Speech Emotion Recognition

Máté Gedeon, Blanka Kövér

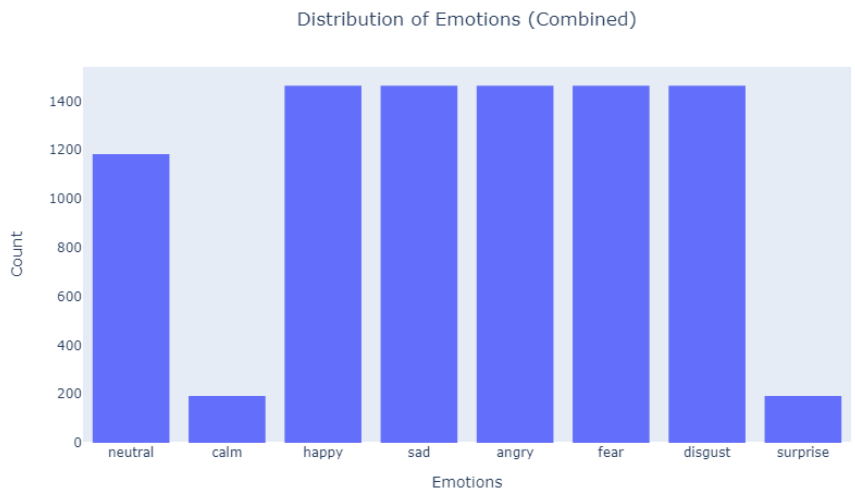
November 27, 2024

Introduction

- Speech Emotion Recognition (SER) is a fairly new field (earliest *arxiv* publication is from 2012)
- The goal is to identify and categorize human emotions through speech
- Challenges:
  - variability in speech
  - subtle differences in emotional cues
  - subjectivity

About the Dataset

- 2 Datasets Used:
  - **RAVDESS** (Ryerson Audio-Visual Database of Emotional Speech and Song): 1440 samples, 24 actors speaking 2 lexically-matched statements in a neutral North American accent across 8 emotions (neutral, calm, happy, sad, angry, fearful, disgust, surprised)
  - **CREMA-D** (Crowd-sourced Emotional Multimodal Actors Dataset): 7,442 clips,91 actors from different age groups, genders, and ethnicities, 6 emotion labels (anger, disgust, fear, happy, neutral, and sad)
- Both datasets are labeled with emotional categories
- The histogram below displays the distribution of emotion classes.



Related Work

- **Related articles**:
  - Speech emotion recognition using deep 1D & 2D CNN LSTM networks
  - Probing Speech Emotion Recognition Transformers for Linguistic Knowledge
  - AST: Audio Spectrogram Transformer
- **Kaggle Notebook**:
  - Widely used with over 6000 copies, showing community interest.
  - Includes data preparation, feature engineering (162 dimensions) and NN model
  - Good initial point, but there are notable areas for improvement:
    - Limited data preprocessing and augmentation
    - Basic NN model architecture without tuning

Novelty in Our Approach

- **Data Augmentation**: Adding diversity to the dataset for robust learning
- **Dimensionality Reduction**:
  - Applying PCA to reduce dimensions
- **Modeling approach** (3 different models):
  1. **Baseline Model** using AutoML for baseline comparison
  2. **Neural Network-Based Model**:
     - Improvements in architecture for better performance
     - Hyperparameter tuning
  3. **Transformer-Based Model**:
     - Inspired by Vision Transformers (ViTs) adapted for audio input

Results

- **PCA**:
  - 80% of the variance is kept with 32 PC
  - 90% of the variance is kept with 56 PC
- **Baseline Model**: Baseline performance established with AutoML
  - validation accuracy: 50.5%, test accuracy: 51.2%
- **Neural Network Model**: Initial testing and hyperparameter tuning
  - validation accuracy: 50.4%, test accuracy: 49.7%
  - PCA did not improve performance
- **Transformer Model**: Basic implementation completed
  - validation accuracy: 70.6%, test accuracy: 68.8%

Confusion Matrices

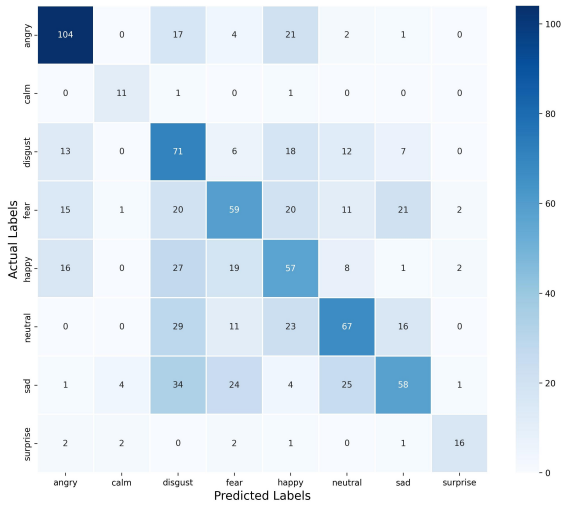


Figure: Baseline Model

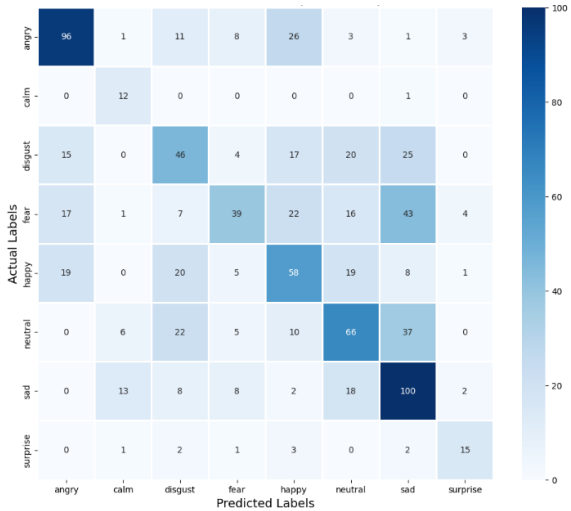


Figure: NN Model

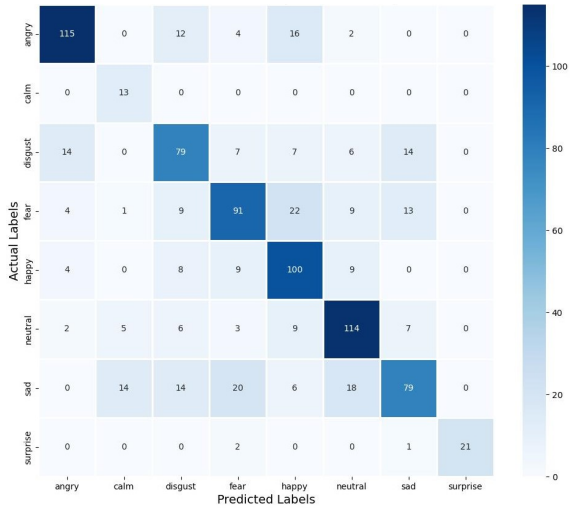


Figure: Transformer Model

Performance Metrics

Class	Precision	Recall	F1 Score
angry	0.69	0.70	0.69
calm	0.61	0.85	0.71
disgust	0.36	0.56	0.44
fear	0.47	0.40	0.43
happy	0.39	0.44	0.41
neutral	0.54	0.46	0.49
sad	0.55	0.38	0.45
surprise	0.76	0.67	0.71

Table: Baseline Model

Class	Precision	Recall	F1 Score
angry	0.83	0.77	0.80
calm	0.39	1.00	0.57
disgust	0.62	0.62	0.62
fear	0.67	0.61	0.64
happy	0.62	0.77	0.69
neutral	0.72	0.78	0.75
sad	0.69	0.52	0.60
surprise	1.00	0.88	0.93

Table: Transformer Model

Conclusion & Future Work

- **Conclusion**
  - The neural network approach is the most time consuming and worst in results
  - The Transformer model easily outperforms the other two
- **Future Work Possibilities**
  - Further optimization of Transformer model
  - Application