Speech Emotion Recognition using Acoustic Models

---

1. Introduction

Speech Emotion Recognition (SER) is the task of identifying the emotional state of a speaker from speech signals. Emotions are conveyed through acoustic cues such as pitch, energy, speaking rate, and spectral characteristics. Automatic recognition of emotions from speech has applications in areas such as human–computer interaction, healthcare, customer support systems, and affective computing.

This work focuses on building and evaluating a speech emotion recognition system using a given dataset consisting of emotional speech recordings. Both zero-shot pretrained acoustic modeling and supervised acoustic modeling approaches are explored to analyze their effectiveness on domain-specific emotional speech data.

---

2. Dataset Description

The dataset consists of audio recordings collected from multiple well-known emotional speech corpora, including RAVDESS, CREMA-D, SAVEE, and TESS. The recordings are provided in .wav format and are organized into emotion-specific directories.

The dataset contains samples belonging to seven emotion classes:

- Angry
- Happy
- Sad
- Neutral
- Fearful
- Disgusted
- Surprised

The audio files are labeled based on their parent directory names, enabling reliable extraction of emotion labels without ambiguity.

---

3. Exploratory Data Analysis (EDA)

Exploratory Data Analysis was performed to understand the distribution and characteristics of the dataset. Analysis of class distribution showed that the dataset is reasonably balanced across the seven emotion categories, making it suitable for supervised learning without the need for aggressive rebalancing techniques.

Additionally, analysis of audio durations revealed that most recordings fall within a short and consistent duration range. This consistency simplifies feature extraction and model training, as minimal padding or truncation is required. Overall, the dataset is well-structured for speech emotion recognition tasks.

---

4. Literature Review

Traditional speech emotion recognition systems rely on handcrafted acoustic features such as Mel-Frequency Cepstral Coefficients (MFCCs), pitch, and energy features combined with classical machine learning classifiers such as Support Vector Machines (SVMs) and Gaussian Mixture Models (GMMs). These approaches are effective when features are carefully designed and the dataset size is moderate.

Recent advances in deep learning and self-supervised learning have led to the development of pretrained acoustic models such as Wav2Vec2 and HuBERT. These models learn rich speech representations from large-scale unlabeled audio data and can be applied to downstream tasks either in a zero-shot manner or through fine-tuning. Such models have demonstrated strong generalization capabilities across various speech-related tasks, including emotion recognition.

---

5. Approaches Studied

Two acoustic-model-based approaches were studied in this work:

5.1 Zero-Shot Pretrained Acoustic Model

A HuBERT-based pretrained speech emotion recognition model was evaluated in a zero-shot setting, meaning no training or fine-tuning was performed on the target dataset. This approach serves as a baseline to assess how well pretrained speech representations generalize to a new emotional speech domain.

5.2 Supervised Acoustic Model (MFCC + SVM)

In the supervised approach, Mel-Frequency Cepstral Coefficients (MFCCs) were extracted from the audio recordings and used as input features for a Support Vector Machine (SVM) classifier. The model was trained using labeled data from the given dataset, allowing it to learn domain-specific acoustic patterns associated with different emotions.

---

6. Hypothesis

It was hypothesized that supervised training of an acoustic model on the domain-specific speech emotion dataset would outperform a zero-shot pretrained model. Additionally, it was expected that a lightweight supervised model using handcrafted acoustic features could achieve competitive performance despite having lower architectural complexity.

---

7. Experimental Setup

The dataset was split into training, development, and test sets using stratified sampling to preserve emotion class distributions across splits. The development set was used for model selection, while the test set was reserved for final evaluation.

Evaluation metrics included:

- Accuracy

- Precision
- Recall
- F1-score

These metrics provide a comprehensive assessment of model performance, particularly in multi-class emotion classification tasks.

---

8. Findings and Results

8.1 Zero-Shot Evaluation Results

The zero-shot HuBERT-based model demonstrated limited performance on the target dataset. While the model is capable of capturing general emotional characteristics from speech, its performance is constrained by domain mismatch between the pretraining data and the target dataset. Additionally, differences in label space required explicit label alignment for meaningful evaluation.

Despite these limitations, the zero-shot model serves as a useful baseline, highlighting the performance achievable without any task-specific training.

8.2 Supervised MFCC + SVM Results

The supervised MFCC-based SVM model significantly outperformed the zero-shot baseline across all evaluation metrics. By leveraging labeled data from the target dataset, the model was able to learn emotion-discriminative acoustic features more effectively.

Confusion matrix analysis showed that the model performs well on high-arousal emotions such as Angry and Disgusted, while some confusion remains between acoustically similar emotions such as Neutral and Sad. These results indicate that supervised learning is more effective for capturing dataset-specific emotional characteristics.

8.3 Comparative Analysis

A direct comparison between the two approaches reveals a clear performance difference. The zero-shot pretrained model offers ease of deployment and does not require training, but suffers from limited adaptability to the target domain. In contrast, the MFCC + SVM model achieves higher accuracy and better overall performance with significantly lower computational complexity.

---

9. Advantages and Drawbacks

9.1 Zero-Shot Pretrained Acoustic Model (HuBERT)

Advantages:

- No training required on the target dataset
- Leverages rich representations learned from large-scale unlabeled speech data
- Useful as a quick baseline model

Drawbacks:

- Performance limited by domain mismatch
- Requires label-space alignment for correct evaluation
- High computational and memory requirements
- Limited adaptability to dataset-specific acoustic patterns

9.2 MFCC + SVM (Supervised Acoustic Model)

Advantages:

- Computationally efficient and lightweight
- Performs well with limited labeled data
- Easy to train, interpret, and debug
- Effectively captures emotion-related spectral features

Drawbacks:

- Relies on handcrafted features
- Does not explicitly model temporal dependencies in speech
- Scalability is limited for very large datasets

---

10. Conclusion

This work demonstrates that supervised acoustic modeling significantly outperforms zero-shot pretrained approaches for speech emotion recognition on the given dataset. While zero-shot models provide a convenient baseline, their performance is limited by domain mismatch and lack of task-specific adaptation. The MFCC + SVM approach offers an effective balance between performance and computational efficiency, making it well-suited for domain-specific emotion recognition tasks.