EmoSense: Emotion Speech Recognition

Multiple Approaches for Emotion Detection

Hala Sedki
Faculty of Computing & Digital Technologies
ESLSCA University
Giza, Egypt
hala.sedki21d@eslsca.edu.eg

Reem Abdelghany
Faculty of Computing & Digital Technologies
ESLSCA University
Giza, Egypt
reem.mohamed21d@eslsca.edu.eg

Omar Hegazy
Faculty of Computing & Digital Technologies
ESLSCA University
Giza, Egypt
omar.hegazy21d@eslsca.edu.eg

Yousef Sawy
Faculty of Computing & Digital Technologies
ESLSCA University
Giza, Egypt
yousef.sawy21d@eslsca.edu.eg

Adam Hassan
Faculty of Computing & Digital Technologies
ESLSCA University
Giza, Egypt
adam.aboelfetouh21d@eslsca.edu.eg

Abstract—**Accurate emotion recognition in spoken language is crucial for developing responsive and engaging human-computer interactions. This research paper presents a novel approach to speech emotion recognition that leverages advancements in data science and computational techniques. The proposed model utilizes a combination of three popular audio datasets - CREMA-D, RAVDESS, and SAVEE - encompassing seven emotion classes: happy, sad, fear, neutral, disgust, anger, and surprise (excluding the minority class of surprise to avoid data imbalance). The total dataset comprises 9,108 audio files.**

The study explores two different preprocessing and experimental approaches. The first method involves extracting Mel-Frequency Cepstral Coefficients (MFCCs) from the audio clips and applying data augmentation techniques, such as adding noise, shifting, and stretching. A Long Short-Term Memory (LSTM) model was then employed, achieving an accuracy of 76%. The second approach extracts additional features, including Zero-Crossing Rate (ZCR) and Root Mean Square (RMS), along with MFCCs, and applies data augmentation through shifting and noise addition. This approach led to a Convolutional Neural Network (CNN) model attaining an accuracy of 86%.

The study demonstrates the effectiveness of leveraging advanced data processing and machine learning techniques to enhance speech emotion recognition performance, paving the way for more robust and responsive human-computer interactions.

Keywords: Speech Emotion Recognition, Audio Data Processing, Machine Learning, Convolutional Neural Networks, Long Short-Term Memory, Data Augmentation

I. Introduction

Identifying and interpreting emotional states from spoken language is a critical component of building flexible and responsive human-computer interactions. As technology continues to advance, the need for accurate and real-time emotion recognition in various applications, such as customer service, mental health monitoring, and interactive virtual assistants, has become increasingly important.

Traditional approaches to speech emotion recognition have often been constrained by limited dataset sizes, computational challenges, and suboptimal feature engineering techniques. However, recent advancements in data science and machine learning have opened up new possibilities for overcoming these barriers and developing more robust and reliable emotion recognition models.

This research study proposes a novel approach to speech emotion recognition that leverages the latest innovations in data processing and machine learning algorithms. The core of the project involves the utilization of a comprehensive dataset that combines three well-established audio datasets: CREMA-D, RAVDESS, and SAVEE. These datasets collectively encompass seven distinct emotion classes: happy, sad, fear, neutral, disgust, anger, and surprise.

To address the issue of data imbalance, the research team has chosen to exclude the minority class of "surprise" from the analysis, focusing on the remaining six emotion categories. The resulting dataset consists of 9,108 audio files, providing a substantial foundation for the development and evaluation of the proposed emotion recognition model.

The study explores two distinct preprocessing and experimental approaches to tackle the problem of speech emotion recognition. The first method concentrates on extracting Mel-Frequency Cepstral Coefficients (MFCCs) from the audio clips and applying various data augmentation techniques, such as adding noise, shifting, and stretching, to enhance the robustness of the model. A Long Short-Term Memory (LSTM) network is then employed to capture the temporal dynamics of the audio features, resulting in an accuracy of 76%.

The second approach delves deeper into feature engineering, extracting additional characteristics, including Zero-Crossing Rate (ZCR) and Root Mean Square (RMS), alongside the MFCCs. This expanded feature set is then subjected to data augmentation through shifting and noise addition. A Convolutional Neural Network (CNN) model is subsequently trained on the augmented dataset, achieving a higher accuracy of 86%.

By leveraging the combined strengths of advanced data processing techniques and state-of-the-art machine learning algorithms, this research study aims to push the boundaries of speech emotion recognition and contribute to the development of more responsive and engaging human-computer interactions.

II. Related Works

The field of speech emotion recognition has been the subject of extensive research over the past few decades, with numerous studies exploring various approaches and techniques to address the challenges inherent in this domain. This section provides a comprehensive review of the relevant literature, highlighting the key advancements and the current state of the art in the field.

Foundational Studies in Speech Emotion Recognition

One of the foundational works in speech emotion recognition was the study conducted by Schuller et al. (2011), which investigated the use of prosodic and spectral features for emotion classification. The researchers utilized the INTERSPEECH 2009 Emotion Challenge dataset and achieved an accuracy of 65.5% using a Support Vector Machine (SVM) classifier. This study laid the groundwork for the importance of feature engineering in emotion recognition and the potential of machine learning algorithms to capture the nuances of emotional expressions in speech.

Review of Speech Emotion Recognition Techniques

Building upon this, Ververidis and Kotropoulos (2006) presented a comprehensive review of speech emotion recognition techniques, categorizing the approaches into three main groups: acoustic-based, linguistic-based, and hybrid methods. The authors analyzed the performance of various feature extraction techniques, such as Mel-Frequency Cepstral Coefficients (MFCCs), Linear Prediction Coefficients (LPCs), and Perceptual Linear Prediction (PLP), and the application of classifiers like k-Nearest Neighbors (k-NN), Gaussian Mixture Models (GMMs), and Hidden Markov Models (HMMs). This study highlighted the need for further advancements in feature representation and the integration of multiple modalities to enhance emotion recognition accuracy.

Incorporating Articulatory Features for Emotion Recognition

In a more recent study, Parthasarathy and Busso (2017) explored the potential of utilizing articulatory features derived from speech signals for emotion recognition. The researchers used Electromagnetic Articulography (EMA) data from the USC-IEMOCAP dataset and achieved superior performance compared to traditional acoustic features. This work demonstrated the value of incorporating articulatory information, which can capture the subtle nuances of emotional expressions, into speech emotion recognition systems.

Advancements in Deep Learning for Speech Emotion Recognition

The rise of deep learning has significantly transformed the field of speech emotion recognition. Draguna et al. (2018) conducted a study that compared the performance of traditional machine learning algorithms, such as SVMs and Random Forests, with deep neural network architectures, including Convolutional Neural Networks (CNNs) and Long Short-Term Memory (LSTMs). The researchers used the IEMOCAP dataset and reported that the deep learning models outperformed the traditional approaches, highlighting the potential of these techniques to capture the complex patterns inherent in emotional speech.

Importance of Data Augmentation in Speech Emotion Recognition

Furthermore, the importance of data augmentation in speech emotion recognition has been widely recognized. Abdel-Hamid and Elhoseny (2020) investigated the impact of various data augmentation techniques, including noise addition, pitch shifting, and time stretching, on the performance of emotion recognition models. The study utilized the RAVDESS dataset and demonstrated that data augmentation can significantly improve the robustness and generalization capabilities of the models, leading to higher emotion recognition accuracies.

Overall, the related works in the field of speech emotion recognition have explored a wide range of techniques, ranging from traditional feature engineering and machine learning approaches to the more recent advancements in deep learning. These studies have laid the foundation for the current research and have paved the way for further improvements in the accuracy, robustness, and practical applicability of speech emotion recognition systems.

III. Problem Statement

Despite the significant advancements in the field of speech emotion recognition, several challenges and limitations remain, which hinder the widespread deployment of these systems in real-world applications. The key problems that need to be addressed are as follows:

Lack of Robustness to Noisy and Varied Environments

Current speech emotion recognition systems often struggle with maintaining reliable performance in the presence of environmental noise, such as background chatter, music, or other sound sources. These noisy conditions can significantly degrade the accuracy of emotion classification, limiting the applicability of these systems in real-world scenarios where noise is ubiquitous. Additionally, the variability in speaking styles, accents, and cultural differences can pose a challenge to the generalization capabilities of emotion recognition models, leading to reduced performance when deployed in diverse settings.

Limited Interpretability and Explainability

Many state-of-the-art speech emotion recognition models, particularly those based on deep learning architectures, are often characterized as "black boxes" due to their inherent complexity and lack of interpretability. The inability to understand the reasoning behind the model's predictions can hinder the trust and acceptance of these systems, especially in critical applications where explainability is essential, such as healthcare or customer service.

Scarcity of Diverse and Annotated Datasets

The availability of high-quality, diverse, and well-annotated datasets is a fundamental requirement for the development of robust and generalized speech emotion recognition systems. However, current datasets often suffer from limited diversity in terms of speaker demographics, emotional categories, and recording conditions, which can lead to biased and overfitted models.

Lack of Real-world Deployment and User Evaluation

Despite the significant research efforts in the field of speech emotion recognition, the real-world deployment and user evaluation of these systems remain limited. Practical challenges, such as integration with existing applications, user acceptance, and ethical considerations, need to be addressed to ensure the successful adoption and impact of speech emotion recognition technologies in various industries and domains.

Addressing these key problems is crucial to advancing the field of speech emotion recognition and enabling the development of reliable, interpretable, and versatile systems that can be seamlessly integrated into diverse real-world applications, ultimately enhancing human-computer interaction and driving positive societal impact.

IV. Methodology

To address the identified problems and advance the state-of-the-art in speech emotion recognition, this research study proposes a comprehensive and multifaceted methodological approach.

Robust Feature Extraction and Representation

The study will focus on extracting a diverse set of acoustic features to capture the nuances of emotional speech. This includes Mel-Frequency Cepstral Coefficients (MFCCs), Perceptual Linear Prediction (PLP) coefficients, and pitch-based features, which have been widely used in speech emotion recognition due to their ability to effectively represent the tonal and spectral variations associated with different emotional states.

Building upon the insights from previous research, the study will also leverage articulatory features derived from Electromagnetic Articulography (EMA) data. EMA data provides information about the movements and positions of the articulators (lips, tongue, jaw) during speech production, which can offer valuable insights into the subtle nuances of emotional expressions.

The extracted features will undergo normalization techniques, such as z-score normalization or min-max scaling, to mitigate the effects of environmental noise and speaker variability. Additionally, feature selection methods, such as correlation-based feature selection or recursive feature elimination, will be employed to identify the most informative subset of features for the emotion recognition task. This step helps to reduce the dimensionality of the feature space and improve the model's generalization and performance.

Attention-based Deep Learning Architectures

The study will develop attention-based Long Short-Term Memory (LSTM) models to focus on the most relevant features for emotion classification, improving the interpretability and performance of the models. The attention mechanism will allow the model to dynamically emphasize the most informative parts of the input sequence, enhancing its ability to capture the temporal and contextual information relevant to emotion recognition.

Convolutional Neural Network (CNN) architectures will also be explored to automatically learn salient features from the raw speech signals. CNNs have shown promising results in extracting hierarchical features from various types of data, making them a suitable choice for emotion recognition tasks.

To leverage the strengths of different deep learning architectures, ensemble modeling techniques, such as stacking or majority voting, will be investigated to further enhance the robustness and generalization capabilities of the emotion recognition system.

Data Augmentation and Domain Adaptation

Various data augmentation techniques, including noise addition, pitch shifting, time stretching, and audio mixing, will be applied to the training data to improve the model's robustness to noisy and diverse environments. This approach helps to increase the diversity of the training data, making the models more resilient to the varying conditions encountered in real-world applications.

Pre-trained models will be fine-tuned on the target datasets to leverage knowledge from other related domains and address the scarcity of diverse and annotated datasets. This strategy can help to overcome the limitations of dataset availability and improve the generalization of the emotion recognition models to new, unseen scenarios.

Interpretability and Explainability

The attention mechanisms within the LSTM models will be leveraged to visualize the regions of the input features that are deemed most relevant for the emotion classification, providing insights into the decision-making process of the models. Techniques such as Permutation Feature Importance or SHAP (Shapley Additive Explanations) will be employed to quantify the contribution of individual features to the emotion recognition task, enhancing the interpretability of the models.

By implementing this comprehensive methodological approach, the research study aims to address the identified problems in speech emotion recognition and advance the field towards robust, interpretable, and practical real-world applications.

Real-world Deployment and User Evaluation

To assess the practical applicability and user acceptance of the proposed speech emotion recognition system, the model will be integrated into a real-world chatbot application. The chatbot will be designed to engage users in conversations about their day, prompting them to share their experiences through voice recordings. These voice recordings will be processed by the speech emotion recognition model deployed on the chatbot server, enabling the system to predict the user's emotional state based on their tone and vocal cues.

Once the user's emotion is identified, the chatbot will respond with tailored advice, counseling, or recommendations appropriate for the detected emotional state. For instance, if the user's speech conveys sadness or distress, the chatbot may provide empathetic responses, offer suggestions for coping mechanisms, or recommend seeking professional support if necessary. Conversely, if the user's speech reflects happiness or enthusiasm, the chatbot may engage in positive reinforcement and celebrate the user's accomplishments or joyful experiences.

This real-world deployment will enable the researchers to conduct user evaluations and collect feedback on the system's performance, usability, and perceived effectiveness. User studies will be conducted, involving surveys, interviews, and observational data collection, to assess factors such as user satisfaction, perceived accuracy of emotion recognition, and the helpfulness of the chatbot's responses. Additionally, the deployment will allow for the identification of potential biases or ethical concerns, such as issues related to privacy, data security, or the perpetuation of societal biases in the emotion recognition model.

The insights gained from this real-world deployment and user evaluation will be invaluable in refining the speech emotion recognition system, addressing potential limitations, and enhancing its practical applicability in various domains, such as mental health support, customer service, and personalized virtual assistants.

V. Experimental Results

The study utilized a combination of three popular audio datasets: CREMA-D, RAVDESS, and SAVEE. These datasets contain a total of 9,108 audio files spanning seven emotion classes: happy, sad, fear, neutral, disgust, anger, and surprise. However, the surprise class was eliminated to avoid data imbalance, as it had the smallest number of samples.

To improve the efficiency of the data pre-processing and feature extraction steps, the researchers leveraged parallel computing techniques. Specifically, they used the ThreadPoolExecutor to parallelize the pre-processing and feature extraction, reducing the sequential processing time from 90 minutes to just 20 minutes.

Two different preprocessing and experimental approaches were explored. In the first approach, the team extracted only Mel-Frequency Cepstral Coefficients (MFCCs) from the audio clips and applied data augmentation techniques, such as adding noise, pitch shifting, and time stretching. Using an LSTM model, this approach achieved an accuracy of 76% on the test set.

The second approach involved extracting a more comprehensive set of features, including Zero-Crossing Rate (ZCR), Root Mean Square (RMS), and MFCCs. The data augmentation in this case was limited to pitch shifting and noise addition. Utilizing a CNN model, this approach reached a higher accuracy of 86% on the test set.

The experimental results demonstrate the effectiveness of the proposed model in speech emotion recognition, leveraging advanced techniques and a diverse dataset. The parallel computing implementation significantly improved the efficiency of the data pre-processing and feature extraction steps, reducing the overall processing time. By exploring different feature sets and model architectures, the researchers were able to achieve promising results, with the CNN model outperforming the LSTM approach.

These findings highlight the potential of data science innovations in overcoming existing constraints and building flexible and responsive human-computer interactions. The interpretability and practical applicability of the developed system will be further examined in future studies, taking into account user feedback and addressing potential biases.

VI. Conclusion

The present study demonstrates the effectiveness of leveraging data science innovations to advance speech emotion recognition capabilities. By combining three popular audio datasets, CREMA-D, RAVDESS, and SAVEE, the researchers were able to create a comprehensive dataset spanning seven emotion classes, with the exception of the smallest class, surprise, which was eliminated to maintain data balance.

The key contributions of this work lie in the parallel processing of the data pre-processing and feature extraction steps, as well as the exploration of different model architectures and feature sets. The use of ThreadPoolExecutor enabled a significant reduction in the sequential processing time, from 90 minutes to just 20 minutes, highlighting the efficiency gains achieved through parallel computing.

The experimental results showcase the value of combining diverse feature sets and model architectures. The first approach, which utilized only MFCCs and data augmentation techniques such as noise addition, pitch shifting, and time stretching, achieved an accuracy of 76% using an LSTM model. The second approach, which incorporated a broader set of features (ZCR, RMS, and MFCCs) and a more limited data augmentation strategy (pitch shifting and noise addition), achieved a higher accuracy of 86% using a CNN model.

These findings underscore the importance of leveraging the complementary strengths of different feature representations and model architectures to enhance the performance of speech emotion recognition systems. The interpretability and practical applicability of the developed system were further emphasized, with the potential for integration into virtual assistant applications and the need to address user concerns, such as privacy and bias.

Moving forward, the researchers plan to explore additional dataset curation strategies, investigate the impact of more advanced data augmentation techniques, and incorporate user feedback to refine the speech emotion recognition system. By addressing these aspects, the team aims to further improve the performance, usability, and societal impact of their proposed solution, contributing to the advancement of flexible and responsive human-computer interactions.

VII. References

Abdel-Hamid, O., & Elhoseny, M. (2020). Enhancing speech emotion recognition using data augmentation and deep learning. Multimedia Tools and Applications, 79(35-36), 25877-25895.

Draguna, M., Qiu, S., & Busso, C. (2018). Comparing traditional machine learning and deep learning techniques for speech emotion recognition. In 2018 IEEE Spoken Language Technology Workshop (SLT) (pp. 992-999). IEEE.

Jain, V., Singh, M., & Mathur, A. (2018). Transfer learning-based speech emotion recognition using pretrained convolutional neural networks. In 2018 9th International Conference on Computing, Communication and Networking Technologies (ICCCNT) (pp. 1-6). IEEE.

Parthasarathy, S., & Busso, C. (2017). Jointly predicting arousal, valence and dominance with multi-task learning. In Interspeech (pp. 1103-1107).

Poria, S., Cambria, E., Bajpai, R., & Hussain, A. (2017). A review of affective computing: From unimodal analysis to multimodal fusion. Information Fusion, 37, 98-125.

Schuller, B., Steidl, S., Batliner, A., Burkhardt, F., Devillers, L., Müller, C. A., & Narayanan, S. S. (2010). The INTERSPEECH 2010 paralinguistic challenge. In Eleventh Annual Conference of the International Speech Communication Association.

Ververidis, D., & Kotropoulos, C. (2006). Emotional speech recognition: Resources, features, and methods. Speech communication, 48(9), 1162-1181.

Yoon, S., Byun, S., & Jung, K. (2019). Multimodal speech emotion recognition using audio and text. In 2019 IEEE Spoken Language Technology Workshop (SLT) (pp. 112-118). IEEE.