Group-26

Project Title: Speech Emotion Recognition using Deep Learning

Team members:

Full Name	**ASU ID**
Aashvik Chennupati	1225259971
Ravi Tej Chaparala	1230035172
Jaya Shankar Maddipoti	1230911684
Hemanth Pasula	1225551795
Dinesh Koushik Deshapathi	1225344913

Problem Statement:
Speech Emotion Recognition (SER) is a crucial component in human-computer interaction systems. The problem we aim to investigate is to develop a deep learning model capable of accurately recognizing emotions from speech audio recordings, including happiness, sadness, anger, and more. This project is motivated by the increasing importance of SER in applications like virtual assistants, mental health monitoring, and customer service analysis.

Related Work:

[1] Deng, J., et al. (2013). "A Survey of Speech Emotion Recognition." ACM Transactions on Multimedia Computing, Communications, and Applications. https://dl.acm.org/journal/tomm

[2] Sharma, A., et al. (2019). "Deep Learning for Emotion Recognition on Small Datasets using Transfer Learning." ACM Multimedia. https://2022.acmmm.org/

[3] Chen, J., Zhang, Y., & Zhao, D. (2023). Speech emotion recognition for scam detection using deep learning. In 2023 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP) (pp. 3845-3849). IEEE. https://ieeexplore.ieee.org/document/8805181

[4] Zhang, Z., Li, Y., & Gao, X. (2023). Scam detection using speech emotion recognition and text analysis. Neural Computing & Applications, 35(3), 2587-2600. https://link.springer.com/article/10.1007/s00521-023-08470-8

Initial Hypothesis:

Our main research question is: Can a deep learning model trained on a diverse dataset accurately recognize emotions from speech audio recordings? We hypothesize that our model will achieve state-of-the-art performance in terms of accuracy and F1 score, surpassing existing approaches.

Dataset(s):

We will split the dataset into a training set (80%), a validation set (10%), and a test set (10%). Preprocessing steps will involve extracting Mel-Frequency Cepstral Coefficients (MFCCs) from audio recordings and normalizing the data.

Dataset Source (Link and Reference)	https://www.kaggle.com/datasets/dmitrybabko/speech-emotion-recognition-en
Number of Instances	2076(Ravdess), 7442(Crema), 480(Savee), 2800(Tess)
Number of Features	Typically ranges from 300 to 400
Class Distribution (# Instances in Each Class, if Applicable)	Crema(anger-1026, disgust-981, fear-1017, happiness-1069, neutral-2349, sadness-1000), Ravdess(anger-252, calm-252, disgust-252, fear-252, happiness-252, sadness-252), Savee(anger-60, disgust-60, fear-60, happiness-60, neutral-120, sadness-60), Tess(anger-390, disgust-400, fear-400, happiness-390, neutral-600, sadness-620)
Dataset Splits	Training: (80%) Validation: (10%) Testing: (10%)
Preprocessing Steps	Normalize the audio data. This can be done by dividing the signal by its maximum amplitude. This will ensure that all of the audio clips have the same volume level. Remove silence from the beginning and end of the audio clips. This can be done by detecting the silence and trimming the clips accordingly. Resample the audio clips to a consistent sampling rate. This will make it easier to compare the features extracted from different audio clips. Extracting the audio features that we are interested in. We will use the functions to extract a variety of different audio features, such as ZCR, energy, RMS, entropy of energy, SC, SF, SR, chroma Stft,

	mel-spectrogram, and MFCC. Standardize the extracted features. This can be done by subtracting the mean and dividing by the standard deviation of each feature. This will make the features more comparable and will improve the performance of machine learning models.

Method(s):

Our project involves main objective: Speech Emotion Recognition (SER). To address SER effectively, we will employ a below approach. We are going to do Scam Detection through Voice Analysis as an extension:

Speech Emotion Recognition (SER):

Emotion Dataset:

We will curate a diverse and extensive dataset of speech audio recordings encompassing a wide range of emotions, including happiness, sadness, anger, and more. This dataset will serve as the foundation for our SER task. We found information about dataset that can be found in dataset section.

Convolutional Neural Network (CNN):

Leveraging the success of CNNs in image and audio analysis, we will design a specialized CNN architecture tailored for SER. This CNN model will be optimized for extracting relevant features from Mel-frequency cepstral coefficients (MFCCs), a well-established technique in SER.

TensorFlow and Keras Implementation:

To facilitate model development and experimentation, we will implement the SER model using TensorFlow and Keras. This combination offers a powerful and flexible deep learning framework.

Enhancing Model Performance:

Our approach aims to enhance the novelty and performance of the SER model. This involves experimenting with various CNN architectures, conducting hyperparameter tuning, and applying data augmentation techniques to better capture emotional cues in speech recordings.

Transfer Learning:

We will explore the potential of transfer learning by utilizing pretrained audio models. This strategy enables us to leverage knowledge gained from large-scale audio datasets and may lead to significant performance improvements.

Comparison with State-of-the-Art:
To validate the effectiveness and innovation of our SER model, we will conduct thorough performance comparisons with existing state-of-the-art SER models. This evaluation will demonstrate our model's ability to accurately recognize emotions from speech audio recordings.

By addressing SER, our project aims to contribute to the advancement of human-computer interaction systems, mental health monitoring, and user security in applications such as virtual assistants and customer service analysis.

Evaluation:
To quantitatively measure our model's performance, we will use the following metrics:

- Standard metrics:
    1. Accuracy
    2. F1 score
    3. Confusion matrices
- Subjective metrics:
    1. Emotion-specific accuracy and recall
    2. Weighted accuracy
    3. ROC AUC score

We will also conduct a subjective evaluation of the model's predictions by having human evaluators rate the emotional content of selected audio clips. This will provide insights into the model's real-world applicability and its ability to distinguish between different emotions accurately.

We will compare our model's performance with baseline models and state-of-the-art SER models using statistical tests to demonstrate its superiority. Additionally, we will explore domain-specific metrics, such as emotion-specific accuracy and recall, to assess the model's ability to distinguish between different emotions accurately.

Rationale:
The metrics listed above are designed to evaluate the model's performance on a variety of tasks, including emotion recognition, and generalization to real-world data. By evaluating the model's performance on these metrics, we can gain a comprehensive understanding of its strengths and weaknesses. This information can be used to improve the model's performance and to assess its suitability for real-world use in sentiment analysis applications.

Management Plan:

Role	Full Name	ASU ID
Data Architect	Aashvik Chennupati	1225259971
Software Architect	Ravi Tej Chaparala	1230035172
Experiment Architect	Jaya Shankar Maddipoti	1230911684
Project Manager	Hemanth Pasula	1225551795
Domain Expert	Dinesh Koushik Deshapathi	1225344913

Project Management Team:

Project Manager (Hemanth Pasula):
Role: Coordinate, track progress, and ensure communication.
Responsibilities: Organize team activities, monitor project timelines, maintain effective team communication, and lead report and presentation coordination.

Development and Research Team:

Software Architect (Ravi Tej Chaparala):
Role: Design and manage code, ensuring quality.
Responsibilities: Architect the software system, maintain code quality, and collaborate with team members.

Experiment Architect (Jaya Shankar Maddipoti):
Role: Design experiments and evaluate methods.
Responsibilities: Design protocols, develop metrics, and evaluate different approaches.

Data Architect (Aashvik Chennupati):
Role: Collect and prepare data.
Responsibilities: Collect, manage, and preprocess data for analysis.

Domain Expert (Dinesh Koushik Deshapathi):
Role: Provide domain-specific knowledge.
Responsibilities: Offer insights and guidance related to the project's domain.
Communication and Collaboration:

Project Reporting:
Project Manager (Hemanth Pasula) will lead the reporting process, with all team members contributing to project reports and presentations.

Extension:

Once the above project output is achieved, we are going to work on its extension which is Scam Detection through Voice Analysis. Below is the information about it.

In addition to recognizing basic emotions from speech audio recordings, the project aims to develop a deep learning model capable of identifying fraudulent callers who are attempting to scam innocent individuals. This system will classify whether a caller's voice is authentic or not, providing a valuable tool for fraud prevention and user protection.

Scam Detection through Voice Analysis:

Scam Call Dataset:
In parallel, we will assemble a dataset comprising recorded scam phone calls and authentic calls while prioritizing privacy and compliance with legal regulations.

Voice Feature Analysis:
To detect fraudulent calls, we will investigate voice features and emotional cues present in scam calls. These cues may exhibit unique patterns that distinguish them from authentic communications.

Deep Learning Model for Scam Detection:
We will design and train a deep learning model tailored for the specific task of identifying fraudulent calls through voice analysis. This model will be equipped to differentiate between authentic and scam calls, contributing to user protection and fraud prevention.

Real-time Integration:
Implementing the scam detection model in real-time voice call systems, it will analyze incoming calls and promptly alert users about the authenticity of the call, providing an additional layer of security.

To evaluate the model's performance on scam detection, we will use a dataset of labeled speech recordings from both genuine callers and scammers. The dataset will be representative of the real-world data that the model will be used on, such as financial scams or customer service scams.