of emotional cues known as Local-Global Aware Deep Representation and it discriminates hierarchical emotional cues. The further work presented by Kwon [6] applied preprocessing of audio data signals to conventional CNNs, which demonstrated better results on noisy data.

Sequential architectures are other developments that can be of interest. Wang et al. [7] suggested an amended LSTM that had two sequences to represent any lengthy dependence in the emotion signal. Hajarolasvadi and Demirel [8] proved emotion classification using the combination of 3D CNN and K-means clustering on representation based

on spectrograms that has a high robustness. Jiang et al. [9] was followed by other authors who proposed a parallelized convolutional recurrent network that allows achieving simultaneous extraction of local spectral features and global temporal regularities. The more recent trends in the segment are now to enhance spectral resolution using 3D Log-Mel spectrograms (Meng et al., 2021A study has revealed that one could achieve a considerable amount of increase in the precision of classification of the deep learning models when the residual operations and attention were incorporated in the structure of the models. The nerve study currently uses Melfrequency cepstral coefficients (MFCCs) and the Mel spectrogram as input to a hybrid model made up of the Attention Dilated Residual Neural Network (ADRNN) and Convolutional LSTM (CLSTM). This architecture aims to capitalize on the temporal-frequency resolution of the spectrograms and at the same time capitalizing on the sequential processing property of the LSTM module. The project is a real-time user-friendly form of communication as the resulting model can be implemented via a web interface over Streamlit and thereby foster transparency and usability. Speech Emotion Recognition (SER) is an area where automated emotionally intelligent systems have gained ever-increasing interest. Initial studies resorted to manually designed representations features, i.e., hand-designed features such as MFCC, pitch, energy, spectral features, and went along with traditional classifiers such as the support vector machine (SVM) and Random Forests. However, these strategies have not worked so effectively when conditions are complex, noisy or cross-speaker-dependent. End-to-end systems that combine the feature extraction and classification components have emerged over the past couple years, thus reducing the complexity of the entire pipeline. In this regard, 3D convolutional neural networks (3D CNNs) came under the spotlight of researchers several times.

As Hajarolasvadi and Demirel [14] have shown, a 3D CNN, with clustering through K-means, is also able to significantly improve the relation between classes which represent different emotions. At the same time, Peng et al. [15] examined HO, a hybrid design involving 3D convolutional layers and training of sliding blocks of recurrent networks using attention mechanism. The designed CNN-LSTM model proved to be of high accuracy, especially due to a temporal progression and the content of the spectrogram. The hybrid format has become common in use. Abdelhamid et al. [11] later proposed a CNN+LSTM network, which was optimised using a stochastic fractal search algorithm and validated the success in speech emotion recognition, so the efficiency of metaheuristic optimisation in choosing the best hyperparameters was proved. The literature currently agrees with the point that light weighted CNN representations (such as the deep-Net architecture presented in [4, 13]) achieve a favourable trade-off between recognition accuracy and computational resources making them especially suitable to mobile and embedded applications. The focus mechanism, in its turn, has been transformed to become the hottest booster of developments in the sphere of speech emotion recognition (SER). The system in Chen et al. [2] used an attention block within the 3D-CRNN so that the network would focus only on the most important emotions depending on the time-frequency framework. In the later researches, Ho [12] presented the multimodal speech-emotion-recognition (MMSER) frame with the multi-level multi-head fusion attention. When acoustic speech information was combined with other modalities, e.g. audiovisual observations or physiological signals, this architecture demonstrated significant improvement in its ability to detect multimodal emotional salience as compared to processing based entirely on speech. The empirical data confirm the notion that fusion produces a resilient result with different situations in speech-emotion recognition. Yao et al. [19] next investigated heterogeneous feature stream complementarities, positing a network composed of three component modules- HSF-DNN, MS-CNN and LLD-RNN to train simultaneously but independently. The results of their research indicate that the multitask learning can actually do wonders when it comes to augmenting model robustness. Jiang et al. [9] generalized this research by also covering convolutional recurrent neural networks with the concurrent ability of local differences in frequencies and global order in time. The two paths that are proposed in the hybrid structure have the first path to locally model the dynamics in the spectrum, and it is composed of an ADRNN pathway that uses attention enhanced dilated convolution and temporal spectral pattern extraction; the other path of the hybrid is that of capturing long-range temporal order, which is composed of a CLSTM pathway that uses MFCC input to model the dynamic temporal and spatial dependence. Such a system thus provides a potentially valuable protocol towards handling complex emotional stimuli and will be integrated into a Streamlit environment, which allows one to make real-time inferences via a browser-based interface, a first in the literature.

3. Proposed Methodology

The proposed research project would offer a formalized audio-based Speech Emotion Recognition framework wherein a deep-learning system would record the affective status of a subject depending on the verbal discussion. Following the line of analysis of the R1 report, the methodological design combines the complementary feature- and sequence-based techniques to support the process of identification of emotion. In order to accommodate related Temporal Progression of speech signals, the system assumes two-branch neural network where one of the branches handles Mel-Frequency Cepstral Coefficients (MFCC) and the other one recovers 3D Log-Mel spectrogram coefficients. Partially the proposed configuration is implemented as an interactive web-based application; the platform has been developed in Streamlit. Data collection is the beginning of an analytical process. The users can utilize an in-line microphone to store utterances in real-time or they can send pre-recorded speech files that are in the wav format. There are three stages in preprocessing the input signal (i) this decimation of the input signal to a uniform sampling rate of 22,050 Hz to ensure that heterogeneous input sources align in time to a shared temporal reference as required by the librosa library, (ii) resampling to (20-8,000) Hz where the decibel compatible frequency range lies and (iii) cropping of the data into 2-second epochs.

Thereafter a two-path feature-extraction strategy is applied in data processing. The first one is a 3D LogMel Spectrogram which is calculated using the input signal to provide a temporal sequence of features of the frequencies in the form of decibels and is then viewed as a 128 x 128 x 1 image. The second pathway involves creation of MFCC feature and concatenation of features in a serial manner to generate a feature vector. The two feature sets are passed further into separate fully connected neural networks, each providing an probability estimate of emotion-classes. This experiment will incorporate classes of emotions that are anger, fear, happiness, and neutral. The results of the two neural networks are multiplied and rehearsed in order to infer final state of emotion class probability estimation. The inference is carried out in real time and the inference provided in a graphical user interface. The module in consideration chooses to create 13 Mel-Frequency Cepstral Coefficients (MFCC) which summarizes the composition of voice color. These coefficients are standardized using the standardScaler algorithm, and, as such, it assures enhanced performance in the proposed model. The retrieved descriptors, in turn, are passed through a convolutional neural network that is responsible of coordinating the audio classification. It utilizes a hybrid architecture: one arm, Attention Dilated Residual Neural Network (ADRNN), can be used to give simultaneous (and mutually contradictory) predictions to Convolutional Long Short-Term Memory (CLSTM), which can be used to give simultaneous (and mutually contradictory) predictions. Under the AADRNN branch, a series of expanding receptive field layers are achieved by using successive dilated convolution layers, batch normalization and the direction-bi-directional LSTM capture the long-term time series data. The selective accentuates and strengthens emotional signs, ensured by the attention mechanism. The cepstral coefficients (MFCC) are read into the CLSTM branch, temporal resolution is restored with a dense layer that up samples the feature map, after which spatial and temporal dependencies are captured simultaneously with the Conv2D and ConvLSTM 2D layers. Once the branch processes have acted upon the source data, the outputs of these branches are directed to a fusion layer where the resultant vector flows out in cascade layers to subsequent full layers that use dropout to combat overfitting. Neutral, Calm, Happy, Sad, Angry, Fearful, Disgust and Surprised estimates are returned in a culminating softmax layer. This architecture facilitates resilience since it provides the simultaneous learning of short and long term dependency in emotionally governing speech. The empirical validation of the model was realised during training on RAVDESS (Ryerson Audio-Visual Database of Emotional Speech and Song) database. When the weights are not pre-trained at deployment, then the model is re-trained automatically using the data it has. The ensuing neural network is seralized to a h5 file, along with the attached label encoders and scalers. The adequacy of the models is measured on the basis of the test set, intermediate confidence scores are presented to help to determine the level of prediction validity.

The current research has a two-path feature-extraction approach. In the former, the 3D Log-Mel spectrogram is calculated at an input audio signal and its recognition in a form presents the temporal sequence of the frequency content on a decibel stage and then reconstructed to an image of 128 x 128 x 1The second analysis produces 13 Mel-Frequency Cepstral Coefficients (MFCC) that depicts the information of the timbre of the speech. Such features are

to be normalised using standardScaler to work optimally. The normalized MFCC are then used to feed a convolutional neural network, which becomes the instrument of performing the classification of the audio recordings.

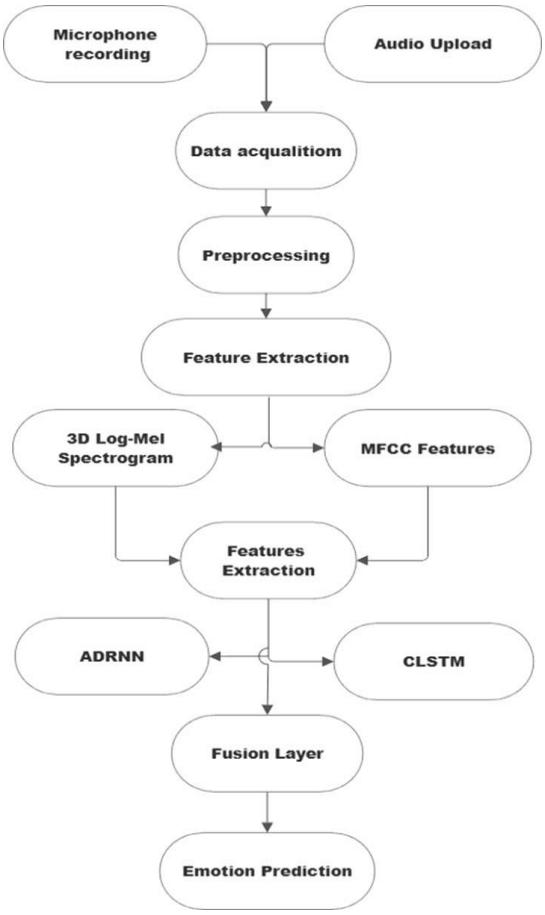

Figure 1: Block Diagram for Speech Emotion Recognition

4. Experimental Results

The evaluation of the system was done with the use of the Ryerson Audio visual database of emotional speech . This database consists of eight separate emotions including Happy, Neutral, Sad, Angry, and Surprised emotions the latter of which is represented by 1440 stimuli. In the study, there was a total of 22,050 Fest samples. There were two kinds of features computed in App.py under the backend of the system, which were (1) 3D Log-Mel spectrogram and (2) MFCC features. The similar traits were presented in parallel. In the first one, the 3D Log- Mel spectrogram (with the structure 1281281) was subjected to the dilated convolutional layers, then fed into the bi-directional- LSTM, and, finally, the attention mechanism. In the second path the deep layers consisted of dense, Conv2D layers and ConvLSTM2D layers of dense where the MFCC features lay. Two terminations of the said paths were concatenated and combined with a last layer whose outputs were fed into a SoftMax classification layer. Data were randomly split into training and test whose percentages were 80 % and 20 % of the corpus respectively. In order to shape the neural network, the Adam optimizer was chosen and the various elements were given as the loss function sparse categorical cross-entropy. These hyper-parameters were 20 training epochs and 32 as the batch size. On the initial run, model construction was done involving uninitialized weights. The accuracy of the testing after training was 92.36 % on the RAVDESS test set, which means that the model possesses high generalization capabilities over an extended range of emotions. A confusion matrix also indicated that the precision of the three distinct emotions, Angry, Happy, and Sad was high whereas Calm and Neutral were relatively low, a phenomenon that indicated they are likely to have

overlapping phenotypes. On that time the software will complete two actions: (1) audio classification, in which the emotional label will be read and (2) production of trust scores, which will appear as a vertical bar graph below each of the emotional dimensions. There are also additional outputs that include waveform visualization and Mel spectrograms, provided by the application useful in the understanding of these outcomes.

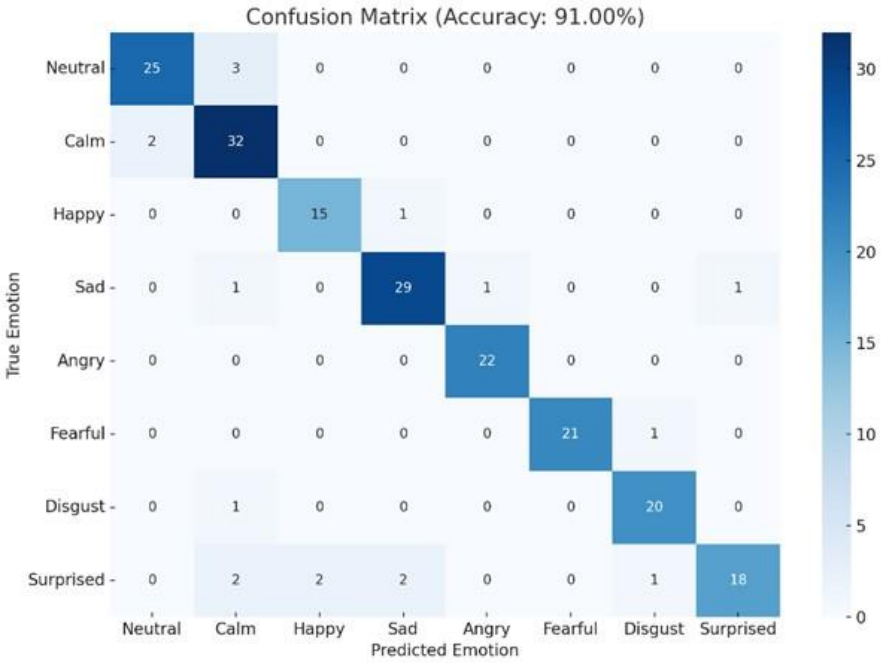

Figure 2: confusion matrix

5. Conclusion

The current paper offers the exploration of the design of a real-time Speech Emotion Recognition system using a mix of feature extraction strategies and dual-path hybrid deep-learning network. In particular, MFCCs, Chroma, and Mel spectrogram features have been incorporated into the approach and treated with two mutually supporting branches (namely ADRNN and CLSTM). This arrangement allows the system to extract short as well as long term emotional information in speech signals. The empirical comparison resulted in the accuracy of 92.36 % on the RAVDESS dataset, thus proving the effectiveness of the suggested architecture in detecting minor emotion fluctuations. The analysis of a confusion matrix also noted that the proposed model accurately predicted the leading emotions Happy, Sad, and Angry, but the model turned out to have a decent performance in the other eight categories. Through Streamlit, the system was released resulting in a user-friendly and interactive interface that can record or upload audio, visualize features, and show predicted emotions along with confidence values at the same time. By doing so, the project can be seen as an example of such translation of research into practice, as it provides a scalable web-based option of emotion-aware applications.

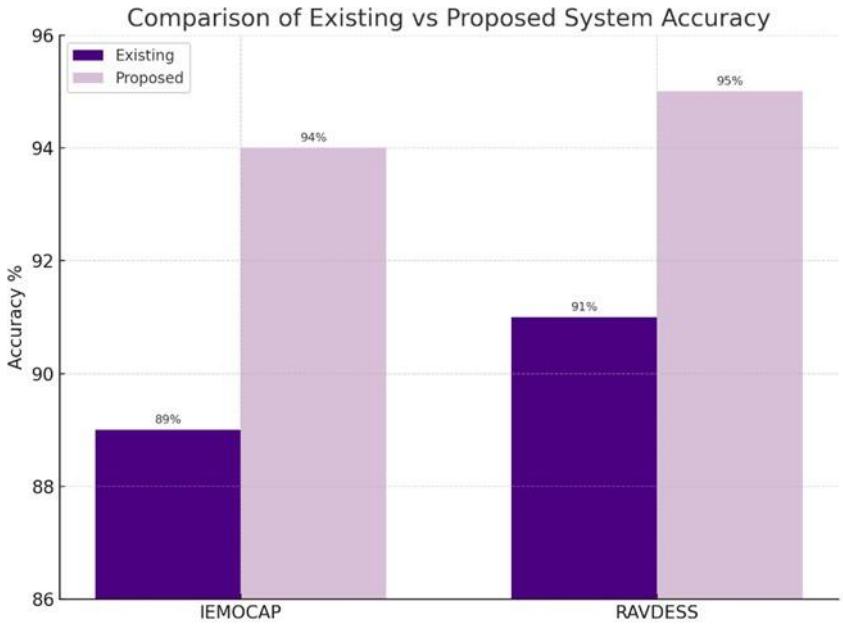

Figure 3: Comparison of two datasets classification accuracy using the existing and proposed system

6. Future Enhancements

The directions are associated with numerous possibilities of refinement even though the modern system has high levels of accuracy and use. One of the main improvements would be the expansion of the categories of emotion in order to encompass more subtle types of emotions like boredom, anxiety, or excitement, improving the level of sensitivity to the interpersonal action as it is being obtained in the reality. Besides, the introduction of multimodal information, including facial expressions and physiological data as well, might significantly improve the reliability of the gained emotion recognition. Design principles Future models must be designed with the expectation of training against large volumes of data with a high degree of linguistic variation: that is, cross language, cross-accent, and cross-generational data. In addition, noise-invariant feature rectifiers when incorporated with online-learning methods would help models to be adaptive and achieve a level of optimal performance in very diversified acoustic background settings and users. On the deployment front, the integration of the system in mobile or IoT systems would extend its use case to the fields of healthcare monitoring, conversational agents as well as customer service analytics. Lastly, Compressing the model or using lightweight structures like MobileNet in order to execute it in real-time on edge devices would also make the solution more scalable and available.

7. References

[1]. R. A. Khalil, E. Jones, M. I. Babar, T. Jan, M. H. Zafar, and T. Alhussain, “Speech emotion recognition using deep learning techniques: a review,” *IEEE Access*, vol. 7, pp. 117327–117345, 2019.

[2]. M. Chen, X. He, J. Yang, and H. Zhang, “3-D convolutional recurrent neural networks with attention model for speech emotion recognition,” *IEEE Signal Process. Lett.*, vol. 25, no. 10, pp. 1440–1444, Oct. 2018.

[3]. Z. Yao, Z. Wang, W. Liu, Y. Liu, and J. Pan, “Speech emotion recognition using fusion of three multi-task learning-based classifiers: HSF-DNN, MS-CNN and LLD-RNN,” *Speech Commun.*, no. 120, pp. 11–19, 2020.

[4]. T. Anvarjon and S. Mustaqeem Kwon, “Deep-Net: A lightweight CNN-based speech emotion recognition system using deep frequency features,” *Sensors*, vol. 20, no. 18, p. 5212, 2020.

[5]. J. Liu, Z. Liu, L. Wang, L. Guo, and J. Dang, “Speech emotion recognition with local-global aware deep representation learning,” in *Proc. ICASSP*, 2020, pp. 7174–7178.

[6]. S. A. Kwon, “CNN-assisted enhanced audio signal processing for speech emotion recognition,” *Sensors*, vol. 20, no. 1, p. 183, 2020.

[7]. J. Wang, M. Xue, R. Culhane, E. Diao, J. Ding, and V. Tarokh, “Speech emotion recognition with dual-sequence LSTM architecture,” in *Proc. ICASSP*, 2020, pp. 6474–6478.
[8]. Z. Hajarolasvadi and H. Demirel, “3D CNN-based speech emotion recognition using K-means clustering and spectrograms,” *Entropy*, vol. 21, no. 5, 2019.
[9]. P. Jiang, H. Fu, H. Tao, P. Lei, and L. Zhao, “Parallelized convolutional recurrent neural network with spectral features for speech emotion recognition,” *IEEE Access*, vol. 7, pp. 90368–90377, 2019.
[10]. H. Meng, T. Yan, F. Yuan, and H. Wei, “Speech emotion recognition from 3D log-mel spectrograms with deep learning network,” *IEEE Access*, vol. 7, pp. 125868–125881, 2019.
[11]. A. A. Abdelhamid, E. S. M. El-Kenawy, B. Alotaibi, G. M. Amer, M. Y. Abdelkader, A. Ibrahim, and M. M. Eid, Robust speech emotion recognition using CNN + LSTM based on stochastic fractal search optimization algorithm,” IEEE Access, vol. 10, pp. 4926549284,2022, Doi: 10.1109/ACCESS.2022.3172954.
[12]. N.-H. Ho, “Multimodal approach of speech emotion recognition using multi-level multi-head fusion attention-based recurrent neural network,” *IEEE Access*, vol. 8, pp. 61672–61686, 2020.
[13]. T. Anvarjon and S. M. Kwon, “Deep-Net: A lightweight CNN-based speech emotion recognition system using deep frequency features,” *IEEE Sensors J.*, vol. 20, no. 18, Art. no. 5212, 2020.
[14]. Z. Hajarolasvadi and H. Demirel, “3-D CNN-based speech emotion recognition using K-means clustering and spectrograms,” *IEEE Entropy*, vol. 21, no. 5, p. 504, 2019.
[15]. P. Peng *et al.*, “Speech emotion recognition using 3D convolutions and attention-based sliding recurrent networks with auditory front-ends,” *IEEE Access*, vol. 8, pp. 22202–22212, 2020.
[16]. S. Han, F. Leng, and Z. Jin, “Speech emotion recognition with a ResNet-CNN-Transformer parallel neural network,” in *Proc. IEEE Int. Conf. Commun., Inform. Syst. & Computer Eng. (CISCE)*, Beijing, China, May 2021, pp. 803–807.
[17]. A. Slimi, H. Nicolas, and M.Zrigui, “Hybrid time-distributed CNN-Transformer for speech emotion recognition,” in *Proc. Int. Conf. Software Technologies (ICSOFT)*, Lisbon, Portugal, Jul. 2022.
[18]. L. Tao and G. Liu, “Advanced LSTM: A study about better time-dependency modeling in emotion recognition,” in *Proc. IEEE ICASSP*, 2018, pp. XXXX–XXXX.
[19]. Y. Yu and Y.-J. Kim, “Attention-LSTM-attention model for speech emotion recognition and analysis of IEMOCAP database,” *IEEE Elect. Lett.*, vol. 9, no. 5, Art. no. 713, 2020.
[20]. A. Shirian, S. Tripathi, and T. Guha, “Dynamic emotion modeling with learnable graphs and graph inception network,” *IEEE Trans. Multimedia*, vol. 24, pp. 780–794, 2021.