Analysis of Speech Emotion: A Survey of SER and FER with Multilingual Dialects

Arundhati Wani[1], Rujuta Kulkarni[2], Ananthan Nair[3], Vishvjita Savkare[4] and Ms. Punam Chavan[5]

[1,2,3,4] Arundhati Wani, Marathwada Mitra Mandal's College of Engineering, Pune

[5]Assistant Professor, Marathwada Mitra Mandal's College of Engineering, Pune

Abstract- With applications in customer service, healthcare, and entertainment, emotion identification is essential to enhancing human-computer interaction. Either Speech Emotion Recognition (SER) or Facial Emotion Recognition (FER) systems have been a major component of traditional approaches to emotion analysis. However, more resilient, multi-modal techniques that integrate speech and facial expressions for a thorough comprehension of human emotions are required due to the growing complexity of real-world applications. With an emphasis on how these systems might be combined for improved emotion recognition, this survey investigates the synergies between SER and FER technologies. An overview of the state of research in SER and FER is given in this study, with a focus on multi-modal systems that take dialect and cultural quirks into account. We hope to provide future research directions that could result in more precise and culturally sensitive emotion identification systems by reviewing current developments, difficulties, and possible solutions.

Keywords- emotion recognition; facial emotion recognition; dialects; multi-modal system

1.INTRODUCTION

For the purpose of improving human-computer interaction in domains such as virtual assistants, customer service, mental health monitoring, and entertainment, emotion identification technology is essential. Speech Emotion Recognition (SER), which evaluates speech cues, and face Emotion Recognition (FER), which looks at face expressions, are the two primary techniques for emotion recognition. Even though each approach is promising on its own, when combined, they can produce a more thorough understanding of human emotions.

Multilingual and dialectal variances, on the other hand, provide difficulties since they impact both the accuracy of SER due to variations in pronunciation and intonation and the efficacy of FER due to expressions that are culturally specific. This survey investigates the combination of SER and FER methods, emphasizing the effects of language variations and dialects on emotion recognition. We will examine the state of the art, point out problems, and offer recommendations for future work on flexible and language-aware emotion detection systems.

2.LITERATURE REVIEW

With a focus on both speech and facial emotion recognition (SER and FER), the paper investigates several approaches for emotion recognition across multiple domains. It addresses deep learning methods for speech-based emotion recognition, including Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs), stressing the significance of acoustic characteristics like pitch, energy, and prosody. The work shows increases in emotion detection accuracy by fusing feature selection with various deep learning techniques. It also emphasizes the use of feature-based methods for dialect identification in machine translation and speech recognition applications. The study also discusses dialect awareness and transformation techniques, as well as the ethical and societal ramifications of dialect management in natural language processing (NLP).

The study emphasizes the promise of Vision Transformers (ViTs), which outperform or match CNN-based models in facial emotion recognition. ViTs use self-attention mechanisms to gather both local and global information in facial images. The paper examines real- world issues, such as lighting and occlusion, in using FER technologies and compares feature selection techniques like Support Vector Machines (SVM) and Auto-Encoders (AE), concentrating on acoustic characteristics for emotion categorization. In the latter section of the paper, the use of EEG signals for real-time emotion recognition is explored. It is found that deep learning models perform better at managing complicated brain data than typical machine learning techniques. Upcoming paths encompass refining models for instantaneous applications, augmenting data, and incorporating deep learning methodologies to enhance human- computer interactions.

3.LITERATURE SURVEY

Table 1 Literature Survey

Sr. No.	Title	Author	Abstract	Gaps
1.	"Multilingual Speech Emotion Recognition With Multi-Gating Mechanism And Neural Architecture Search"	Zihan Wang, Qi Meng, HaiFeng Lan, XinRui Zhang, KeHao Guo, Akshat Gupta	The study presents the Speech Emotion Recognition (SER) technique, which combines a CNN + Bi-LSTM + self-attention architecture with language-specific and multi-domain models. It makes use of multi-gating with Neural Architecture Search (NAS) and contrastive auxiliary loss.	• Despite the improved performance of deep learning algorithms, involuntary facial movements still make it difficult to make delicate, fleeting facial expressions.
2.	"Facial Emotion Recognition Using Conventional Machine Learning And Deep Learning Methods: Current Achievements, Analysis And Remaining Challenges"	Amjad Rehman Khan	This study discusses the state-of- the-art in Facial Emotion Recognition (FER), with a focus on the contributions of deep learning techniques like CNNs and RNNs. It emphasizes how crucial it is to use IoT sensors and deep learning for better FER in security and healthcare.	• DL based FER methods require significant computational resources for both training and testing phases which may not be feasible in real time applications or on devices with limited processing power. • The limitations of 2-D data in addressing variations in facial poses and subtle behaviors, but the effectiveness of 3-D datasets for wide range of expressions still requires further exploration.
3.	"Deep Learning Techniques For Speech Emotion Recognition From Database To Models"	Babak Joze Abbaschian , Daniel Sierra-Sosa and Adel Elmaghraby	In the current paper, standard machine learning techniques for Speech Emotion Recognition (SER) are reviewed after an evaluation of deep learning approaches using accessible datasets. The research concludes with a multi-aspect evaluation of realistic neural network techniques in SER.	• It helps to comprehend the dialog response effectively when one is aware of one's feelings during the discussion. As of right now, this aspect of human-computer interaction remains unsolved, and there is no universal answer outside of a small number of applications. • Observing the variations in pragmatic approaches among speakers of distinct dialects offers a crucial sociological viewpoint on dialectal diversity. But NLP can not now adequately account for these.
4.	"Natural Language Processing For Dialects Of A Language: A Survey"	Aditya Joshi, Raj Dabre, Diptesh Kanojia, Zhuang Li, Haolan Zhan, Gholamreza Haffari, Doris Dippold	Language dialects are a key component of datasets that are examined in this survey. The performance degradation of NLP models for dialectic datasets and its implications for language technology equity sparked this work, which examines prior research on NLP in dialects in terms of datasets and methodology.	• With the use of target language translations of English queries, generative AI claims multilingual competence in several languages. Verifying multilingualism by translating sentences is erroneous. Therefore, it hasn't been put into practice yet. • For linguistic dialects spoken by historically marginalized communities, like the African-American community, natural language processing (NLP) may

				not work as effectively.
5.	"A Study On A Speech Emotion Recognition System With Effective Acoustic Features Using Deep Learning"	Sung-Woo Byun and Seok-Pil Lee	This study suggested a feature combination for a recurrent neural network model that can enhance emotion recognition performance. To determine the best mix of acoustic variables that influence speech emotion, statistical analysis was carried out.	• The usage of multiple languages in a single audio frame for classification of emotions is an unexplored area thus far. • The classification of emotions into a wider overlapping spectrum has not been taken into consideration.
6.	"VITFER: Facial Emotion Recognition With Vision Transforms"	Aayushi Chaudhari, Chintan Bhatt, Achyut Krishna and Pier Luigi Mazzeo	In this work, transformers and the ResNet-18 model were used to classify facial expression recognition (FER). This study compares the Vision Transformer model with state-of-the-art models on hybrid datasets and looks at how well it does in this task.	• The unavailability of optimal dataset poses an obstacle for a trained model. • The usage of multiple languages in a single audio frame for classification of emotions is an unexplored area thus far. • The classification of emotions into a wider overlapping spectrum has not been taken into consideration.
7.	"Speech Emotion Recognition With Deep Learning"	Hadhami Aouani and Yassine Ben Ayed	The main issues with machine learning are listed in this study. In addition to those, deep learning- based tools have begun to provide SER with an advantage. Prior to using deep learning extensively, SER depended on HMM and GMM procedures.	• It notes the limited range of features used, which may not fully capture emotional complexity, and highlights issues with model generalization across different datasets. • Key research gaps include the lack of multimodal approaches integrating visual and auditory signals, the need for real-time processing in interactive applications, neglect of cultural diversity in emotional expressions, and the absence of standardized evaluation metrics for study comparisons.
8.	"Facial Emotion Recognition: A Survey And Real-World User Experiences In Mixed Reality"	Dhwani Mehta, Mohammad Faridul Haque Siddiqui and Ahmad Y. Javaid	This study looks at how well virtual assistants and smart speakers comprehend their users, particularly when it comes to identifying phrases with questionable meanings. Translation services between languages can be provided by the same program.	• Limitations include difficulties in maintaining specific emotions during detection and the need for further improvements in real-time emotion recognition.
9.	"M1m2: Deep Learning Based Real-Time Emotion Recognition From Neural Activity"	Sumya Akter, Rumman Ahmed Prodhan, Tanmoy Sarkar Pias, David Eisenberg and Jorge Fresneda Fernandez	This essay categorizes feelings and shortcomings related to learning other languages and identifies prosody based on basic frequencies, speaking speed, duration, and intensity.	• Despite high accuracy, ethical concerns around emotional recognition and data privacy are not fully addressed. Additionally, the impact of different EEG acquisition methods and environmental factors on accuracy remains unexplored. • The suggested models' ability to be validated across several datasets or original people is limited by their dependence on a single dataset (DEAP). • The paper does not discuss the potential for real-time applications beyond the experimental context.
10.	"Speech Emotion	Chen Jie	To increase the accuracy of	• The paper acknowledges

	Recognition Based On Convolutional Neural Network”		emotion detection in speech, a model for speech emotion recognition that makes use of CNNs and altered MFCC characteristics is put forth.	inconsistencies in previous works, particularly regarding speaker- dependent versus speaker- independent results, suggesting that more rigorous validation methods are necessary.
11.	“Speech Signal-Based Modelling Of Basic Emotions To Analyse Compound Emotion: Anxiety”	Rathi Adarshi Rammohan, Jeevan Medikonda, Dan Issac Pothiyil	An autoencoder- and neural network-based speech-based emotion identification system is intended to recognize the basic emotions that make up anxiety, which is a compound emotion consisting of fear, anger, and sadness.	• Key gaps include the challenge of extracting universally distinguishable acoustic features across different languages, regions, and genders, which affects emotion recognition accuracy. • The study notes the limitations posed by a small dataset size, emphasizing the need for a larger, validated database of basic emotions and anxiety signals.
12.	“Research On Multi-Modal Mandarin Speech Emotion Recognition Based On SVM”	Chen Caihua	In order to investigate speech-based multimodal emotion identification, this work addresses signal preprocessing, feature extraction, fusion techniques, and SVM classification, ensuring that the emotional features extracted are legitimate.	• It does not address the potential limitations or challenges associated with the integration of various modalities, which could affect the robustness of the emotion recognition system. • The findings' generalizability may be limited by their reliance on a particular dataset, such the Berlin Speech Emotion Database.
13.	“Large Language Model-Based Emotional Speech Annotation Using Context And Acoustic Feature For Speech Emotion Recognition”	Jennifer Santoso, Kenkichi Ishizuka, Taiichi Hashimoto	In order to improve speech emotion recognition (SER) systems' performance and outperform human labeling, this study suggests adopting large language models (LLMs) for emotion annotation in speech.	• It focuses mainly on the linguistic context without fully addressing the limitations of acoustic feature extraction in noisy or spontaneous speech environments. • The reliance on LLMs also introduces concerns about computational cost and scalability, especially when processing large volumes of real-time data. • The paper does not explore how the approach performs across diverse languages or culturally varying emotional expressions.
14.	“Phonetic Anchor-Based Transfer To Facilitate Unsupervised Cross-Lingual Speech”	Shreya G. Upadhyay1, Luz Martinez-Lucas2, Bo- Hao Su1, Wei-Cheng Lin2, Woan-Shiuan Chien1 , Ya-Tse Wu1, William Katz3, Carlos Busso2, Chi- Chun Lee1	This work proposes a phonetic anchor-based approach for cross- lingual speech emotion recognition (SER) with a 58.64% unweighted average recall (UAR) to achieve better performance in Taiwanese Mandarin and American English.	• There is a lack of focus on the phonetic commonalities that could serve as anchors for effective domain adaptation. The paper also notes that existing models often fail to account for corpus-wise variability, leading to suboptimal performance in cross- lingual settings.
15.	“Feature Comparison For Speech Emotion Recognition On Hindi Language”	Surbhi Khurana, Amita Dev, Poonam Bansal	This study presents a phonetic anchor-based method for cross- lingual speech emotion recognition (SER) that enhances performance in American English and	• The paper does not explore the impact of dialectal variations or cultural nuances in emotional expression within the Hindi-speaking population, which could affect the generalization of the results.

			Taiwanese Mandarin, achieving a 58.64% unweighted average recall (UAR).	• It lacks exploration of advanced feature extraction techniques that could improve performance in noisy or spontaneous speech environments.
16.	“CNN-BIGRU Speech Emotion Recognition Based On Attention Mechanism”	Liyan Zhang, Yetong Wang, Jiaxin Du, Xinyu Wang	Using zero crossing rate, short-term energy, and MFCC features as inputs to a convolutional neural network with an attention mechanism and BiGRU, this study improves the accuracy of speech emotion recognition, surpassing CNN-BiLSTM and CNN-GRU models on the CASIA Chinese sentiment corpus.	• Its limitations lie in the relatively small and specific datasets used, which may not generalize well to real-world applications. • The model lacks exploration of how noise and overlapping emotions in spontaneous speech affect accuracy.
17.	“Audio And Text Sentiment Analysis Of Radio Broadcasts”	Naman Dhariwal, Sri Chander Akunuri, Shivama, And K.Sharmila Banu	In order to improve real-time sentiment analysis in the media sector and uncover trends in public opinion, this research suggests a bifurcate and mix computational strategy for sentiment analysis of audio data. It does this by using natural language processing (NLP) techniques and tools to extract sentiments from radio broadcasts.	• It does not address the challenges of multimodal fusion, particularly how discrepancies between text and audio signals (e.g., sarcasm or ambiguous tones) impact sentiment detection. • The dataset is limited to radio broadcasts, which may not fully capture diverse speaking styles or emotional complexities found in other media.
18.	“Creation And Analysis Of Emotional Speech Database For Multiple Emotions Recognition”	Ryota Sato, Ryohei Sasaki, Norisato Suga, Toshihiro Furukawa	In order to improve speech emotion recognition (SER) by addressing the simultaneous presentation of emotions in utterances, this study provides a fresh emotional speech database with 2,025 examples, including 1,525 with multiple emotions and various intensities.	• The database's coverage in terms of languages, cultures, and emotional depth is limited. The paper also does not extensively address how well the database performs in real-world, noisy environments, or how well it can generalize to spontaneous speech, which is often less controlled than the recordings used.
19.	“Breaking The Silence: Whisper-Driven Emotion Recognition In AI Mental Support Models”	Xinghua Qu, Zhu Sun, Shanshan Feng, Caishun Chen, Tian Tian	In contrast to conventional text- based approaches, this study offers a customized voice-based emotional support conversation system that uses large language models (LLMs) to evaluate vocal inputs for improved emotional insight.	• It focuses on whispers alone and does not explore how different vocal characteristics (e.g., volume, pitch variation) in non-whispered speech could impact the model’s effectiveness.
20.	“Leveraging Speech PTM, Text LLM, And Emotional TTS For Speech Emotion Recognition”	Ziyang Ma, Wen Wu, Zhisheng Zheng, Yiwei Guo, Qian Chen, Shiliang Zhang, Xie Chen	This work presents successful data augmentation strategies on the IEMOCAP dataset to increase speech emotion recognition (SER) performance. It does this by utilizing the data2vec model, GPT- 4 for text generation, and Azure TTS for voice synthesis.	• It does not address the potential limitations in integrating these different models, especially in terms of computational complexity and latency. • The paper does not delve into how these models handle low-resource languages or emotions that are less explicitly expressed in speech or text.

4.METHODOLOGY

The Speech Emotion Recognition (SER) project is being implemented using a comprehensive process that integrates user interface development, natural language processing (NLP), and deep learning. First, a wide range of tagged audio samples representing different emotional states, including happiness, sadness, rage, and surprise, are carefully gathered into a dataset. The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS) dataset is publically accessible repositories from which this dataset may be retrieved. Following collection, the data is carefully preprocessed to remove noise, normalize it, and extract key audio characteristics as pitch, energy, spectrograms, and Mel-frequency cepstral coefficients (MFCCs). These characteristics are essential since they encompass the acoustic qualities required to identify emotions in speech.

The research then focuses on deep learning model creation and deployment. The ability of some architectures, including Convolutional Neural Networks (CNNs), to efficiently capture both spatial and temporal patterns in audio data led to their selection. Well-known frameworks for model development include PyTorch, TensorFlow, and Keras. Techniques like transfer learning, which enable the inclusion of pre-trained models that have already learnt pertinent features from big datasets, are used to improve model performance. Moreover, the training set is purposefully expanded through the use of data augmentation techniques, which enhances the model's capacity to generalize to previously unobserved data.

NLP techniques are incorporated into the methodology to assess the speech's semantic content concurrently with audio processing. This entails recording spoken words into text, which is subsequently analyzed to extract linguistic elements that could offer further context about the speaker's emotional state. Transformer-based models, like DistilBERT or BERT, are used to capture the nuances of language, allowing for a deeper comprehension of emotions that go hand in hand with the auditory characteristics.

The methodology incorporates natural language processing (NLP) techniques to evaluate the semantic content of the speech simultaneously with audio processing. This involves transcribe spoken words into a text document that is then examined to extract linguistic components that may provide further context regarding the speaker's emotional state. The subtleties of language are captured by transformer-based models, such as DistilBERT or BERT, which provide a richer understanding of emotions that are closely related to the auditory features.

A user-friendly interface is created using frameworks like Flask, Django, or Streamlit to encourage user involvement. With tools for recording and viewing data, this interface makes it simple for users to enter speech or audio samples. Waveforms and spectrograms are examples of visualization techniques that aid in improving user comprehension of both the audio data and the model's predictions.

After development is finished, the model is put into use on scalable cloud computing platforms such as Google Cloud or AWS, which enables real-time apps to use it. The purpose of REST APIs is to facilitate smooth communication between the frontend and backend so that different apps, like chatbots or customer support systems, can make use of the SER features.

Lastly, the project places a strong emphasis on continuous improvement by adding user feedback, updating the model with fresh data on a regular basis, and keeping an eye on performance metrics to make sure it works well in practical situations. This iterative process makes the SER system more reliable and user-focused by improving the model's accuracy and adjusting it to changing linguistic and emotional responses.

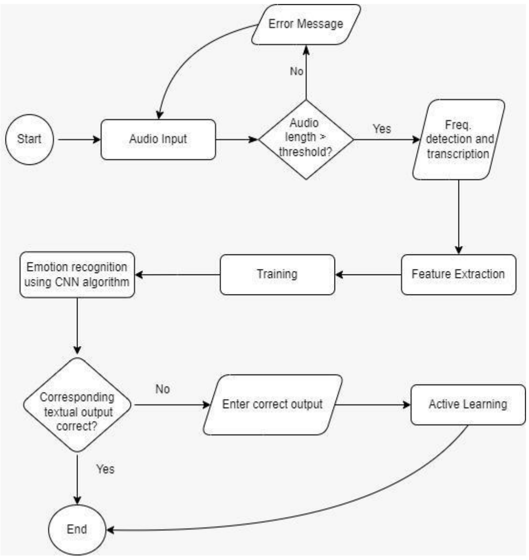
Error Message
No
Start
Audio Input
Audio length > threshold?
Yes
Freq. detection and transcription
Feature Extraction
Training
Emotion recognition using CNN algorithm
Corresponding textual output correct?
No
Enter correct output
Active Learning
Yes
End

Figure 1 System Architecture

5.FUTURE SCOPE

The majority of study papers and the overall survey done have problems with using many languages and dialects in the same audio frame. This presents a significant challenge for speech emotion recognition and offers room for development.

6.CONCLUSION

Though they confront difficulties due to linguistic and cultural variety, emotion recognition technologies such as Speech Emotion Recognition (SER) and Facial Emotion Recognition (FER) offer intriguing insights into human emotions. FER might misread expressions across cultural boundaries, while SER can identify emotions in speech but has trouble with language variances. By utilizing both verbal and visual clues, a multimodal method that combines both enhances accuracy. Subsequent investigations ought to concentrate on dialect-specific SER and culturally adapted FER models in order to improve inclusivity and practical performance.

REFERENCES

[1] Zihan Wang, Qi Meng, HaiFeng Lan, XinRui Zhang, KeHao Guo, Akshat Gupta, “Multilingual Speech Emotion Recognition With Multi-Gating Mechanism and Neural Architecture Search” , arXiv 2022

[2] Amjad Rehman Khan, “Facial Emotion Recognition Using Conventional Machine Learning and Deep Learning Methods: Current Achievements, Analysis and Remaining Challenges, Information” , MDPI 2022

[3] Babak Joze Abbaschian , Daniel Sierra-Sosa and Adel Elmaghraby, “Deep Learning Techniques For Speech Emotion Recognition From Database To Models” , Sensors, MDPI 2021

[4] Aditya Joshi, Raj Dabre, Diptesh Kanojia, Zhuang Li, Haolan Zhan, Gholamreza Haffari, Doris Dippold, “Natural Language Processing For Dialects Of A Language: A Survey” , arXiv 2024

[5] Sung-Woo Byun and Seok-Pil Lee, “A Study On A Speech Emotion Recognition System With Effective Acoustic Features Using Deep Learning” , Applied Sciences, MDPI 2021

[6] Aayushi Chaudhari, Chintan Bhatt, Achyut Krishna and Pier Luigi Mazzeo, “ViTFER: Facial Emotion Recognition With Vision Transforms” , Applied System Innovation, MDPI 2022

[7] Hadhami Aouani and Yassine Ben Ayed, “Speech Emotion Recognition With Deep Learning” , ScienceDirect 2020

[8] Dhwani Mehta, Mohammad Faridul Haque Siddiqui and Ahmad Y. Javaid, “Facial Emotion Recognition: A Survey And Real-world User Experiences In Mixed Reality” , Sensors, MDPI 2018

[9] Sumya Akter, Rumman Ahmed Prodhan, Tanmoy Sarkar Pias, David Eisenberg and Jorge Fresneda Fernandez, “M1M2: Deep Learning Based Real- Time Emotion Recognition From Neural Activity, Sensors” , MDPI 2022

[10] Chen Jie, “Speech Emotion Recognition Based On Convolutional Neural Network” ,IEEE 2021

[11] Rathi Adarshi Rammohan, Jeevan Medikonda, Dan Issac Pothiyil, “Speech Signal-Based Modelling of Basic Emotions to Analyse Compound Emotion: Anxiety”, IEEE 2020

[12] Chen Caihua ,“Research on Multi-modal Mandarin Speech Emotion Recognition Based on SVM”, IEEE 2019

[13] Jennifer Santoso, Kenkichi Ishizuka, Taiichi Hashimoto, “Large language model-based emotional speech annotation using context and acoustic feature for speech emotion recognition”, IEEE 2024

[14] Shreya G. Upadhyay, Luz Martinez-Lucas, Bo-Hao Su, Wei-Cheng Lin, Woan-Shiuan Chien , Ya-Tse Wu, William Katz, Carlos Busso, Chi-Chun Lee, “Phonetic anchor-based transfer to facilitate unsupervised cross-lingual speech”, IEEE 2023

[15] Surbhi Khurana, Amita Dev, Poonam Bansal, “Feature Comparison for Speech Emotion Recognition on Hindi Language”, IEEE 2023

[16] Liyan Zhang, Yetong Wang, Jiaxin Du, Xinyu Wang ,“CNN- BiGRU Speech Emotion Recognition Based on Attention Mechanism”, IEEE 2023

[17] Naman Dhariwal, Sri Chander Akunuri, Shivama, And K.Sharmila Banu, “Audio and Text Sentiment Analysis of Radio Broadcasts”, IEEE 2023

[18] Ryota Sato, Ryohei Sasaki, Norisato Suga, Toshihiro Furukawa, “Creation and Analysis of Emotional Speech Database for Multiple Emotions Recognition”, IEEE 2020

[19] Xinghua Qu, Zhu Sun, Shanshan Feng, Caishun

Chen, Tian Tian, “Breaking the Silence: Whisper-Driven Emotion Recognition in AI Mental Support Models” , IEEE 2024

[20] Ziyang Ma, Wen Wu, Zhisheng Zheng, Yiwei Guo, Qian Chen, Shiliang Zhang, Xie Chen, “Leveraging Speech PTM, Text LLM, and Emotional TTS For Speech Emotion Recognition”, IEEE 2024