Emotion Recognition using Deep Learning Models

Candiate Number: 40781, 49663, 49583, 46930 [1]

Abstract

A comprehensive comparison of deep learning-based speech emotion recognition (SER) systems is presented, focusing on both traditional signal-derived features and self-supervised representations. A unified dataset comprising 12,162 emotion-annotated audio samples was constructed by merging four publicly available corpora—RAVDESS, TESS, SAVEE, and CREMA-D—spanning seven emotion categories. Four feature extraction methods were examined: Mel-Frequency Cepstral Coefficients (MFCC), Log Mel Spectrogram, HuBERT, and Wav2Vec 2.0. Multiple neural architectures, including convolutional and attention-based models, were evaluated for each feature type.Self-supervised approaches outperformed traditional methods, with Wav2Vec 2.0 achieving the highest validation accuracy (80.11%), followed by HuBERT (74.89%), MFCC-based CNNs (70.16%), and Log Mel Spectrogram (68.39%). All models significantly surpassed a public Kaggle baseline using CNNs on MFCCs (49.02%). Emotions with higher arousal, such as anger, were more easily recognized, while subtler emotions remained more difficult. Grad-CAM visualizations on MFCC and Log Mel models revealed attention to semantically meaningful regions. Feature-level analysis of HuBERT and Wav2Vec 2.0 showed that Wav2Vec 2.0 dispersed emotional cues across latent dimensions, while HuBERT concentrated information in a smaller set of highly variable features. These results demonstrate the effectiveness of contextualized embeddings and support their use in robust speech-only emotion recognition systems.

1. Introduction

Speech Emotion Recognition (SER) aims to automatically identify human emotions from speech, a key component in building emotionally intelligent systems. Its importance spans a wide range of real-world applications, including affect-aware virtual assistants, mental health diagnostics, customer service analytics, and social robotics. As voice-based interfaces become increasingly prevalent, the ability to understand emotional cues from speech is crucial for enabling more natural and human-centric interactions.

Although SER has broad potential, it remains challenging due to variations in speech, speaker differences, and the lack of large labeled emotional datasets. The choice of acoustic features also greatly affects model performance. Traditional features like MFCCs offer compact spectral information but often miss important temporal and contextual cues. In contrast, self-supervised models like HuBERT and Wav2Vec 2.0 learn rich context from large amounts of unlabeled data. Given these differences, a systematic comparison of feature extraction strategies is important to evaluate their effectiveness for SER.

This study evaluates how different acoustic feature extraction methods—MFCC, Log Mel Spectrogram, HuBERT, and Wav2Vec 2.0—combined with different deep learning architectures affect SER performance. A unified dataset is constructed by merging four widely-used emotional speech corpora: RAVDESS, TESS, SAVEE, and CREMA-D. This combined dataset offers increased speaker and emotion diversity, serving as a robust benchmark for cross-corpus evaluation.

Multiple models tailored to each feature type are trained and evaluated using validation accuracy, with the best of each retained. Compared to a publicly available Kaggle CNN-MFCC baseline (49.02%), all models show clear improvements. The top-performing model, based on partially fine-tuned Wav2Vec 2.0 features, achieved 80.11%, outperforming HuBERT (74.89%) and MFCC-based CNNs (70.16%). These results underscore the superiority of self-supervised representations for SER and the importance of feature-architecture alignment.

The performance of our emotion classification model is further assessed using per-class recall. The analysis reveals that high-arousal, for instance, Surprise and Angry are the easiest to detect. Conversely, Happy, Fear, and Sad are often confused for each other, due to the similar and less obvious acoustic patterns. Despite the class imbalance, the model achieves reliable performance among all emotions.

This work presents a large-scale evaluation of traditional and self-supervised features for SER, providing insights into effective feature-model combinations for building robust speech-only emotion recognition systems.

2. Related Work

Speech Emotion Recognition (SER) has been extensively studied due to its wide applications in virtual assistants, healthcare, and human-computer interaction. Early approaches relied on handcrafted features such as MFCCs, pitch, energy, and formants, combined with classifiers like GMMs, SVMs, and HMMs.

With the rise of deep learning, researchers explored CNNs, RNNs, and attention-based models. Khalil et al. (Khalil et al., 2019) provided a thorough review of SER pipelines, highlighting the shift from traditional to deep learning approaches. They emphasized challenges such as spontaneous speech, cross-corpus generalization, and data imbalance. Zhao et al. (Zhao et al., 2019) showed that 2D-CNN-LSTM outperformed individual CNN or LSTM on the Berlin dataset with 95% accuracy. Similarly, Parry et al. (Parry et al., 2019) evaluated cross-corpus performance of CNN-LSTM on SAVEE, RAVDESS, and TESS with varying results. Dolka et al. (Dolka et al., 2021) used a basic ANN on MFCC features, while Asiya and Kiran (Asiya & Kiran, 2021) emphasized the importance of large datasets, reporting 89% accuracy using CNNs on RAVDESS and TESS. Singh et al. (Singh et al., 2023) proposed a CNN-LSTM-Attention architecture, tested on a combination of RAVDESS, SAVEE, and TESS, and achieved 90.19% accuracy.

Recent works have also leveraged self-supervised feature extractors. Hsu et al. (Hsu et al., 2021) introduced HuBERT, a masked prediction model that learns contextual speech representations without labels. Pepino et al. (Pepino et al., 2021) demonstrated the effectiveness of wav2vec2 embeddings for SER, outperforming traditional spectral features. Morais et al. (Morais et al., 2022) found similar results across IEMOCAP and CREMA-D, reinforcing the value of pre-trained models.

As a baseline, we adopted the publicly available Kaggle notebook "Audio Emotion — Part 3 - Baseline model" (Lok, 2021), which uses the same integrated dataset comprising RAVDESS, TESS, SAVEE, and CREMA-D. The baseline model is a 1D CNN with four convolutional blocks (256 → 128 → 64 filters), each followed by ReLU, batch normalization, and dropout. Max-pooling is applied after every two layers, and the output is flattened and fed into a softmax classifier. This model achieves 49.02% accuracy on the combined dataset. To improve upon this, we evaluate multiple feature types (MFCCs, Log Mel Spectrograms, HuBERT, Wav2Vec 2.0) and model architectures including CNNs, CNN-LSTMs, and attention-based networks.

3. Data Description

This study utilises four widely-used publicly available datasets for speech emotion recognition (SER): SAVEE, RAVDESS, CREMA-D, and TESS. Together, these corpora provide a diverse and comprehensive coverage of emotional expressions across speakers, genders, and recording conditions.

- **SAVEE**: Contains recordings from 4 male speakers, expressing 7 emotions: *neutral*, *anger*, *disgust*, *fear*, *happiness*, *sadness*, and *surprise*.
- **RAVDESS**: Includes speech from 24 professional actors (12 male, 12 female) vocalizing two predefined statements across 8 emotions, including *calm*, *neutral*, *happy*, *sad*, *angry*, *fearful*, and *surprised*.
- **CREMA-D**: Comprises recordings from 91 speakers (balanced across gender and ethnicity), expressing 6 categorical emotions with 4 intensity levels, using 12 standardized utterances. This dataset presents the most acoustic and emotional diversity.
- **TESS**: Features 2 female speakers performing utterances across 7 emotions based on Ekman's six universal categories plus *neutral*.

3.1. Combined Dataset Statistics

After merging these four datasets, we obtained a comprehensive corpus for further recognition tasks. The combined dataset contains a total of 12,162 audio samples, each labelled with one of the seven primary emotions: Anger, Disgust, Fear, Happiness, Sadness, Surprise, and Neutral.

Label	Count
happy	1923
fear	1923
disgust	1923
angry	1923
sad	1923
neutral	1895
surprise	652

Table 1. Number of samples per emotion label

As shown in the table above, there exists a degree of class imbalance across different emotional categories, especially for the surprise category. This imbalance needs to be carefully addressed during model training to avoid biased performance towards majority classes.

3.2. Data Preprocessing

Figure 1 demonstrates that the first 0.5 seconds of most recordings are silent, and the majority of samples are shorter

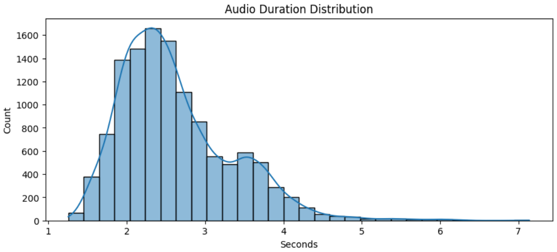

Figure 1. Audio Length Distribution

than 4.5 seconds. These observations suggest the presence of consistent silence at the beginning of each utterance and a need to unify input duration. While most audio files are sampled at 16 kHz, some use other rates such as 24,414 Hz, 44,100 Hz, or 48,000 Hz, which may introduce inconsistencies in downstream processing. Additionally, RMS energy analysis revealed that angry utterances exhibit the highest median energy and variance, followed by surprise, whereas neutral and sad display the lowest and most stable energy levels.

Based on these findings, we established a standardized preprocessing pipeline. The initial 0.5 seconds of each recording were removed, and the remaining audio was truncated to 4 seconds. Recordings shorter than 4 seconds were zero-padded to ensure uniform input length. All files were resampled to 16 kHz to ensure compatibility with pretrained models such as wav2vec2. Audio was processed using `librosa` and `soundfile`, and corrupted or unreadable files were excluded. After preprocessing, the dataset was randomly split into training and validation sets, and feature extraction was conducted separately for each subset to preserve evaluation integrity.

4. Architecture

We extracted features using four methods: Mel Frequency Cepstral Coefficient (MFCC), Mel Spectrogram, HuBERT embedding, and Wav2Vec 2.0, where only the best models for each method is reported, except for MFCC, where another comparable model is also reported.

4.1. MFCC with Energy Feature

MFCC is a widely used feature extraction method in audio processing. After removing unwanted noise, the signal is divided into short frames and transformed to the frequency domain using the Fourier Transform. The spectrum is then filtered through Mel-scale filters, converted to a logarithmic scale, and finally passed through a Discrete Cosine Transform (DCT) to produce n cepstral coefficients (Habib et al., 2021; Abdul & Al-Talabani, 2022). In this study, we use $n = 40$.

In addition to MFCCs, an energy feature is incorporated to capture the overall intensity of speech signals, which is often correlated with emotional expression. The energy is calculated by computing the root mean square (RMS) value for each frame, representing the signal's power over time.

To determine suitable architectures for MFCC-based emotion classification, a series of neural network structures were systematically evaluated. These included shallow and deep convolutional networks, CNN-LSTM hybrids, and attention-augmented CNN-BiLSTM models. The number of convolutional blocks ranged from one to five, with dropout rates from 0 to 0.5 and varying levels of L2 regularization. The CNN+BiLSTM+Attention model will be referred to as **”MFCC2”** in the upcoming sections. Among all configurations, a Deep CNN architecture consistently achieved the best validation performance and was therefore selected as the final best model for MFCC input.

The Deep CNN model is composed of four stacked convolutional blocks designed to extract hierarchical spatial patterns from MFCC inputs. In the subsequent sections, this model will be referred to as the ”**MFCC1**”. The architecture includes:

- **Input:** 41 × 216 × 1
- **Four convolutional blocks:** Conv2D (filters = 32, 64, 128, 256), kernel size 3×3, with ReLU + BatchNorm + MaxPooling2D (2×2)
- **Dropout rates:** 0, 0, 0.1, and 0.2 respectively per block
- **Fully connected layers:** Flatten + Dense (256 units, ReLU, L2 regularization $\lambda = 0.001$) + Dropout (0.3) + Dense (7 units, softmax)

4.2. Log Mel Spectogram

Log Mel Spectrogram feature extraction is similar to MFCC but omits the DCT step, resulting in a feature shape of (128, 126). We tested various architectures, including LSTM, multi-head attention, and local feature learning blocks (LFLBs) (Zhao et al., 2019). Notably, the best-performing model for MFCC also achieved the highest accuracy among all models using Log Mel Spectrogram. This model will be referred to as **”MS”**. The architecture is as follows.

- **Input:** 128 × 126 × 1
- **Four convolutional blocks:** Conv2D (filters = 32, 64, 128, 256), kernel size 3×3, with ReLU + BatchNorm + MaxPooling2D (2×2)
- **Dropout rates:** 0, 0, 0.1, and 0.2 respectively per block

- **Fully connected layers:** Flatten + Dense (256 units, ReLU, L2 regularization $\lambda = 0.001$) + Dropout (0.3) + Dense (7 units, softmax)

4.3. HuBERT model

HuBERT (**H**idden-**U**nit BERT) is a self-supervised speech representation learning model based on the Transformer architecture, proposed by Facebook AI Research. It leverages large-scale unlabeled speech data to learn rich and generalizable speech patterns and has been successfully applied to tasks such as ASR (Automatic Speech Recognition) and SER (Speech Emotion Recognition).

The training involves two key stages:

1. **Offline Clustering:** Extract low-level acoustic features and use K-means to create pseudo labels.
2. **Masked Prediction Training:** Predict hidden unit labels for masked time steps, capturing long-range dependencies.

The architecture includes:

- **Convolutional Feature Encoder** for initial processing.
- **Transformer Encoder** for contextual modeling.

Compared to handcrafted features like MFCC, HuBERT learns more expressive and semantically meaningful representations (Hsu et al., 2021).

For each audio input, HuBERT outputs a feature sequence of approximately $[200, 768]$, where 200 time steps result from 4-second audio at 16 kHz with subsampling.

To improve efficiency and reduce redundancy, we apply dimensionality reduction by computing mean and standard deviation across the temporal dimension, transforming the sequence into a compact $[1536]$-dimensional vector.

A lightweight **MLP + Self-Attention** architecture is designed with this kind of feature, and this model will be referred to as **"HuBERT"** in the subsequent sections.

- **Input Layer**: $[1536]$ feature vector.
- **Dense Layer**: 512 units, L2 regularization, BatchNorm, LeakyReLU, Dropout (0.4).
- **Residual Connection**: After an additional 512-unit Dense layer.
- **Attention Mechanism**: Self-Attention layer to enhance feature interactions.
- **Dense Layers**: 128 and 64 units with ReLU and Dropout (0.4, 0.3).
- **Output Layer**: Softmax over emotion classes.

This approach balances efficiency and robustness, though it sacrifices detailed temporal information. Additional full features with CNN+LSTM were explored but did not yield better results.

4.4. Wav2Vec 2.0

Wav2Vec2.0 is a self-supervised speech representation learning model based on a Convolutional Encoder followed by a Transformer-based context network, proposed by Facebook AI Research (Baevski et al., 2020). It is trained on large-scale unlabeled audio using a contrastive objective that encourages the model to distinguish true latent representations from distractors at masked time steps. The model has been widely adopted for downstream tasks such as SER due to its ability to extract expressive features directly from raw audio waveforms without requiring handcrafted inputs or manual labels. The training involves the following key steps.

1. **Feature Encoding:** A convolutional encoder processes raw audio into latent representations.
2. **Contrastive Masked Prediction:** Random time steps are masked, and the model is trained to correctly identify true future representations among a set of distractors.

The architecture includes:

- Convolutional Feature Encoder for subsampling raw waveform inputs.
- Transformer Context Network to model long-range dependencies over time.

For each audio input, Wav2Vec2.0 outputs a sequence of frame-level hidden states with shape approximately [200, 768], where 200 time steps correspond to a 4-second audio segment sampled at 16 kHz with subsampling.

Among several configurations we explored, including static pooling of hidden states and temporal modeling via recurrent and attention-based architectures, the best results were achieved by fine-tuning the last 5 Transformer encoder layers of the pretrained model (Experiment 9 in the notebook). Earlier layers remained frozen to preserve general acoustic representations and reduce overfitting risk. We used global average pooling across the time dimension to aggregate the contextual embeddings, followed by a lightweight fully-connected classification head. The model architecture is as follows (referred to as **Wav2Vec 2.0** in the next sections).

- **Input:** Raw audio waveform, 64,000 samples (4 seconds at 16 kHz)
- **Base Model:** facebook/wav2vec2-base (pretrained)
- **Unfrozen Layers:** Last 5 Transformer encoder blocks
- **Output Features:** [200, 768] frame-wise embeddings
- **Pooling:** Global Average Pooling across time
- **Output Features:** Two fully connected layers with 128 and 64 units and a ReLU activation. Each layer is followed by a Dropout layer with a rate of 0.4 and 0.3, respectively. Finally, a softmax classification layer is added.

5. Training Methods

As the labels are one-hot encoded, all the models are trained using the Categorical Cross-Entropy Loss. Our main evaluation metric is the validation accuracy. To address overfitting, we have attempted the following data augmentation techniques: (1) Gaussian random noise with $\mu = 0$ and $\sigma^2 \in [0.001, 0.003]$ (2) Pitch shift between 2 semitones (3) Time stretching with a stretch factor between 0.9 and 1.1. However, there was no significant improvement in the validation accuracy. Therefore, we removed these procedures from our final models and notebooks. We restored the best weights with the highest validation accuracy using ModelCheckPoint. The training hyperparameters varies from model to model, and are described in the following sections, and the key parameters are summarized in Table 2. As a baseline result, we use the result from a kaggle notebook by Eu Jin Lok (Lok, 2021), who applied simple deep CNN model on MFCC features and achieved a validation accuracy of 49% in classifying 7 emotions on the same dataset.

5.1. MFCC

Both models were trained using the Adam optimizer (LR = 0.001) and a batch size of 32. Learning rate scheduling was applied via ReduceLROnPlateau (factor = 0.5, patience = 3), and early stopping (patience = 3) was used to prevent overfitting.

5.2. Log Mel Spectrogram

The data is first batched to a size of 128. The best models was trained using the Adam optimizer (LR = 0.0001). Similar to MFCC, a learning rate scheduling (factor = 0.5, patience = 3) and early stopping (patience = 5) was also utilized.

5.3. HuBERT

The model is optimized using the AdamW optimizer with a learning rate of 0.0001, and trained with categorical cross-entropy loss incorporating label smoothing (rate = 0.08). The batch size used in this model is 128. To allow for a finer adjustment on the plateau, we apply a learning rate scheduling with factor = 5 and patience = 5. We also specify an early stopping criteria with patience = 5.

Table 2. Compiler Parameters. MS, OPT, LR, BS refer to the Log Mel Spectrogram feature extraction, Learning Rate, and Batch Size, respectively.

MODEL	OPT	LR	BS	EPOCHS
MFCC	ADAM	1E-3	32	50
MS	ADAM	1E-4	128	70
HUBERT	ADAMW	1E-4	128	80
WAV2VEC	ADAMW	1E-4	32	30

Table 3. Classification Model Results. MFCC and MS1 refer to different feature extraction methods; Wav2Vec2 is pretrained and fine-tuned.

MODEL	ACCURACY
BASELINE	49.02%
MFCC1	70.16%
MFCC2	66.95%
MS	68.39%
HUBERT	74.89%
WAV2VEC2	**80.11%**

5.4. Wav2Vec 2.0

We selected AdamW (LR = 0.0001) as an optimizer, and included label smoothing with a rate of 0.08 in the model training. Only the early stopping criteria with patience = 5 is applied, and the batch size is set to 32.

6. Numerical Results and Interpretation

6.1. Model Results

Table 3 illustrates that Wav2Vec 2.0 model achieved the highest validation accuracy, followed by HuBERT, MFCC with energy features, and Log-Mel Spectrogram. Together with a non-overfitting process for HuBERT, it highlights the superiority of self-supervised representations over traditional feature extraction methods for emotion classification tasks. Our results are similar to (Meghanani & Hain, 2024)'s study, who found that the performance in downstream tasks can be ordered as HuBERT > Wav2vec 2.0 > MFCC > Log Mel Spectrogram, but the only difference is that Wav2vec 2.0 performs better than HuBERT in our experiment.

The Wav2Vec 2.0 model achieved the highest accuracy of 80.11%, outperforming HuBERT and traditional feature-based models, in contrast to prior findings (Hsu et al., 2021).

While Wav2Vec 2.0 lacks HuBERT's clustering-based pretraining (Baevski et al., 2020), its ability to capture continuous acoustic patterns proved advantageous for emotion recognition. The results highlight the effectiveness of partial fine-tuning, as selectively updating the top Transformer layers yielded significant performance gains over frozen embeddings.

The HuBERT model achieved 74.89% accuracy without signs of overfitting, consistent with its capacity to learn rich acoustic and linguistic representations (Hsu et al., 2021). Its clustering-based pretraining facilitates the capture of contextual and phonetic features beneficial for SER. However, it slightly underperformed Wav2Vec 2.0, possibly due to the use of the smaller base model (90M parameters) and limited alignment between phonetic abstractions and emotional cues. Nevertheless, even a shallow classifier model with CNN and RNN can achieve a substantial accuracy.

For MFCC-based features, the Deep CNN achieved the highest validation accuracy (70.16%) among MFCC configurations, outperforming the CNN-BiLSTM model (66.95%) despite greater complexity in the experiments. Its superior performance is attributed to convolutional networks' ability to extract local and hierarchical patterns from MFCCs, whose static 2D structure limits temporal modeling. Progressive dropout and deeper architecture further enhanced generalization.

Given the similarity between MFCC and Log Mel Spectrogram features, their performance was compared using the same model architecture. The Log Mel Spectrogram model achieved 68.39% accuracy, slightly below the MFCC counterpart. While preserving time-frequency information, its highly correlated values may introduce noise and speaker-specific artifacts, making learning less efficient and more prone to overfitting.

Compared to the baseline CNN model using MFCCs (49.02% accuracy) (Lok, 2021), all proposed methods showed substantial improvements. The MFCC-based Deep CNN exceeded the baseline by over 17 percentage points, while Log Mel Spectrogram also improved. Notably, self-supervised models Wav2Vec 2.0 and HuBERT achieved the highest gains, highlighting the advantage of advanced representations for SER.

It is also worth noting that some datasets are easier to distinguish than others. The TESS dataset achieves nearly 100% validation accuracy. A possible reason is that it only contains female recordings, which have been reported to be easier to classify in SER tasks relative to male recordings (Lausen & Schacht, 2018). On the other hand, the overall performance tends to deteriorate as the CREMA-D data is included. This could be attributed to the greater diversity in the speakers' geographical backgrounds introduced in this dataset, whereas the other datasets primarily feature native speakers.

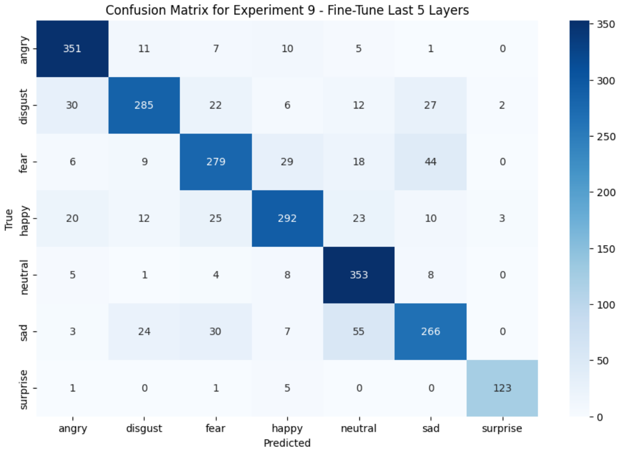

Figure 2. Confusion Matrix for Wav2Vec 2.0 Model

Table 4. Recall for each Emotion

Emotion	Recall
Angry	91.17%
Disgust	74.22%
Fear	72.47%
Happy	75.84%
Neutral	93.14%
Sad	69.09%
Surprise	94.62%

6.2. Confusion Matrix

This section discusses the confusion matrix of the Wav2Vec 2.0 model, which achieved the highest validation accuracy. The confusion matrix is shown in Figure 2. Overall, the model performs well in classifying the emotion, as indicated by the concentration of values in the diagonal entries of the confusion matrix.

However, our data exhibits a class imbalance as discussed in Table 1, and it is difficult to judge which emotions are easiest to detect from the values in the diagonal entries. To address this, we conducted a supplement recall analysis, and the results are shown in Table 4.

Despite the initial concern about the model performance in detecting Surprise due to the limited sample size, it emerged as the easiest emotion to detect, followed by Neutral, and Angry. The high recalls for Surprise and Angry were expected, as these emotions are typically associated with high arousal. Neutral, on the other hand, is somewhat surprising given the more moderate signal patterns observed in Figure 3. In contrast, Happy, fear, and Sad are more challenging

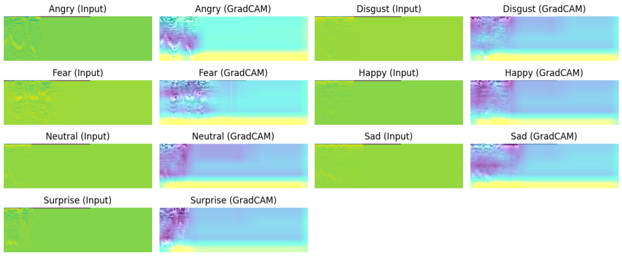

Figure 3. GradCAM Results on MFCC Features

to detect and they are in fact often mistaken for each other, possibly due to their more subtle and overlapping acoustic features. Nevertheless, the model achieves reasonable recall across all classes, indicating robust performance in emotion classification.

6.3. Sanity Check

6.3.1. MFCC and Log Mel Spectrogram

We first performed sanity checks using explainable AI (XAI) tools on the best MFCC and Log Mel Spectrogram models (”MFCC1” and ”MS”), which is an XAI technique that can visualize parts of an input image that are most influential in a neural network’s prediction. Grad-CAM computes gradients of the predicted class score with respect to feature maps in a chosen layer. We selected the last convolutional layer as proposed by (Selvaraju et al., 2019).

We sampled one observation from each emotion label to perform Grad-CAM. Figure 3 visualizes the raw MFCC inputs and their corresponding Grad-CAM outputs, while Figure 3 displays the same for the Log Mel Spectrogram features. In MFCC, the top row represents energy and has a much larger value range than other coefficients, resulting in a dark band and otherwise flat coloring due to color scaling. In contrast, Log-Mel inputs exhibit clearer structure with visible variation across time and frequency.

All inputs are padded to a fixed 4-second duration, resulting in the uniformly colored right regions. Grad-CAM’s attention being focused on the left for both MFCC and Log Mel Spectrogram is consistent with the actual signal. High-arousal emotions like angry and fear exhibit strong vertical patterns and receive strong Grad-CAM activation, suggesting the model associates dynamic spectral changes with emotional expression. Subtler emotions like Neutral and Happy show more diffuse attention.

Overall, these figures demonstrate that both models are capable of focusing on meaningful regions in the input. While Log-Mel Spectrogram provides more interpretable inputs and produces more localized Grad-CAM responses, it has a lower validation accuracy than MFCC. Possibly due to the irrelevant noises preserved in Log Mel Spectrogram.

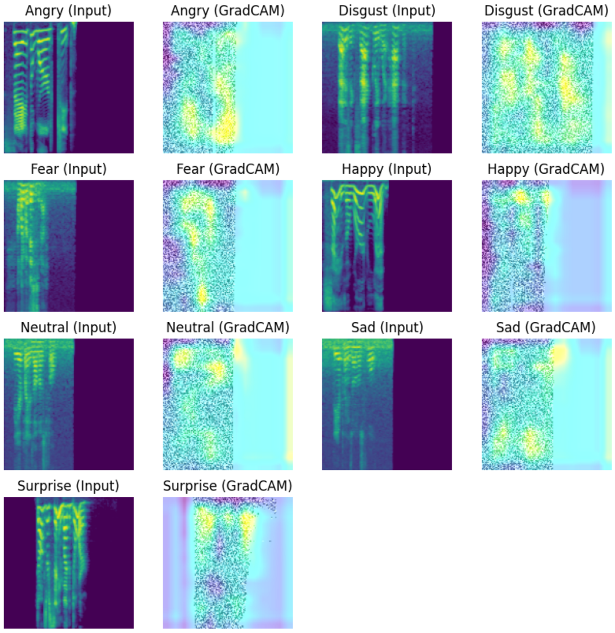

Figure 4. GradCAM Results on Log Mel Spectrogram Features

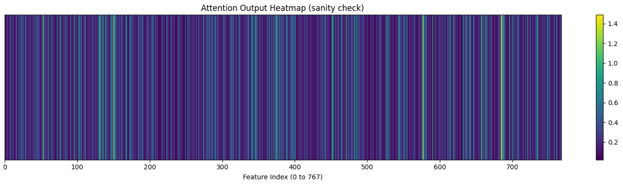

Figure 5. Attention-like Output Heatmap

6.3.2. Wav2Vec 2.0

An attention-like heatmap was generated by selecting a random audio sample from the test set, then averaging the absolute values of the final-layer feature activations across temporal dimension. This ”attention-like” heatmap offers a quick sanity check, showing which of the 768 latent dimensions the model relies on most strongly for that instance. The resulting heatmap (Figure 5) revealed a broadly distributed activation pattern, with moderate energy spread across nearly the entire feature space and no clusters that represent dominant dimensions.

This activation suggests that the model encodes emotion-related information in a distributed manner, leveraging a wide array of features rather than depending on a small specific subset. This implies the model captures rich emotional

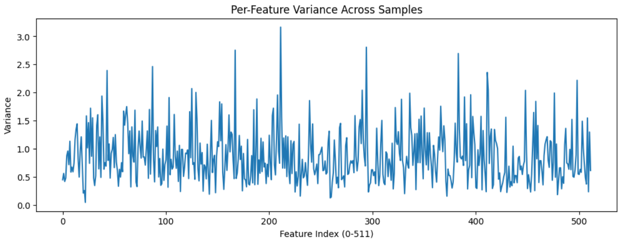

Figure 6. Attention Layer Output Feature Variance (HuBERT)

cues that span multiple acoustic and contextual dimensions and does not overfit to to isolated traits. This supports that Wav2Vec2's deep contextual embeddings are effectively adapted during partial fine-tuning to extract subtle emotion-relevant information from raw waveforms.

6.3.3. HuBERT

Variance for each output feature from the attention layer across samples is calculated and displayed in Figure 6. It can help identify which features are most informative for the model. In the attention layer output features, based on the plot, some have very high variance meaning they vary significantly across samples and may carry useful signal. The lists of top-10 most and least variable features help highlight which dimensions are most and least informative.

7. Conclusion

This study systematically investigated the impact of different acoustic feature extraction methods, including MFCC, Mel Spectrogram, HuBERT, and Wav2Vec 2.0, combined with various deep learning architectures, on SER. We explored different feature extraction pipelines, data augmentation strategies, model architectures, and training procedures, ultimately identifying the best model for each feature type. Our results highlight the importance of aligning feature selection and model design with the characteristics of the combined dataset.

Specifically, the best-performing model was based on Wav2Vec 2.0 features, achieving an accuracy of 80.11%, followed by HuBERT at 74.89%, MFCC1 at 70.16%, Mel Spectrogram (MS) at 68.39%, MFCC2 at 66.95%, and the baseline at 49.02%. However, it is worth noting that the Wav2Vec 2.0-based model exhibited moderate overfitting, and both MFCC and Mel Spectrogram models also showed signs of overfitting. In contrast, the HuBERT-based models did not show obvious overfitting, demonstrating more stable generalization behavior across the experiments.

The model demonstrates strong overall performance, particularly in identifying high-arousal emotions like Surprise and Angry. Even Neutral is accurately recognized, indicating clear acoustic consistency. Although subtle emotions such as Happy, Fear, and Sad remain more challenging, the results show that the model handles emotional diversity effectively despite class imbalance.

While MFCCs effectively capture spectral features, they lack temporal dynamics critical to emotional speech. Future work should incorporate delta features and explore architectures better suited for sequential data, such as Temporal Convolutional Networks or transformer-based models like the Audio Spectrogram Transformer. Robustness evaluation under multilingual, accented, or noisy conditions is also recommended. For self-supervised models like Wav2Vec 2.0 and HuBERT, scaling to larger datasets and applying advanced augmentation could further improve performance. Additionally, dimensionality reduction methods such as PCA may help manage high-dimensional embeddings; in our experiments, PCA on HuBERT features yielded results comparable to mean+std pooling, with the latter preferred for efficiency.

References

Abdul, Z. K. and Al-Talabani, A. K. Mel frequency cepstral coefficient and its applications: A review. *IEEE Access*, 10:122136–122158, 2022. doi: 10.1109/ACCESS.2022.3223444.

Asiya, S. and Kiran, V. Speech emotion recognition-a deep learning approach. *International Journal of Engineering Research & Technology*, 10(12):1–5, 2021.

Baevski, A., Zhou, H., Mohamed, A., and Auli, M. wav2vec 2.0: A framework for self-supervised learning of speech representations. *CoRR*, abs/2006.11477, 2020. URL https://arxiv.org/abs/2006.11477.

Dolka, A. et al. Speech emotion recognition using ann on mfcc features. *International Journal of Advanced Computer Science and Applications*, 12(6):456–460, 2021.

Habib, M., Faris, M., Qaddoura, R., Alomari, M., Alomari, A., and Faris, H. Toward an automatic quality assessment of voice-based telemedicine consultations: A deep learning approach. *Sensors*, 21:3279, 05 2021. doi: 10.3390/s21093279.

Hsu, W.-N., Bolte, B., Tsai, Y.-H. H., Lakhotia, K., Salakhutdinov, R., and Mohamed, A. Hubert: Self-supervised speech representation learning by masked prediction of hidden units. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2021.

Khalil, R. A., Jones, E., Babar, M. I., Jan, T., Zafar, M. H., and Alhussain, T. Speech emotion recognition using deep learning techniques: A review. *IEEE Access*, 7:117327–117345, 2019. doi: 10.1109/ACCESS.2019.2936124.

Lausen, A. and Schacht, A. Gender differences in the recognition of vocal emotions. *Frontiers in Psychology*, 9:882, 06 2018. doi: 10.3389/fpsyg.2018.00882.

Lok, E. J. Audio emotion - part 3: Baseline model. https://www.kaggle.com/code/ejlok1/audio-emotion-part-3-baseline-model, 2021. Accessed: 2025-05-01.

Meghanani, A. and Hain, T. Improving acoustic word embeddings through correspondence training of self-supervised speech representations, 2024. URL https://arxiv.org/abs/2403.08738.

Morais, E., Hoory, R., Zhu, W., Gat, I., Damasceno, M., and Aronowitz, H. Speech emotion recognition using self-supervised features. *arXiv preprint arXiv:2202.03896*, 2022.

Parry, J., Schuller, B., and Cummins, N. Analysis of deep learning architectures for cross-corpus speech emotion recognition. In *INTERSPEECH*, pp. 1651–1655, 2019.

Pepino, L., Riera, P., and Ferrer, L. Emotion recognition from speech using wav2vec 2.0 embeddings. *arXiv preprint arXiv:2104.03502*, 2021.

Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., and Batra, D. Grad-cam: Visual explanations from deep networks via gradient-based localization. *International Journal of Computer Vision*, 128 (2):336–359, October 2019. ISSN 1573-1405. doi: 10.1007/s11263-019-01228-7. URL http://dx.doi.org/10.1007/s11263-019-01228-7.

Singh, J., Saheer, L. B., and Faust, O. Speech emotion recognition using attention model. *International Journal of Environmental Research and Public Health*, 20(6): 5140, 2023. doi: 10.3390/ijerph20065140.

wen Yang, S., Chi, P.-H., Chuang, Y.-S., Lai, C.-I. J., Lakhotia, K., Lin, Y. Y., Liu, A. T., Shi, J., Chang, X., Lin, G.-T., Huang, T.-H., Tseng, W.-C., tik Lee, K., Liu, D.-R., Huang, Z., Dong, S., Li, S.-W., Watanabe, S., Mohamed, A., and yi Lee, H. Superb: Speech processing universal performance benchmark, 2021. URL https://arxiv.org/abs/2105.01051.

Zhao, J., Mao, X., and Chen, L. Speech emotion recognition using deep 1d & 2d cnn lstm networks. *Biomedical Signal Processing and Control*, 47:312–323, 2019.

A. Statement about individual contributions

Table 5. Statement about individual contributions

Task	**Percentage**	**Contributor(s)**
Data Collection and Preprocessing	10%	All members
Literature Review	10%	All members
MFCC-related model, training and evaluation	15%	49663
MS-related model, training and evaluation	15%	49583
HuBERT-related model, training and evaluation	15%	40781
Wav2Vec2-related model, training and evaluation	15%	46930
Notebooks re-organization	10%	All menbers (46930 contributed more)
Report Writing	10%	All menbers (49583 contributed more)
Overall	**100%**	**Each member contributed equally**