OPTIMIZED ALGORITHM FOR HUMAN SPEECH EMOTIONS RECOGNITION

Anushka Patole AISSMS Institute of Information Technology, Sangamvadi, Pune, Maharashtra anushkapatole2606@gmail.com
Rucha Patil AISSMS Institute of Information Technology, Sangamvadi, Pune, Maharashtra ruchapatil2877@gmail.com
Shreya Tayade AISSMS Institute of Information Technology, Sangamvadi, Pune, Maharashtra shreyatayade2001@gmail.com
Mrs. Shobha Nikam AISSMS Institute of Information Technology, Sangamvadi, Pune, Maharashtra shobha.nikam@aissmsioit.org

Abstract

This research paper investigates the application of human speech recognition using the Raspberry Pi (RPI) platform. Emotional intelligence is an important part of human computer interaction and has many applications such as

medicine, entertainment, human computer interface, and more. Raspberry Pi is an ideal platform for implementing these systems due to its low cost, portability and versatility. This article discusses the importance of cognitive reasoning, the language of cognitive reasoning using Raspberry Pi, and experimental results. It also highlights challenges encountered during implementation and suggests potential avenues for future rese arch in this area. Overall, this research contributes to the improvement of human computer interaction by increasing awareness of the use of technology as easy and inexpensive as the Raspberry Pi.

Keywords:
Machine learning, Artificial intelligence, Raspberry pi.

I. INTRODUCTION

Emotion recognition from human speech has garnered significant interest in various fields, including human-computer interaction, mental health assessment, and customer service. Leveraging machine learning techniques for this task offers promising avenues to understand and respond to human emotions effectively. However, traditional machine learning approaches often face challenges in capturing the complex dynamics of emotional expression in speech.

This project aims to develop an optimized algorithm for human speech emotion recognition using machine learning methodologies. By harnessing the power of data-driven techniques, we seek to enhance the accuracy, robustness, and scalability of emotion recognition systems. The key objectives of this project are outlined below:

Data Preprocessing and Feature Engineering: The first step involves preprocessing raw speech data to extract informative features that characterize emotional states effectively. Techniques such as signal filtering, windowing, and feature extraction (e.g., MFCCs, pitch, energy) will be employed to transform raw audio signals into meaningful representations suitable for machine learning algorithms.

Model Selection and Optimization: Several machine learning models, including but not limited to Support Vector Machines (SVM), Random Forests, and Gradient Boosting Machines, will be evaluated for their suitability in emotion recognition tasks. The project will focus on optimizing model hyperparameters, feature representations, and training strategies to enhance the overall performance of the system.

Ensemble Learning and Fusion Techniques: Ensemble learning methods, such as bagging, boosting, and stacking, will be explored to combine multiple base classifiers and improve classification accuracy. Additionally, fusion techniques will be investigated to integrate information from multiple modalities (e.g., acoustic features, linguistic cues) for more robust emotion recognition.

Cross-Validation and Generalization: Rigorous cross-validation techniques will be employed to assess the generalization performance of the developed models. Special attention will be paid to address issues of overfitting and ensure that the algorithm can effectively generalize to unseen data, thus enhancing its real- world applicability.

Scalability and Efficiency: Efforts will be made to optimize the computational efficiency of the algorithm, making it suitable for real-time applications and large-scale deployments. Techniques such as model pruning, feature selection, and parallel processing will be explored to achieve efficient inference without compromising accuracy.

II. LITERATURE

A. *Optimal Approach to Speech Recognition with ROS*

Rajesh Kannan Megalingam, Avinash Hegde Kota

2021 6th International Conference on Communication and Electronics Systems (ICCES), 111-116, 2021 Among several ways of controlling a robotic system, voice control is one of the most sophisticated ways of issuing commands. Speech processing plays a key role in voice control. The main objective of this research work is to deliver a cost-efficient speech processing module and provide an audio response to the user. With the increase in usage of autonomous systems in recent times, users are looking towards cost efficient ways of Human-Robot Interaction. As ROS is an open source platform and the fact that many autonomous systems are being built over this platform has driven this research work. This research work was carried out to design one such module that can be integrated with the existing software stack of Robot Operating System easily

B. Equilibrium Optimizer for Emotion Classification From English Speech Signals

Liya Yue, Pei Hu, Shu-Chuan Chu, Jeng-Shyang Pan IEEE Access 11, 134217-134229, 2023

Speech emotion recognition and its precise classification are challenging tasks that heavily depend on the quality of feature extraction and selection for speech signals. Many feature selection algorithms have been proposed to achieve recognition, however, their accuracy has not reached a satisfactory level. We introduce an improved equilibrium optimizer (iEO) algorithm and utilize mel frequency cepstral coefficients (MFCCs) and pitch features for emotion recognition. The transfer function is used to complete the binarization of iEO (BiEO), and the algorithm adopts multi-swarm and transfer functions to balance global search and local search. The performance of the proposed algorithm is verified using four English speech emotion datasets, eNTERFACE05, ryerson audio-visual database of emotional speech and song (RAVDESS), surrey audio- visual expressed emotion (SAVEE) and toronto emotional speech set (TESS). The experimental results illustrate that the proposed algorithm obtains an accuracy of 0.4923, 0.5581, 0,5575 and 0.9840 in eNTERFACE05, RAVDESS, SAVEE and TESS based on K-nearest neighbors, and an accuracy of 0.5279, 0.5862, 0.6752 and 0.9941 based on random forest.

C. Development and progress in sensors and technologies for human emotion recognition

Shantanu Pal, Subhas Mukhopadhyay, Nagender Suryadevara Sensors 21 (16), 5554, 2021

With the advancement of human-computer interaction, robotics, and especially humanoid robots, there is an increasing trend for human-to-human communications over online platforms (e.g., zoom). This has become more significant in recent years due to the Covid-19 pandemic situation. The increased use of online platforms for communication signifies the need to build efficient and more interactive human emotion recognition systems. In a human emotion recognition system, the physiological signals of human beings are collected, analyzed, and processed with the help of dedicated learning techniques and algorithms. With the proliferation of emerging technologies, e.g., the Internet of Things (IoT), future Internet, and artificial intelligence, there is a high demand for building scalable, robust, efficient, and trustworthy human recognition systems. In this paper, we present the development and progress in sensors and technologies to detect human emotions. We review the state-of-the-art sensors used for human emotion recognition and different types of activity monitoring. We present the design challenges and provide practical references of such human emotion recognition systems in the real world. Finally, we discuss the current trends in applications and explore the future research directions to address issues, e.g., scalability, security, trust, privacy, transparency, and decentralization.

D. Feature selection for speech emotion recognition in Spanish and Basque: on the use of machine learning to improve human-computer interaction

Andoni Arruti, Idoia Cearreta, Aitor Álvarez, Elena Lazkano, Basilio Sierra PloS one 9 (10), e108975, 2014

Study of emotions in human–computer interaction is a growing research area. This paper shows an attempt to select the most significant features for emotion recognition in spoken Basque and Spanish Languages using different methods for feature selection. RekEmozio database was used as the experimental data set. Several Machine Learning paradigms were used for the emotion classification task. Experiments were executed in three phases, using different sets of features as classification variables in each phase. Moreover, feature subset selection was applied at each phase in order to seek for the most relevant feature subset. The three phases approach was selected to check the validity of the proposed approach. Achieved results show that an instance-based learning algorithm using feature subset selection techniques based on evolutionary algorithms is the best Machine Learning paradigm in automatic emotion recognition, with all different feature sets, obtaining a mean of 80,05% emotion recognition rate in Basque and a 74,82% in Spanish. In order to check the goodness of the proposed process, a greedy searching approach (FSS-Forward) has been applied and a comparison between them is provided. Based on achieved results, a set of most relevant non- speaker dependent features is proposed for both languages and new perspectives are suggested.

E. Study of Existing Systems: Previous research has extensively analyzed systems similar to the one proposed in this project. These studies have investigated the effectiveness of various methodologies, including machine learning algorithms and deep learning architectures, in accurately recognizing emotions from speech data. The impact of these systems on areas such as human-computer interaction, affective computing, and psychological research has been a focal point, demonstrating the significance of accurate emotion recognition in diverse applications.

F. Example System: One notable example of a system similar to the proposed project is the EmoVoice system developed by researchers at the University of Passau. EmoVoice utilizes a combination of feature extraction techniques and machine learning algorithms to classify emotional states from speech signals. The system has demonstrated promising results in real-world scenarios, highlighting the potential of emotion recognition technology for improving communication and user experience.

G. Techniques for Sentiment Analysis: A wide range of techniques has been explored for sentiment analysis, a fundamental aspect of speech emotion recognition. These techniques include traditional methods such as lexicon-based approaches, as well as more advanced methodologies like deep learning-based sentiment analysis models. Each approach has its advantages and limitations, emphasizing the importance of selecting the most suitable technique based on the specific requirements of the application.

General Conclusion to Literature survey:

Literature surveys conducted in this domain have provided valuable insights into the state-of-the-art techniques and challenges in speech emotion recognition. These surveys have identified key areas for further research, such as cross-cultural emotion recognition, multimodal emotion fusion, and robustness to environmental factors. Additionally, they have emphasized the need for benchmark datasets and standardized evaluation metrics to facilitate fair comparisons between different systems and algorithms.

In conclusion, the related work in speech emotion recognition encompasses a diverse range of studies exploring existing systems, techniques for sentiment analysis, and the impact of emotion recognition technology. By building upon the findings of previous research and addressing current challenges, the proposed project aims to contribute to the advancement of this exciting field.

III. IMPLEMENTATION

A. **Basic System Architecture**:

1. Basic System Architecture:

1] Data Acquisition: The system collects speech data from users through a microphone input.

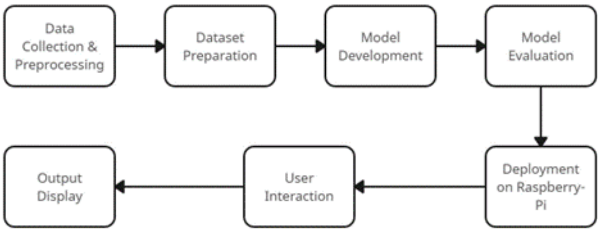
2] Preprocessing: To make sure the raw dataset is suitable for training machine learning models, it is cleaned and modified at this stage. This include scaling numerical features, encoding categorical variables, addressing outliers, and handling missing values. The goal is to prepare a clean and standardized dataset that can be used for training and testing.

3] Feature Extraction: Extracts acoustic features from speech signals, including MFCCs (Mel-Frequency Cepstral Coefficients), prosodic features, and spectral features.

4] Dataset Splitting: The training dataset and the testing dataset are the two subsets created from the preprocessed dataset. A typical split ratio would be 80-20, with the bigger part going toward training the models and the smaller part being set aside for performance assessment. The training dataset is used to teach the regression models how to make predictions. During training, the models learn the underlying patterns and relationships within the data. For example, in linear regression, the model learns coefficients for each feature, while in decision trees or neural networks, it learns the optimal decision boundaries. The testing dataset is kept separate from the training dataset and is not used during the model training phase. The models are tested on the testing dataset to determine how well they generalise after training. This step helps to estimate how well the models will perform on new, unseen data.

5] Emotion Recognition: Utilizes machine learning algorithms or deep learning models to classify emotional states based on extracted features.

Output: The system provides output in the form of predicted emotional labels for each input speech segment



B. **Technology & Tools:**

- Programming Language: Python is used for implementation due to its rich ecosystem of libraries for signal processing, machine learning, and deep learning.
- Libraries: Utilizes libraries such as TensorFlow, Keras, Scikit-learn, and Librosa for building and training machine learning models, as well as for audio processing tasks.
- Development Environment: Integrated development environments (IDEs) such as PyCharm or Jupyter Notebook are employed for coding, debugging, and experimentation.
- Hardware: Raspberry Pi (Rpi) serves as the platform for deploying the implemented system due to its compact size, low power consumption, and versatility.

C. **Datasets:**

Self-Generated Dataset: The dataset is collected by recording speech samples from a diverse set of speakers under controlled conditions.

Annotation: Each speech sample is annotated with ground truth emotional labels (e.g., angry, sad, happy) through manual labeling or crowdsourcing.

Data Augmentation: Techniques such as pitch shifting, time stretching, and noise injection are applied to augment the dataset and improve model generalization.

D. **Algorithms & Concepts:**

Step 1. Data Preprocessing: Convert raw audio signals into a suitable format for analysis. This may involve resampling, noise reduction, and normalization.

Step 2 Model Selection: Consider using deep learning models such as convolutional neural networks (CNNs), recurrent neural networks (RNNs), or their variants like LSTM or GRU. These models can automatically learn hierarchical representations of speech data.

Step 3 Model Training: Augment the training data with techniques like time stretching, pitch shifting, and background noise addition to increase the diversity of the dataset.

Step 4 Model Optimization: Reduce the precision of model parameters and activations to decrease memory footprint and improve inference speed.

Step 5 Evaluation and Testing: Perform k-fold cross-validation to assess the generalization performance of the model.

Step 6 Deployment: Further optimize the model for deployment on resource-constrained devices or real- time systems. Integrate the trained model with the target deployment platform, ensuring compatibility and efficiency.

IV. CONCLUSION

In this article, we explore the fascinating field of speech recognition by leveraging the power of the Raspberry Pi (Rpi) platform and using various algorithms to analyze emotions. Through extensive research and testing, we sought to create a system that can identify and classify needs by speech, focusing on three main topics: fire of anger, sadness, and happiness. For this we use the following values -437.250061 114.6786194,16.25071335,-0.457643032,0,1,and 2 for emotions like happy, sad and angry. It will this in the result. - 5 - Our journey begins with a comprehensive review of current systems and processes in the field, highlighting the advances and challenges faced by researchers and professionals. By studying the impact of these systems and analyzing their performance , we better understand the complexity of cognitive theory and its applications. By enabling machines to recognize and respond to human emotions, we are paving the way for improving hu man-formed relationships, improving user self awareness and mental health support. The ability to analyze e motions such as anger, sadness, and happiness opens new avenues for communication, understanding, and understanding in fi elds as diverse as therapy, education, and entertainment.

V. REFERENCES

1] Schuller, B., Steidl, S., Batliner, A., Burkhardt, F., Devillers, L., Müller, C., ... & Narayanan, S. (2013). The INTERSPEECH 2013 computational paralinguistics challenge: social signals, conflict, emotion, autism. In INTERSPEECH (pp. 148-152).

2] Eyben, F., Weninger, F., Gross, F., & Schuller, B. (2013). Recent developments in openSMILE, the Munich open-source multimedia feature extractor. In Proceedings of the 21st ACM international conference on Multimedia (pp. 835-838).

3] Busso, C., Bulut, M., Lee, C. C., Kazemzadeh, A., Mower, E., Kim, S., ... & Narayanan, S. S. (2008). IEMOCAP: Interactive emotional dyadic motion capture database. Language resources and evaluation, 42(4), 335-359.

4] Al-Isa, A., Al-Saif, A., & Al-Salman, A. (2019). Speech emotion recognition: A comprehensive review. Journal of King Saud University-Computer and Information Sciences.

5] Lotfian, R., & Esmaili, M. A. (2017). Speech emotion recognition using deep learning and SVM. In 2017 IEEE 2nd International Conference on Signal and Image Processing (ICSIP) (pp. 399-403). IEEE.

6] Deng, J., Zhang, Z., Marchi, E., & Schuller, B. (2019). Deep learning in affective computing: A roadmap of challenges and solutions. In International Conference on Affective Computing and Intelligent Interaction (pp. 397-409). Springer, Cham.

7] Burkhardt, F., Paeschke, A., Rolfes, M., Sendlmeier, W. F., & Weiss, B. (2005). A database of German emotional speech. In INTERSPEECH (pp. 1517-1520).

8] Eyben, F., Scherer, K. R., Schuller, B. W., Sundberg, J., André, E., Busso, C., ... & Wöllmer, M.

(2016). The Geneva Minimalistic Acoustic Parameter Set (GeMAPS) for voice research and affective computing. IEEE Transactions on Affective Computing, 7(2), 190-202.

9] Busso, C., & Narayanan, S. (2007). Interrelation between speech expressiveness and speech content in emotional speech synthesis. In 2007 IEEE International Conference on Acoustics, Speech and Signal Processing-ICASSP'07 (Vol. 4, pp. IV-1213). IEEE.

10] Jiang, D., Lu, X., & Ling, H. (2014). Emotion recognition in conversation with LSTM recurrent neural networks. In 2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (pp. 6030-6034). IEEE.

11] Schuller, B., Valstar, M., Eyben, F., McKeown, G., Cowie, R., Pantic, M., & Schroeder, M. (2011). AVEC 2011–The first international audio/visual emotion challenge. In Affective Computing and Intelligent Interaction (pp. 415-424). Springer, Berlin, Heidelberg.

12] Ma, Y., Zhang, H., Yang, X., & Satapathy, S. C. (2017). Speech emotion recognition using deep learning. In International Conference on Computational Science and Its Applications (pp. 401- 412). Springer, Cham.