International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 10 Issue: 11 | Nov 2023 www.irjet.net p-ISSN: 2395-0072

Literature Review On: "Speech Emotion Recognition Using Deep Neural Network"

Samarth Adkitte[1],Prof. Vina Lomte[2], Mansi Fale[3] ,Vaibhavi Kudale[4],Shivani Ugile[5]

[1-5] Department of Computer Engineering, RMD Sinhgad School of Engineering, Warje,Pune-58,India

---

Abstract - *Emotions play crucial role in human communication, influencing how information is conveyed and understood. Recognizing emotions from speech is a challenging yet essential task with various applications, including human-computer interaction, sentiment analysis, and mental health assessment. This paper presents an overview of Speech Emotion Recognition (SER) using Deep Neural Network techniques. Additionally, we review various evaluation metrics commonly used in SER research, such as accuracy, F1-score, and confusion matrices, to assess the performance of SER models objectively. Overall, this research contributes to growing body of knowledge in SER and serves as a comprehensive resource for researchers, practitioners, and developers interested in leveraging machine learning techniques to recognize emotions from speech signals.*

Key Words: Convolutional Neural Network(CNN), Emotion Detection , Speech Emotion Recognition, Deep Learning, Mental Health Assessment

1. INTRODUCTION

In the past few years, lots of countries around the world - both developed and improving - have been doing research on recognizing speech to get emotions. The searchings and improvements in speech emotion recognition technology research demonstrate that the benefits and limitations of recognition have a substantial impact on speech recognition outcomes . The effect of emotions not only influences individual behavior but also have broader implications for social media interactions. By emotions people can express their thoughts and they can communicate easily. Emotion recognition has the potential to facilitate machines' understanding of human emotions. In human emotional interactions, speech and facial expressions often possess shared thematic and temporal characteristics. Consequently, speech and facial expressions have become a topic of developing interest in the field of emotion computing research. In many scenarios, it's essential to recognize covered emotions because they correspond to the actual emotional state. In the field of emotion recognition, the fundamental concept is to accurately get an individual's emotional state and look forward to understanding the current scenario of his/her mind, but sometimes it may not be able to possess the actual state of the emotion. In 1969, the existence of micro emotional expressions was first identified. These subtle facial expressions disclose concealed emotions of individuals, providing insight into their genuine and real emotional state even when they attempt to conceal it. The detection of micro expressions can assist in accurately understanding an individual's emotional state, even when they are aim to mask or hide their emotions. Recently, the progress made in the development of conversational agents has become increasingly well known in recent years. This has shined a surge of research interest in computationally modelling emotion recognition in conversations. The arousal/valence model is the most widely used model for identifying and classifying emotions. This model submits that emotions are distributed within a two dimensional space, consisting of both arousal and valence. The arousal axis ranges from calm to excited, while the valence axis ranges from positive to negative. When both the arising and valence levels are negative, it recommends a depressed emotional state. Conversely, when both levels are positive, it indicates a feeling of excited state. The emotions represented on both axes include fear, anger, sadness, happiness, disgust, acceptance, anticipation, and surprise, and can be explained accordingly . There are two approaches to Speech Emotion Recognition (SER). The first involves estimating the location of emotions in either two or three dimensions, while the second involves classifying emotions into specific categories such as "sadness," "joy," and "sorrow," among others. While conventional SER methods typically aim to evaluate a single emotion, it is common for multiple emotions to be explained in human speech, with varying degrees of intensity . Most strategies that analyze a speaker's words' high and low pitches are essential in identifying or recognizing a person's emotions. Words like "great" and "awesome," for instance, indicate that someone is pleased with something. A person is upset when they say, "Stay quiet" or "Get it off," on the other hand. Emotional voice conversion (EVC) is a voice conversion (VC) technique used to transfer the emotional style of one person to another. EVC has a wide range of potential applications, including text-to-speech (TTS) systems. It is important to note that emotional voice conversion differs from traditional speech voice conversion in different ways. While speech voice conversion is primarily related with changing speaker identity, emotional voice conversion is concentrated on transferring the emotional state of a speaker. Finding the best selective speech representation for the task at hand and a lack of data are two major issues in the field of Speech Emotion Recognition (SER).

2. LITERATURE SURVEY:

The research paper talks about how computers can understand human emotions from the sound of our voice. There are two important steps: first, the computer needs to pick out important information from the sound. Second, it uses a special program called a classifier to figure out what emotion the person is feeling. Right now, scientists are using advanced methods like deep learning to make this process work better. There are also high-level organized methods for speech signals and some of them are using low-level features for speech emotion recognition. In this research paper, we reviewed 20 research papers and most of them are based upon the DNN and CNN from model building of speech emotions present in the speech signal. It was found that many achievements and research have been made in different countries and in different universities on SER, while (SER) has some potential benefits, and advantages that it recognizes the true emotions of the person possess but also poses certain limitations that can hinder its accuracy in recognizing emotions. and to overcome these effects the technology of Deep and shallow neural network is proposed [3][5]. Emotions have an essential function in daily life, thus interactions between people and computers must be harmonic. [1].A person's significant emotional state cannot be accurately assessed by a conventional standard emotion identification system. Hence, emotions are performed on the concealed emotional speech signals, the signals that are synthesized by the standard emotion speech signal. In short, the fundamental principle is to comprehend the person's actual emotional condition. [2]. The model states that the emotions are distributed into two phases first having range from calm to excited and other having range from positive to negative [6]. The relevant emotional indicators found in the utterances have been carefully chosen and retrieved by the SER system. [7]. The analysis of the dataset has revealed that songs are generally characterized by a more pronounced expression of emotions, such as joy, sadness, anger, and fear, while speeches tend to exhibit a more restrained emotional content, and this can be concluded by applying different methods on a dataset [7]. A human being may have multiple emotions at a time to understand this concept an emotional speech database is used. After performing statistical analysis on database, it was found that most samples contain multiple emotions [8]. For continuous emotion recognition convolutional neural networks, deep neural networks are used. It takes the characteristics from the raw signal to obtain situational data details. Ryota Sato et.all, [11] proposed that textual information for emotion identification and acoustic features for speech recognition both are used by the (DNN) deep neural networks to obtain hidden feature representation for both modalities. These features are then interlinked to classify the emotion of the speaker [11]. [13] In this research paper it trains the model with the emotions of users using self-referential features. It analyses the emotions of users and calculates the accuracy of the model to predict the neutral value of all the data.

3.Algorithmic Survey

Table 1: Algorithmic Survey of Research Studies

Sr.No.	Publication Detail	Algorithm used	Dataset	Accuracy	Reasearch Gap Identified
1.	Speech Emotion Recognition using Dialogue Emotion Decoder and CNN Classifier	SVM,CNN,DNN	Kaggle	90%	To explore novel approaches to improve the accuracy and robustness of speech emotion recognition systems.
2.	Improved Cross-Corpus Speech Emotion Recognition Using Deep Local Domain Adaptation	CNN	Kaggle	85%	To overcome these challenges and make SER systems more robust and adaptable across different emotional speech datasets.
3.	Analysis of Concealed Anger Emotion in a Neutral Speech Signal	CNN,SVM	Kaggle	89%	Based on the assessment of the likelihood calculated from their respective T scores, the monitoring and analysis of the progression of emotion is achieved
4.	A Parallel-Model Speech Emotion Recognition Network Based on Feature Clustering	CNN,DNN	Kaggle	89%	To calculate the F-Emotion value of the extracted speech emotion features for each emotion category.
5.	Speech Emotion Recognition Based on Self-Attention Weight Correction for Acoustic and Text Features	CNN	Kaggle	87%	To emphasize the Speech segments with low CM as segments with a higher probability of containing emotion in the acoustic feature.

6.	Human–Robot Collaboration Using Sequential-Recurrent-Convolution-Network-Based Dynamic Face Emotion and Wireless Speech Command Recognitions	CNN	Kaggle	90**%**	To make it more effective and applicable to achieve the corresponding feature vector of facial emotion.
7.	Disentangled Speaker Embedding for Robust Speaker Verification	CNN,DNN	Kaggle	90%	To produce synthetic samples that share the common representations with the original data.
8.	Creation and Analysis of Emotional Speech Database for Multiple Emotions Recognition	CNN	Kaggle	87%	Responses of evaluators are non-independence, that the utterances in the database quite often have multiple emotions, and that intensity values of emotions are required.
9.	An Affective Service based on Multi-Modal Emotion Recognition, using EEG enabled Emotion Tracking and Speech Emotion Recognition	CNN	Kaggle	85%	To make it more effective and applicable to human interaction derives from emotional awareness.
10.	Multi-Classifier Interactive Learning for Ambiguous Speech Emotion Recognition	MCIL	Kaggle	89%	To construct ambiguous labels of emotion, which can better represent ambiguous emotion.
11.	Speech Emotion and Naturalness Recognitions With Multitask and Single-Task Learnings	DNN	Kaggle	89%	To be improved in the future work may also include a balancing strategy to improve the model performance, as well as mapping continuous scores to ordinal labels.
12.	Jointly Predicting Emotion, Age, and Country Using Pre-Trained Acoustic Embedding	CNN,RNN	Kaggle	87%	To tackle the limitations of the current SSLs, such as a smaller size of pre-training data, a large number of emotion categories, and unbalanced data
13.	3D Convolutional Neural Network for Speech Emotion Recognition With Its Realization on Intel CPU and NVIDIA GPU	CNN	Kaggle	85%	To expedite the 3D CNN on the datasets by providing lower runtimes than the CPU executions.
14.	Hybrid LSTM-Transformer Model for Emotion Recognition From Speech Audio Files	CNN,DNN,RNN	Kaggle	86%	To overcome the problem of data shortage in training and testing datasets.
15.	Autoencoder With Emotion Embedding for Speech Emotion Recognition	CNN,RNN,SVM	Kaggle	89%	To help the model extract deep attention features. In addition, the use of text information can be a measure to further improve the accuracy of SER.
16.	Transfer Subspace Learning for Unsupervised Cross-Corpus Speech Emotion Recognition	KNN,SVM	Kaggle	85%	To improve the performance.With the development of deep learning techniques, its strong nonlinear representation ability will help bridging the source and target domains.

17.	Learning Deep Binaural Representations With Deep Convolutional Neural Networks for Spontaneous Speech Emotion Recognition	CNN	Kaggle	87%	To integrate different binaural representations. explore other advanced fusion methods such as graph-based fusion graph-based fusion with metric learning
18.	Multimodal Emotion Recognition With Temporal and Semantic Consistency	CNN	Kaggle	89%	To make it more effective and Adding more emotions to the system since this system can identify only 8 emotions. .
19.	Speech Emotion Recognition using Machine Learning	KNN	Kaggle	86%	To make it more effective and applicable the emotional speech database which contains multiple emotions and intensities labels
20	Speech Emotion Recognition using Deep Learning Techniques:A Review	CNN	Kaggle	84%	To make it more effective and Adding more emotions to the system since this system can identify only 8 emotions.

4.Technological Survey

Table 2: Technological Survey

Years	2022-23	2021-22	2020-21	2019-20	2018-19
Methodology used	-Transformer-based architecture for capturing contextual information in multilingual speech. -Online learning techniques for real-time emotion recognition.	-Support Vector Machines (SVM) for classification. -Convolutional Neural Networks for feature extraction. Recurrent Neural Networks for modeling temporal patterns.	-Convolutional Neural Networks (CNNs) for feature extraction from audio spectrograms. Long Short-Term Memory (LSTM) networks for modeling temporal dependencies. -Transformer-based architectures for sequence-to-sequence modeling of audio data.	-Combination of deep learning models (e.g., CNNs, LSTMs) with hand-crafted features .Ensemble models to combine predictions from different feature sets. -Transfer learning techniques to adapt models from one dataset to another.	-Deep neural networks, including Convolutional Neural Networks (CNNs) for acoustic feature extraction and Long Short-Term Memory (LSTM) networks for temporal modeling.

5. CONCLUSION

Speech Emotion Recognition (SER) is an important term as it helps to recognize the emotions from Humans as proposed in the work. Speech Emotion Recognition (SER) using Machine Learning is a dynamic and rapidly evolving field with significant potential to revolutionize human- computer interaction, mental health care, entertainment, and various other industries. SER systems aim to automatically detect and classify human emotions in spoken language, providing valuable insights into emotional states and enabling more empathetic and context-aware technology. It helps to build the communications between Humans and machines. In this paper, we came up with a new way to recognize emotions using Convolutional Neural Networks (CNN). Our technique takes input data and uses these tools to tell us what emotions the person is feeling. With the help of research papers, work try to identify the gaps present in the existing literature survey also, concluded that CNN may give best result as TRUE and FALSE emotions by applying the several tasks as listed above like pre-processing, feature extraction and classifier. As researchers and practitioners continue to innovate in this field, the possibilities for SER are boundless, offering promising prospects for the future of technology and human communication.

For future directions, Investigate the latest deep learning architectures being used for SER, such as transformers, graph neural networks, or more advanced recurrent and convolutional networks. Analyze how these architectures have improved the accuracy and robustness of emotion recognition.For future of real world applications ,Investigate the practical applications of SER in various fields, including mental health support, customer service, human-computer interaction, and virtual reality.

REFERENCES

[1] Linqin Cai, Jiangong Dong, Min Wei, "Multi-Modal Emotion Recognition from Speech and Facial Expression Based on Deep Learning" 2020 Chinese Automation Congress (CAC) — 978-1-7281-7687-1/20/ IEEE, pp. 5726-5729.M. Young, The Technical Writer's Handbook. Mill Valley, CA: University Science, 1989.

[2] Vamsi Vijay Mohan Dattada, Dr M Jeevan, "Analysis Of Concealed Anger Emotion In A Neutral Speech Signal" University of Warwick.

[3] Jian Wang, Zhiyan Han, "Research on Speech Emotion Recognition Technology based on Deep and Shallow Neural Network" Proceedings of the 38th Chinese Control Conference July 27-30, 2019, Guangzhou, China, pp. 3555-3558

[4] H. Xu, H. Zhang, K. Han, Y. Wang, Y. Peng, and X. Li, "Learning alignment for multimodal emotion recognition from speech," in Proc. INTERSPEECH, 2019, pp. 3569.3573.

[5] B. W. Schuller, "Speech emotion recognition: Two decades in a nutshell, benchmarks, and ongoing trends," Commun. ACM, vol. 61, no. 5, pp. 90–99, 2021

[6] S. Gupta et al., "Speech emotion recognition using svm with thresholding fusion," in Proc. Int. Conf. Signal Process. Integr. Netw., 2021, pp. 570–574.

[7] S. Bhosale, R. Chakraborty, and S. K. Kopparapu, "Deep encoded linguistic and acoustic cues for attention based end to end speech emotion recognition," in Proc. Int. Conf. Acoust., Speech Signal Process., 2020, pp. 7189–7193.

[8] Zhao, Y. Zheng, Z. Zhang, H. Wang, Y. Zhao, and C. Li, "Exploring spatio- temporal representations by integrating attention-based bidirectional-LSTM- RNNs and FCNs for speech emotion recognition," in Proc. Interspeech, 2020,pp. 272–276

[9] M. El Ayadi, M. S. Kamel, and F. Karray, " Survey on speech emotion recognition: Features, classification schemes, and databases," Pattern Recognition, vol.44, no.3, pp.572– 587, 2021.

[10] B. W. Schuller, "Speech emotion recognition: two decades in a nutshell, bench- marks, and ongoing trends," Communications of the ACM, vol.61, no.5, pp.90–99, 2019

[11] CHENGHAO ZHANG AND LEI XUE," Autoencoder With Emotion Embedding for Speech Emotion Recognition", IEEE Access (Volume: 9), 30 March 2021, 10.1109/ACCESS.2021.3069818

[12] LI-MIN ZHANG , GIAP WENG NG , YU-BENG LEAU 2 AND HAO YAN," A Parallel-Model Speech Emotion Recognition Network Based on Feature Clustering", IEEE Access (Volume: 11), 11 July 2023, 10.1109/ACCESS.2023.3294274

[13] BAGUS TRIS ATMAJA AND AKIRA SASOU," Evaluating Self-Supervised Speech Representations for Speech Emotion Recognition", IEEE Access (Volume: 10), 28 November 2022, 10.1109/ACCESS.2022.3225198

[14] MOHAMMAD REZA FALAHZADEH , EDRIS ZAMAN FARSA , ALI HARIMI , ARASH AHMADI AND AJITH ABRAHAM," 3D Convolutional Neural Network for Speech Emotion Recognition With Its Realization on Intel CPU and NVIDIA GPU", IEEE Access (Volume:10), 26 October 2022, 10.1109/ACCESS.2022.3217226.

[15] NA LIU , BAOFENG ZHANG , BIN LIU , JINGANG SHI, LEI YANG , ZHIWEI LI, AND JUNCHAO ZHU," Transfer Subspace Learning for Unsupervised Cross-Corpus Speech Emotion Recognition", IEEE Access (Volume: 9), 02 July 2021, 10.1109/ACCESS.2021.3094355

[16] SHIQING ZHAN , AIHUA CHEN , WENPING GUO , YUELI CUI , XIAOMING ZHAO , AND LIMEI LIU," Learning Deep Binaural Representations With Deep Convolutional Neural Networks for Spontaneous Speech Emotion Recognition", IEEE Access (Volume: 8), 23 January 2020, 10.1109/ACCESS.2020.2969032.

[17] BAGUS TRIS ATMAJA, AKIRA SASOU AND MASATO AKAGI," Speech Emotion and Naturalness Recognitions With Multitask and Single-Task Learnings", IEEE Access (Volume: 10), 07 July 2022, 10.1109/ACCESS.2022.3189481

[18] FELICIA ANDAYANI , LAU BEE THENG , MARK TEEKIT TSUN AND CASLON CHUA," Hybrid LSTM-Transformer Model for Emotion Recognition From Speech Audio Files", IEEE Access (Volume: 10), 31 March 2022 , 10.1109/ACCESS.2022.3163856.

[19] JENNIFER SANTOSO , TAKESHI YAMADA , KENKICHI ISHIZUKA , TAIICHI HASHIMOTO , AND SHOJI MAKINO," Speech Emotion Recognition Based on Self-Attention Weight Correction for Acoustic and Text Features", IEEE Access (Volume: 10), 03 November 2022, 10.1109/ACCESS.2022.3219094

[20] Ryota Sato, Ryohei Sasaki, Norisato Suga &Toshihiro Furukawa," Creation and Analysis of Emotional Speech Database for Multiple Emotions Recognition", in Proc. conf. Oriental COCOSDA, 07 November 2020.