iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 11 **Issue:** XI **Month of publication:** November 2023

DOI: https://doi.org/10.22214/ijraset.2023.56463

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 11 Issue XI Nov 2023- Available at www.ijraset.com

Survey of Audio - Facial Emotion Decoder System

Shweta Sondawale[1], Batul Chinikamwala[2], Srushti Dangde[3], Sakshi Salaskar[4], Arpita Shinde[5]
Computer Science Department, Sinhgad Academy of Engineering

Abstract: ***Rapid advances in cognitive technology have helped facilitate the seamless transition of human-computer interactions. This study offers a method to fuse voice and facial expressions, consider the addition of emotional information between voice and face, and overcome the limitations of unimodal thinking with a single thought. This survey paper provides a comprehensive overview of the advancements in SER systems, highlighting the evolution of this technology over the years and its applications across various domains. The core of this survey paper explores the methodologies and techniques employed in speech emotion recognition, including feature extraction, machine learning algorithms, deep learning architectures, and database resources.***
Keywords: human-computer interaction, speech emotion recognition, multimodal emotion recognition, feature extraction, multivariate analysis

I. INTRODUCTION

Emotions are a problem in daily life. By sharing their feelings, people can communicate better and understand each other. Machines can understand the human mind through emotional recognition, which has many important applications. For example, thinking in human-computer interaction (HCI) helps robots provide recommendations based on the user's emotions to increase the effectiveness of HCI. Emotions are a fundamental aspect of human communication and expression, playing a pivotal role in our daily lives. Recognizing and understanding emotions are not only essential for human-human interaction but also increasingly crucial in the context of human-computer interaction, artificial intelligence, and affective computing. Speech, as one of the most natural and prevalent means of communication, carries a wealth of emotional information. Consequently, Speech Emotion Recognition (SER) systems have emerged as a key technology for discerning and responding to emotional cues in spoken language.

This survey paper provides an extensive examination of the field of Speech Emotion Recognition, offering a comprehensive overview of its development, methodologies, applications, and challenges. As emotions influence our thoughts, behaviors, and decisions, the ability of machines to accurately recognize and respond to emotional cues in speech has transformative potential across diverse domains. In an age where technology is increasingly intertwined with our lives, SER offers an avenue for machines to comprehend and appropriately respond to human emotions.

II. EXISTING SYSTEMS

1) Emotional services based on multisensory perception using EEG-assisted emotional monitoring and speech recognition - Joseph Weizenbaum created ELIZA, a computer for language analysis, in 1964 and 1966. Without any built-in structure, ELIZA interacted and mimicked discussions with users to give them the impression that the software understood them. One of ELIZA's most well-known scripts, "DOCTOR," emulated the "Person-Centered Therapy" approach by asking open-ended questions in response to user input.

2) Attention Driven Fusion for Multi-Modal Emotion Recognition - When using integrated auditory and text modalities to identify emotions, deep learning has proven to be a potent substitute for hand-crafted techniques. Prior to applying attention, fusion, and classification, baseline systems simulate emotion information in text and audio modes separately utilizing Deep Convolutional Neural Networks (DCNN) and Recurrent Neural Networks (RNN). In this research, we describe a deep learning-based method for fusing and using textual and audio data to classify emotions. To extract acoustic features from raw audio, we first apply a SincNet layer based on parameterized sinc functions with band-pass filters, followed by a DCNN.

3) Multi-Modal Emotion Recognition by Fusing Correlation Features of Speech-Visual - This letter suggests a multi-modal emotion recognition approach by fusing correlation characteristics of speech-visual variables in order to successfully fuse speech and visual features. First, two-dimensional convolutional neural networks (2D-CNN) and three-dimensional convolutional neural networks (3D-CNN) are used to extract voice and visual data, respectively. Second, a feature correlation analysis technique used in multi-modal fusion processes the voice and visual features. Additionally, the feature correlation analysis technique applies the class information of speech and visual characteristics, which can successfully combine speech and visual features and enhance the performance of multi-modal emotion identification.

Support vector machines (SVM) bring multi-modal speech and visual emotion recognition to a close.

4) Efficient Speech Emotion Recognition Using Multi-Scale CNN And Attention - Speech emotion recognition is a difficult task. Bi-directional recurrent neural neptwork (Bi-RNN) and attention mechanisms have become the de facto method for speech emotion recognition thanks to recent advances in deep learning. These methods extract and attend multi-modal features—such as audio and text - before fusing them for subsequent emotion classification tasks. In this article, we suggest a straightforward yet effective neural network design to utilize speech's acoustic and lexical information. The suggested framework obtains hidden representations for both audio and text using multi-scale convolutional layers (MSCNN).

5) Multi-Modal Emotion Recognition from Speech and Facial Expression Based on Deep Learning - The rapid advancement of emotion identification helps to realize a highly harmonious experience of human-computer interaction. This research proposed a method that integrates speech and facial expression characteristics, taking into consideration the complementarity of the emotional information of speech and facial expressions and overcoming the single modal emotion detection constraint of single emotional features.

III. COMPARISON

TABLE I
COMPARISON OF EXISTING SYSTEMS

Sr No	Topic	Advantages	Disadvantages
1.	Emotional services based on multisensory perception using EEG	Enhanced Emotional Insights	Technological Limitations
2.	Attention Driven Fusion for Multi-Modal Emotion Recognition	Improved Accuracy	Bias and Fairness Concerns
3.	Multi-Modal Emotion Recognition by Fusing Correlation Features of Speech-Visual	Contextual Understanding	Computational Complexity
4.	Efficient Speech Emotion Recognition Using Multi-Scale CNN And Attention	Interpretable Features	Interpretable Features
5.	Multi-Modal Emotion Recognition from Speech and Facial Expression Based on Deep Learning	Improved User Experience	Limited to Audio and Visual Data

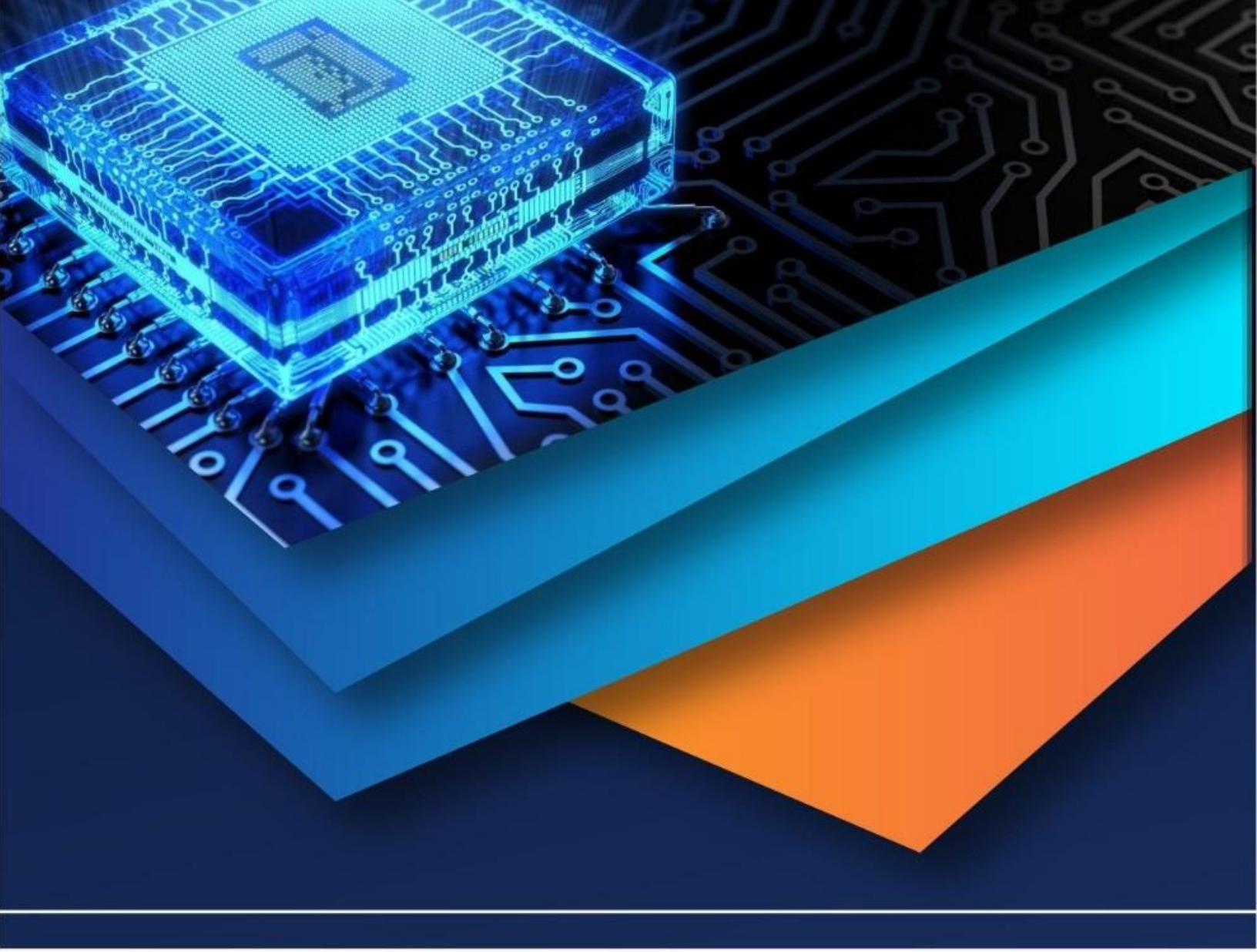
IV. CONCLUSION

This survey paper has provided a comprehensive overview of the field of speech emotion recognition, highlighting the significant progress and developments made in recent years. The growing interest and importance of speech emotion recognition in applications ranging from human-computer interaction to mental health support.

REFERENCES

[1] C. Busso, M. Bulut, C. Lee, et al. “IEMOCAP: interactive emotional dyadic motion capture database,” Language Resources & Evaluation., vol.42, no. 4, 2008.
[2] G. Theodoros, P. Gianni “PyAudioAnalysis: An open-source python library for audio signal analysis,” PLOSONE., vol, 10, no, 12, December 2015.
[3] J. Cai et al., “Feature-Level and Model-Level Audiovisual Fusion for Emotion Recognition in the Wild,” 2019 IEEE Conference on Multimedia Information Processing and Retrieval (MIPR), San Jose, CA, USA, 2019, pp. 443-448.
[4] S. Poria, I. Chaturvedi, E. Cambria and A. Hussain, “Convolutional MKL based multimodal emotion recognition and sentiment analysis,” 2016 IEEE 16th International Conference on Data Mining (ICDM), Barcelona, 2016, pp. 439-448.
[5] A. Zadeh, M. Chen, S. Poria, E. Cambria, L.-P. Morency, “Tensor fusion network for multimodal sentiment analysis”, 2017 Conference on Empirical Methods in Natural Language Processing (EMNLP), Copenhagen, Denmark, 2017 pp. 1114–1125
[6] N. Majumder, D. Hazarika, A. Gelbukh, E. Cambria, S. Poria, “Multimodal sentiment analysis using hierarchical fusion with context modeling,” Knowledge Based Systems., vol, 161, pp.124-133, December. 2018.
[7] Y. Kim and E. M. Provost, “Leveraging inter-rater agreement for audio-visual emotion recognition,” 2015 International Conference on Affective Computing and Intelligent Interaction (ACII), Xi'an, 2015, pp. 553-559
[8] W.-J. Yan, Q. Wu, Y.-J. Liu, S.-J. Wang, and X. Fu, “Casme database: a dataset of spontaneous micro-expressions collected from neutralized faces,” in 2013 10th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG), China, April 22-26, 2013, pp. 1–7.
[9] P. Ekman and W. V. Friesen, “Nonverbal leakage and clues to deception,” in Psychiatry. Taylor & Francis, 1969, vol. 32, no. 1, pp. 88–106.

[10] V. Rodellar, D. Palacios, E. Bartolom, and P. Gmez, "Vocal fold stiffness estimates for emotion description in speech," in International Conference on Bio-inspired Systems and Signal Processing, Spain, January 11-14, 2013, pp. 112–119.
[11] C. M. Hurley, "Do you see what I see? Learning to detect micro expressions of emotion," in Motivation and Emotion, 2012, vol. 36, no. 3, pp. 371–381.



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY