IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.53
Volume 4, Issue 1, September 2024

The Role of Deep Learning in Speech-Based Emotional Intelligence Systems

Shaik Karimunnisa Begum[1] and Dr. Kusum Rajawat[2]
Research Scholar, Department of Computer Science[1]
Research Guide, Department of Computer Science[2]
Sunrise University, Alwar, Rajasthan, India

Abstract: *Pattern recognition and natural language processing are emphasizing voice emotion recognition. Deep learning algorithms have improved spoken emotion identification in recent years. Some speech emotion recognition research lacks a broad comparison of deep learning models and methodologies. This makes identifying the best tactics and their pros and cons difficult. Thus, this work reviews deep learning methods for voice emotion identification in detail. The method is a comparative literature assessment of relevant articles on data collecting and deep learning. EMO-DB, RAVDESS, TESS, CREMA-D, IEMOCAP, and Danish Emotional Speech Databases will be explored. German-speaking EMO-DB and Danish-speaking Danish Emotional Speech Database are the only datasets not in English. These datasets mostly returned basic emotions like happiness, grief, neutrality, disgust, surprise, and fury. Deep learning has enhanced spoken emotion recognition, according to the study. The CNN-RNN combination is a complicated deep learning model that extracts acoustic information to reliably recognize speech emotion. This will effect contact center analytics, emotional computing, human-computer interaction, psychological research, and clinical diagnostics.*

Keywords: Deep learning, deep neural network, convolutional neural network, Speech emotion recognition using recurrent neural networks

I. INTRODUCTION

Speech emotion recognition has gone from a specialty to a vital aspect of human-computer interaction (Chen et al., 2017). This change was partly caused by speech recognition accuracy improvements. Voice-activated devices and virtual assistants are also popular. Google Assistant, Siri, Cortana, and Alexa are examples of speech-based technology that has become commonplace. Users may send messages, play music, make reminders, and operate smart home equipment with voice commands alone since these virtual assistants understand and respond to speech.

Lane & Georgiev (2015) found that direct vocal contact eliminates the need for input devices and allows listeners to respond naturally to computers. A proposed system aims to allow Deep learning approaches for vocal emotion identification are the study's focus. Comparative literature reviews of deep learning algorithms and datasets are employed. This work is composed as follows. Section II explains the study's purpose. Section II summarizes the method. Section III of this study discusses deep learning speech emotion recognition algorithms. Section IV covers voice emotion recognition datasets. Section V presents findings and discussion, while Section VI presents conclusions.

Spoken conversation systems are being used in contact center talks, car driving systems, and medical applications to assess emotions (Nassif et al., 2019). Despite these advancements, HCI systems encounter several challenges when they go from lab testing to real-world use (Balomenos et al., 2005). Maintaining accuracy and efficiency is challenging when tackling resilience, voice signal unpredictability, ambient noise, and real-time processing.

Speech signals fluctuate in tone, pitch, and pronunciation, making it difficult to design algorithms that can consistently assess and grasp the intended message. Environmental noise complicates the operation by disrupting the input signal. Real-time processing adds complexity, requiring systems to react properly and quickly without sacrificing performance. These challenges must be addressed to produce reliable speech processing systems that can handle everyday conversation. So, to improve machine-based emotion recognition, we must focus on solving these issues. Determining an individual's emotional state is a particular job that may be used to standardize any emotion model. Discrete emotion

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.53
Volume 4, Issue 1, September 2024

models are essential among the numerous models used to classify emotions. Discussed emotions include boredom, anger, surprise, disgust, fear, pleasure, neutrality, grief, and joy (Vogt et al., 2008). Another prominent model uses a continuous three-dimensional space including arousal, valence, and potency.

Deep learning, a burgeoning topic of machine learning, has garnered attention lately (Schmidhuber, 2015).

Deep learning has several benefits over standard Speech Emotion Recognition (SER) approaches. These include the ability to detect complicated patterns and attributes without human feature extraction or modification. Deep learning systems can handle unlabeled data well and find key traits. Deep neural networks (DNNs) have feedforward structures with hidden layers between input and output. CNNs and DNNs handle pictures and videos well. RNNs and LSTMs are good for language-related classification tasks including natural language processing and voice emotion recognition (Schmidhuber, 2015). These models are effective classifiers but have limitations. CNNs learn features from high-dimensional input data well, but they need a lot of storage capacity and tend to learn from minor biases and variations. Similar to LSTM-based RNNs, they represent long-range continuous text data and manage variable input data well.

Objectives Of The Study

This study's goal is to do a thorough analysis of deep learning methods for SER. In addition to identifying possible obstacles and future research areas, the study seeks to provide useful information on the most recent state-of-the-art methods and advancements in the domains.

The project seeks to accomplish the following precise objectives:

- To investigate several deep learning techniques used in speech emotion recognition, such as CNNs, RNNs, and their variants.
- To determine the shortcomings and difficulties that the current deep learning approach in SER faces;
- To evaluate the efficacy of several deep learning techniques in terms of accuracy and generalization capacities.

This extensive overview's primary goal is to provide researchers with a complete grasp of the status of deep learning methods for speech emotion detection at the moment. By accomplishing this goal, the research hopes to provide a solid basis for further developments in this quickly changing industry.

II. METHODS

This study uses literature analysis of relevant publications. This technique surveys and evaluates deep learning studies on speech-based emotion identification. This strategy involves methodically obtaining relevant articles, understanding and analyzing their content, and evaluating the sources' quality and reliability. This paper examines contemporary deep learning speech emotion identification research. Conferences, scientific publications, and other media provide relevant content. To understand data, methods, and results, each publication was extensively evaluated after collection. The research methodology, sample size, and statistical analysis were considered while assessing the articles' validity and reliability. This method aims to teach Deep Learning approaches for voice emotion recognition and the latest findings that may be utilized to further this field.

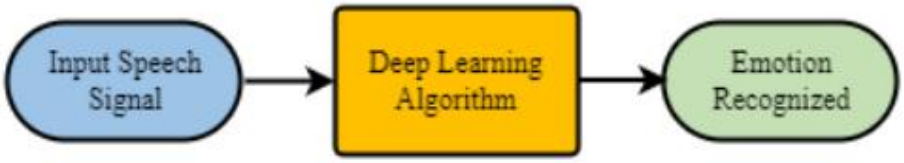

Fig. 1. Deep Learning Flow Mechanism

Input Speech Signal

Voice Input employs deep learning to discern voice emotions from speech sound signals. We can capture human voices speaking or producing noises with different emotions. Nassif et al. (2019) state that speech signals provide various types of information, including speaker recognition, emotion detection, health assessment, voice detection, accent detection, and age detection.

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.53
Volume 4, Issue 1, September 2024

Speaker recognition has two basic components. Speaker identification and authenticity are two types of speaker recognition. Speaker identification aims to identify the registered speaker of an utterance. This is useful in public and media settings. Interactions with government or district organizations, radio station calls, insurance agencies, and recorded transactions are examples (Reynolds, n.d., 2002).

Emotion recognition analyzes speech cues to recognize unknown emotions. Emotion detection has two branches: identification and verification. The initial branch seeks undiscovered feelings. In contrast, the second branch compares the input audio signal to several emotion models to determine the best match to confirm an emotion. Emotion detection determines whether an emotion is known or unknown. Emotion detection may capture speakers' emotions during contact center discussions and provide feedback to operators for monitoring (Petrushin, 2020). Other uses include voicemail sorting her messages by caller sentiment and recognizing people with emotional speech patterns (like gladness) in suspicious contexts. This involves doing.

Voice signals are humanity's most sophisticated and quickest form of communication, according to Swain et al. (2018). Complex signal processing systems, networking, and various signalling units deliver message, speaker, and language information. Recently, significant study has focused on translating human voice into word strings. Though breakthroughs have been made, robots cannot understand the speaker's emotional state and cannot discern particular emotions. Speech emotion identification is an emerging academic topic that analyzes speech signals to identify emotions. The goal is to make voice the most efficient human-machine communication. However, robots' low intelligence makes speech recognition difficult. Researchers are trying to overcome this issue and improve robots' voice recognition.

Deep Learning Algorithm

Deep learning algorithms, or deep neural networks, are based on artificial neural networks. Many hidden layers make these networks "deep" compared to standard neural networks, which may have hundreds of layers. Deep learning algorithms have outperformed standard machine learning algorithms, hence research has switched to them. This pattern also appears in his SER. These techniques eliminate feature extraction and selection by letting the deep learning process find the necessary characteristics. Akçay & Oğuz (2020) highlight CNNs and RNNs as popular deep learning algorithms in SER research.

Recurrent Neural Networks (RNNs)

Data sequencing and emotion classification are done using RNNs like LSTM and GRU. LSTM-RNN is used in speech emotion detection research. RNNs have limited short-term memory, but the LSTM design helps them access long-term memory. Gate RNNs like LSTM RNN manage long-term dependencies. Special "LSTM cells" feature inner repetitions as well as RNN outer repeats. LSTM RNNs include gate units with extra parameters and sigmoidal nonlinearities in addition to the input/output mechanism. LSTM (Long Short-Term Memory) networks' gating units control information flow by deciding whether to store, accept input, output, or delete. A typical LSTM cell has three gates: main, forget, and remember. The opening and shutting of these gates allows LSTM cells to manage information storage and input/output time (Akçay & Oğuz, 2020).

Lee and Tashev (2015) present RNN (prop.)-ELM, advancing speech emotion detection. The proposed technique improves accuracy by 12%, from 52.13% to 63.89%. From 57.91% to 62.85%, the system improves 5% in two cases. The research presents a novel voice emotion detection framework using a recurrent neural network (RNN) and a powerful learning approach. This unique technique solves the problems of collecting substantial contextual impacts in emotional speech and managing emotional labeling confusion. Recurrent neural networks and a maximum likelihood-based learning technique provide the suggested approach important insights into emotion detection and significant advances in the area.

Emotion detection systems help develop human behavior informatics and create effective human-machine interaction systems, according to Ho et al. (2020). These systems effectively and reliably handle human behavioral data, allowing natural communication. We present a multimodal speech emotion detection method using recurrent neural networks and self-versus-multiple-headed attention processes. She uses word embedding of MFCC audio signals and text data in this system. Training these functions concurrently in the time domain yields good performance on the IEMOCAP,

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.53
Volume 4, Issue 1, September 2024

MELD, and CMU-MOSEI datasets. However, this scenario may be improved. Instead of integrating modalities later, synchronize audio and text data at a lower level to strengthen their link. Perceptual linear prediction (PLP), chroma, and prosody may also be included. For text data, she selected emotionally important terms and filtered out irrelevant material. Through pre-training and refining, domain-her matching approaches are being used to transfer speech recognition expertise to emotion detection.

One year later, Byun & Lee (2021) developed a Korea emotional speech database for speech emotion analysis. The study used recurrent neural networks to combine features to improve emotion identification. Language-based emotion classification was performed in this research using the Korean Emotion Language Database. Speech fragments into shorter periods, and identifying emotions in those intervals improved accuracy to 83.81%. This was compared to a single long LSTM model's 75.51% speech emotion recognition accuracy. Thus, the research showed that shorter gaps improve speech analysis emotion recognition.

Deep Neural Networks (DNNs)

Attention processes may be used with Deep Neural Networks (DNNs) for voice emotion recognition. For emotion categorization, DNN architectures concentrate on important speech signals using attention methods. Was planned. Lieskovská et al. (2021) analyzed current SER advancements and examined how various attentional processes affect her SER system's performance. They employed an attentional DNN model for the investigation. The researchers examined how various attentional processes influenced her SER system using IEMOCAP, Emo-DB, and RECOLA, three popular datasets. Integration of attentional processes considerably increased SER system performance, according to the research. This implies that attentional processes enhance his SER's accuracy and effectiveness.

Scheidwasser-Clow et al. (2022) created the SER Adaption Benchmark (SERAB) to evaluate Deep Neural Network (DNN)-based utterance-level SER methods. SERAB (SER Analysis Benchmark) evaluates SER techniques' performance and generalization. SERAB is a comprehensive framework for comparing DNN-based speech and emotion representations because to their fast advancement. It comprises tasks in numerous languages, dataset sizes, and sentiment categories and delivers accurate performance and generalization estimates. Many recent baselines were used to evaluate SERAB's efficacy. All metrics showed that BYOL (Bootstrap Your Own Latent) techniques performed best. BYOL-A models pre-trained on Audio Set (BYOL-S) speech samples showed a 3% accuracy boost over the first technique. These assessment results may provide standards for emerging approaches like CvT-based methods presented in this paper.

Convolutional Neural Networks (CNNs)

A CNN, or Shift-Invariant Artificial Neural Network (SIANN), has specialized filters or regions in its hidden layer. The CNN's shift-invariant architecture lets it detect patterns and features in input data regardless of their spatial position. These filters or zones target certain input signal properties. According to (Hubel & Wiesel, 1968), the visual neural cortex is a specialized structure that processes visual information. This thought inspired their method. That adapts to input signal characteristics. Speech emotion detection relies on CNNs to extract features, analyze time-frequency patterns, model spatial information, use transfer learning, include multi-modal data, use attention processes, and participate in ensemble modeling. These apps improve emotion identification accuracy and efficacy. Convolutional Neural Networks (CNNs) may learn characteristics from complex input data. This function learns characteristics from even little changes and distorted appearances, therefore it requires a lot of memory during development. For this reason, CNNs normally include a convolutional layer followed by down sampling. Weng et al. (1993) suggested a convolutional layer with numerous filter banks whose weights are changed by backpropagation.Bertero & Fung (2017) developed a convolutional neural network that can recognize anger, pleasure, and melancholy with 66.1% accuracy. The researchers compared the trained network to His SVM's fundamental features. We utilized a crowdsourced dataset of his TED presentations, manually labeled by a student, to train and assess our technique. Theano was used to implement CNN. We also trained a linear SVM using INTERSPEECH 2009 Emotion Challenge characteristics for comparison analysis.

Zhao et al. (2019) developed a method for obtaining deep emotional traits to properly recognize speech emotion. This was done using CNNs and LSTM models. The researchers found that constructed networks performed well in voice

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.53
Volume 4, Issue 1, September 2024

emotion identification after extensive testing. These promising results demonstrate their method's ability to capture and understand spoken language's delicate emotional signals. Zhao et al.'s study advances emotion detection and lays the basis for future advances.

Mekruksavanich et al. (2020) used a one-dimensional convolutional neural network. In Thai language datasets, they classified negative emotions with 96.60% accuracy. The researchers also applied their approach to SAVEE, RAVDESS, TESS, Cream-D, and Thai datasets.

In the same year, Anvarjon et al. (2020) developed a lightweight model with low computing cost and good detection accuracy to provide an innovative and efficient speech emotion detection (SER) technique. Did. They selected his CNN method for this. Researchers employed IEMOCAP and Emo DB language records to test the proposed approach. They found that their CNN-based His-SER system beats the existing state-of-the-art in detection accuracy. This research might improve SER and improve real-world solutions.

Auto Encoder

Autoencoders can learn features and reduce dimensionality for voice emotion detection. Autoencoders improve speech emotion identification by using unsupervised learning, feature learning, dimensionality reduction, and denoising. Autoencoders are useful for extracting relevant features and enhancing emotion detection algorithms, especially with minimal labeled data. In their 2020 research, Aouani & Ayed introduced the Harmonic to Noise Rate (HNR) to improve emotion identification. This new feature was integrated with MFCC coefficients, ZCR, and TEO. They used an Auto-Encoder, a sophisticated dimension reduction approach, to integrate. They used the RML emotion database, a popular emotion identification benchmark, to test their algorithm. Comparing their method to other emotion identification systems, the findings were impressive. They also found that auto-encoder dimension reduction increased identification rates. That work demonstrated the possibility of using the Harmonic to Noise Rate, along with other features including an Auto-Encoder, to enhance SER. They contribute much to the development of strong and accurate emotion analysis and comprehension systems.

The researchers tested Zhang & Xue (2021)'s technique using IEMOCAP and EMODB, two popular datasets. The study showed that the suggested model outperformed existing spoken emotion detection algorithms. The research also found that the method improved classification accuracy. classification rate of 71% for the IEMOCAP dataset and 95.6% for his EMODB HE. These findings suggest that latent expression and acoustic feature-based spoken emotion identification systems may improve accuracy.

The Combination of Deep Learning Technique

Tarunika et al. (2018) employed DNNs and k-NNs to recognize emotional content in speech, especially in the setting of a frightened frame of mind. They trained and tested their models using deep learning and their own sound database. Their study made significant contributions to palliative care systems by increasing emotion identification, which may improve the well-being of patients.

Pandey et al. (2019) tested deep learning methods for speech-based emotion detection and classification. That work used CNNs and LSTMs to record emotion using mel and magnitude spectrograms and MFCC to assess the impact. EMODB and IEMOCAP datasets were utilized for experiments. The log-mel spectrogram shows that the CNN+LSTM architecture performs well.

Yao et al. (2020) integrated DNNs, CNNs, and RNNS to create a framework. They used IEMOCAP databases to do this. Weighted pooling in neural networks with an attention mechanism focused on emotionally relevant input segments. Thus, this method improved task accuracy. This novel system might improve emotional content analysis in numerous applications.

Datasets For SER

Speech emotion recognition (SER), like other deep learning applications, requires a good training dataset. This includes manual annotation of samples by human agents, however since emotions are subjective, people may perceive and classify emotional voices differently. Another individual may see wrath as enthusiasm. We need a dependable system that lets several people assess and label samples and confidently choose the right label to remove this uncertainty.

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.53
Volume 4, Issue 1, September 2024

Simulated, semi-natural, and natural language voice emotion databases exist. A simulated dataset is created by trained speakers reading the same text with various emotions. Nature-inspired collections invite viewers to feel diverse emotions. However, human listeners emotionally annotate real datasets from television stations, YouTube videos, and contact centers (Douglas-cowie et al., 2000).

Table: The classifications of databases that containing emotional speech

No.	Database	Language	Emotion
1.	EMO-DB Dataset	Use German Language	Sadness, Happiness. Boredom Disgust. Anger, Neutral
2.	Danish Emotional Speech Databases	Use Danish Language	Sadness, Anger, Neutral Surprise, joy
3.	RAVDESS Dataset	Use English Language	Sad, Happy, Angry, Surprised Fearful, Calm, Disgusted Neutral
4.	TESS Dataset	Use English Language	Pleasantly, Angry, Surprised Happy, Fearful, Disgusted, Sad Neutral
5.	CREMA-Dataset	Use English Language	Sadness, Happiness, Anger, Disgust, Neutral, Fear
6.	IEMOCAP Dataset	Use English Language	Happiness, Anger, Sadness Frustation

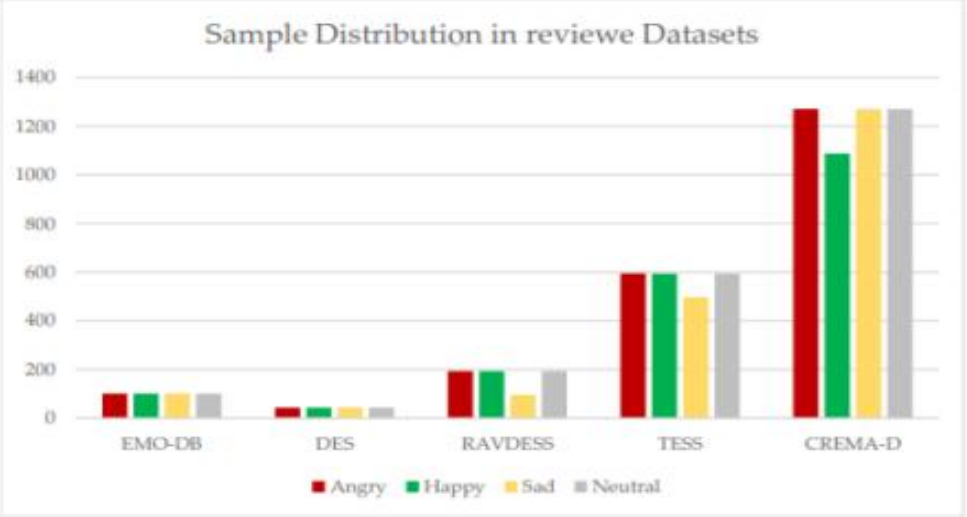

Fig. 2. Comparison the distribution of comparable emotions across the analyzed simulated datasets Source : (Abbaschian et al., 2021)

The paragraph summarizes the highest accuracy for each dataset, the system training characteristics, the approaches used, and, where appropriate, the layer count in each approach in the table below:

Table A comparison result of literature review (Abbaschian et al., 2021)

Research Title	Methodology and Number of Layers	Features	Dataset and Accuracy
Article by Zhao et al. "Speech Emotion Recognition Using Deep 1D and 2D CNN LSTM Networks" (2019) focuses on deep learning models, especially the application of his 1D and 2D CNN LSTM networks to speech emotion recognition tasks.	This paper proposes using DCNN with sequence length 5 and LSTM networks for the task at hand.	The paper explores the use of Pulse Code Modulation (PCM) and Log- Mel Spectrogram representations for the analysis of speech data.	In this study, the EMO-DB dataset achieved 95.33% accuracy and the IEMOCAP dataset achieved 86.16% accuracy.

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.53
Volume 4, Issue 1, September 2024

An article by Xie et al. titled "Speech Emotion Classification Using Attention-Based LSTMs" (2019) focuses on the application of attention-based long-short-term memory (LSTM) models to speech emotion classification tasks.	LSTM, DNN/5	The study examined various acoustic features for speech analysis, including MFCC and Zero Crosing	In this study, the eENTERFACE dataset achieved 89.6% accuracy and the GEMEP dataset achieved 57% accuracy. On the other hand, the CASIA dataset achieved 92.8% accuracy.
Article by Chatziagapi et al. The book entitled Data Augmentation Using His GANs for Speech Emotion Recognition (2019) focuses on the application of generative adversarial networks (GANs) for data augmentation in the context of speech emotion recognition.	This paper uses deep convolutional neural network (DCNN) architectures, specifically VGG19 model and GAN/19, for the present task.	The feature extraction process involves extracting 128 MFCCs	In this study, the IEMOCAP dataset achieved 53.6% accuracy and the Feel-25k dataset achieved 54.6% accuracy.
Article by Sahu et al. "Enhancing Speech Emotion Recognition Using Generative Adversarial Networks" (2018) explores the use of generative adversarial networks (GANs) to improve the performance of speech emotion recognition systems.	In the current task, the research uses a combination of Generative Adversarial Networks (GANs) and SVMs	This work uses the 1582- dimensional openSMILE feature space.	In this study, the IEMOCAP dataset achieved an accuracy of 60.29%.
Article titled "Adversarial Machine Learning and Speech Emotion Recognition": "Using Generative Adversarial Networks for Robustness" Latif et al. (2018) describes the application of generative adversarial networks (GANs) to improve the robustness of speech emotion recognition systems in the context of adversarial machine learning. are investigating.	The paper suggests the utilization of LSTM networks and GANs with a factor of 2 for the task at hand.	The paper focuses on the use of eGeMAPS features for speech analysis.	In this study, the aibo dataset achieved an accuracy rate of 64.86%, while the IEMOCAP dataset achieved an accuracy rate of 53.76%.
Article entitled "Unsupervised Learning Approach to Feature Analysis for Automatic Speech Emotion Recognition" by Eskimez et al. (2018) consider an unsupervised learning approach to feature analysis in the context of automatic speech emotion recognition.	The paper proposes use dof Convolutional Neural Networks (CNN) in combination with Variational Autoencoders (VAE) with different configurations, including sequence lengths of 5, 6, 4, 10, and 5, for the task at hand.	Studies analyzing speech data use logmel spectrograms.	In this study, the IEMOCAP dataset achieved an accuracy of 48.54%.
Article titled "A Variable Autoencoder for Learning Speech Emotion Latent	The paper investigates the utilization of		

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.53
Volume 4, Issue 1, September 2024

Expressions": A preliminary study by Latif et al. (2017) explores the application of variational autoencoders (VAEs) to learning latent representations of vocal emotion.	Variational Autoencoders (VAEs) in combination with two layers of LSTM networks with a sequence length of 4 for the task at hand.	The Log-Mel Spectrogram representation is utilized in the study for analyzing speech data.	In this study, the IEMOCAP dataset achieved an accuracy of 64.93%.

III. FINDINGS AND DISCUSSION

This research emphasizes HCI in voice emotion recognition. Voice emotion recognition is crucial to many applications, and we study its challenging implementation hurdles. This research then sets broad aims, including exploring deep learning methods for speech emotion detection, analyzing them, thoroughly assessing their performance, and highlighting their limits and obstacles. To accomplish these aims, this technique analyzes relevant papers' literature to get a complete knowledge. This study reviews deep learning technologies in SER.

Deep learning can automatically recognize complicated audio patterns without feature extraction or parameter adjustment, making it potential for SER. CNNs, RNNs, and their derivatives have performed well in image, video, and speech-based classification tasks. We must admit that these models have limits. CNNs learn features from tiny variances and distortions, which may be problematic, whereas LSTM-based RNNs need a lot of storage.

The study carefully explains deep learning techniques, focusing on RNNs and CNNs and their benefits in voice emotion identification. CNNs excel in extracting salient characteristics from complicated input data, whereas RNNs with LSTM cells access and use long-term memory well. The publication cites several important research showing how deep learning methods increase SER performance.

The study emphasizes the potential of deep learning algorithms, particularly RNNs with LSTM architecture and CNNs, in SER. Attentional processes are thought to enhance speech emotion recognition systems. Many research have shown that deep learning models can reliably recognize emotions from speech, necessitating the creation of benchmark frameworks to thoroughly test their performance and generalization capabilities.

IV. CONCLUSION

This article analyzes SER deep learning methods in detail. Recent research has thoroughly examined DNNs, RNNs, CNNs, AEs, and their combinations. These methods classify natural emotions including melancholy, happiness, neutrality, anger, contempt, fear, and boredom. The model is simpler to train and more efficient with shared weights. Deep learning approaches have many drawbacks, including their multi-layered design, lower efficiency in processing changing input data, and the potential of over learning layer-wise knowledge. Deep learning models like DNNs and RNNs have complicated internal architectures that make training, fine-tuning, and optimization difficult. Hyperparameter adjustment, computing resources, and longer training cycles are needed for numerous layers and coupled nodes.

Complex structures are difficult to manage and optimize. These models assume similar distributions for training and test data. However, large changes in the input data distribution, such as new classes or input features, might hamper model adaptation. Thus, deep learning models may fail to generalize to new examples, reducing efficiency and performance. This constraint may be solved by fine-tuning or retraining the model with new data.

This paper assesses deep learning methods' efficacy and limits and suggests ways to improve SER systems. Speech emotion recognition affects human-computer interaction, affective computing, psychology, and healthcare. It allows natural interactions, tailored therapies, and better emotional well-being measurement. Multimodal fusion, transfer learning, explain ability, long-term dynamics modeling, real-world applications, and ethics are future research objectives. These directions attempt to integrate numerous modalities, solve domain shift, improve transparency, capture temporal dynamics, assess practical value, and assure responsible development. These improvements will boost system performance and optimize SER's versatility.

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.53
Volume 4, Issue 1, September 2024

REFERENCES

[1]. Abbaschian, B.J., Sierra-Sosa, D., & Elmaghraby, A. (2021). Deep learning techniques for speech emotion recognition, from databases to models. In *Sensors (Switzerland),* 21(4), 1-27. MDPI AG. https://doi.org/10.3390/s21041249

[2]. Akçay, M.B. & Oğuz, K. (2020). Speech emotion recognition: Emotional models, databases, features, preprocessing methods, supporting modalities, and classifiers. *Speech Communication*, *116*(June 2019), 56–76. https://doi.org/10.1016/j.specom.2019.12.001

[3]. Anvarjon, T., Mustaqeem, & Kwon, S. (2020). Deep-net: A lightweight cnn-based speech emotion recognition system using deep frequency features. *Sensors (Switzerland)*, *20*(18), 1–16. https://doi.org/10.3390/s20185212 Aouani, H. & Ayed, Y.B. (2020). Speech Emotion Recognition with deep learning. *Procedia Computer Science*,*176*, 251–260. https://doi.org/10.1016/j.procs.2020.08.027

[4]. Balomenos, T., Raouzaiou, A., Ioannou, S., Drosopoulos, A., Karpouzis, K., & Kollias, S. (2005). Emotion analysis in man-machine interaction systems. *Lecture Notes in Computer Science*, 3361, 318–328. https://doi.org/10.1007/978-3-540-30568-2_27

[5]. Bertero, D. & Fung, P. (2017). A first look into a convolutional neural network for speech emotion detection. In Proceedings of the IEEE *International Conference on Acoustics, Speech, and Signal Processing* (ICASSP) (pp. 5115-5119). Clear Water. http://ieeexplore.ieee.org/document/7953131/

[6]. Byun, S.W. & Lee, S.P. (2021). A study on a speech emotion recognition system with effective acoustic features using deep learning algorithms. *Applied Sciences (Switzerland)*, *11*(4), 1–15. https://doi.org/10.3390/app11041890

[7]. Chatziagapi, A., Paraskevopoulos, G., Sgouropoulos, D., Pantazopoulos, G., Nikandrou, M., Giannakopoulos, T., Katsamanis, A., Potamianos, A., & Narayanan, S. (2019). Data augmentation using GANs for speech emotion recognition. *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, *2019-Septe*, 171–175. https://doi.org/10.21437/Interspeech.2019-2561

[8]. Chen, M., Zhou, P., & Fortino, G. (2017). Emotion Communication System. *IEEE Access*, *5*, 326–337. https://doi.org/10.1109/ACCESS.2016.2641480

[9]. Douglas-cowie, E., Cowie, R., & Schröder, M. (2000). A New Emotion Database: Considerations, Sources and Scope. *In*, 39–44.

[10]. Eskimez, S. E., Duan, Z., & Heinzelman, W. (2018). Unsupervised learning approach to feature analysis for automatic speech emotion recognition. In Proceedings of the 2018 IEEE *International Conference on Acoustics, Speech and Signal Processing* (ICASSP) (pp. 5099-5103). http://ieeexplore.ieee.org/document/8462417/

[11]. Ho, N. H., Yang, H. J., Kim, S. H., & Lee, G. (2020). Multimodal Approach of Speech Emotion Recognition Using Multi-Level Multi-Head Fusion Attention-Based Recurrent Neural Network. *IEEE Access*, *8*, 61672–61686. https://doi.org/10.1109/ACCESS.2020.2984368

[12]. Hubel, D. H. & Wiesel, T. N. (1968). Receptive fields and functional architecture of monkey striate cortex. *The Journal of Physiology*, *195*(1), 215–243. https://doi.org/10.1113/jphysiol.1968.sp008455

[13]. Latif, S., Rana, R., & Qadir, J. (2018). *Adversarial Machine Learning And Speech Emotion Recognition: Utilizing Generative Adversarial Networks For Robustness*. 1–7. http://arxiv.org/abs/1811.11402

[14]. Latif, S., Rana, R., Qadir, J., & Epps, J. (2017). *Variational Autoencoders for Learning Latent Representations of Speech Emotion: A Preliminary Study*. http://arxiv.org/abs/1712.08708

[15]. Lee, J. & Tashev, I. (2015). High-level feature representation using recurrent neural network for speech emotion recognition. *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, *2015-Janua*, 1537–1540. https://doi.org/10.21437/interspeech.2015-336

[16]. Lieskovská, E., Jakubec, M., Jarina, R., & Chmulík, M. (2021). A review on speech emotion recognition using deep learning and attention mmechanism. *Electronics (Switzerland)*, 10(10). https://doi.org/10.3390/electronics10101163

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.53
Volume 4, Issue 1, September 2024

[17]. Nassif, A. B., Shahin, I., Attili, I., Azzeh, M., & Shaalan, K. (2019). Speech Recognition Using Deep Neural Networks: A Systematic Review. *IEEE Access*, 7, 19143–19165. https://doi.org/10.1109/ACCESS.2019.2896880

[18]. Pandey, S.K., Shekhawat, H.S., & Prasanna, S.R.M. (2019). Deep learning techniques for speech emotion recognition: A review. *2019 29th International Conference Radioelektronika, RADIOELEKTRONIKA 2019 - Microwave and Radio Electronics Week, MAREW* 2019. https://doi.org/10.1109/RADIOELEK.2019.8733432

[19]. Petrushin, V.A. (2000). Emotion recognition in speech signal: Experimental study, development, and application. *6th International Conference on Spoken Language Processing, ICSLP* 2000, *Icslp*, 6–9. https://doi.org/10.21437/icslp.2000-791

[20]. Reynolds, D.A. (n.d.). *2002 Reynolds D - An overview of automatic speaker recognition.pdf*.

[21]. Sahu, S., Gupta, R., & Espy-Wilson, C. (2018). *On Enhancing Speech Emotion Recognition using Generative Adversarial Networks*. http://arxiv.org/abs/1806.06626

[22]. Scheidwasser-Clow, N., Kegler, M., Beckmann, P., & Cernak, M. (2022). Serab: a Multi-Lingual Benchmark for Speech Emotion Recognition. *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings*, *2022-May*, 7697–7701. https://doi.org/10.1109/ICASSP43922.2022.9747348

[23]. Schmidhuber, J. (2015). Deep Learning in neural networks: An overview. *Neural Networks*, *61*, 85–117. https://doi.org/10.1016/j.neunet.2014.09.003

[24]. Swain, M., Routray, A., & Kabisatpathy, P. (2018). Databases, features and classifiers for speech emotion recognition: a review. *International Journal of Speech Technology*, 21(1), 93–120. https://doi.org/10.1007/s10772-018-9491-z

[25]. Tarunika, K., Pradeeba, R.B., & Aruna, P. (2018, October 16). Applying Machine Learning Techniques for Speech Emotion Recognition. *2018 9th International Conference on Computing, Communication and Networking Technologies, ICCCNT 2018*. https://doi.org/10.1109/ICCCNT.2018.8494104

[26]. Vogt, T., André, E., & Wagner, J. (2008). Automatic recognition of emotions from speech: A review of the literature and recommendations for practical realisation. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, *4868 LNCS*, 75–91. https://doi.org/10.1007/978-3-540-85099-1_7

[27]. Xie, Y., Liang, R., Liang, Z., Huang, C., Zou, C., & Schuller, B. (2019). Speech Emotion Classification Using Attention-Based LSTM. *IEEE/ACM Transactions on Audio Speech and Language Processing*, 27(11), 1675– 1685. https://doi.org/10.1109/TASLP.2019.2925934

[28]. Yao, Z., Wang, Z., Liu, W., Liu, Y., & Pan, J. (2020). Speech emotion recognition using fusion of three multi-task learning-based classifiers: HSF-DNN, MS-CNN and LLD-RNN. *Speech Communication*, *120*, 11–19. https://doi.org/10.1016/j.specom.2020.03.005

[29]. Zhang, C. & Xue, L. (2021). Autoencoder with emotion embedding for speech emotion recognition. *IEEE Access*, *9*, 51231–51241. https://doi.org/10.1109/ACCESS.2021.3069818

[30]. Zhao, J., Mao, X., & Chen, L. (2019). Speech emotion recognition using deep 1D & 2D CNN LSTM networks.

[31]. *Biomedical Signal Processing and Control*, *47*, 312–323. https://doi.org/10.1016/j.bspc.2018.08.035