Speech Emotion Recognition: A review on current methodologies and challenges

Prof. Tanuja Zende [1]
Assistant Professor [1] ,

Dr. Ramachandra. V. Pujeri [2]
Pro-VC[2] ,

Dr. Suvarna. Pawar[3]
Professor[3]

MIT Art Design and technological University Pune,INDIA
tanuja.zende@mituniversity.edu.in [1] sriramu.vp@mituniversity.edu.in [2] suvarna.pawar@mituniversity.edu.in [3]

Abstract

Recently, we have witnessed a rapidly evolving field in Speech emotion recognition [SER] with significant inferences for human-computer interaction [HCI], affective computing and healthcare. In this review paper, we aim to provide a complete investigation of current methodologies and challenges in SER. Our investigation revolves around the progress of feature extraction techniques, classification algorithm and datasets used in the arena. Furthermore, we have also kept our attention closely on recent trends in deep learning-based approaches and the application of multimodal data for enhanced emotion recognition. Key challenges such as dataset bias, domain adaptation and model interpretability that researchers encounter in SER are also major part of our interest. By synthesizing current progressions and identifying impending research directions, we aim to contribute to the ongoing progress of accurate and vigorous SER systems.

KEYWORDS

Speech Emotion Recognition, Deep Learning, Feature Extraction, Datasets, Challenges, Machine Learning.

INTRODUCTION

With applications accommodating HCI, sentiment analysis and mental health assessment off late, SER has played a vital role. A better understanding and response to user needs by deciphering the emotions transported through speech is essential for developing empathetic systems. The accuracy and robustness of emotion recognition systems across varied linguistic and cultural contexts was possible due to various methodological evolution in SER. In our work, the intent is to provide a comprehensive outline of current methodologies and challenges in SER, thereby exploring advancements in classification algorithms, datasets and feature extraction techniques. To improve emotion recognition accuracy and depth, we also delve into recent inclinations in deep learning approaches and integration of multimodal data sources.

This review seeks to contribute to the ongoing development of emotion-aware systems that can revolutionize HCI and enable more nuanced emotion understanding in various domains by synthesizing the dominant research findings thereby highlighting key advancements. Bridging the gap between emotion recognition research and practical applications is our main goal, hence, nurturing the creation of intellectual systems skilled with distinguishing and responding to human emotions successfully.

MOTIVATION

HCI, sentiment analysis, and mental health assessment by enabling systems to adapt to user's emotional states is all possible now with the help of SER. Improvement in deep learning and multimodal data integration bid openings for systems with more accurate and culturally diverse emotion recognition. To put forth the current challenges faced by SER, current emotion recognition and propose future research directions, creating intelligent systems that can understand and respond to human emotions effectively in various applications is our major view in this work.

OBJECTIVES

We will be Able to

1. Provide a wide-ranging overview of existing methodologies and challenges in SER.
2. Synthesize recent advancements in feature extraction techniques, classification algorithms, and multimodal data integration.
3. Contribute to the growth of accurate and vigorous emotion recognition systems.
4. Identify research gaps, such as dataset bias and model interpretability.
5. Propose future research directions to advance the field of speech emotion recognition.
6. Bridge the gap between emotion recognition research and practical applications.
7. Foster the creation of intelligent systems capable of perceiving and retorting to human emotions effectually.

SCOPE

Explore advancements in feature extraction techniques and classification algorithms in speech emotion recognition. Discuss the integration of multimodal data sources for more nuanced emotion recognition. Analyze recent trends in deep learning-based approaches for emotion recognition. Address challenges such as dataset bias and model interpretability in emotion recognition systems. Identify potential future research directions to enhance the correctness and robustness of emotion recognition. Focus on practical applications of emotion recognition in HCI and sentiment analysis. Aim to contribute to the development of emotion-aware arrangements that can effectively comprehend and reply to human emotions in various contexts.

DATASET

RAVDESS (Ryerson Audio-Visual Database of Emotional Speech and Song): Contains speech recordings of artists portraying various emotions, including neutral, calm, happy, sad, angry, fearful, disgust, and surprised.

IEMOCAP (Interactive Emotional Dyadic Motion Capture): Consists of recordings of actors engaged in scripted and improvised dialogs, expressing a range of emotions such as anger, happiness, sadness and neutral.

SAVEE (Surrey Audio-Visual Expressed Emotion): Features speech recordings of male speakers expressing emotional states such as - happiness, sadness, anger, and disgust.

EmoDB (Emotional Database): Includes German speech recordings of actors uttering specific sentences with emotional categories like - anger, boredom, disgust, fear, happiness, and sadness.

MSP-IMPROV (Multimodal Spontaneous Personality-IMPROV): Contains spontaneous speech data with emotional annotations, covering emotions like joy, sadness, anger, and neutral expressions.

CREMA-D (Crowd-Sourced Emotional Multimodal Actors Dataset): Comprises audiovisual recordings of actors performing emotional scenarios, covering various emotions like: anger, disgust, fear, happiness, sadness and surprise.

LITERATURE SURVEY

Recent advancements in speech emotion recognition have been highlighted in several cutting-edge research papers. Zhang et al. (2021) introduced Transformer networks for end-to-end emotion recognition in speech, achieving state-of-the-art results. Wang et al. (2022) explored adversarial learning to enhance the robustness of emotion recognition systems. Li et al. (2021) proposed the use of graph convolutional networks to capture dependencies in speech features. Chen et al. (2022) investigated self-supervised learning for emotion representation learning. Zhou et al. (2021) delved into federated learning for privacy-preserving emotion recognition. Gupta et al. (2022) studied multimodal fusion techniques for cross-cultural emotion recognition. These papers showcase the latest advancements in speech emotion recognition, addressing challenges such as robustness, privacy, cross-cultural variations, and multimodal integration for improved emotion classification accuracy

Recent IEEE papers in SER have delved into various innovative approaches. Liu et al. (2022) conducted a survey on attention mechanisms role in improving accuracy of emotion recognition. Wang et al. (2022) inspected the usefulness of transfer learning for acclimating emotion recognition models to new datasets and domains. Zhang et al. (2022) aimed to work on noisy environments using deep learning and developing robust emotion recognition models. Chen et al. (2022) examined emotion-aware speech synthesis using generative adversarial networks. Li et al. (2022) premeditated the influence of data augmentation techniques in refining the oversimplification as well as robustness of emotion recognition models. The mentioned IEEE papers all contribute to evolving the field of SER by addressing challenges like emotion-aware synthesis, attention mechanisms, transfer learning, robust recognition in noisy environments and data augmentation for refining the accuracy and performance.

METHODOLOGY

1. Feature Extraction Techniques:

- Discuss traditional features like MFCCs, prosody, and spectral features.
- Explore advanced feature extraction methods such as deep learning-based features.

2. Classification Algorithms:

- Review popular machine learning algorithms used for SER.
- Discuss the effectiveness of deep learning models namely CNNs, RNNs, and Transformers

CURRENT TRENDS

Several key areas are of major interest that attract the attention of SER. One of the prominent trends is the amalgamation of multimodal information by merging audio and visual cues for more accurate classification of emotions. Next, improving emotion recognition performance can be achieved by involving the investigation of deep learning technique that includes the use of graph convolutional networks, transformer and generative adversarial networks. Also, it is observed that many researchers are investigating self-supervised and transfer learning methods thereby leveraging unlabeled data hence developing models capable of new emotion recognition. Data security in emotion recognition systems have gained traction by using Privacy-preserving approaches, such as federated learning. However, the effects of noisy environment to develop a robust emotion recognition model that also addresses cross-cultural variations are now the emerging trends. A diverse and dynamic landscape in SER research incorporate these current trends.

CHALLENGES

SER is also impacted by many challenges, prompting the direction of current research efforts. Labeled data may be viewed as one of the major challenges, mainly for underrepresented emotions or specific cultural contexts, obstructing the development of robust and generalizable emotion recognition models. Next, as individuals, the variability and subjectivity of emotional expression, may vary from person to person, culture to culture etc. thereby exhibiting emotions differently. This may lead to ambiguity in emotion classification. Further complications are encountered in emotion recognition tasks due to the issue of cross-cultural variations necessitating in developing models that are sensitive to diverse emotional expressions globally. Also, the robustness of emotion recognition systems in noisy environments poses a significant concern, as the acoustic interference or background noise can impact the accuracy of classifying the emotions. Another pressing concern is observed when it comes to privacy and data security in emotion recognition models, since the use of personal data for emotional analysis is done. Evolving the field of SER and developing more reliable and effective emotion recognition technologies needs the accountability of these challenges incurred.

FUTURE SCOPE

SER holds promising opportunities towards innovation and advancement. Integration of explainable artificial intelligence (AI) techniques to improve the interpretability of emotion recognition models maybe regarded as a key area, hence, making it easy for the users to understand how decisions are made. Another exciting avenue for future exploration maybe considered as development of context-aware emotion recognition systems which have the ability to adapt to the user preferences and situational context. In real-world applications, namely, mental health monitoring, HCI and personalized systems, exploration of affective computing offers substantial potential for improving user experience. Collective research across numerous disciplines, such as psychology, linguistics and computer science, may help to further enrich emotion recognition models by integrating a deeper understanding of human emotions and behaviors. Embracing ethical considerations, such as ensuring impartiality, transparency and accountability in emotion recognition technologies, will be regarded as crucial for nurturing responsible innovation. Therefore, in conclusion, the future of SER is self-assured to revolutionize emotional analysis, HCI and personalized services, thereby offering range of aids and opportunities for societal bearing.

REFERENCES

[1] Liu, S., et al. (2022). "Exploring Attention Mechanisms for Enhancing Speech Emotion Recognition Accuracy." IEEE Transactions on Affective Computing

[2] Wang, Y., et al. (2022). "Transfer Learning for Emotion Recognition in Speech Signals." IEEE Transactions on Audio, Speech, and Language Processing.

[3] Zhang, H., et al. (2022). "Robust Emotion Recognition in Noisy Environments Using Deep Learning." IEEE Signal Processing Letters.

[4] Chen, Q., et al. (2022). "Emotion-aware Speech Synthesis using Generative Adversarial Networks." IEEE Transactions on Multimedia.

[5] Li, W., et al. (2022). "Enhancing Speech Emotion Recognition with Data Augmentation Techniques." IEEE International Conference on Acoustics, Speech, and Signal Processing.