Emotion Detection through Human Verbal Expression Using Deep Learning Techniques

Rudra Tiwari
Department of Computer Science
Maulana Azad National Institute of Technology
Bhopal, India
rudratiwari2901@gmail.com

Ayushi Prajapati
Department of Computer Science
Maulana Azad National Institute of Technology
Bhopal, India
ayuprajapati2807@gmail.com

Saravana Chandran
Computer Science
Maulana Azad National Institute of Technology
Bhopal, India
saravanachandranc@gmail.com

Divyansh Agrawal
Computer Science
Maulana Azad National Institute of Technology
Bhopal, India
mragrawal9012@gmail.com

Akhtar Rasool
Computer Science Department
Maulana Azad National Institute of Technology
Bhopal, India
akki262@gmail.com

Abhishek Jadhav
Computer Science Department
Maulana Azad National Institute of Technology
Bhopal, India
abhishekjadhav.cs@gmail.com

Abstract—**Detecting emotions from speech signals is a crucial yet intricate aspect of Human-Computer Interaction (HCI). In the field of speech emotion recognition (SER) literature, different methods are applied to identify emotions in speech signals. These include techniques involving analysis of speech and its classification. Recently, Deep Learning methodologies are proposed as a substitute for conventional SER approaches. Spectrogram and Mel-Frequency Cepstral Coefficients (MFCC) are examples of speech features that preserve low-level characteristics associated with emotions. At the same time, textual information aids in capturing semantic meaning. Both types of features contribute to different aspects of emotion detection. In this study, we delved into the enhancement of model learning through Convolutional Neural Network (CNN) based architectures and dataset augmentation. This was particularly crucial because datasets can exhibit variability, sometimes requiring additional measures to improve model performance.**

I. INTRODUCTION

. An essential component of human communication, emotion shapes our interactions and interpersonal relationships. Emotion analysis and recognition from speech is an essential part of social robots, mental health monitoring, human-computer interaction, and many other applications in various domain. Reliable and accurate speech-to-emotion recognition can open the door to more responsive and sympathetic technology, improving our daily interactions and quality of life. The domain of recognizing emotions from speech has transitioned from a specialized domain, into a critical component of Human-Computer Interaction (HCI) [1]–[3]. These setups seek to enhance the inherent engagement between humans and machines by enabling voice interactions, as opposed to relying on conventional input devices to comprehend spoken content, making it more intuitive for human users to engage [4]–[6]. Various applications for this technology exist, including dialogue systems for verbal languages in call centre interactions, integrated vehicle driving systems, and employing the emotional patterns in speech for medical applications [7]. However, as HCI systems make the shift from laboratory testing to real-world applications, there remain several unresolved challenges that need to be effectively addressed [8]–[10]. Consequently, substantial efforts are necessary to tackle these issues and enhance the precision of emotion recognition through machine-based methods. The process of SER typically comprises two main steps: feature extraction and feature classification [11]. Within the domain of speech analysis and processing, researchers have formulated various types of features, like source-based excitation features, prosodic characteristics, and components related to the vocal tract, and a combination of features. [12]. In the subsequent phase, these extracted features are subjected to classification using both linear and non-linear classifiers. Since signals generated from speech are generally regarded as non-stationary, classifiers with non-linear characteristics are often preferred for SER tasks. There are several non-linear classifiers that are commonly used in SER, including the Gaussian Mixture Model (GMM) and the Hidden Markov Model (HMM) [13]. Deep Learning has become a prominent focus of study within the machine learning domain and has gained growing recognition in recent times [14]. Deep Learning techniques applied to speech emotion recognition (SER) offer several advantages when compared to traditional methods. These advantages include capability of automatically discerning complex structures and features without requiring manual extraction of features and meticulous adjustment, and the ability to directly derive low-level features from raw data and the capacity to manage unlabeled data.

The utilization of Deep Belief Networks (DBN) given by Y. Kim, H. Lee, and E. M. Provost [32] and W.L.Zheng, J.Zhu, Y.Peng [33] has demonstrated a considerable enhancement compared to baseline models [26]-[31] that don't incorporate deep learning techniques. Subsequently, Han et al. [34] introduced a DNN-ELM (Deep Neural Network - Extreme Learning Machine) approach to obtain high-level features from unprocessed data and employed a neural network with solitary hidden layer for segment-level SER, resulting in a modest increase in accuracy.

Zheng et al. [35] employed spectrograms in combination with Deep Convolutional Neural Networks (CNN) for their SER approach, while Fayek et al. [36] explored data augmentation in conjunction with a Deep Neural Network (DNN) for enhanced results. Lee et al. [24] adopted a bi-directional Long Short-Term Memory (LSTM) model trained on feature sequences, resulting in an accuracy of 62.8% in recognizing emotions on the IEMOCAP dataset (Interactive Emotional Dyadic Motion Capture dataset) [37], representing a substantial improvement over the DNN-ELM approach [34].

Jin et al. [25] experimented with the combination of representations based on acoustic and lexical features and achieved an accuracy of approximately 69.2% on a 4-class IEMOCAP dataset. Additionally, Satt et al. [38] utilized Mel-scale spectrograms in conjunction with a deep Convolutional Neural Networks and a combination of CNN and LSTM to

XXX-X-XXXX-XXXX-X/XX/$XX.00 ©20XX IEEE

attain improved outputs on the IEMOCAP dataset. These advancements in deep learning techniques have significantly enhanced the performance of SER systems.

The organization of the paper is as follows: first, we have the literature survey which examines the prior research on the topic. The next two sections are the Methodology and Model Used. The final section of the paper analyzes the Work Completed and provides a thorough Result Analysis before coming to a close with major conclusions and recommendations for the future in the Conclusions section.

II. Literature Survey

A. Background

Affective computing, human-computer interaction, and healthcare are just a few industries that have made emotion detection from speech signals a hot topic in recent years. Sentiment analysis, virtual assistants, mental health assessments, and other applications will all function more effectively if humans can automatically identify and comprehend the emotions expressed in their speech. In order to recognise emotions, early research in this field concentrated on manual annotation and basic acoustic features [41]. But as deep learning and machine learning techniques have developed, more advanced methods have also appeared, providing greater robustness and accuracy. The work of [40], which introduced the utilization of Support Vector Machines (SVM) with acoustic features to classify emotions in speech, is one of the seminal studies that paved the way for modern speech emotion detection. This study inspired more research and proved that machine learning algorithms could be used to recognise emotions [17]. CNNs and recurrent neural networks (RNNs) have been used for the analysis of acoustic features due to the quick development of deep learning models [42]. As a result of these models' impressive advancements in identifying complex patterns in speech signals, emotion classification accuracy has increased [43]. Additionally, researchers have been able to train models that are better able to handle real-world variation in speech and emotional expression thanks to the adoption of larger and more diverse datasets, such as the IEMOCAP dataset [44]. Still, there are several difficulties in the speech emotion recognition domain. Managing gender and cultural differences in emotional expression is one of the main challenges [45] Research on creating models that are responsive to these differences is still ongoing perceptual and acoustic distinctions in nonverbal emotional vocalizations between authentic and intentionally produced expressions. Furthermore, there are still many obstacles to overcome, including those pertaining to noisy environments, speaker variability, and the requirement for labelled data for training [46]. Transfer learning and multi-modal emotion recognition have garnered interest recently as viable approaches to enhance the robustness and generalization of emotion detection models (arXiv:2202.08974 [cs.SD]). These methods improve speech-to-emotion recognition by utilizing information from other domains or modalities [47]. Because Convolutional Neural Networks (CNNs) can extract intricate spectral features from audio signals, they have become increasingly popular within the scope of speech emotion recognition. Early research by Grazina et al. Examining 2D Feature Spaces for Speech Recognition Using Deep Learning showed that 2D CNNs could successfully extract acoustic features for the purpose of classifying emotions. Their exceptional accuracy when applying CNNs to speech log-mel spectrogram images has created new opportunities for deep learning in this field [48]. The performance of speech-emotion recognition systems has also been enhanced by transfer learning, which makes use of information from pre-trained models. For deriving high-level features from spectrogram images, [49] implemented transfer learning by utilizing pre-trained CNN models on extensive image datasets. According to Zhang et al., this method produced notable increases in the accuracy of emotion recognition, demonstrating the potential of transfer learning in the field. By adding depth-wise separable convolutions, which decreased the model's complexity and increased its efficiency while maintaining high accuracy, Zhao et al.'s recent work from 2019 [21] expanded the use of CNNs in speech emotion recognition even further, which used transfer learning to fine-tune a CNN pre-trained on a sizable image dataset, is a noteworthy example. Transfer learning's cross-domain utility was highlighted by the model's improved performance in emotion classification tasks after it was fine-tuned using speech data [50].

B. Critical Analysis

In a recent study, Hazarika et al.[51] showed how well transfer learning works with the well-liked natural language processing model BERT (Bidirectional Encoder Representations from Transformers) to classify emotions in speech text. The utilization of pre-trained models from related domains was demonstrated by this approach [51]. Residual Networks (ResNets) can capture hierarchical features and manage deep networks well, they have found use in the field of speech emotion detection. Residual Nets were first developed for image classification tasks. ResNets have demonstrated encouraging results in enhancing the performance of SER models, despite being an uncommon choice at first. [52] investigated how Residual Networks could be modified for the activity of speech emotion recognition(SER). They used a modified version of the ResNet architecture to process speech signal acoustic feature extraction. Their model's ability to recognise complex emotional cues in speech was greatly enhanced by the ability to train much deeper networks using the residual connections in ResNets [52]. The work by Tran et al. (2022) [53], which suggested the use of a hybrid model that combines ResNets with Transformers for speech emotion recognition, is another significant development in the field. Their method produces better emotion recognition performance by efficiently capturing both local and global features in the acoustic data [53]. In addition, Basodi et al.'s (2020) [54] study addressed issues with efficiently training deep networks by introducing a novel training strategy for deep ResNet-based models in SER [54]. Improved SER capabilities have become achievable by the use of ResNets in speech emotion detection, which represents a shift in approach towards utilising the advantages of deep residual networks to increase model accuracy and generalisation. As indicated by the cited works, using ResNets for speech emotion recognition is a promising direction that will enable researchers to leverage ResNets' strong feature extraction capabilities and hierarchical modelling.

III. Methodology

A. Traditional Techniques

Emotional recognition systems, which assess digitized speech, generally comprise three essential elements: signal preprocessing, feature extraction, and classification [15]. The process begins with acoustic preprocessing, involves the processes of noise reduction and segmentation to ascertain significant units within the signal [16]. Feature extraction is then employed to recognize relevant characteristics within the signal, and classifiers are used to map these extracted feature vectors to specific emotions. In this section, we will provide an in-depth exploration of the processing of speech signals, the extraction of features, and the classification process [17]. Additionally, we will address the distinctions between natural and acted speech, as these differences are pertinent to the topic [18], [19].

Figure 1 illustrates an elementary setup used for speech-based emotion recognition. Firstly, the system performs speech enhancement to remove noisy components from the input signal. The next phase comprises two key components: feature extraction and feature selection. This stage involves extracting the necessary features from the preprocessed speech signal and selecting relevant features from the extracted set. Typically, the process of extracting and selecting features relies on the examination of speech signals in both frequency and temporal domains. In the third stage, a variety of classifying algorithms, like GMM(Gaussian Mixture Models) and HMM(Hidden Markov Models), are deployed for the classification. Finally, different emotions are recognized based on the results of feature classification.

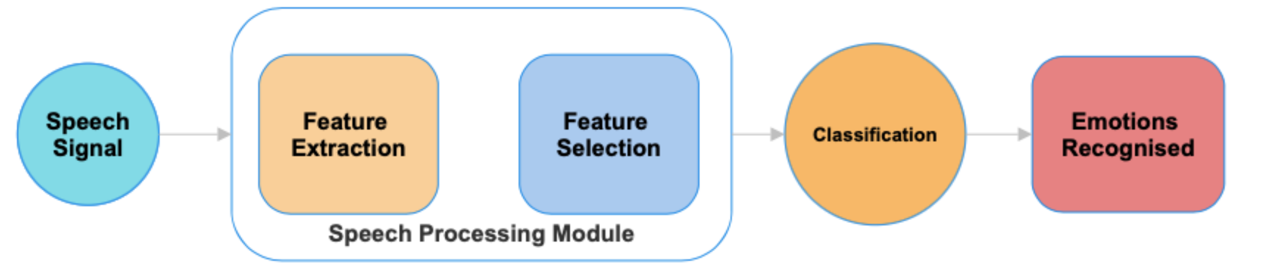


fig 1

B. Datasets Used

The evaluation of these classification systems hinges on two critical factors: the quality of the databases used and the attained efficiency. The methods and intention for collecting speech databases can vary significantly, depending on the goals of developing such systems. Table 1 presents the attributes of various openly accessible databases containing emotional speech, that have been utilized in this study, offering insight into the diversity of resources used in emotion recognition research.

Table 1	**Dataset Details**		
	Dataset	***Emotions***	***Size***
1	TESS(Toronto Speech Emotional Set)	Happy, sad, angry, neutral, fearful, surprised and disgusted	2000
2	RAVDESS (Ryerson Audio-Visual Database of Emotional Speech and Song)	Happy, sad, angry, neutral, fearful, surprised and disgusted	2184
3	CREMA-D(Crowd-Sourced Emotional Multimodal Actors Dataset)	Disgust, Anger, fear, happiness, sadness, and neutral	7000
4	SAVEE(Surrey Audio-Visual Expressed Emotion)	Happy, sad, angry, neutral, fearful, surprised and disgusted	4000

C. Input Data Preprocessing

Data Preprocessing: Once the dataset is collected, the next step is to preprocess the audio data. This includes converting the audio files to a common format, such as WAV or MP3, and standardizing the audio quality across all recordings. Additionally, feature extraction is performed on the audio files, where the speech signal is transformed into a set of features that can be used by CNN for training and testing. The commonly used feature extraction techniques for speech recognition include Mel Frequency Cepstral Coefficients (MFCC), Linear Predictive Coding (LPC), and Perceptual Linear Prediction (PLP).

Spectrograms are commonly used in SER because they capture both temporal and spectral information in speech signals. Emotional information in speech signals is often conveyed through changes in pitch, volume, and rhythm, which can be detected from the frequency content of the signal. Spectrograms also provide a compact and informative way to visualize the frequency content of speech signals, which can aid in the interpretation of the features extracted from them.

Data augmentation is an important step in the training of CNNs. This step involves creating additional samples from the existing data by applying various transformations such as shifting, scaling, and adding noise. The process is done to enhance the diversity of the training data and to prevent overfitting. Image augmentation is a commonly employed technique in machine learning to generate additional training data and enhance the generalization capability of models used in image analysis. In the context of training classification models on spectrograms, image augmentation has been adopted to boost accuracy and resilience. One approach involves utilizing image augmentation methods such as rotation, flipping, and elastic deformation to augment the dataset. The researchers reported that employing image augmentation techniques led to an improvement in the model's classification accuracy and reduced overfitting to the training data.

D. Feature Extraction

A spectrogram is a visual depiction of the spectrum of frequencies of a signal as it varies with time. In the context of SER, spectrograms are commonly utilized for extracting features from speech signals that can be employed to train machine learning models.

To generate a spectrogram, a speech signal is first divided into small segments (typically 20-30 ms in duration) called frames. Each frame is subsequently converted into the frequency domain through a mathematical process called the Fast Fourier Transform (FFT). The resulting frequency spectrum is depicted as a two-dimensional image, where time is plotted on the x-axis and frequency on the y-axis. The brightness or intensity of each pixel in the image signifies the magnitude of the frequency component at the corresponding time and frequency.

Spectrograms are particularly useful when combined with different methods for extracting features, such as Mel Frequency Cepstral Coefficients (MFCCs) and Prosodic Features, which capture additional information about the speech signal that can be used to improve classification performance. Figure 2 illustrates an example of the wave plot and spectrogram of a male actor with fear.

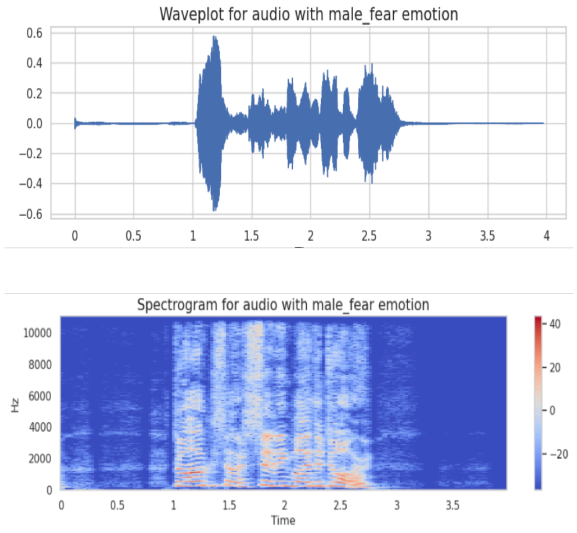


Figure 2

IV. Model Used

A. Model 1: Resnet-34 with transfer learning

ResNet-34 is a CNN architecture that was first presented in the paper "Deep Residual Learning for Image Recognition" by He et al. in 2015. It is a modification of the ResNet architecture, recognized for its effectiveness in training deep convolutional neural networks (CNNs). ResNet-34 is a 34-layer CNN that is composed of multiple building blocks called residual blocks. Every residual block is composed of multiple convolutional layers, followed by batch normalisation and ReLU activation layers. The residual blocks are connected to each other in a way that enables the network to acquire knowledge of residuals (i.e., the differences between the input and the output of the block) instead of the outputs themselves. This makes it easier for the network to learn and helps to prevent the vanishing gradient problem, which can occur when training very deep networks. Figure 3 shows an example of the architecture of ResNet-34 model.

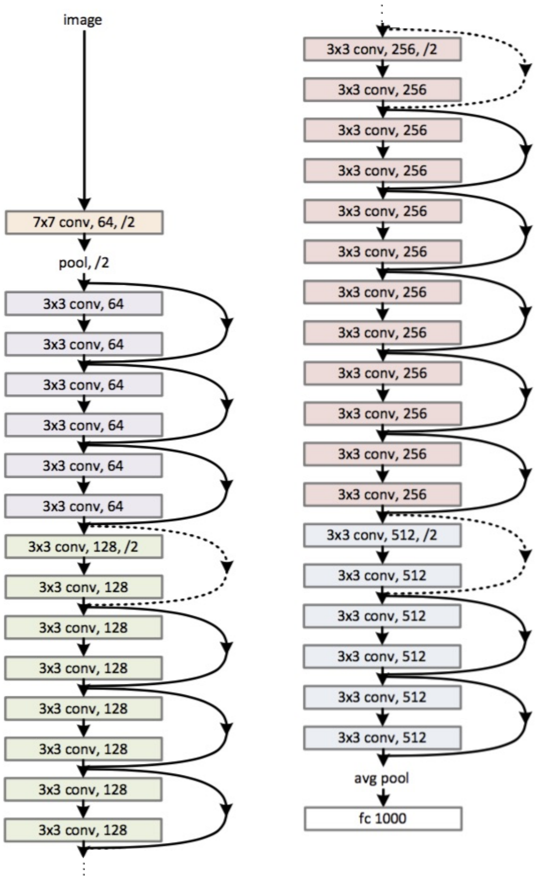
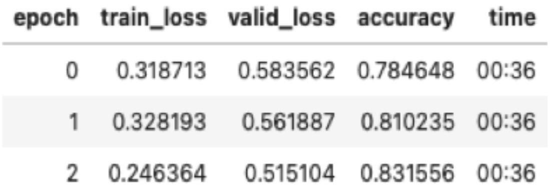


Figure 3

ResNet-34 has been extensively employed for tasks related to classifying images and has attained leading-edge performance outcomes on several benchmark datasets. It has also been used as a starting point for transfer learning, where the pre-trained ResNet-34 model is fine-tuned on a different dataset for a specific task.

With this model, only two Datasets were used, TESS and RAVDESS, for which the model was able to predict the emotions with 83.1% accuracy. Fig 4 shows the learning rate and accuracy with the number of epochs.



epoch	train_loss	valid_loss	accuracy	time
0	0.318713	0.583562	0.784648	00:36
1	0.328193	0.561887	0.810235	00:36
2	0.246364	0.515104	0.831556	00:36

Figure 4

B. Model 2

In this model, we introduce a novel Convolutional Neural Network (CNN) architecture, specifically conceived and developed to cater to the distinctive characteristics of our dataset. Diverging from the conventional approach of employing pre-existing architectures such as ResNet34, like Model 1, The Model 2 is the culmination of a meticulous design process. The architecture is strategically layered and configured to excel in extracting salient features and performing accurate classification tasks pertinent to our research objectives. The deliberate calibration of the network's layers, coupled with precise activation and normalization techniques, signifies a deliberate move towards a customized, application-specific neural network design that enhances the predictive capabilities of our model.

Figure 5 shows the architecture of the CNN Model in detail. For this model, we have used all 4 datasets mentioned above, ie, SAVEE, TESS, RAVDESS, CREMA-D.

V. Work Done and Result Analysis

In our research, we amalgamated four prominent datasets - RAVDESS, TESS, SAVEE, and CREMA-D - to enrich the volume and variability of our training and testing corpora. This integration aims to harness the comprehensive emotional spectrum encapsulated within the collective datasets, thereby enabling our model to learn from a more representative sample of human emotions.

Data augmentation serves as a cornerstone of our strategy to mitigate the model's overfitting. By simulating a more extensive dataset through the generation of transformed replicas of existing data, we enhance the generalization ability of our Convolutional Neural Network (CNN). The augmentation techniques employed include geometric transformations and photometric alterations that not only diversify the training set but also significantly reduce the dimensional footprint of the input images. Post-augmentation, the datasets were divided into training and validation subsets, ensuring a balanced representation of every emotion category. The labeling process was automated by extracting annotations from the folder structure, thereby streamlining the preparation phase for subsequent model training.

Our custom CNN architecture was designed with a focus on capturing the nuanced expressions of emotions. The training phase involved exposing the network to a multitude of augmented images, which the CNN processed to discern and recognize patterns correlating with specific emotional states.

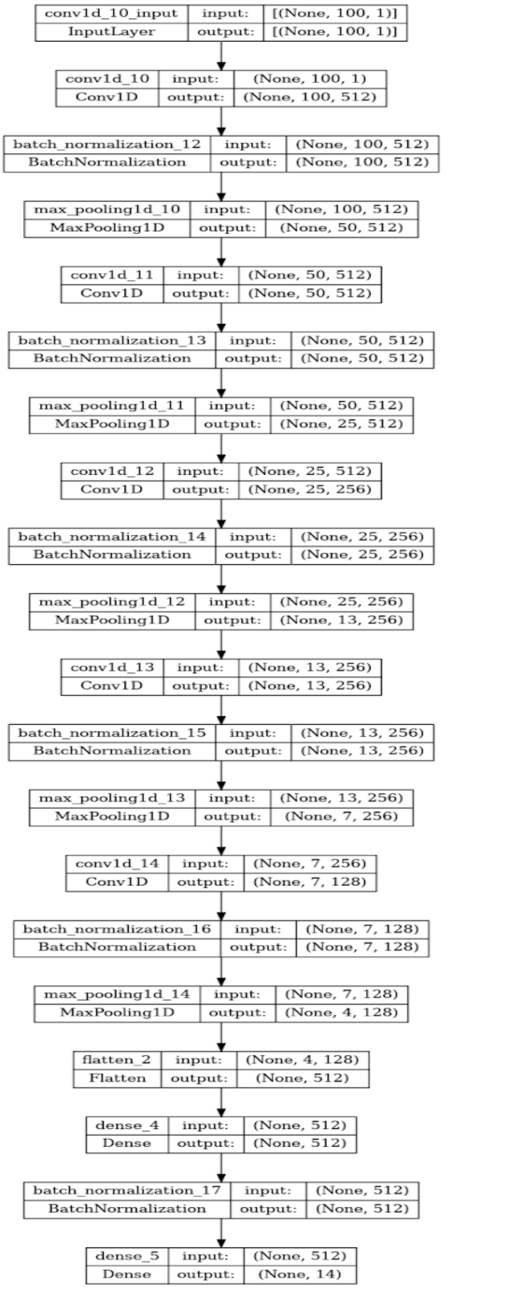

conv1d_10_input | input: | [(None, 100, 1)]
InputLayer | output: | [(None, 100, 1)]
conv1d_10 | input: | (None, 100, 1)
Conv1D | output: | (None, 100, 512)
batch_normalization_12 | input: | (None, 100, 512)
BatchNormalization | output: | (None, 100, 512)
max_pooling1d_10 | input: | (None, 100, 512)
MaxPooling1D | output: | (None, 50, 512)
conv1d_11 | input: | (None, 50, 512)
Conv1D | output: | (None, 50, 512)
batch_normalization_13 | input: | (None, 50, 512)
BatchNormalization | output: | (None, 50, 512)
max_pooling1d_11 | input: | (None, 50, 512)
MaxPooling1D | output: | (None, 25, 512)
conv1d_12 | input: | (None, 25, 512)
Conv1D | output: | (None, 25, 256)
batch_normalization_14 | input: | (None, 25, 256)
BatchNormalization | output: | (None, 25, 256)
max_pooling1d_12 | input: | (None, 25, 256)
MaxPooling1D | output: | (None, 13, 256)
conv1d_13 | input: | (None, 13, 256)
Conv1D | output: | (None, 13, 256)
batch_normalization_15 | input: | (None, 13, 256)
BatchNormalization | output: | (None, 13, 256)
max_pooling1d_13 | input: | (None, 13, 256)
MaxPooling1D | output: | (None, 7, 256)
conv1d_14 | input: | (None, 7, 256)
Conv1D | output: | (None, 7, 128)
batch_normalization_16 | input: | (None, 7, 128)
BatchNormalization | output: | (None, 7, 128)
max_pooling1d_14 | input: | (None, 7, 128)
MaxPooling1D | output: | (None, 4, 128)
flatten_2 | input: | (None, 4, 128)
Flatten | output: | (None, 512)
dense_4 | input: | (None, 512)
Dense | output: | (None, 512)
batch_normalization_17 | input: | (None, 512)
BatchNormalization | output: | (None, 512)
dense_5 | input: | (None, 512)
Dense | output: | (None, 14)

Figure 5

Upon testing, the model demonstrated an accuracy rate of 93.4% on the test data, underscoring its proficiency in emotion recognition. Table 2 presents an in-depth comparative examination of conventional algorithms and Deep Learning algorithm, concerning the classification of various emotions. [20].

Table 2	**Algorithm Comparison**			
	Algorithm	***Anger***	***Happy***	***Sad***
1	K - Nearest Neighbour	93%	55%	77%
2	Linear Discriminant Analysis	68%	49%	72%
3	SVM(Support Vector Machine)	74%	70%	93%
4	Regularized Discriminant Analysis	83%	73%	97%
5	Convolutional Neural Network	96%	94%	93%

A. Comparison With State-of-the-art-methods

This article provides a framework for choosing from a variety of pre-trained models that categorize emotions in human speech using machine learning. In Table 2, the model's outcomes are contrasted with a previously suggested technique for Emotion Detection. Table 3 below shows that, in terms of evaluation metrics (such as accuracy), our proposed model is superior to other cutting-edge models.

Table 3	**Result Comparison**		
Author	***Techniques Used***	***Accuracy***	***Proposed Model's Accuracy***
J. Zhao et. al [21]	CNN & DBN	91.6%	93.4%
E. Lakomkin et. al [22]	RNN & CNN	83.2%	93.4%
P. Tzirakis et. al [23]	CNN & ResNet50	78.7%	93.4%

VI. Conclusion

Within this study, we have introduced a suite of Convolutional Neural Network (CNN) architectures tailored to analyze both speech features and transcriptions. The innovative design of these models has been pivotal in surpassing the accuracy benchmarks set by contemporary state-of-the-art methodologies. The synthesized dataset, an amalgamation of RAVDESS, TESS, SAVEE, and CREMA-D, when employed for training, culminated in a significant elevation of emotion detection accuracy to 93.4%. This increment of nearly 2% marks a notable advancement in the domain of automated emotion recognition.

The utility of the proposed models extends beyond mere academic interest, presenting practical implications for the development of emotionally aware artificial intelligence. The enhanced capability to discern emotion and sentiment from speech equips these models to substantially improve the dynamics of human-computer interaction. Applications such as conversational agents, social robotics, and interactive entertainment systems stand to benefit profoundly from this heightened emotional intelligence. The incorporation of our models within these applications promises to pave the way for more intuitive, responsive, and empathetic user experiences.

The implications of our research suggest a trajectory towards increasingly sentient machine communication, opening avenues for future investigations to refine these models further. The potential to integrate such emotionally cognizant systems into daily technology offers a glimpse into a future where machines understand not just commands, but the emotional context that underpins human communication.

References

1. Björn W. Schuller. 2018. Speech emotion recognition: two decades in a nutshell, benchmarks, and ongoing trends. Commun. ACM 61, 5 (May 2018), 90–99. https://doi.org/10.1145/3129340
2. M. Shamim Hossain and Ghulam Muhammad. 2019. Emotion recognition using deep learning approach from audio–visual emotional big data. Inf. Fusion 49, C (Sep 2019), 69–78. https://doi.org/10.1016/j.inffus.2018.09.008
3. M. Chen, P. Zhou and G. Fortino, "Emotion Communication System," in *IEEE Access*, vol. 5, pp. 326-337, 2017, doi: 10.1109/ACCESS.2016.2641480.
4. Nicholas D. Lane and Petko Georgiev. 2015. Can Deep Learning Revolutionize Mobile Sensing? In Proceedings of the 16th International Workshop on Mobile Computing Systems and Applications (HotMobile '15). Association for Computing Machinery, New York, NY, USA, 117–122. https://doi.org/10.1145/2699343.2699349

5. J. G. Rázuri, D. Sundgren, R. Rahmani, A. Moran, I. Bonet, and A. Larsson, 'Speech emotion recognition in emotional feedbackfor Human-Robot Interaction', International Journal of Advanced Research in Artificial Intelligence (IJARAI), vol. 4, no. 2, pp. 20–27, 2015..
6. D. Le and E. M. Provost, "Emotion recognition from spontaneous speech using hidden MARKOV models with deep belief networks," in Proc. IEEE Workshop Autom. Speech Recognition. Understand., Dec. 2013, pp. 216–221.
7. A. B. Nassif, I. Shahin, I. Attili, M. Azzeh, and K. Shaalan, "Speech recognition using deep neural networks: A systematic review," IEEE Access, vol. 7, pp. 19143–19165, 2019.
8. S. Lalitha, A. Madhavan, B. Bhushan, and S. Saketh, "Speech emotion recognition," in Proc. Int. Conf. Adv. Electron. Comput. Com- mun. (ICAECC), Oct. 2014, pp. 1–4.
9. K. R. Scherer, "What are emotions? And how can they be measured?" Social Sci. Inf., vol. 44, no. 4, pp. 695–729, 2005.
10. T.Balomenos, A.Raouzaiou, S.Ioannou, A.Drosopoulos, K.Karpouzis, and S. Kollias, "Emotion analysis in man-machine interaction systems," in Proc. Int. Workshop Mach. Learn. Multimodal Interact. Springer, 2004, pp. 318–328.
11. S. G. Koolagudi and K. S. Rao, "Emotion recognition from speech: A review," Int. J. Speech Technol., vol. 15, no. 2, pp. 99–117, 2012.
12. M. El Ayadi, M. S. Kamel, and F. Karray, "Survey on speech emotion recognition: Features, classification schemes, and databases," Pattern Recognit., vol. 44, no. 3, pp. 572–587, 2011.
13. A. D. Dileep and C. C. Sekhar, "GMM-based intermediate matching kernel for classification of varying length patterns of long duration speech using support vector machines," IEEE Trans. neural Netw. Learn. Syst., vol. 25, no. 8, pp. 1421–1432, Aug. 2014.
14. J. Schmidhuber, "Deep learning in neural networks: An overview," Neu- ral Netw., vol. 61, pp. 85–117, Jan. 2015.
15. T. Vogt and E. André, "Comparing feature sets for acted and spontaneous speech in view of automatic emotion recognition," in Proc. IEEE Int. Conf. Multimedia Expo (ICME), Jul. 2005, pp. 474–477.
16. C.N. Anagnostopoulos, T. Iliou, and I. Giannoukos, "Features and classifiers for emotion recognition from speech: A survey from 2000 to 2011," Artif. Intell. Rev., vol. 43, no. 2, pp. 155–177, 2015.
17. A. Batliner, B. Schuller, D. Seppi, S. Steidl, L. Devillers, L. Vidrascu, T. Vogt, V. Aharonson, and N. Amir, "The automatic recognition of emotions in speech," in Emotion-Oriented Systems. Springer, 2011, pp. 71–99.
18. E. Mower, M. J. Mataric, and S. Narayanan, "A framework for automatic human emotion classification using emotion profiles," IEEE Trans. Audio, Speech, Language Process., vol. 19, no. 5, pp. 1057–1070, Jul. 2011.
19. J. Han, Z. Zhang, F. Ringeval, and B. Schuller, "Prediction-based learning for continuous emotion recognition in speech," in Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP), Mar. 2017, pp. 5005–5009.
20. M. Sidorov, S. Ultes, and A. Schmitt, "Emotions are a personal thing: Towards speaker-adaptive emotion recognition," in Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP), May 2014, pp. 4803–4807.
21. J. Zhao, X. Mao, and L. Chen, "Speech emotion recognition using deep 1D & 2D CNN LSTM networks," Biomed. Signal Process. Control, vol. 47, pp. 312–323, Jan. 2019.
22. E. Lakomkin, M. A. Zamani, C. Weber, S. Magg, and S. Wermter, "On the robustness of speech emotion recognition for human-robot interaction with deep neural networks," in Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst. (IROS), Oct. 2018, pp. 854–860.
23. P. Tzirakis, G. Trigeorgis, M. A. Nicolaou, B. Schuller, and S. Zafeiriou, "End-to-end multimodal emotion recognition using deep neural networks," IEEE J. Sel. Topics Signal Process., vol. 11, no. 8, pp. 1301–1309, Dec. 2017.
24. J. Lee and I Tashev. High-level feature representation using recurrent neural network for speech emotion recognition. In INTERSPEECH, 2015.
25. Q. Jin, C. Li, S. Chen, H. Wu . "Speech emotion recognition with acoustic and lexical features." in IEEE International Conference on Acoustics, Speech and Signal Processing 2015:4749-4753.
26. V. Dimitrios, and C. Kotropoulos. "Fast and accurate sequential floating forward feature selection with the Bayes classifier applied to speech emotion recognition." Signal Processing 88.12, pp. 2956-2970, 2008.
27. X. Mao, L. Chen, and L. Fu. "Multi-level Speech Emotion Recognition Based on HMM and ANN." WRI World Congress on Computer Science and Information Engineering, 225-229, 2009.
28. S. Ntalampiras and N. Fakotakis. "Modeling the Temporal Evolution of Acoustic Parameters for Speech Emotion Recognition." IEEE Transactions on Affective Computing 3.99, pp. 116-125, 2012.
29. H. Hu, M. -X. Xu and W. Wu, "GMM Supervector Based SVM with Spectral Features for Speech Emotion Recognition," *2007 IEEE International Conference on Acoustics, Speech and Signal Processing - ICASSP '07*, Honolulu, HI, USA, 2007, pp. IV-413-IV-416, doi: 10.1109/ICASSP.2007.366937.
30. D. Neiberg, K. Laskowski, and K. Elenius. "Emotion Recognition in Spontaneous Speech Using GMMs." INTERSPEECH, 2006.
31. C. -H. Wu and W. -B. Liang, "Emotion Recognition of Affective Speech Based on Multiple Classifiers Using Acoustic-Prosodic Information and Semantic Labels," in IEEE Transactions on Affective Computing, vol. 2, no. 1, pp. 10-21, Jan.-June 2011, doi: 10.1109/T-AFFC.2010.16.
32. Y. Kim, H. Lee and E. M. Provost, "Deep learning for robust feature generation in audiovisual emotion recognition," 2013 IEEE International Conference on Acoustics, Speech and Signal Processing, Vancouver, BC, Canada, 2013, pp. 3687-3691, doi: 10.1109/ICASSP.2013.6638346.
33. W. -L. Zheng, J. -Y. Zhu, Y. Peng and B. -L. Lu, "EEG-based emotion classification using deep belief networks," 2014 IEEE International Conference on Multimedia and Expo (ICME), Chengdu, China, 2014, pp. 1-6, doi: 10.1109/ICME.2014.6890166.
34. K. Han, D. Yu, and I. Tashev. Speech emotion recognition using deep neural networks and extreme learning machines. In INTERSPEECH, 2014.
35. W. Q. Zheng, J. S. Yu and Y. X. Zou, "An experimental study of speech emotion recognition based on deep convolutional neural networks," 2015 International Conference on Affective Computing and Intelligent Interaction (ACII), Xi'an, China, 2015, pp. 827-831, doi: 10.1109/ACII.2015.7344669.
36. H. M. Fayek, M. Lech and L. Cavedon, "Towards real-time Speech Emotion Recognition using deep neural networks," 2015 9th International Conference on Signal Processing and Communication Systems (ICSPCS), Cairns, QLD, Australia, 2015, pp. 1-5, doi: 10.1109/ICSPCS.2015.7391796.
37. Busso, C., Bulut, M., Lee, CC. *et al.* IEMOCAP: interactive emotional dyadic motion capture database. *Lang Resources & Evaluation* **42**, 335–359 (2008). https://doi.org/10.1007/s10579-008-9076-6
38. A. Satt, S. Rozenberg, R. Hoory. "Efficient Emotion Recognition from Speech Using Deep Learning on Spectrograms" in INTERSPEECH, Stockholm, 2017.
39. Ekman, P. (1972). Universal and Cultural Differences in Facial Expression of Emotions. In J. Cole (Ed.), Nebraska Symposium on Motivation (pp. 207-283). Lincoln: University of Nebraska Press.
40. B. Schuller, G. Rigoll and M. Lang, "Speech emotion recognition combining acoustic features and linguistic information in a hybrid support vector machine-belief network architecture," 2004 IEEE International Conference on Acoustics, Speech, and Signal Processing, Montreal, QC, Canada, 2004, pp. I-577, doi: 10.1109/ICASSP.2004.1326051.
41. M. Tahon and L. Devillers, "Towards a Small Set of Robust Acoustic Features for Emotion Recognition: Challenges," in IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 24, no. 1, pp. 16-28, Jan. 2016, doi: 10.1109/TASLP.2015.2487051
42. W. Lim, D. Jang and T. Lee, "Speech emotion recognition using convolutional and Recurrent Neural Networks," 2016 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA), Jeju, Korea (South), 2016, pp. 1-4, doi: 10.1109/APSIPA.2016.7820699
43. R. W. Picard, E. Vyzas and J. Healey, "Toward machine emotional intelligence: analysis of affective physiological state," in IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 23, no. 10, pp. 1175-1191, Oct. 2001, doi: 10.1109/34.954607.
44. Busso, C., Bulut, M., Lee, CC. et al. IEMOCAP: Interactive emotional dyadic motion capture database. Lang Resources & Evaluation 42, 335–359 (2008). https://doi.org/10.1007/s10579-008-9076-6
45. Scherer, K. R., & Wallbott, H. G. (1994). Evidence for universality and cultural variation of differential emotion response patterning. Journal of Personality and Social Psychology, 66(2), 310–328.)

46. (Li et al., 2017). Transfer learning and multi-modal emotion recognition have garnered interest recently as viable approaches to enhance the robustness and generalization of emotion detection models (arXiv:2202.08974 [cs.SD]).

47. Junfeng Zhang, Lining Xing, Zhen Tan, Hongsen Wang, Kesheng Wang, Multi-head attention fusion networks for multi-modal speech emotion recognition, Computers & Industrial Engineering,).

48. H. Meng, T. Yan, F. Yuan and H. Wei, "Speech Emotion Recognition From 3D Log-Mel Spectrograms With Deep Learning Network," in IEEE Access, vol. 7, pp. 125868-125881, 2019, doi: 10.1109/ ACCESS.2019.2938007).

49. Zhang, L., Wang, L., Dang, J., Guo, L., Guan, H. (2018). Convolutional Neural Network with Spectrogram and Perceptual Features for Speech Emotion Recognition. In: Cheng, L., Leung, A., Ozawa, S. (eds) Neural Information Processing. ICONIP 2018. Lecture Notes in Computer Science(), vol 11304. Springer, Cham. https://doi.org/10.1007/978-3-030-04212-7_6

50. S.H. Shabbeer Basha, Sravan Kumar Vinakota, Viswanath Pulabaigari, Snehasis Mukherjee, Shiv Ram Dubey, AutoTune: Automatically Tuning Convolutional Neural Networks for Improved Transfer Learning, Neural Networks

51. Devamanyu Hazarika, Soujanya Poria, Roger Zimmermann, Rada Mihalcea, Conversational transfer learning for emotion recognition, Information Fusion,

52. S. Han, F. Leng and Z. Jin, "Speech Emotion Recognition with a ResNet-CNN-Transformer Parallel Neural Network," 2021 International Conference on Communications, Information System and Computer Engineering (CISCE), Beijing, China, 2021, pp. 803-807, doi: 10.1109/CISCE52179.2021.9445906

53. M. Tran and M. Soleymani, "A Pre-Trained Audio-Visual Transformer for Emotion Recognition," ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Singapore, Singapore, 2022, pp. 4698-4702, doi: 10.1109/ ICASSP43922.2022.9747278

54. S. Basodi, C. Ji, H. Zhang and Y. Pan, "Gradient amplification: An efficient way to train deep neural networks," in Big Data Mining and Analytics, vol. 3, no. 3, pp. 196-207, Sept. 2020, doi: 10.26599/ BDMA.2020.9020004.