Cross-Lingual Self-Supervised Learning is All You Need for Speech Emotion Recognition

Xinyuan Tong

Abstract

This work explores Speech Emotion Recognition (SER) using traditional machine learning approaches and advanced self-supervised speech models. Starting from Logistic Regression and Random Forest classifiers trained on hand-engineered acoustic features, we establish initial baselines. We then incorporate self-supervised pre-trained models, including wav2vec 2.0 and a cross-lingually pre-trained model. By fine-tuning these models on a combined dataset (RAVDESS, CREMA-D, TESS, and SAVEE), we achieve substantial performance gains. Our best model attains a test accuracy of 85.38% and an F1 score of 85.35%. This improvement highlights the effectiveness of cross-lingual pre-training in capturing universal speech representations, enhancing robustness and accuracy under diverse conditions. The results underscore the value of self-supervised, multilingual pre-training for advancing the state-of-the-art in SER.

1. Introduction

Human speech is a rich and multifaceted channel of communication, conveying not only linguistic information but also the speaker's emotional state through subtle variations in prosody, tone, and pitch. The automatic recognition of these emotional cues, commonly referred to as Speech Emotion Recognition (SER), has wide-ranging applications, including improving human-computer interaction, enhancing mental health monitoring tools, supporting driver safety systems, and delivering more empathetic customer service experiences. Despite its potential, robust SER remains challenging due to variability in language, accents, recording conditions, and the inherent complexity of emotional expression.

Recent advances in self-supervised learning and large-scale pre-training have significantly improved the performance of SER systems. Models that learn from unlabeled speech data, such as **wav2vec 2.0** (Baevski et al.), have demonstrated strong capabilities in extracting meaningful representations directly from raw audio. Further, cross-lingual pre-trained models like **wav2vec2-large-xlsr-53**(Conneau et al.) leverage multilingual training data to learn even more universal and robust speech representations, making them well-suited to real-world conditions where emotional cues must be reliably detected across diverse speakers and contexts.

In this project, we explore a progression of approaches: from traditional machine learning methods using hand-engineered acoustic features to fine-tuned self-supervised pre-trained speech models. After establishing baselines using Logistic Regression and Random Forest classifiers, we fine-tuned wav2vec 2.0 and subsequently employed the **wav2vec2-large-xlsr-53** model. By leveraging this cross-lingual pre-training, we achieved state-of-the-art performance on combined emotional speech datasets (RAVDESS, CREMA-D, TESS, and SAVEE), culminating in a final test accuracy of approximately 85.38% and a test F1 score of about 85.35%. These results underscore the value of self-supervised, cross-lingual representations in advancing SER, enhancing both accuracy and robustness.

2. Related Work

Early SER approaches relied heavily on hand-engineered acoustic features, such as MFCCs, pitch, and formant frequencies, and traditional classifiers like Support Vector Machines, Gaussian Mixture Models, and Hidden Markov Models (Akçay & Oğuz, 2020). While these methods achieved moderate success, they often struggled to generalize across languages, noisy environments, and diverse speaker traits, limiting their applicability in real-world scenarios.

The emergence of deep learning techniques and larger, more diverse emotional speech corpora prompted researchers to adopt end-to-end models. Luna-Jiménez et al. (Luna-Jiménez et al.) proposed a multimodal emotion recognition system that leveraged pre-trained Wav2Vec 2.0 representations, showing that fine-tuning fully on target emotional datasets improved accuracy. Sadok et al. (Sadok et al.) introduced VQ-MAE-S, a self-supervised model operating in a discrete latent space, achieving state-of-the-art results on popular datasets like RAVDESS and EmoDB. These studies highlight that pre-trained, self-supervised speech models can capture rich emotional representations, outperforming traditional, hand-engineered baselines.

More recent efforts have examined multilingual and cross-lingual training strategies. For instance, Conneau et al. (Conneau et al.) introduced XLSR models trained across multiple languages, facilitating robust and language-agnostic speech representations. Incorporating these representations into SER tasks can help overcome limitations arising from linguistic variation and limited training data. This cross-lingual perspective is particularly relevant when aiming for high performance across diverse speakers and recording conditions, as it improves model robustness and generalization.

Our approach builds directly on these insights. By fine-tuning the **wav2vec2-large-xlsr-53-english**(Grosman, 2021) model, a cross-lingually pre-trained variant of wav2vec 2.0, we leverage its capability to learn universal acoustic features from a wide variety of speech sources. Unlike earlier works that focused primarily on monolingual pre-trained models, we show that cross-lingual training data enhances the model's adaptability, capturing subtle emotional nuances even in challenging conditions. Our results surpass those achieved by both traditional methods and earlier self-supervised models, confirming the value of cross-lingual pre-training and fine-tuning for improving SER accuracy and robustness.

3. Datasets and Evaluation

We utilized four datasets for this project: RAVDESS, CREMA-D, TESS, and SAVEE.

3.1. Datasets

3.1.1. Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS)

(Livingstone & Russo) contains 7,356 audio-visual files from 24 professional actors (12 female, 12 male). The actors vocalize two lexically matched statements in a neutral North American accent. The dataset includes expressions of calm, happy, sad, angry, fearful, surprise, and disgust emotions in speech, and calm, happy, sad, angry, and fearful emotions in song.

3.1.2. Crowd-Sourced Emotional Multimodal Actors Dataset (CREMA-D)

(Luna-Jiménez et al.) comprises 7,442 clips from 91 actors (48 male, 43 female) aged between 20 and 74, representing diverse races and ethnicities. Actors spoke from a selection of 12 sentences presented using one of six different emotions: anger, disgust, fear, happy, neutral, and sad.

3.1.3. Toronto Emotional Speech Set (TESS)

(Pichora-Fuller & Dupuis) includes 2,800 recordings from two female actors (aged 26 and 64 years) who spoke 200 target words in the carrier phrase "Say the word ______" using seven emotions: anger, disgust, fear, happiness, pleasant surprise, sadness, and neutral.

3.1.4. Surrey Audio-Visual Expressed Emotion (SAVEE)

(Jackson & ul haq) consists of recordings from four male actors in seven emotional states: anger, disgust, fear, happiness, sadness, surprise, and neutral. The dataset contains 480 British English utterances.

These datasets provide a diverse range of emotional speech samples, facilitating the development and evaluation of robust SER models.

By combining these datasets, we obtained a more comprehensive dataset with the following distribution of emotional classes:

- Angry: 1,863 samples
- Calm: 192 samples
- Disgust: 1,863 samples
- Fearful: 1,863 samples
- Happy: 1,863 samples
- Neutral: 1,583 samples
- Sad: 1,923 samples
- Surprised: 652 samples

3.2. Evaluation Metrics

For evaluation, we used Accuracy and F1 Score as metrics. Accuracy measures the proportion of correctly classified instances among the total instances:

$$\text{Accuracy} = \frac{\text{Number of Correct Predictions}}{\text{Total Number of Predictions}} \tag{1}$$

The F1 score is the harmonic mean of precision and recall, providing a balance between the two:

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

Where Precision is the proportion of true positives among all positive predictions, and Recall is the proportion of true positives among all actual positive cases. These metrics help evaluate not only the overall performance but also how well the model distinguishes between different emotion classes.

4. Methods

We began by establishing baseline using basic machine learning methods, specifically Logistic Regression and Random Forest classifiers. These models were trained on traditional acoustic features extracted from the speech signals, such as Mel-Frequency Cepstral Coefficients (MFCCs), chroma features, and spectral contrast.

4.1. Data Preprocessing and Feature Extraction

For data preprocessing, we resampled all audio files to a target sampling rate of 16 kHz and normalized the audio signals. To enhance the model's robustness to real-world conditions, we employed data augmentation techniques, including adding noise, time stretching, pitch shifting, and reverberation.

The feature extraction process involved:

- Extracting MFCCs (40 coefficients).
- Computing chroma features.
- Calculating spectral contrast.

These features were concatenated to form the final feature vector for each audio sample.

By using the t-SNE method, we reduced the dimensionality of the extracted features to two dimensions, allowing us to visualize the distribution of emotional classes in a 2D space. The t-SNE visualization, shown in Figure 1, provides insight into how well-separated the different emotions are in the feature space. This visualization helps us understand the effectiveness of the feature extraction process and identify potential areas where emotions may overlap or cluster closely, indicating potential classification challenges.

4.2. Baseline Models

- **Logistic Regression Model**: A linear classifier that models the probability of a sample belonging to a particular class. We used L2 regularization and optimized the model using stochastic gradient descent.
- **Random Forest Classifier**: An ensemble of decision trees that can capture nonlinear relationships and interactions between features. We set the number of estimators to 100 and used Gini impurity as the splitting criterion.

We implemented the models using **scikit-learn** and trained them on the combined dataset consisting of the four datasets mentioned earlier.

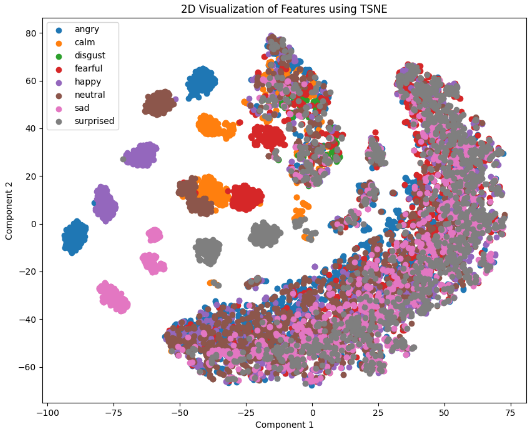

Figure 1. Feature visualization using t-SNE

4.3. Advanced Method with Pre-trained Model

Building upon the baseline models, we developed an advanced method using the wav2vec 2.0 Base model (Baevski et al.) as the base. We added a linear layer for classification and fine-tuned the entire model on the combined dataset.

The wav2vec 2.0 model leverages self-supervised learning to learn representations directly from raw audio data. Fine-tuning involves adjusting the pre-trained model weights to better suit the SER task.

4.4. Advanced Method with Cross-Lingual Pre-trained Model

Building upon the baseline models and the initial wav2vec 2.0 Base model, we further developed an advanced method using the **wav2vec2-large-xlsr-53-english** model (Grosman, 2021). This model leverages unsupervised cross-lingual representation learning, enabling it to capture more complex and nuanced features from raw audio data.

The **wav2vec2-large-xlsr-53-english** model is based on Facebook's XLSR-53 (Conneau et al.), which extends the wav2vec 2.0 architecture to a multilingual setting. XLSR-53 is pre-trained on raw audio data from 53 different languages using self-supervised learning techniques. This cross-lingual pre-training allows the model to learn universal speech representations that are not limited to a single language, enhancing its ability to generalize across diverse speech patterns and acoustic variations.

Key advantages of using a cross-lingual model include:

- **Richer Feature Representations**: Exposure to multiple languages enables the model to capture a wider variety of phonetic and prosodic features, which can be beneficial for distinguishing subtle emotional cues in speech.

- **Enhanced Robustness**: The diverse training data helps the model become more resilient to noise, accents, and other variations, leading to improved performance in real-world scenarios.

- **Transfer Learning Efficiency**: Leveraging knowledge from multiple languages can enhance the model's ability to adapt to specific tasks with limited labeled data, such as Speech Emotion Recognition (SER).

By fine-tuning the XLSR-53 model on our combined emotional speech dataset, we aim to exploit these advantages to achieve superior SER performance compared to models pre-trained on a single language.

5. Experiments

5.1. Baseline Models

We evaluated the baseline models with and without data augmentation.

5.1.1. Without Data Augmentation

The results without data augmentation were:

- Logistic Regression: Accuracy = 53.64%, F1 Score = 53.21%

- Random Forest: Accuracy = 60.42%, F1 Score = 59.79%

5.1.2. With Data Augmentation

After applying data augmentation techniques, the results improved:

- Logistic Regression: Accuracy = 67.95%, F1 Score = 67.88%

- Random Forest: Accuracy = 68.68%, F1 Score = 68.44%

5.2. Fine-tuning wav2vec 2.0 and Hyperparameter Optimization

We fine-tuned the wav2vec 2.0 Base model to develop a robust Speech Emotion Recognition (SER) system. Initially, the model was trained using the following hyperparameters:

- Batch size: 64

- Learning rate: 1×10^{-4}

- Epochs: 10

- Gradient accumulation steps: 4

The validation results across epochs for the initial setup are shown in Table 1.

Table 1. Validation Results for Initial wav2vec 2.0 Fine-Tuning

Epoch	Validation Loss	Validation Accuracy
1	1.3481	0.5670
2	1.0899	0.7593
3	0.7116	0.7554
4	0.5441	0.7867
5	0.4431	0.7877
6	0.3081	0.8062
7	0.2850	0.8112
8	0.2119	0.8181
9	0.1710	0.8211
10	0.1444	0.7862

Using the best model from epoch 9, we evaluated it on the test set and achieved a Test Accuracy of 80.80% and a Test F1 Score of 80.92%. While these results were promising, the training process suggested potential areas for improvement, particularly in reducing validation loss fluctuations and further improving accuracy.

To optimize performance, we experimented with the following updated hyperparameters:

- Batch size: 32 (reduced to increase gradient updates per epoch and improve convergence)

- Learning rate: 1×10^{-4} (kept consistent to avoid overfitting or instability)

- Epochs: 20 (increased to allow the model to fully utilize the training data)

- Gradient accumulation steps: 4

The updated validation results are presented in Table 2.

The model from epoch 19 yielded the best performance on the validation set. When evaluated on the test set, it achieved:

- Test Accuracy: 81.54%

- Test F1 Score: 81.46%

5.3. Advanced Model with Cross-Lingual Pre-trained Model

To further enhance the performance of our Speech Emotion Recognition system, we employed the **wav2vec2-large-xlsr-53-english** model. This model benefits from cross-lingual

Table 2. Validation Results for Optimized wav2vec 2.0 Fine-Tuning

Epoch	Validation Loss	Validation Accuracy
1	1.0209	0.6567
2	0.8070	0.7205
3	0.6323	0.7867
4	0.6393	0.7912
5	0.7181	0.7723
6	0.6136	0.8216
7	0.6619	0.8067
8	0.6066	0.8186
9	0.6812	0.8142
10	0.8501	0.7877
11	0.6991	0.8226
12	0.6654	0.8201
13	0.8650	0.7828
14	0.7575	0.8211
15	0.7452	0.8236
16	0.7283	0.8196
17	0.7090	0.8191
18	0.8337	0.8201
19	0.7212	0.8336
20	0.7962	0.8226

pre-training, which provides a more comprehensive feature extraction capability compared to the wav2vec 2.0 Base model.

We first fine-tuned the XLSR model with a batch size of 32 across multiple epochs. We recorded validation loss and accuracy at the end of each epoch. Table 3 shows the validation performance. While the validation accuracy increased significantly from the baseline, we observed some fluctuations in later epochs.

Table 3. Validation Results (Batch Size = 32) for XLSR Model

Epoch	Validation Loss	Validation Accuracy
1	0.9569	0.6428
2	0.5766	0.7932
3	0.5552	0.8221
4	0.4952	0.8371
5	0.4763	0.8545
6	0.6375	0.8082
7	0.5376	0.8445
8	0.5451	0.8485
9	0.5661	0.8426

We chose the checkpoint at Epoch 8 (Validation Accuracy = 0.8485) as a promising starting point. To further refine performance, we adjusted the batch size to 64 and resumed training for several more epochs. The results of this tuning phase are shown in Table 4.

Table 4. Validation Results (Batch Size = 64) Fine-Tuning from Epoch 8 Checkpoint

Epoch (continued)	Validation Loss	Validation Accuracy
1	0.6309	0.8401
2	0.5959	0.8500
3	0.6010	0.8590
4	0.6076	0.8490
5	0.6448	0.8545

From this second stage of fine-tuning, we selected the model at Epoch 3 (overall representing the third epoch after switching to batch size 64), which achieved a Validation Accuracy of 0.8590, the highest among these trials.

5.4. Final Evaluation on Test Set

We evaluated the final chosen model on the held-out test set. The final results are shown below:

- Test Accuracy: 85.38%
- Test F1 Score: 85.35%

Table 5. Classification Report on Test Set (Final XLSR Model)

Class	Precision	Recall	F1-score	Support
angry	0.92	0.89	0.91	280
disgust	0.75	0.96	0.84	28
fearful	0.83	0.88	0.86	279
happy	0.81	0.77	0.79	280
neutral	0.90	0.82	0.86	280
sad	0.86	0.91	0.89	237
surprised	0.79	0.80	0.80	289
calm	0.94	0.99	0.97	98
accuracy	0.8538 (1771 samples)			
macro avg	0.85	0.88	0.86	1771
weighted avg	0.86	0.85	0.85	1771

Table 5 shows the classification report on the test set, and Figure 2 presents the confusion matrix.

5.5. Models Performance Comparison

Table 6 summarizes the test accuracy and F1 scores achieved by the various models evaluated in this study. The results highlight a clear progression in performance as we move from baseline methods employing hand-engineered acoustic features to advanced models leveraging self-supervised and cross-lingual representation learning.

Table 6. Performance Comparison of Different Models on Test Set

Model	**Test Accuracy (%)**	**Test F1 Score (%)**
Logistic Regression (No Aug.)	53.64	53.21
Logistic Regression (Aug.)	67.95	67.88
Random Forest (No Aug.)	60.42	59.79
Random Forest (Aug.)	68.68	68.44
wav2vec 2.0 Base	81.54	81.46
wav2vec2-large-xlsr-53-english	**85.38**	**85.35**

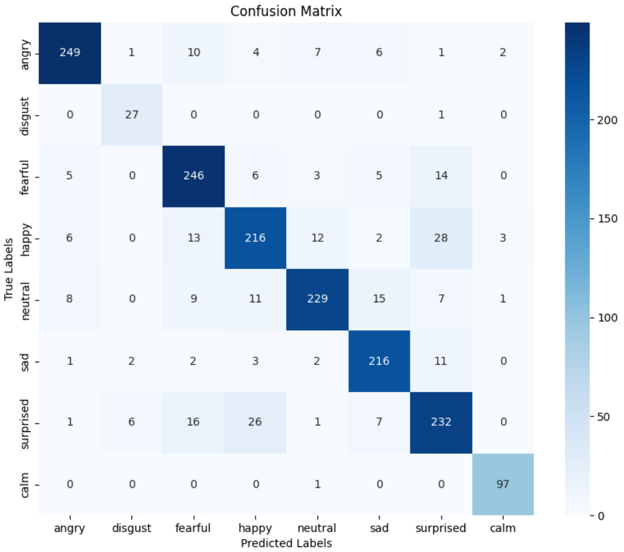

Figure 2. Confusion Matrix of the Final XLSR Model on the Test Set

The baseline models (**Logistic Regression** and **Random Forest**), trained on hand-engineered acoustic features (MFCCs, chroma, and spectral contrast), achieved moderate results. Data augmentation improved their accuracy and F1 scores, illustrating their sensitivity to training data diversity. Among the baselines, Random Forest slightly outperformed Logistic Regression, likely due to its ability to model complex, nonlinear relationships.

The introduction of the **wav2vec 2.0 Base** model marked a significant improvement over the baselines. By leveraging self-supervised learning directly on raw audio data, wav2vec 2.0 captured richer and more robust speech representations. This end-to-end approach allowed the model to learn nuanced features that were not fully captured by hand-engineered methods. Moreover, its pre-training on large amounts of unlabeled speech enabled the model to become more robust to noise and variability in the audio signals, resulting in substantially higher accuracy and F1 scores.

Building on this success, the **wav2vec2-large-xlsr-53-english** model further pushed performance boundaries. Its cross-lingual pre-training with data from 53 languages enabled it to learn universal acoustic representations, enhancing its ability to distinguish between subtle emotional cues. This comprehensive exposure to diverse speech patterns and phonetic variations improved robustness and generalization, leading to the highest accuracy and F1 scores in our experiments.

In summary, the progression from baseline models to wav2vec 2.0 and then to the cross-lingual **wav2vec2-large-xlsr-53-english** model demonstrates the value of self-supervised and cross-lingual representation learning for Speech Emotion Recognition. The advanced models effectively overcame limitations of the baselines, such as susceptibility to noise and reliance on less expressive hand-engineered features, thus achieving state-of-the-art performance in accurately classifying emotions from speech.

6. Discussion

6.1. Error Analysis

To gain deeper insights into the model's performance, we conducted a manual error analysis on the misclassified samples. A total of 259 samples were incorrectly classified. The analysis revealed the following patterns:

1. **Emotion Overlap**: Emotions such as ”Happy” and ”Surprised” often get confused due to their similar prosodic features. For instance, high energy and elevated pitch levels characteristic of both emotions can lead to misclassification.

2. **Limited Sample Size**: The ”Disgust” category, with only 28 samples, posed a significant challenge. The limited data availability likely hindered the model's ability to generalize, resulting in a lower precision of 0.75 but a high recall of 0.96.

3. **Subtle Emotional Cues**: Emotions with more subtle or nuanced vocal expressions, such as ”Calm,” were handled exceptionally well, whereas ”Fearful” and ”Sad”

showed high recall but moderate precision, indicating occasional confusion with other emotions.

6.2. Unexpected Findings

One surprising observation was the model's exceptional performance on the ”Calm” and ”Angry” emotions, achieving precision and recall rates above 0.90. This suggests that certain emotions with distinct vocal signatures are more readily captured by the model. Conversely, the high recall but lower precision for ”Disgust” indicates that while the model is adept at identifying true ”Disgust” instances, it occasionally over-predicts this emotion, likely due to its limited representation in the training data.

Additionally, the model maintained robust performance despite the presence of background noise and speaker variability, underscoring the benefits of the cross-lingual pretraining in enhancing feature robustness.

6.3. Future Improvements

Based on the error analysis, future work could focus on:

- **Enhancing Data Diversity:** Increasing the number of samples for underrepresented emotions like *Disgust* to improve the model's precision.
- **Noise Robustness:** Implementing advanced noise reduction and augmentation techniques to mitigate the impact of background noise and speaker variability.
- **Fine-Grained Feature Analysis:** Exploring additional acoustic features or employing attention mechanisms to better distinguish between similar emotions.
- **Multi-modal Integration:** Incorporating visual cues or textual data to complement the audio-based emotion recognition, potentially improving overall accuracy and robustness.

7. Conclusion

In this project, we investigated a spectrum of approaches for Speech Emotion Recognition, from conventional machine learning models relying on engineered acoustic features to advanced self-supervised, cross-lingually pre-trained speech encoders. Through iterative experimentation, we found that baseline methods, Logistic Regression and Random Forest, offered moderate performance that improved somewhat with data augmentation but remained constrained by their reliance on limited, hand-crafted features.

Adopting wav2vec 2.0 and, ultimately, the **wav2vec2-large-xlsr-53-english** model dramatically improved SER outcomes. By harnessing the power of large-scale, cross-lingual pre-training, we achieved a test accuracy of 85.38% and an F1 score of 85.35%. This represents a substantial gain over our baselines and highlights the importance of robust feature representations learned from diverse, unlabeled speech data. The cross-lingual model's ability to capture universal speech patterns proved particularly valuable, enabling more effective emotion classification across different speakers, emotional states, and environmental conditions.

Through this project, we learned several key lessons. First, the quality and diversity of learned representations matter greatly. Self-supervised models trained on large, multilingual corpora are more resilient to variability and noise. Second, fine-tuning pre-trained models, rather than relying solely on hand-engineered features, unlocks richer emotional cues embedded in raw audio waveforms. Finally, while we made significant strides, challenges remain in handling underrepresented emotion categories and fine-grained emotional distinctions. Future research may involve gathering more balanced datasets, integrating multimodal signals, and exploring more sophisticated model architectures or attention mechanisms to further advance the field of SER.

8. Code and Data

The code and datasets used in this project are publicly available on GitHub. You can access the repository at the following link:

https://github.com/JustinTong0323/467final

References

Akçay, M. B. and Oğuz, K. Speech emotion recognition: Emotional models, databases, features, preprocessing methods, supporting modalities, and classifiers. *Speech Communication*, 116:56–76, 2020. ISSN 0167-6393. doi: https://doi.org/10.1016/j.specom.2019.12.001. URL https://www.sciencedirect.com/science/article/pii/S0167639319302262.

Baevski, A., Zhou, H., Mohamed, A., and Auli, M. wav2vec 2.0: A framework for self-supervised learning of speech representations. URL http://arxiv.org/abs/2006.11477.

Conneau, A., Baevski, A., Collobert, R., Mohamed, A., and Auli, M. Unsupervised cross-lingual representation learning for speech recognition. URL http://arxiv.org/abs/2006.13979.

Grosman, J. Fine-tuned XLSR-53 large model for speech recognition in English. https://huggingface.co/jonatasgrosman/wav2vec2-large-xlsr-53-english, 2021.

Jackson, P. and ul haq, S. Surrey audio-visual expressed emotion (SAVEE) database.

Livingstone, S. R. and Russo, F. A. The ryerson audio-visual database of emotional speech and song (RAVDESS): A dynamic, multimodal set of facial and vocal expressions in north american english. 13(5): e0196391. ISSN 1932-6203. doi: 10.1371/journal.pone.0196391. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5955500/.

Luna-Jiménez, C., Kleinlein, R., Griol, D., Callejas, Z., Montero, J. M., and Fernández-Martínez, F. A proposal for multimodal emotion recognition using aural transformers and action units on RAVDESS dataset. 12 (1):327. ISSN 2076-3417. doi: 10.3390/app12010327. URL https://www.mdpi.com/2076-3417/12/1/327. Number: 1 Publisher: Multidisciplinary Digital Publishing Institute.

Pichora-Fuller, M. K. and Dupuis, K. Toronto emotional speech set (TESS). URL https://borealisdata.ca/dataset.xhtml?persistentId=doi:10.5683/SP2/E8H2MF.

Sadok, S., Leglaive, S., and Séguier, R. A vector quantized masked autoencoder for speech emotion recognition. URL https://arxiv.org/abs/2304.11117v1.