TDL Project Problem statement

Team Number : 17

AUDIO EMOTIONS RECOGNIZER USING DEEP LEARNING

Team name and SRN:

- Anirudh Sai Lanka (PES2UG21CS073)
- Atif Shaik (PES2UG21CS105)
- B Karthik (PES2UG21CS111)
- Abhishek (PES2UG21CS020)

Outline

- Submitted Dataset details
- Problem Statement
- Possible Outcome
- Model Proposed
- Novelty
- How will the model be validated?

Submitted Dataset

- Our team has submitted a dataset containing voice notes of roughly 6-8 seconds.
- Some notes are smaller while some are larger to add a little variation to the model's training process.
- These voice notes are basically going to showcase 7 different human emotions for the model to learn from.
- Each one of us has submitted 10 voice notes each, and if necessary, we will be taking in more voice samples in the future for training and testing purposes (to achieve better accuracies)

Problem Statement

- Our aim is to be able to judge the emotion of the user using just his voice.
- We are going to make a model which will be capable of predicting the emotional state of the user using the provided voice notes.
- We will be using well known libraries like librosa, scipy, sounddevice etc and their descriptions will be mentioned in a dedicated slide.

Possible Outcomes

- **High Accuracy Model:**
  You could develop a model that accurately predicts the emotions of individuals from their voice notes, achieving high accuracy on test data.

- **Real-time Emotion Recognition:**
  If your model is efficient, it could be implemented in real-time applications such as call centers, where emotions of callers could be analyzed to provide better customer service.

Model Proposed

- Give reasons for selecting the model.

1. **Automatic Feature Extraction : A significant advantage of CNNs is their ability to automatically learn features directly from the input data (MFCCs in this case).**
2. **Capturing Spatial Relationships in Audio : CNNs excel at recognizing patterns in spatially-related data. While audio seems sequential, the extracted features (MFCCs) represent the frequency spectrum at a specific time instance**
3. **Hierarchical Feature Learning : CNNs have a layered architecture where initial layers extract simple features like edges or lines, while subsequent layers combine these to form more complex, higher-level features.**

Novelty

- We plan to integrate CNN+LSTM approach, shortcomings of one model will be handled by other.
- Very few papers have incorporated this type of Architecture.
- Our custom voices used to train a model on this kind of architecture could add a feature to novelty.

How will the model be validated ?

- Evaluation Metrics
  - F1 - Score
  - Recall
  - Precision
- Data-Driven Validation
  - Training Set (70-80%)
  - Validation Set (10-15%)
  - Testing Set (10-20%)
- Human Evaluation
  - Conduct human evaluation studies where people listen to audio samples and compare their perceived emotions with the model's predictions. This can provide valuable insights into how well the model aligns with human judgments of emotion in audio

Thank You