IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

SPEECH EMOTION RECOGNITION

[1]Kumari Deepika, [2]Ashutosh Kumar Singh, [3]Kumar Satyarth [4]Shivaksh Vyas, [5]Umesh M

[1,2,3,4]Student, [5]Assistant Professor

Information Science & Engineering,

RNS Institute of Technology, Bengaluru, India

Abstract: Attention has been drawn to Speech Emotion Detection (SED) due to its crucial role in understanding human emotional states through speech signals. This extensive survey paper explores the historical progression, current methodologies, and future prospects of SED. It explores early approaches, emphasizes the emergence of advanced machine learning techniques, and examines the challenges in accurately discerning emotions from speech. The survey provides an in-depth analysis of notable datasets, evaluation metrics, and cutting-edge models used in SED. Additionally, it outlines potential improvements and upcoming trends that are poised to shape the future landscape of this field. This survey consolidates a vast amount of information, offering insights into the past, present, and promising directions for advancing SED.

I. INTRODUCTION

In the realm of human communication, emotions hold a crucial role, acting as the cornerstone of our interactions and expressions. Understanding and interpreting emotions conveyed through speech is fundamental for our social engagements, enabling us to grasp subtleties, intentions, and feelings embedded within spoken language. Speech Emotion Recognition (SER) stands at the intersection of various fields like signal processing, machine learning, psychology, and linguistics. Its aim is to decipher and comprehend the emotional content inherent in spoken communication.

The importance of SER lies in its potential to strengthen human-machine interaction and amplify the effectiveness of various applications. By discerning emotional cues in speech, systems can better cater to users' emotional states, thus refining user experiences across domains such as human-computer interaction, healthcare, entertainment, customer service, and more. For instance, in mental health applications, SER can assist in identifying and monitoring emotional distress through speech patterns, contributing to early intervention and personalized treatment strategies. Similarly, in customer service, recognizing customer emotions can elevate service quality by customizing responses to their emotional needs.

Recognizing emotions from speech involves intricate analyses of acoustic features, linguistic cues, prosodic elements, and contextual information. Researchers and practitioners in this field utilize various techniques, spanning from traditional signal processing algorithms to advanced machine learning and deep learning models. These approaches aim to extract, analyze, and interpret features from speech signals, enabling the classification or regression of emotions into discrete categories or continuous dimensions.

Despite significant advancements, SER faces multifaceted challenges. Variability in emotional expressions across cultures, languages, and individuals poses a substantial hurdle. Moreover, the scarcity of labelled emotional speech data and the inherent ambiguity in emotional cues in speech further complicate accurate detection.

This survey paper aims to comprehensively explore the landscape of SER by investigating its historical evolution, methodologies, techniques, evaluation metrics, applications, challenges, recent advancements, and future prospects. By ana- lyzing existing literature and methodologies, it aims to present the current state-of-the-art, address challenges, and suggest avenues for Speech Emotion Recognition's future research and development.

II. Background and Literature Survey

2.1 Historical Development of Speech Emotion Recognition

The historical development of speech emotion recognition spans numerous pivotal moments that have profoundly influenced the evolution of this field. Initially, early efforts concentrated on fundamental acoustic features like pitch, intensity, and spectral characteristics to identify emotional cues within speech. The 1970s and 1980s witnessed the emergence of rule- based systems that utilized predefined algorithms to categorize emotions based on acoustic properties.

In the 1990s, machine learning emerged as a transformative force, enabling the utilization of classifiers such as neural networks and support vector machines to enhance accuracy. The 2000s signalled a transition towards multimodal emotion recognition, combining speech with facial expressions and gestures for a more comprehensive grasp of emotions.

The advent of deep learning in the 2010s brought significant advancements to speech emotion recognition, harnessing neural networks to automatically extract intricate features and achieve unparalleled levels of accuracy. Presently, the field is progressing with the integration of context-aware models that take into account linguistic content, speaker characteristics, and situational context to further improve the accuracy of emotion detection in speech.

2.2 Relevant Theories, Methodologies, and Techniques

Comprehending the advancements in the field of speech emotion recognition requires a grasp of the foundational theories, methodologies, and techniques involved. Psychological theories such as dimensional models (e.g., Valence-Arousal-Dominance), appraisal theories, and discrete emotion theories form the theoretical basis for identifying and categorizing emotions in speech.

Methodologies encompass a broad spectrum of approaches, including various feature extraction techniques like Mel- Frequency Cepstral Coefficients (MFCCs), prosodic features, and spectral features. These techniques play a crucial role in capturing acoustic cues that indicate different emotions present in speech.

Machine learning algorithms such as Support Vector Machines (SVM), Hidden Markov Models (HMM), and deep learning architectures like Convolutional Neural Networks (CNN) and Recurrent Neural Networks (RNN) are pivotal in both modelling and classifying emotional features extracted from speech signals.

Moreover, the amalgamation of multiple modalities (speech, facial expressions, gestures) and context-aware strategies significantly contributes to refining the accuracy of speech emotion detection systems by capturing a more holistic view of emotional cues. Integrating and comprehending these diverse theories, methodologies, and techniques play a pivotal role in the continual improvement of speech-emotion detection systems.

2.3 Key Research Papers, Methods, and Models

Numerous influential research works have significantly impacted the realm of speech emotion recognition. Notably, the studies conducted by Scherer et al. (2003) exploring the acoustic correlates of emotions, along with the groundbreaking work by Deng et al. (2013) introducing deep learning for feature extraction, have wielded substantial influence. Methodologies such as the utilization of Hidden Markov Models (HMMs) for sequence modelling and models like Long Short-Term Memory (LSTM) networks for capturing temporal dependencies have demonstrated remarkable effectiveness in tasks related to emotion detection.

Moreover, the COMPASS dataset curated by Lee et al. (2019) has emerged as a pivotal benchmark for assessing speech emotion recognition systems. This dataset has played a crucial role in establishing standardized evaluation metrics and fostering comparative analyses among different models.

The evolution of speech emotion recognition signifies a progression from basic acoustic analyses to the adoption of sophisticated machine learning and deep learning techniques. The integration of diverse theories and methodologies from psychology, linguistics, and signal processing has been instrumental in propelling the field forward. Seminal research works, innovative methodologies, and benchmark datasets have significantly contributed to advancing both the comprehension and application of speech emotion recognition across various domains.

III. Speech Emotion Recognition Techniques

3.1 Method of Extracting Features

• Acoustic Features

– Mel-Frequency Cepstral Coefficients (MFCCs): Represent the short-term power spectrum of a sound, particularly effective in capturing spectral characteristics. This method maps the short-term power spectrum onto the Mel scale, emphasizing features that are perceptually relevant.

– Prosodic Features: Encompass various elements including pitch, intensity, duration, and formant frequencies. These elements capture the rhythm, stress, and intonation of speech, contributing significantly to its emotional content.

– Spectral Centroid, Bandwidth, and Energy: Numerical descriptors characterizing the spectral properties of a signal. Centroid represents the "center of mass" of the spectrum, bandwidth indicates the frequency range, and energy measures the signal power. These features provide insights into the frequency distribution and overall signal strength.

• Linguistic Features

– Word-based Analysis: Involves analyzing the usage of emotion-specific vocabulary, identifying words or phrases associated with particular emotions. This analysis correlates specific words with emotional states.

– Syntax and Semantics: Focuses on understanding how sentence structure and semantic meaning contribute to emotional context. Analyzing word arrangement and meaning aids in recognizing emotional expressions.

– Pragmatic Elements: Considers contextual information that influences emotional expression, including non-verbal cues, cultural context, and situational factors. This aspect involves understanding the pragmatic factors that influence emotional speech.

3.2 Techniques in Machine Learning

• Statistical Models

– Gaussian Mixture Models (GMMs): Used to model probability distributions of features. These models represent each emotion category as a mix of Gaussian distributions, aiding in probabilistic classification.

– Support Vector Machines (SVMs): Effective in high-dimensional feature spaces, SVMs aim to find the best hyperplane to separate different emotion classes. They work well for both linear and non- linear classification tasks.

– Decision Trees and Random Forests: Decision trees segment the feature space based on feature values. Random forests combine multiple decision trees to enhance accuracy and generalize better by capturing feature interactions.

• Deep Learning Models

– Convolutional Neural Networks (CNNs): Specialized in extracting hierarchical features from spectrograms via convolutional filters. CNNs autonomously learn pertinent features from raw input data, aiding in emotion detection.

– Recurrent Neural Networks (RNNs): Apt for capturing temporal dependencies in sequences, making them efficient for analyzing sequential speech data. RNNs consider previous information when processing each input.

– **Long Short-Term Memory Networks (LSTMs):** Specifically designed to manage long-range dependencies in sequences using a memory cell structure. LSTMs excel in capturing context and temporal nuances in speech.
– **Transformer-based Architectures:** Utilize attention mechanisms to comprehend context across input sequences. This enables the model to focus on relevant portions of the input data for emotion detection.

3.3 Comparative Analysis

• Strengths:

– Machine Learning:
∗ **Interpretability:** Machine learning models such as decision trees or linear models like SVMs offer transparent decision-making processes. These models provide explanations, enabling users to comprehend the rationale behind specific classifications.
∗ **Works Well with Limited Data:** Traditional machine learning algorithms can yield reasonable results even with limited datasets. They are less reliant on extensive data compared to deep learning methods, proving effective with smaller datasets.

– Deep Learning:
∗ **Feature Learning:** Deep learning excels in autonomously learning intricate representations or features from raw data. Neural networks, especially architectures like CNNs and RNNs, possess the capability to discern complex patterns and relationships inherent in data, reducing reliance on manual feature engineering.
∗ **Complex Pattern Recognition:** Deep learning models exhibit superior performance in recognizing intricate patterns, handling nonlinear relation- ships, and capturing subtle nuances in complex data, such as speech signals.

– Feature Extraction:
∗ **Tailored Feature Selection:** Feature extraction methods facilitate the selection of specific features most pertinent to the given task. It offers flexibility in choosing and customizing discriminative features essential for identifying emotional cues in speech.

• Limitations:

– Machine Learning:
∗ **Complexity in Relationships:** Traditional ma- chine learning models might struggle when faced with highly complex or nonlinear relationships between features and emotions. They may fail to capture intricate patterns, impacting performance in specific scenarios.

– Deep Learning:
∗ **Demands for Data and Resources:** Deep learning models, especially those with multiple layers, necessitate large datasets for effective training and generalization. Additionally, training deep neural networks requires substantial computational re- sources, rendering them computationally expensive.
∗ **Lack of Interpret-ability:** Due to their intricate architectures, deep learning models often function as 'black boxes,' lacking transparency in decision- making processes. Understanding their reasoning or internal mechanisms behind predictions can pose challenges.

– Feature Extraction:
∗ **Sensitive to Variability and Noise:** Feature extraction methods might struggle with variability in emotional expressions and noise within data. They may not generalize well across diverse patterns or effectively handle noisy data.

IV. DATASETS, EVALUATION METRICS AND EXPERIMENTATION STRATEGIES

4.1 Datasets in Speech Emotion Recognition

• **Emo-DB (Emotional Database):** A frequently utilized German dataset that contains acted emotional speech recordings performed by 10 actors. This database covers seven emotions: anger, boredom, disgust, fear, happiness, sadness, and neutrality.

• **IEMOCAP (Interactive Emotional Dyadic Motion Capture):** This database captures dyadic interactions, encompassing emotional speech data derived from both scripted and improvised conversations between actors. It provides a rich multi-modal portrayal of emotions.

• **RAVDESS (Ryerson Audio-Visual Database of Emotional Speech and Song):** An extensive dataset consisting of acted speech performed by professional actors. It covers eight emotions: calm, happy, sad, angry, fearful, surprised, disgusted, and neutral emotions, presented in both speech and song.

• **SAVEE (Surrey Audio-Visual Expressed Emotion):** A British English database containing acted emotional speech by male speakers. It includes recordings that represent seven emotions: anger, disgust, fear, happiness, sadness, surprise, and neutrality.

• **CREMA-D (Crowd-sourced Emotional Multimodal Actors Dataset):** A diverse dataset comprise speech recordings from professional actors. This dataset spans emotions across various modalities, offering audio-visual emotional content.

4.2 Evaluation Metrics in Speech Emotion Recognition

• **Accuracy:** Assesses the overall accuracy of the extracted information, determining how well the parser correctly identifies and categorizes elements such as education, work experience, and skills.

• **Precision and Recall:** Precision measures the accuracy of specific information extraction (e.g., skills) concerning the total predicted instances. Recall evaluates the parser's ability to capture all relevant instances of a particular information category.

• **F1-Score:** Represents the harmonic mean of precision and recall, offering a balanced evaluation of the parser's overall performance in information extraction.

• **Confusion Matrix:** Offers a tabular representation com- paring model predictions to true values, facilitating an understanding of classification performance across various emotions.

• **Arousal-Valence Space:** A dimensional representation frequently utilized in SER, depicting emotions in a two- dimensional space (arousal and valence). Model predictions can be assessed by comparing predicted emotion positions to actual positions in this space.

• **Receiver Operating Characteristic (ROC) Curve and Area Under the Curve (AUC):** Commonly employed in binary classification problems, ROC curves plot the true positive rate against the false positive rate. AUC quantifies model performance by representing the area under the ROC curve.

• **Mean Opinion Score (MOS):** A subjective evaluation metric where human annotators rate the quality of emotion detection results, offering a qualitative assessment of model performance.

• **Cross-validation:** A technique used to evaluate model performance by partitioning the dataset into subsets for training and testing multiple times. This approach minimizes variability and ensures more dependable performance estimates.

4.3 Experimentation Strategies

• **Model Fine-tuning**

– Model fine-tuning involves adapting or adjusting a pre-existing model, often previously trained on a large dataset or a related task, to better perform a specific task, such as recognizing emotions from speech. This process entails updating the model's parameters using a smaller, task-specific dataset.

∗ **Process:**

· **Pre-trained Model:** Start with a model that has already been trained on a substantial dataset, leveraging its learned representations.

· **Task-specific Data:** Utilize a dataset specifically annotated for emotion recognition.

· **Iterative Optimization:** Adjust the model's parameters through multiple training iterations on the emotion recognition dataset.
· **Backpropagation:** Update the model's weights by minimizing a loss function that measures the difference between predicted and actual emotion labels.

– **Purpose:** Fine-tuning enables the model to adapt its learned features to the nuances of the emotion recognition task, thereby enhancing its performance in recognizing emotions from speech data.

• **Masking Strategies**
– Masking involves concealing or modifying specific parts of the input data to enhance model robustness, promote generalization, or augment the training pro- cess.

∗ **Frame-based Masking:**
· **Concept:** Segmenting input data into frames, such as short time segments in audio.
· **Application:** Masking or altering specific patches during training.
· **Purpose:** Encouraging the model to learn robust representations by focusing on various segments of the input data.

∗ **Patch-based Masking:**
· **Concept:** Dividing the input into patches or sections.
· **Application:** Concealing or modifying specific frames or portions of frames during model training.
· **Purpose:** Augmenting data to improve the model's ability to recognize emotions in speech amidst diverse contexts or noise.

– **Purpose:** Masking strategies aid models in learning more generalized features, reducing reliance on specific patterns and improving their capability to recognize emotions across diverse speech contexts.

• **Hyperparameter Tuning**
– Hyperparameters are settings external to the model architecture that influence its learning process. Tuning these parameters involves exploring various combinations to find optimal settings for the model's performance.

∗ **Parameters Tuned:**
· **Encoder Depth:** Number of layers in the neural network's encoding component.
· **Masking Ratio:** Proportion of data masked during training.
· **Token Sizes:** Size or length of input tokens or segments.

∗ **Strategy:**
· **Grid Search or Random Search:** Systematically or randomly exploring a range of hyper- parameter values.
· **Evaluation**: Assessing model performance with different settings using validation datasets.
· **Optimization:** Identifying the configuration that yields the best emotion recognition performance.

– **Purpose:** Hyperparameter tuning aims to determine the most effective configuration for the model's learning process, maximizing its ability to accurately and efficiently recognize emotions from speech data.

V. CONCLUSIONS

There's a strong focus on utilizing diverse datasets sourced from multiple platforms, including acted subsets, podcasts, and languages like Farsi. Larger datasets provide a more comprehensive understanding of emotions and contribute significantly to model training and assessment.

The choice of features and models plays a crucial role in emotion recognition. Techniques such as MFCC, deep learning architectures (CNNs, LSTMs, Transformers), attention mechanisms, and fusion of contextual information have exhibited potential. However, performance often varies depending on the features used and the complexity of the models.

Understanding emotional context across different time- frames and hierarchical structures in speech is vital. Models like TIM-Net and SpeechFormer++ underscore the significance of capturing multi-scale contextual information, intra- and inter-unit features, and leveraging implicit relationships in speech to enhance performance.

Transformer architectures have proven highly effective in speech emotion recognition, surpassing previous methods. Their ability to capture implicit linguistic information, handle linguistic and paralinguistic cues, and generalize across domains suggests a potential shift towards pre-trained, transformer-based models in the field.

Approaches that combine different methodologies, like VQ- MAE-S, showcasing the combination of VQ-VAE with MAE for speech representation learning, have demonstrated enhanced performance. Fusion methods and adaptation techniques show promise in improving SER accuracy.

The incorporation of contextual, multimodal, and domain- specific information enhances emotion recognition. Context- dependent domain adversarial neural networks highlight the benefits of considering contextual and multimodal data for improved emotion recognition, particularly in low-resource conditions.

Persistent challenges include disentangling emotion from speech content, generalizing across diverse corpora, effectively handling contextual information, and leveraging unlabeled data. Future research might concentrate on disentangling emotional and content information, integrating semantic and lexical details, and evaluating unlabeled samples from varied sources to further enhance SER systems.

REFERENCES

[1] C. M. Lee, S. S. Narayanan, "Towards detecting emotions in spoken dialogs," IEEE transactions on speech and audio processing, Vol. 13, No. 2, March 2022.
[2] S. S. Narayanan, "Toward detecting emotions in spoken dialogs," IEEE Trans. Speech Audio Process., vol. 13, no. 2, pp. 293–303, Mar. 2021.
[3] A. Nogueiras, A. Moreno, A. Bonafonte, Jose B. Marino, "Speech Emotion detection Using Hidden Markov Model," Eurospeech, 2021.
[4] Speech Emotion detection Using CNN, Article In International Journal Of Psychosocial Rehabilitation, June 2020, Harini Murugan
[5] ZT Liu et al., "Speech emotion detection based on an improved brain emotion learning model," Neurocomputing, vol. 309, pp. 145-156, 2018.
[6] KP Seng et al., "A combined rule-based and machine learning audio- visual emotion detection approach," IEEE Transactions on Affective Computing, vol. 9, no. 1, pp. 3-13, 2018.
[7] S Deb and S. Dandapat, "Multiscale amplitude feature and significance of enhanced vocal tract information for emotion classification," IEEE Transactions on Cybernetics, vol. 49, no. 3, pp. 802-815, 2018.
[8] S Zhang et al., "Speech emotion detection using an enhanced kernel isomap for human-robot interaction," International Journal of Advanced Robotic Systems, vol. 10, no. 2, pp. 1-7, 2017.
[9] S. Kuchibhotla et al., "An optimal two-stage feature selection for speech emotion detection using acoustic features," International Journal of Speech Technology, vol. 19, pp. 657, 2016.
[10] S Chen et al., "Speech emotion classification using multiple kernel Gaussian process," APSIPA 2016, pp. 1-4, 2016.