TLSWi Emotion Recognition

Project Report

Team:

Ezinne Chinaza Anyaegbudike

Venkata Sreeharsha Asam

Humphrey Ojebor

Supriya Pallamparthi

Bing Zhai

(Academic Supervisor)

Overview

A pivotal aspect of effective human communication is emotion recognition. This AI-driven emotion recognition project represents a cutting-edge endeavor in the field of artificial intelligence (AI) and human-computer interaction. Leveraging advanced machine/deep learning algorithms, this project aimed to design, implement, and deploy a robust solution capable of discerning different emotional states from audio input in real-time.

Introduction

Artificial intelligence (AI) has revolutionized various aspects of our lives, and its potential for understanding human emotions is becoming increasingly evident. AI emotion recognition involves utilizing machine learning algorithms to identify and classify human emotions from various vocal modalities in communication. While it has been extensively explored for adults, its application to children presents a unique set of challenges and opportunities.

Children's emotional expressions are often subtle and complex, making them more difficult to accurately recognize using traditional emotion recognition techniques. Additionally, children's emotional development is dynamic and influenced by various factors, including age, cultural background, and individual experiences. These complexities necessitate the development of this research tailored specifically for children.

In today’s ever evolving technologically driven world, the ability to understand and interpret human emotions through artificial intelligence has become a transformative pursuit. Emotion recognition not only enhances our understanding of a person’s experiences but also opens avenues for personalized interactions and interventions.

TLSWi’s initiative to understand and recognize different emotions that kids who have gone through trauma or other life-transforming situations exhibit as they talk about their experiences, is a brilliant way to understand these kids and how they truly feel about each experience they narrate. The findings of the project could potentially lead to better understanding and support for their emotional development, enhance interventions, and tailor educational approaches.

Literature Review.

As many authors and mental health specialists, like Judith Herman and William Faulkner, have acknowledged, trauma can have a severe psychological impact that may not go away with time (Straussner et al, 2014). According to psychology, "trauma" is an emotionally upsetting, unpleasant, or disturbing event that frequently has long-term detrimental effects on the mind, body, and nervous system (Straussner et al, 2014).

To recognize these traumas and its long-lasting effect in individuals, the emotion recognition model was designed. Human emotion can be identified in a variety of ways, including spoken communication, body posture, facial expressions, and gestures. Human emotion recognition also makes use of a wide range of physiological characteristics, including skin resistance, blood pressure, muscle activity, temperature, and heart rate (Youddha and Shivani, 2022). One of the main research questions in human emotion detection from speech is how to determine the speaker's emotional state from a given speech signal. The ability to identify emotions in speech has drawn increasing attention from researchers in recent decades due to its applications in a variety of real-world scenarios, including call centre conversations, automated response systems, online tutoring, spoken dialogue systems, pain recognition, depression diagnosis, and many more (Youddha and Shivani, 2022). Without linguistic knowledge, a typical speech emotion recognition system extracts the emotional states from the speech signal (Moataz et al, 2011). Another research question in speech emotion recognition is how to best choose a feature set from speech signals (Zhen-Tao et al, 2018).

In this work, we developed an effective framework for recognizing speech emotions, integrating unique classifiers to distinguish between five typical emotional states: happy, neutral, sad, angry and disgust. Machine learning and deep learning models such as SVM, logistic regression and CNN will be deployed to model the dataset. The emotional speech recognition techniques will be able to handle large amounts of speech data in real time and extract speech features that systematically convey attitude and emotion (Dimitrios and Constantine, 2006).

Methodology

This emotion recognition project employed a comprehensive methodology encompassing data preparation, model development, and evaluation.

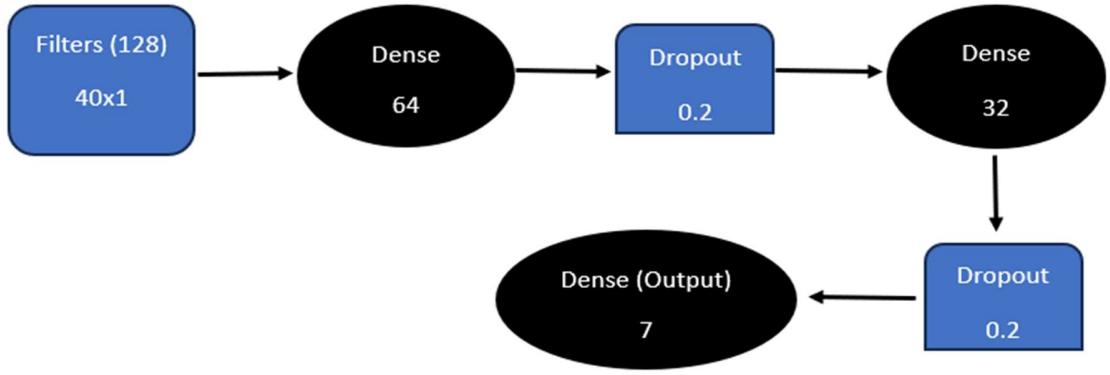
Data Sources and Labelling: The project made use of the suggested *TESS Toronto Emotional Speech* dataset from Kaggle. This set contains 2800 audio samples of seven different emotions (angry, sad, happy, pleasant surprise, disgust, fear, neutral), with all emotions having equal sample size of 400 each.

Data Preprocessing: The audio recordings were preprocessed to enhance the quality and consistency of the data. This included data visualization and feature extraction. There was no need for noise reduction or speech segmentation because each voice sample was less than 5 seconds in length. The Mel-Frequency Cepstral Coefficients (MFCCs) feature extraction method was used to extract 40 features from each sample, because this technique captures spectral features. And finally, the preprocessed data was split into train and test set, allowing 80% of the data to be trained, and using the remainder for the testing.

Model Development: This project explored convolutional neural networks (CNNs), long-short term memory (LSTM), multilayer perceptron (MLP), and logistic regression. These models were each fed with the preprocessed dataset and on the bases of the same hyperparameters. These

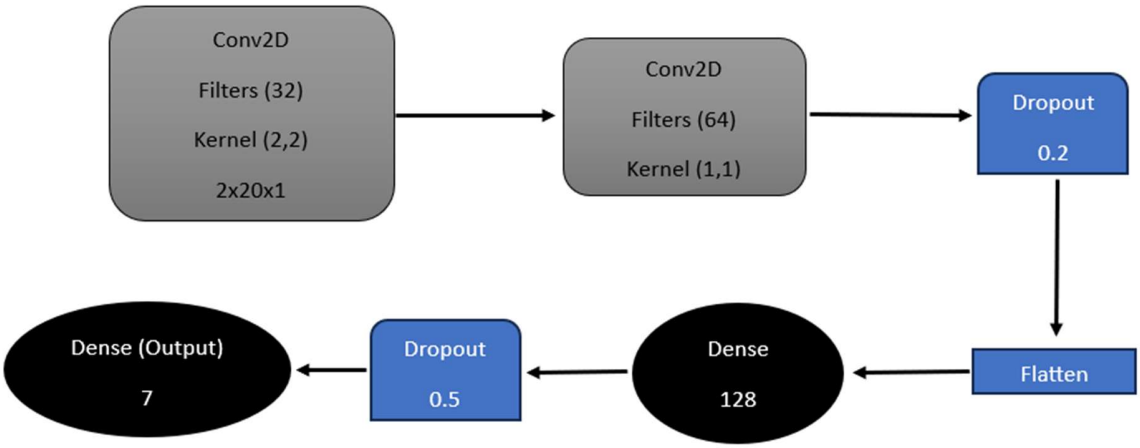
trained models were evaluated using a held-out test set to assess their accuracy and generalization ability. Various evaluation metrics, such as accuracy, precision, and recall, were used to quantify the performance of the models.

The **LSTM** architecture adopted is shown below:



The **MLP** architecture adopted is shown below:

The **CNN** architecture adopted is shown below:



Model Deployment: The best-performing AI emotion recognition model was deployed on a server, to back the user interface.

Model Findings

The validation set used in this analysis consists of 840 audio sample with 120 of each emotion label, this implies that the test set is balanced and at such even the accuracy of the analysis can be accepted as a good evaluator of the performance, resulting in the same level of performance on the precision, recall and f1 score.

Model Result Evaluation:

Model	Accuracy	Precision	Recall	F1 Score
CNN	**0.99**	**0.99**	**0.99**	**0.99**
LSTM	**0.96**	**0.96**	**0.96**	**0.96**
MLP	**0.97**	**0.97**	**0.97**	**0.97**
Logistic Regression	**0.98**	**0.98**	**0.98**	**0.98**

User Interface (UI)

At the start of the project, it was proposed and planned that the best performing machine or deep learning algorithm will be utilized in the creation of the user interface. However, while executing the UI, it was realized that that due to the pattern of encoding the labels to make it suitable and callable for some of the algorithms, the UI might end up forecasting digits instead of labels. Among all four algorithms employed, logistic regression could bypass the one-hot encoder process, and for this reason the user interface was backed by it, as opposed to the CNN which performed the best among all.

Challenges and Limitations

This project like any, had challenges encountered along the course of execution. The first major obstacle was the lack of access to kid-voiced audio dataset. Artificial intelligence is still garbage-in garbage-out, this means that it can only learn and imitate what it was taught. Though the models performed very well on the test data it was provided, it is very unlikely that it would perform so well when faced with audio recorded from kids. This lack of generalization would mean that there is more work to be done on the project. Also, the dataset used in training the models were made up of voice recordings of less than 5 seconds each. This means that they cannot be expected to forecast emotions from speeches expressed in longer recording and would need to forecast emotions every 5 seconds or less.

There is also the problem of speech-overlapping, which was not tackled in the course of this project. This overlap can happen when there are multiple people speaking at the same time. These models were not trained to recognize this issue, and that means when this happens, the model could pick up the speech of the other person(s) and predict their emotions instead.

Recommendations

Based on the challenges encountered, the models seem to have overfit on the dataset, but could not generalize on unseen data. This poor performance on new data could defeat the main purpose of the project. The project may perform better if it was trained on real audio data generated from real scenarios rather than acted voices. Furthermore, there is a need to try hybrid algorithms, which is an angle this project did not explore.

The graphical user interface that was created is functional but lacks an approach to generating these predictions in real time. This would need to be improved, and it would help if there were a member of the team with app development expertise.

References

Straussner, S.L.A., Calnan, A.J. Trauma Through the Life Cycle: A Review of Current Literature. *Clin Soc Work J* 42, 323–335 (2014). https://doi.org/10.1007/s10615-014-0496-z

Youddha Beer Singh, Shivani Goel, A systematic literature review of speech emotion recognition approaches, Neurocomputing, Volume 492, 2022, Pages 245-263, ISSN 0925-2312, https://doi.org/10.1016/j.neucom.2022.04.028. (https://www.sciencedirect.com/science/article/pii/S0925231222003964)

Moataz El Ayadi, Mohamed S. Kamel, Fakhri Karray, Survey on speech emotion recognition: Features, classification schemes, and databases, Pattern Recognition, Volume 44, Issue 3, 2011, Pages 572-587, ISSN 0031-3203, https://doi.org/10.1016/j.patcog.2010.09.020. (https://www.sciencedirect.com/science/article/pii/S0031320310004619)

Zhen-Tao Liu, Min Wu, Wei-Hua Cao, Jun-Wei Mao, Jian-Ping Xu, Guan-Zheng Tan, Speech emotion recognition based on feature selection and extreme learning machine decision tree, Neurocomputing, Volume 273, 2018, Pages 271-280, ISSN 0925-2312, https://doi.org/10.1016/j.neucom.2017.07.050. (https://www.sciencedirect.com/science/article/pii/S0925231217313565)

Dimitrios Ververidis, Constantine Kotropoulos, Emotional speech recognition: Resources, features, and methods, Speech Communication, Volume 48, Issue 9, 2006, Pages 1162-1181, ISSN 0167-6393, https://doi.org/10.1016/j.specom.2006.04.003. (https://www.sciencedirect.com/science/article/pii/S0167639306000422)