Speech Emotion Recognition

Introduction

Speech Emotion Recognition (SER) is a critical area of study in the field of computational paralinguistics, focusing on identifying the emotional state of a speaker from their vocal expressions. This report extends on our initial project proposal by explaining our methodologies, datasets employed and insights from our research.

The goal is to enhance the understanding and application of SER in practical scenarios, such as improving customer service interactions, aiding in mental health assessments and emergency services.

The primary challenge addressed in this project is the accurate recognition and classification of emotional states from audio samples using machine learning models.
Despite advancements in the field, SER systems often struggle with high variability in speech and the subtleties of emotional expression. This results in low accuracy in real-world applications.

We decided to combine multiple high-quality datasets in an effort to improve the generalizability, while catering our model to a wider range of accents, emotions and audio files.

Datasets

Our study utilized three main datasets:

1. RAVDESS: This dataset includes 1440 audio files from 24 actors expressing emotions like calm, happiness, sadness, anger, fear, surprise, and disgust in a controlled environment.
2. SAVEE: Comprising audio files from four male speakers, this dataset is phonetically balanced and covers emotions such as anger, disgust, fear, happiness, sadness, and surprise.
3. CREMA-D: A diverse dataset featuring 7442 clips from 91 actors, representing a range of emotions and ethnic backgrounds.

Pre-processing Techniques

To standardize the audio files, we took the steps highlighted below:

- Resampling: Ensuring a consistent sampling rate across all datasets.
- Amplitude Normalization: Balancing the volume across clips to prevent volume variations from affecting the model's performance.
- Silence Trimming: Removing non-informative parts of the audio to enhance processing efficiency.
- Duration Handling: Standardizing the length of all clips to three seconds.

Feature Extraction

Key features extracted for analysis included:

- Mel-Frequency Spectrogram: Provides a visual representation of the sound spectrum.
- MFCC (Mel-frequency cepstral coefficients): Captures the timbral aspects of the audio signal.
- Zero-Crossing rate: a measure of times in a given interval/frame that the amplitude of the speech signals passes through a value of zero

Model Implementation Overview

Three models were tested:

1. Baseline Model (Logistic Regression): Utilized MFCCs, Mel-Spectrogram and Zero-crossing rate and achieved an accuracy of 32.6%.
2. CNN Model: Employed Mel Spectrograms, and MFCCs, with a training accuracy of approximately 80% and a validation accuracy of 55%.
3. Pre-trained Transformers (Wav2Vec2): This model showed a validation accuracy of around 47%.

Baseline Model (Logistic Regression)

We started with a baseline Logistic Regression model using MFCCs, Zero Crossing Rate, and Mel-Spectrogram features. We calculated the mean and standard deviation of each sequence within every feature and used an 80/20 train-test split. The model achieved an accuracy score of 32.6%.

CNN Model

We then experimented with a Convolutional Neural Network (CNN) model using Mel Spectrograms and MFCCs as features. The model achieved a training accuracy of ~80% and a validation accuracy of ~55%, which is close to the state-of-the-art accuracy for the RAVDESS dataset.

To address overfitting, we added Dropout layers, which increased the validation accuracy from 40% to 55%. The zig-zag pattern in the accuracy plot indicated overfitting, but the model still achieved a decent result with a relatively small number of parameters.

Pre-Trained Transformers (Wav2Vec2)

Finally, we explored the use of pre-trained Transformers, specifically Wav2Vec2, using the feature extractor provided by Hugging Face. The validation accuracy was around 47%, which is lower than the state-of-the-art for the RAVDESS dataset.

The pre-trained model did not perform well due to the absence of emotion labels, different acoustic features, and variability in expression compared to the pre-training data.

Insights and Analysis

The CNN model demonstrated potential with a training accuracy of 80%; however, it suffered from overfitting as evidenced by the lower validation accuracy. The logistic regression model was less effective, particularly in distinguishing neutral and fearful emotions. The pre-trained transformer model did not perform as expected, likely due to the absence of specific emotion labels and variability in expression.

Conclusions

In this report, we have presented our approach to the Speech Emotion Recognition problem using various machine learning techniques. We have explored different datasets, pre-processing methods, feature extraction techniques, and models to classify emotions from audio data.

Our experiments have provided insights into the performance of these methods and highlighted the challenges faced in this domain. The analysis indicates that while CNNs show promise, they require adjustments to reduce overfitting. Logistic regression appears insufficient for complex emotional classifications for the combined dataset that we tested for. Wav2vec2 didn't perform well because it has been trained to learn from unlabelled data so it doesn't focus on paralinguistic features that are really important for audio emotion recognition. If we used a transformer specifically trained for this task, we would have seen better performance.

Recommendations

- Exploring data augmentation techniques to increase the diversity of the training data.
- Collecting more data to improve model generalization.
- Using models specifically trained for emotion recognition tasks.
- Deploying the models in real-world applications.

Contributions and Links

Github - https://github.com/YashvardhanRanawat7/SER_BA865

CNN Weights and Biases Report - https://api.wandb.ai/links/bostonuyash/ndpx8exq

Hugging Face Model - https://huggingface.co/yranawat/results

Jishnu	Yashvardhan
EDA, Pre-processing, Logistic Regression, CNN	EDA, Feature Extraction, CNN, Wav2vec2

Appendix

1. https://www.youtube.com/watch?v=O04v3cgHNeM
2. https://www.kaggle.com/code/shivamburnwal/speech-emotion-recognition
3. https://wandb.ai/mostafaibrahim17/ml-articles/reports/An-Introduction-to-Audio-Classification-with-Keras--Vmlldzo0MDQzNDUy#:~:text=%EF%BB%BFKeras%20is%20a%20go,both%20beginners%20and%20advanced%20users
4. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC10662716/