WWW.IJRPR.COM

International Journal of Research Publication and Reviews

Journal homepage: www.ijrpr.com ISSN 2582-7421

SPEECH EMOTION RECOGNITION USING DEEP LEARNING

Ayisha Sidhiqa.S[1] , Mrs.B.Zunaitha[2]

[1] UG Student, Department of Computer Science, Sri Krishna Aditya College of Arts and Science, Coimbatore.

[2] Assistant Professor, Department of Computer Science, Sri Krishna Aditya College of Arts and Science, Coimbatore.

ABSTRACT :

Speech emotion recognition (SER) as a Machine Learning problem continues to garner a significant amount of research interest, especially in the affective computing domain. This is due to its increasing potential, algorithmic advancements and applications in real world scenarios. Human speech contains paralinguistic information that can be represented using quantitative features such as pitch, intensity and Mel-Frequency Cepstral Coefficients (MFCC).It is commonly achieved following three steps: data processing, feature selection/ extraction and classification based on the underlying emotional features. The nature of these steps coupled with the distinct features of human speech , underpin the use of ML method for implementation. In this paper we present a systematic review of research that addressed SER task from ML perspective over the last decade, with emphasis on the three SER implementation steps. Different challenges including the tissue of low classification accuracy of speaker-independent experiments and solutions associated with them are discussed in detail.

Keywords— Speech emotion recognition; Machine learning; Speaker-independent experiment; Classification; Audio emotion recognition.

Introduction:

Speech is a natural way of communication among human beings. It provides information about thoughts, feelings, moods, and the context of the speaker's communication. The subtleties of emphasis, tone, phrasing, variations in utterance speed and continuity, and the accompanying physical gestures convey something of the inner life of impulse and feeling. "Speech signal contains a mixture of information, including cues to speaker identity, affect, and lexical and grammatical emphasis for the spoken message. Isolating affective information is complicated". It established body of research focuses on the correlation between speech and emotions.

Due to the need for machines with affective capacities to utilize their talents responsibly, affective computing is an area of research that requires careful and thorough examination. The literature contains various examples of affective computing systems based on image , audio, video , text, physiological signals , and multi-model emotion recognition all of which are powered by machine learning or deep learning models. Speech Emotion Recognition (SER) has emerged as a popular area of study these days, with various applications across various domains.

It can be employed in various areas, such as lie detection and criminal investigations , medical diagnosis and monitoring ,robotic emotion expressions, machine-human interaction systems , call center answering, robotic assistance and helpline systems, theatre performance and interaction enhancements , mental health and fitness analysis in the classroom and online teaching, emotional state recognition of drivers and intelligence assistance including the digital advertisement, online gaming, and customer feedback evaluation.

Noise has a significant impact on speech-emotion recognition systems' accuracy. However, more research is necessary to evaluate them in noisy settings. A survey paper is needed to investigate all aspects of noisy speech emotion recognition to facilitate real-time monitoring development. This paper also considers the obstacles and strategies for noisy speech emotion recognition, the current state-of-the-art models and techniques, a comprehensive review of existing systems, and recommends potential areas for improvement.

Literature Survey:

The research conducted in the field of speech motion recognition is covered in this section. Various authors have employed deep learning and machine learning techniques such as Convolution neural networks (CNN), recurrent neural networks (RNN), support vector machines (SVM), long short-term memory (LSTM), and bidirectional LSTM are used to predict human emotions. The authors of the research [1] used a deep convolution recurrent network with LSTM to analyze the dataset and automatically determine the optimal speech signal representation. The dataset is used on both motion capture markers and audio data from five pairs of components are included in the Iemocap dataset, which was employed. It includes the two DCNNLSTM network algorithms. The deep neural network has been provided for the dumb video, which has no audio. The initial step in analyzing emotion recognition is to identify the 42-dimensional features. Second, a 74.62% accuracy rate was attained utilizing the SVM classification approach. These articles' primary flaw is that they use an audio signal based dataset with additional characteristics. This paper focuses mostly on the features.

Several important insights for enhancing speech emotion recognition (SER) are provided in this paper:

1. **Brain-Inspired Models**: Leveraging neural architectures modeled on human emotion perception.

2. **Multi-Task Learning:** Enhancing accuracy by incorporating auxiliary tasks (e.g., identifying emotional attributes like arousal or valence).
3. **Feature Integration:** Combining implicit and explicit emotional cues for better recognition of complex emotions.
4. **Dataset Utilization:** Insights on using datasets like IEMOCAP to benchmark models.

Existing System:

Machine learning models such as Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs) are the mainstay of current SER systems. Pitch, energy, and other prosodic elements, spectrograms, and Mel-frequency cepstral coefficients (MFCCs) are examples of audio features that these models are good at identifying patterns from. However, their primary purpose is not emotion awareness; rather, it is jobs like computer vision and natural language processing. Although they are employed, methods like hidden Markov models (HMMs) and support vector machines (SVMs) have had little success in addressing the intricate structure of speech emotional cues.

Disadvantages:

- Lack of Emotion Perception Understanding: These models are unable to incorporate human-like emotional perception, including implicit emotional cues like valence or arousal, which results in incorrect recognition in emotionally complicated situations.
- Over-fitting on Small Datasets: Due of the imbalance and limited size of many SER datasets, current models over-fit and perform poorly in practical settings.
- Limited Generalization: Due to differences in speech characteristics, models that were trained on particular datasets frequently have trouble generalizing to new settings or speakers.
- Failure in Multimodal Integration: Emotions are expressed not only through words but also through body language, context, and facial expressions. Current systems are unable to successfully incorporate multimodal information.
- High Computational Cost: Deep learning models, especially RNNs and CNNs, require significant computational resources for training and inference, making them less practical for deployment on low power devices.

Merits of Traditional Method:

- **Interpretability and Simplicity:** SVMs and HMMs are reasonably easy to construct and interpret because to their straightforward design, which makes it evident how judgments are made depending on input features.
- **Efficient Use of Limited Data**: Due to the limited availability of big, labeled emotional speech corpora, these techniques can function well with smaller datasets.

Proposed System:

The study methodology known as Speech Emotion Recognition (SER) uses a methodical approach to recognizing emotions in speech. Data collection is the first step in the process, where speech datasets with a range of speakers, emotions, and languages are gathered. The next step is preprocessing, which improves the quality of the data by segmenting, normalizing speech, and reducing noise. Pitch, energy, spectrograms, Mel Frequency Cepstral Coefficients (MFCCs), and other important characteristics that aid in differentiating between emotions are then retrieved. Models for categorization use these features as inputs. To evaluate these characteristics and precisely categorize emotions, machine learning methods like Support Vector Machines (SVM) or sophisticated deep learning models like Convolutional Neural Networks (CNNs) and Long Short-Term Memory (LSTM) networks are used. Last but not least, performance parameters such as accuracy, precision and to recall. This methodology ensures a reliable and efficient SER system, contributing to advancements in human-computer interaction, healthcare, and communication technologies.

In order to increase emotion extraction from speech data, the study focuses on improving speech emotion recognition (SER) by integrating human-like emotion perception methods. It is anticipated that the suggested system's brain-inspired multi-task learning model will outperform conventional techniques in identifying subtle emotional indicators in speech. The IEMOCAP dataset, a well-known standard in emotion recognition research, is used in this study. The system's ability to increase emotion recognition is demonstrated by the results, which show improvements in un-weighted and weighted accuracy of 2.44% and 3.18%, respectively.

Tools and Framework:

The tools and frameworks used in the experiment likely include popular deep learning libraries like TensorFlow or PyTorch, though the paper doesn't specify the exact tools. It also uses methods like convolutional neural networks (CNNs) and long short-term memory networks (LSTMs), which are common for processing sequential audio data.

Merits:

By resolving some of the major issues in speech emotion recognition, especially with regard to emotion classification accuracy, the system outperforms earlier models in terms of accuracy. The multi-task learning strategy employed in this study enhances the model's generalisation across various emotional settings and enables it to collect a wider variety of emotional signals.

Demerits:

Its reliance on the IEMOCAP dataset, which could not accurately reflect all real-world circumstances, particularly with regard to language or cultural diversity, is one drawback. The computationalcomplexity of deep learning and multi-task learning approaches, which may necessitate substantial resources for deployment and training, is another possible disadvantage. Furthermore, even though the article presents encouraging results, additional validation across a wider range of datasets and real-world applications will be required to prove the system's robustness.

A.Data preprocessing:

In order to guarantee high-quality input for the model, data preprocessing is an essential step in Speech Emotion Recognition (SER). Audio data must be cleaned and prepared for analysis. To improve clarity, background noise is filtered out in the first step, known as noise reduction. Speech normalization is then used to minimize discrepancies between recordings and normalize loudness levels. By breaking up audio files into smaller frames, segmentation makes it easier to do in-depth analysis. Non-informative portions of the audio are removed using techniques like silence removal. Lastly, the data is frequently transformed into feature vectors or spectrograms for additional processing. Preprocessing guarantees that the data is consistent, clean, and prepared for modeling and feature extraction.

B.Dataset:

The Ryerson Audio visual Data base of Emotional Speech and Song (RAVDESS) is a type of SER. It contains a total of 24 professional actors, with 12 male and12 female actors recording voices. It includes 7356 files in total. Happiness, calmness, sadness, anger, surprise, dismay, and terror are all expressed in speech using the phrases "kids are talking by the door" and "dogs are sitting by the door."All of the emotions mentioned above express these two statements. The database includes only video, only audio, and full audio/video. Since we're concentrating on speech recognition, this project solely uses audio.It has a normal and strong emotional intensity. This is seen in Table 1 below.

Dataset	Actors	Instances	Emotions
RAVDESS	24	7356	8

C. Emotion Classification:

The last stage of SER is emotion classification, which involves choosing the ML algorithms and designing the SER model architecture. The model architecture design suggests the kinds of audio features needed and how to use them, bridging the gap between the feature extraction process and the choice of ML algorithms for SER. The choice of algorithms is how the architecture design is put into practice. The kinds of ML algorithms for SER models that have been employed in previous research were examined in this study. This aids in developing recommendations for model architectural design, ML algorithm selection, and a direction for the SER model's design. Table 2 lists the four implicit properties of A–D along with the appropriate classifiers.

Table 2:The Attribute Classification

Attribute	**Parts**	**Label 1**	**Label 0**
A	Frontal Cortex	Happy, angry	Sad,neutral
B	Thalamus	Sad	Happy, angry, neutral
C	Hippocampus	Sad, angry	Happy, neutral
D	Anterior	Happy	Anger, Sad, neutral

Result and discussion :

Improved performance is demonstrated by the results of the suggested speech emotion recognition at a set, the brain-inspired multi-task learning framework produced a 3.18% improvement in WA and a 2.44% improvement in UA when compared to conventional models. These improvements demonstrate how well the algorithm can catch more complex emotional traits. While acknowledging difficulties including computational complexity and the requirement for testing on a variety of datasets to guarantee robustness, the discussion (SER) system, especially in un-weighted accuracy (UA) and weighted accuracy (WA). On the IEMOCAP on highlights the framework's promise for real-world applications.

Brain research is constantly examining the structure of the brain and the fundamental principles of emotions. In order to create a classification of implicit emotional attributes that resemble the brain structure linked to emotions, this work combines the mechanism of the human brain's emotional perception with artificial intelligence's continuous modeling of the human brain. Implicit emotion information is introduced as auxiliary information to perceive emotion through multi-task learning, improving speech emotion detection and proving the effectiveness of the network proposed in this paper. The dataset classifies emotions into nine categories: anger, excitement, happiness, grief, frustration, fear, neutral, surprise, and other. The four main emotions we employ in our research are anger, excitement, happiness, and sorrow.

ACKNOWLEDGEMENT:

I would like to express my sincere gratitude to Sri Krishna Adithya College of Arts and Science and the Department of Computer Science for providing me with the opportunity to undertake and complete this research paper. Their unwavering support, guidance, and encouragement throughout this project have been invaluable.

REFERENCES:

1. J. LeDoux, Rethinking the emotional brain. Neuron 73(4), 653–676 (2012)

2. Huahu, X., Jue, G., & Jian, Y. (2010). Application of speech emotion recognition in intelligent household robot. 2010 International Conference on Artificial Intelligence and Computational Intelligence, 537–541.

3. Liu, G., Cai, S., & Wang, C. (2023). Speech emotion recognition based on emotion perception. EURASIP Journal on Audio, Speech, and Music Processing, 2023(1).

4. Low, L.-S. A., Maddage, M. C., Lech, M., Sheeber, L. B., & Allen, N. B. (2011). Detection of clinical depression in adolescents' speech during family interactions. IEEE Transactions on Bio-Medical Engineering, 58(3), 574–586. Speech emotion recognition based on emotion perception. (n.d.). Gov.ua. Retrieved January 8, 2025