**CS599 BIOMETRICS**

**A red sign with white text

Description automatically generated**

**Final Project: Detection of emotions from speech**

***Abstract***

Speech Emotion Recognition (SER) is an emerging field at the intersection of affective computing and biometric analysis, with profound implications for human-computer interaction, mental health diagnostics, and enhanced customer service experiences. This project embarked on the development of a sophisticated SER system employing Convolutional Neural Networks (CNNs) to provide accurate emotional assessments from spoken language.

The model was trained on a comprehensive assortment of emotional speech data from several well-curated databases, including CREMA-D, TESS, and RAVDESS by integrating advanced digital signal processing and deep learning techniques. These datasets encompass a wide array of emotions, articulated under various acoustic conditions, providing the diversity needed for robust model training.

The CNN architecture was meticulously designed to capture subtle acoustic features that signify emotional states, achieving an overall classification accuracy of 94%. This high accuracy rate signifies a significant breakthrough in automatic emotion recognition, showcasing the model’s ability to interpret emotional nuances from complex speech patterns effectively. The implications of such a system are vast, ranging from enhancing AI-driven communication interfaces to providing tools for psychological analysis and even improving security systems through emotional awareness.

Future directions for this research include expanding the dataset to include more nuanced emotional expressions and background noises to simulate real-world applications more closely. Additionally, experimenting with hybrid models that combine CNNs with other neural network architectures may offer improvements in processing efficiency and accuracy. The goal is to refine the system to such a degree that it can seamlessly integrate into various applications, providing intuitive and responsive user interactions that are sensitive to the user’s emotional context.

**Introduction to Detection of Emotions from Speech:**

Detection of emotions from speech, commonly referred to as Speech Emotion Recognition (SER), is an advanced computational technique that analyzes human speech to identify the speaker's emotional state. This technology leverages the nuances and subtleties in the tone, pitch, tempo, and volume of speech, which are indicative of underlying emotions. SER systems use a variety of audio processing and machine learning methods to extract acoustic features from speech, such as Mel-frequency cepstral coefficients (MFCCs), pitch, and energy levels.

**Why** **Detection of Emotions from Speech:**

SER enhances human and machine interaction, making interactions more natural and intuitive. By understanding emotional cues, systems can provide more empathetic and context-aware responses, improving user satisfaction and engagement. This is particularly valuable in virtual assistants, educational software, and interactive gaming applications.

1. **Mental Health Assessment:** In the healthcare sector, SER offers a noninvasive tool to monitor and assess emotional well-being. It can detect signs of mental health issues such as depression or anxiety based on changes in vocal attributes over time. This capability makes it a valuable component in telehealth platforms, where remote assessment needs to be both efficient and accurate.
2. **Enhanced Customer Service:** Customer service centers utilize SER to gauge customer emotions during interactions, allowing representatives to adjust their approach in real time to handle complaints or inquiries more effectively. This adaptation can lead to better issue resolution and higher customer satisfaction.
3. **Security and Surveillance:** In security contexts, SER can add a layer of safety by detecting stress or aggression in voices, potentially identifying threatening situations before they escalate. This application is helpful in areas like emergency dispatch, public transportation, and other public service areas.
4. **Market Research and Advertising:** Companies can use SER to analyze customer feedback in voice surveys and other interactive marketing tools to better understand consumer reactions and emotions toward products or services. This insight helps businesses tailor their strategies to better meet consumer needs and improve marketing effectiveness.
5. **Legal and Forensic Analysis:** SER can play a role in legal settings by analyzing individuals' emotional states during interrogations or testimonies. Such analysis might provide additional layers of insight into the credibility of the statements being made.

**Tools, Datasets and Libraries:**

1. **Tools:**

**Google Colab:** This project was developed using Google Colab, a cloud-based Python programming environment that offers free access to GPUs and a collaborative workspace. It supports a range of libraries and frameworks, facilitating efficient coding, testing, and sharing of the project.

1. **Datasets:**

**CREMA-D:** The Crowd-Sourced Emotional Multimodal Actors Dataset (CREMA-D) comprises 7,442 clips from 91 actors, each labeled with specific emotions. This dataset is instrumental in training models to recognize a broad spectrum of emotional expressions.

**Toronto Emotional Speech Set (TESS):** TESS features voice recordings from two actresses, expressing seven different emotions, consistently recorded and clearly labeled, making it valuable for refining the model’s emotion classification capabilities.

**RAVDESS:** The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS) includes recordings of 24 professional actors in both neutral and emotionally expressive tones, providing high-quality data for vocal emotion analysis.

1. **Libraries:**

* **Pandas**: Provides robust data structures for data analysis and manipulation.
* **NumPy**: Essential for large-scale mathematical operations on arrays and matrices.
* **Librosa**: Specialized for audio signal processing, particularly useful for extracting audio features.
* **Matplotlib**: Enables the creation of wide-ranging visualizations for data analysis.
* **Seaborn**: Builds on Matplotlib with advanced plot types and stylish aesthetics.
* **Scikit-learn**: Offers straightforward tools for data mining, data analysis, and machine learning.
* **TensorFlow**: A versatile framework for building and training ML models.
* **Keras**: Simplifies neural network construction and training with high-level APIs over TensorFlow.
* **IPython.display & Audio**: Enhances notebook interactivity, including audio data handling.

**Process of the Project:**

1. **Tools & Datasets**
   * Utilized Google Colab for its computational resources and ease of collaboration, with an integrated setup to directly import datasets from Kaggle. This integration facilitated seamless access to primary datasets such as CREMA-D, TESS, and RAVDESS, which offer diverse samples of emotional speech essential for the analysis and training processes.  
       
     - Integrating Kaggle API

A screenshot of a computer

Description automatically generated

* Kaggle API upload to Collab File

A screenshot of a computer

Description automatically generated

* Importing data from data sets

A screenshot of a computer

Description automatically generated

- Pre-Processing Data Frame

A screenshot of a computer

Description automatically generated

1. **Data Visualization and Exploration**
   * Visualizations such as count of emotions, Mel spectrograms, and MFCCs were used to analyze the dataset, helping to understand the balance of emotions and the characteristics of the audio features extracted.

**Count of Emotions:**

* This visualization helps in understanding the balance or imbalance among different emotional categories in the data, which is crucial for training a well-performing model.
* Each bar represents the count of audio samples labeled with emotions such as sadness, neutrality, happiness, anger, fear, surprise, and disgust.

A screenshot of a graph

Description automatically generated

* **Waveplots and Spectrograms** Waveplots were used to visually represent the amplitude variations in audio signals, highlighting changes in loudness over time. In addition, spectrograms were employed to display the changes in frequency spectrums over time, providing a comprehensive overview of the audio's frequency components.

A screenshot of a computer screen

Description automatically generated

* **MFCC**

A screen shot of a screen

Description automatically generated

1. **Data Preprocessing and Augmentation**

Comprehensive preprocessing involved normalization and noise reduction. Augmentation techniques like pitch shifting and time stretching were applied to improve the model’s robustness and ability to generalize.

* **Noise Injection**: Adding varying levels of noise to audio samples to simulate real-world disturbances.
* **Stretching**: Modifying the playback speed of audio, thus changing its duration while maintaining the original pitch.
* **Shifting**: Adjusting the starting point of audio waveforms, which alters the timing of when the sound initiates or concludes.
* **Pitch Modification**: Altering the pitch of audio signals while retaining their underlying meaning.

A screenshot of a computer

Description automatically generatedA screen shot of a computer

Description automatically generatedA screenshot of a computer

Description automatically generatedA screenshot of a computer

Description automatically generated

1. **Feature Extraction**

Feature extraction is essential for analyzing and identifying relationships among various elements in a dataset. Since audio data is not directly interpretable by models, it must be transformed into a format that models can understand through feature extraction.

Features utilized in our implementation include:

* **Zero Crossing Rate**: Measures the rate at which the audio signal changes from positive to negative or back, indicating the frequency of the sound wave.
* **Mel-frequency Cepstral Coefficients (MFCC)**: Captures the short-term power spectrum of a sound, using a linear cosine transform of a log power spectrum on a nonlinear mel scale of frequency.
* **Root Mean Square (RMS) Value**: Represents the square root of the average power of the audio signal, indicating its amplitude.

A screenshot of a computer

Description automatically generated

1. **Data & Feature Preparation**

Prepared the dataset for modeling by structuring features, encoding categorical labels, and splitting data into training and testing sets. Feature scaling was conducted to standardize input data for the neural network.A screenshot of a computer program

Description automatically generated

1. **CNN Model**

Developed a Convolutional Neural Network (CNN) specifically tailored for audio data to classify emotional states based on the extracted features. The model architecture included convolutional layers for feature detection and dense layers for classification.

* **Flatten Layer:**

Converts the three-dimensional output of the last pooling layer into a one-dimensional array to be fed into the dense layers, facilitating the transition from convolutional to fully connected layers.

* **Dense Layers (Fully Connected Layers):**

The network includes a dense layer with 512 units and ReLU activation, followed by a batch normalization step.

* **Output Layer:**

Composed of a dense layer with 7 units corresponding to the emotional categories and employs softmax activation to produce a probability distribution across these categories.

* **Model Compilation:**

Compiled with the Adam optimizer and categorical cross-entropy as the loss function, this setup is optimized for multi-class classification tasks. The primary metric for monitoring during training is accuracy.

**CNN INPUT:**



**CNN OUTPUT:**

A screenshot of a computer program

Description automatically generated

The model features a comprehensive architecture with a total of **11,844,999** parameters, divided into trainable and non-trainable categories. This design is meticulously crafted to efficiently capture and learn from audio data features, enabling precise classification of emotions.

1. **Training and Validation**
   * The model underwent rigorous training and validation, where performance metrics such as accuracy and loss were closely monitored over several epochs to ensure effective learning without overfitting.
   * The model was trained over 46 epochs, showing how the performance improved with each iteration.
   * Accuracy of our model on test data: 93.62 %

A screenshot of a graph

Description automatically generated

1. **Final Validation**
   * Conducted a final assessment to evaluate the model's accuracy and generalization on new data, which confirmed its capability to accurately classify emotions from speech.
   * Reflecting a balance between precision and recall, the F1-scores are consistently high across emotions, suggesting effective harmonic mean performance between precision and recall.

A graph with green squares

Description automatically generated

* + The model reached an impressive overall accuracy of 94%

A screenshot of a computer screen

Description automatically generated

1. **Conclusion**

* Project achieved a notable 94% accuracy in speech emotion recognition, with strength in identifying surprise, neutral, and angry emotions. This performance underscores the model’s capability in extracting and interpreting key audio features effectively.
* Moving forward, to further improve and validate the model’s robustness, Plan to explore additional datasets and experiment with alternative features. This will help enhance the model's adaptability and accuracy across various emotional expressions and real-world scenario.

**YouTube:**

15 Min Video:

<https://youtu.be/XAwQ7SPDXtY>

5 Min Video:

<https://youtu.be/eg7mo_tlFr4>

***References***

* 1. *Elsevier. (2023, August 14). Speech emotion recognition using machine learning - A systematic review. Intelligent Systems with Applications.* [*https://www.sciencedirect.com/science/article/pii/S2667305323000911*](https://www.sciencedirect.com/science/article/pii/S2667305323000911)
  2. *Runner, P. (2020, April 10). Audio signal feature extraction and clustering. Medium. https://medium.com/heuristics/audio-signal-feature-extraction-and-clustering-935319d2225*
  3. *Zhao, Y., & Shu, X. (2023, November 21). Speech emotion analysis using Convolutional Neural Network (CNN) and gamma classifier-based error correcting output codes (ECOC). Nature News.* [*https://www.nature.com/articles/s41598-023-47118-4*](https://www.nature.com/articles/s41598-023-47118-4)
  4. *Mostafaabdlhamed. (2023, July 1). Speech emotion recognition. Kaggle.* [*https://www.kaggle.com/code/mostafaabdlhamed/speech-emotion-recognition-97-25-accuracy/notebook*](https://www.kaggle.com/code/mostafaabdlhamed/speech-emotion-recognition-97-25-accuracy/notebook)
  5. *University of Toronto. (2010). Toronto emotional speech set (TESS). TSpace. https://tspace.library.utoronto.ca/handle/1807/24487*