Emotion Classification Using Speech Signals

Joel Raj K

November 1, 2024

1 Introduction

Emotion classification from speech signals is a critical task with applications in various domains, including human-computer interaction, therapeutic diagnostics, customer service automation, and sentiment analysis. This project aims to classify speech audio files into seven emotional categories: Angry, Happy, Sad, Neutral, Fearful, Disgusted, and Surprised. The dataset consists of 12,798 audio samples, which were split into training, validation, and test sets for model training and evaluation. Given the complexity of emotional classification, multiple models and feature extraction approaches were tested to find an optimal balance between generalization and model complexity.

2 Approach and Pipeline

The development pipeline involved:

1. Initial experiments with large pre-trained models, focusing on the Hubert Large model for extraction and classification. This approach, while robust, resulted in overfitting due to the limited size of our dataset.
2. Adjusting to base versions of pre-trained models, specifically Wav2Vec2, which improved generalization and validation accuracy.
3. Introducing a speech processing technique using spectrogram images, which were subsequently processed with a ResNet18 model to leverage convolutional operations for feature extraction.
4. Planning future work that includes expanding dataset size, implementing data augmentation techniques, and incorporating regularization strategies to further enhance model robustness.

3 Model Architectures and Results

The following sections outline each model's architecture, results, and observations based on accuracy, precision, recall, and f1-score.

3.1 Model 1: Hubert Large Model

The first model employed the Hubert Large pre-trained model to directly extract and classify features. Fine-tuning was applied to this large model to adapt it to our dataset.

- **Architecture:** The model utilizes a transformer-based architecture specifically designed for robust audio feature extraction. The large number of parameters, however, led to overfitting when trained on our 12,000-sample dataset.
- **Results:** Despite its theoretical robustness, the model yielded a test accuracy of 63.07%, with relatively low precision and recall for several emotion categories.

3.2 Model 2: Wav2Vec2 Model for Sequence Classification

Transitioning from Hubert Large, a Wav2Vec2 model was employed. As a smaller model, Wav2Vec2 is particularly effective for feature extraction while being less prone to overfitting.

- **Architecture:** Wav2Vec2's transformer layers were used for speech feature extraction. A linear classifier was added on top, outputting a probability distribution over emotion classes.
- **Results:** This model achieved a test accuracy of 69.63%, marking an improvement in generalization. Precision, recall, and f1-score metrics were also notably better than the first model.

3.3 Model 3: WavLM Model with Enhanced Classifier

In this model, WavLM was used as a feature extractor, followed by a custom neural network classifier comprising fully connected layers with batch normalization and dropout for regularization.

- **Architecture:** The WavLM model outputs feature embeddings, which are then processed by a deep classifier network with multiple dense layers, batch normalization, and dropout layers to improve generalization and reduce overfitting.
- **Results:** This approach yielded the highest accuracy of 80.68%. It showed strong precision, recall, and f1-scores across most emotion classes, indicating effective generalization and robust classification performance.

3.4 Model 4: ResNet18 on Mel Spectrogram Images

Here, audio files were converted into mel spectrogram images, which were then used as inputs to a modified ResNet18 model. This approach leverages the convolutional capabilities of ResNet to process spectrogram data.

- **Architecture:** A ResNet18 model was utilized, with its initial layers frozen to leverage pre-trained weights. The final layer was modified to output seven classes corresponding to our emotion categories.
- **Results:** This model achieved a test accuracy of 72.76%. While it performed moderately well, it was limited in generalization compared to the WavLM-based model.
- **Scope:** This model has a good scope of improvement in performance if we have a larger dataset.

4 Performance Comparison of Models

Table **??** presents a detailed comparison of all four models based on the metrics of precision, recall, f1-score, and support for each emotion category.

Table 1: Comparison of Model Performance Metrics for Emotion Classification

Model	**Emotion**	**Precision**	**Recall**	**F1-Score**	**Support**
Model 1 (Hubert Large)	Angry	0.81	0.72	0.76	325
	Happy	0.57	0.57	0.57	280
	Sad	0.66	0.48	0.55	307
	Neutral	0.57	0.64	0.61	325
	Fearful	0.61	0.79	0.68	269
	Disgusted	0.55	0.61	0.58	325
	Surprised	0.90	0.62	0.73	89
Model 2 (Wav2Vec2)	Angry	0.86	0.82	0.84	325
	Happy	0.64	0.72	0.68	280
	Sad	0.60	0.62	0.61	307
	Neutral	0.74	0.53	0.62	325
	Fearful	0.64	0.86	0.73	269
	Disgusted	0.68	0.62	0.65	325
	Surprised	0.83	0.82	0.82	89
Model 3 (WavLM with Classifier)	Angry	0.89	0.90	0.90	329
	Happy	0.76	0.76	0.76	258
	Sad	0.78	0.68	0.73	315
	Neutral	0.83	0.80	0.82	348
	Fearful	0.83	0.88	0.85	244
	Disgusted	0.73	0.79	0.76	344
	Surprised	0.88	0.93	0.90	82
Model 4 (ResNet18 on Spectrogram)	Angry	0.83	0.84	0.84	329
	Happy	0.68	0.63	0.66	258
	Sad	0.63	0.62	0.63	315

Model	Emotion	Precision	Recall	F1-Score	Support
	Neutral	0.79	0.71	0.75	348
	Fearful	0.75	0.75	0.75	244
	Disgusted	0.64	0.73	0.68	344
	Surprised	0.91	0.98	0.94	82

5 Discussion and Conclusion

The results show that Model 3, utilizing WavLM for feature extraction with a custom classifier, achieved the highest accuracy (80.68%) and performed well across most emotion categories. The Wav2Vec2 model (Model 2) also yielded reasonable accuracy and balanced metrics, while Model 4 (ResNet18 on Spectrogram) performed moderately well but demonstrated limitations in generalization.

These findings underscore the importance of aligning model complexity with dataset size to avoid overfitting. Utilizing WavLM embeddings with a well-regularized classifier appears to be the most promising approach for this dataset.

6 Future Work

Future efforts should focus on:

- **Expanding the Dataset:** A larger dataset could improve model generalization.
- **Data Augmentation:** Implementing audio augmentations (e.g., pitch shift, noise addition) to diversify the data.
- **Regularization Techniques:** Adding techniques such as early stopping, dropout tuning, and batch normalization to further reduce overfitting.

A Evaluation Metrics Formulas

The following metrics were used to evaluate model performance:

A.1 Precision

$$\text{Precision} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Positives}} \tag{1}$$

A.2 Recall

$$\text{Recall} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Negatives}} \tag{2}$$

A.3 F1-Score

$$\text{F1-Score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{3}$$

A.4 Support

Support refers to the number of actual occurrences of each class in the dataset.

References

[1] Microsoft. WavLM: A Unified Pre-trained Model for Speech and Audio Tasks. 2021.

[2] Facebook AI. Wav2Vec2: A Framework for Self-Supervised Learning of Speech Representations. 2020.

[3] Facebook AI. HuBERT: Self-Supervised Speech Representation Learning by Masked Prediction of Hidden Units. 2021.