Efficient Fine-Tuning Approaches on HuBERT for Speech Emotion Recognition on Multiple Labels

Anonymous ACL submission

Abstract

Models like HuBERT have shown significant promise in automatic speech recognition (ASR). In this work, we explore both vanilla fine-tuning and parameter-efficient fine-tuning of the HuBERT model for speech emotion recognition (SER). While most previous research on SER has focused on four basic emotions—happy, sad, angry, and neutral—we extend this by incorporating additional emotions: surprise, fear, disgust, and calm, bringing the total to eight. Our experiments utilize four diverse datasets to enhance the robustness of our findings. Our methodology involves using the Wav2Vec2FeatureExtractor from the HuBERT model to extract features from raw audio files. These features are fed into a sequence classification model built on the HuBERT architecture. We fine-tuned the model in three different approaches -vanilla Finetuning, Parameter efficient finetuning over QKV projection and classifier using LoRA over a combination of several publicly available emotional speech datasets, including RAVDESS, CREMA-D, TESS, and SAVEE. The vanilla fine-tuned method outperforms all fine-tuned approaches overall. However, parameter-efficient approaches are still satisfactory and can be used in case of low resources and limited computational power.

1 Introduction

Speech Emotion Recognition (SER) is an essential aspect of human-computer interaction, significantly contributing to more natural and effective communication systems.(Ramakrishnan and El Emary, 2013) While traditional SER systems primarily focus on basic emotions such as happy, sad, angry, and neutral,(Busso et al., 2004)(Durand et al., 2007) there is a growing need to encompass a broader range of emotions for more comprehensive applications. This research aims to extend the emotional categories to include surprise, fear, disgust, and calm, thereby covering a total of eight distinct emotions.

HuBERT (Hsu et al., 2021), known for its robust feature extraction capabilities (Wu et al., 2024), leverages self-supervised learning to pretrain models on large-scale unlabelled data, which can then be fine-tuned for specific tasks. This study explores both vanilla fine-tuning and parameter-efficient fine-tuning of the HuBERT model to enhance its performance in SER. We used the ported version of S3PRL's Hubert for the SUPERB Emotion Recognition task from hugging face. [1]

The challenge of limited annotated data in SER remains a significant bottleneck especially when compared to the vast datasets available for ASR(Ao et al., 2022). To address this, our experiments utilize a combination of several publicly available emotional speech datasets , including RAVDESS (Ryerson Audio-Visual Database of Emotional Speech and Song)(Livingstone and Russo, 2018), CREMA-D (Crowd-sourced Emotional Multimodal Actors Dataset)(Cao et al., 2014), TESS (Toronto Emotional Speech Set)Pichora-Fuller and Dupuis, 2020, and SAVEE (Surrey Audio-Visual Expressed Emotion). These datasets are split into training, validation and evaluation sets to ensure the robustness and generalization of our findings.

Our methodology involves using the Wav2Vec2FeatureExtractor from the HuBERT model (Yang et al., 2021) to extract features from raw audio files, since it seemed to perform well on previous works.(Pepino et al., 2021) (Chen and Rudnicky, 2023)These extracted features serve as inputs to a sequence classification model built on the HuBERT architecture.(CHAKHTOUNA et al., 2024) The feature extraction process is crucial as it captures the intricate details of speech signals, which are pivotal for accurate emotion recognition. We implement data augmentation techniques, including pitch shifting, time stretching, and

[1] https://huggingface.co/superb/hubert-large-superb-er

noise addition, to artificially expand the dataset and improve model robustness. (Grichkovtsova et al., 2012)Additionally, batch normalization and dropout are employed during training to prevent overfitting and enhance generalization.

In conclusion, this study aims to push the boundaries of SER by leveraging the powerful feature extraction capabilities of the HuBERT model, combined with innovative fine-tuning strategies and comprehensive emotional datasets. The outcomes of this research have the potential to significantly enhance the accuracy and applicability of SER systems in various domains, from customer service interactions to mental health monitoring.

2 Methodology

2.1 HubertModel

The HuBERT (Hidden-Unit BERT) model is a self-supervised learning model designed for speech representation. It operates on a masked prediction framework, where portions of the input audio sequence are masked, and the model is trained to predict these masked sections. This approach leverages hidden units—discrete representations formed by clustering acoustic features—allowing the model to capture various nuances in speech, such as intonation, pitch, and rhythm. Although HuBERT does not directly classify emotions, it learns rich speech features that are invaluable for downstream tasks like emotion recognition. By fine-tuning HuBERT on a labeled emotion dataset, these learned features can be adapted for the specific task of speech emotion detection. In this study, we fine-tuned the HuBERT model using three different approaches and analyzed their performance to enhance the accuracy of emotion classification in speech.

2.2 Fine Tuning

We experimented with three different fine-tuning techniques [2] to adapt the pretrained model to our specific task. Previous works have demonstrated that fine-tuning large models on domain-specific tasks, such as emotion recognition, yields excellent performance.(Cao et al., 2014)(Siriwardhana et al., 2020)(Gao et al., 2023) Moreover, parameter-efficient fine-tuning techniques are particularly advantageous, as they optimize resource and time utilization while delivering effective results. (Lashkarashvili et al., 2024) (Gao et al., 2024)(Li et al., 2023)

2.2.1 Full Fine-Tuning

This approach involves updating all the parameters of the model during training. It allows the model to learn task-specific features but requires more computational resources and training time. The entire model, including the feature extractor, encoder, and classification head, was fine-tuned on our dataset.

2.2.2 Parameter-Efficient Fine-Tuning (PEFT) with LoRA on K, Q, V Projection Layers

This method involves adding low-rank matrices to the key (K), query (Q), and value (V) projection layers in the self-attention mechanism.(Feng and Narayanan, 2023) It significantly reduces the number of trainable parameters while retaining the majority of the pretrained weights. Only the K, Q, and V projection layers were fine-tuned with the Lora technique, keeping the rest of the model parameters frozen.

2.2.3 Parameter-Efficient Fine-Tuning (PEFT) with LoRA on Classifier Layer

This approach focuses on updating only the classification head of the model while keeping the pre-trained feature extractor and encoder layers fixed. It is useful when the amount of labeled data is limited. Only the classification head was fine-tuned to adapt the model to our specific task.

3 Experiment

In our experiment, we performed extensive fine-tuning on a pre-trained Hubert model, focusing on optimizing key parameters for the Q,K,V projection layers and the classifier layer to enhance performance on the target dataset. We utilized various configurations and components in our model training. The optimizer used for training was Adam, with a learning rate of 1e-5. The training was conducted over 50 epochs.

The hardware configuration included an NVIDIA L40 GPU with 46068 MB of memory. Each fine-tuning approach took approximately 2 hours to complete.

The rest of the configuration settings are standard, as the model was sourced from the Hugging Face repository. Additional details can be found in Table 1 The subsequent sections provide detailed insights into the dataset and fine-tuning parameters used.

[2]https://github.com/usc-sail/peft-ser

Layer Type	Input Shape	Output Shape	Param #
Input	[1, 16000]		
HubertFeatureEncoder	[1, 16000]	[1, 512, 49]	3,945,696
Conv1d (Layer 0)	[1, 1, 16000]	[1, 512, 3199]	5,632
Conv1d (Layers 1-4)	[1, 512, 3199]	[1, 512, 199]	3,147,776
Conv1d (Layers 5-6)	[1, 512, 199]	[1, 512, 49]	786,944
FeatureProjection	[1, 512, 49]	[1, 49, 1024]	525,824
HubertEncoderStableLayerNorm	[1, 49, 1024]	[1, 49, 1024]	433,012,992
HubertEncoderLayerStableLayerNorm	[1, 49, 1024]	[1, 49, 1024]	17,958,528 (each)
Projector	[1, 49, 1024]	[1, 49, 256]	262,400
Classifier	[1, 49, 256]	[1, 49, 8]	2,056
Total Parameters			**437,865,352**

Table 1: Summary of the Hubert Model used for Fine-tuning on SER with 8 Emotions

3.1 Datasets

Emotion	Source	Count
Angry	CREMA-D	1271
	RAVDESS	192
	SAVEE	60
	TESS	400
Calm	RAVDESS	192
Disgust	CREMA-D	1271
	RAVDESS	192
	SAVEE	60
	TESS	400
Fear	CREMA-D	1271
	RAVDESS	192
	SAVEE	60
	TESS	400
Happy	CREMA-D	1271
	RAVDESS	192
	SAVEE	60
	TESS	400
Neutral	CREMA-D	1087
	RAVDESS	96
	SAVEE	120
	TESS	400
Sad	CREMA-D	1271
	RAVDESS	192
	SAVEE	60
	TESS	400
Surprise	RAVDESS	192
	SAVEE	60
	TESS	400

Table 2: Count of files for each emotion and source

In our study on speech emotion recognition (SER), we utilized four key datasets to train and evaluate our emotion classification models. Table 2 summarizes the number of files for each emotion and source.

We utilized the following datasets for our experiments: **Toronto Emotional Speech Set (TESS)** The TESS dataset consists of 2,800 high-quality audio recordings from two female actresses, each portraying seven emotions (anger, disgust, fear, happiness, pleasant surprise, sadness, and neutral) across 200 target words.
Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS) The RAVDESS includes 1,440 speech audio files from 24 professional actors (12 female, 12 male), each expressing seven emotions with varying intensity.
Surrey Audio-Visual Expressed Emotion (SAVEE) The SAVEE dataset contains recordings from four male speakers, each delivering 120 utterances across seven emotions. Despite its male-only composition, SAVEE offers high-quality, phonetically balanced sentences that complement the other datasets.
Crowd Sourced Emotional Multimodal Actors Dataset (CREMA-D) The CREMA-D features 7,442 audio clips from 91 diverse actors, spanning multiple races and ethnicities. Each actor delivers sentences in one of six emotions at various intensity levels. The diversity and volume of CREMA-D ensure robust model training and prevent overfitting.

These datasets collectively provide a comprehensive foundation for developing a robust SER model capable of accurately identifying emotions from diverse audio sources.

3.2 Pretrained Model

We utilized the HubertForSequenceClassification model, which is based on the HuBERT architecture. It consists of several components:

- **Feature Extractor:** Utilizes multiple convolutional layers to process raw audio inputs.
- **Feature Projection:** Projects extracted features into a higher-dimensional space.

- **Encoder:** Composed of multiple transformer layers to capture temporal dependencies in the audio sequence.
- **Classification Head:** A final linear layer to map the encoder outputs to class probabilities.

3.3 Audio Preprocessing

The preprocessing steps involved several key operations to prepare the audio data for training:

- **Loading Audio Files:** Audio files were loaded using the *librosa* library, which provides functionality for analyzing and extracting features from audio signals.
- **Resampling:** All audio files were resampled to a uniform sample rate to ensure consistency across the dataset.
- **Feature Extraction:** We used the Wav2Vec2 (Baevski et al., 2020) feature extractor to convert raw audio signals into a sequence of feature vectors. This involved computing mel-frequency cepstral coefficients (MFCCs) and other relevant audio features.
- **Normalization:** The extracted features were normalized to have zero mean and unit variance to facilitate faster convergence during training.
- **Segmentation:** Long audio files were segmented into shorter, fixed-length clips to create a uniform input size for the model.(Rybach et al., 2009)

4 Results

In this section, we present the performance metrics of our finetune experiments, including full finetuned of Hubert abbreviated as Hubert(FT), Parameter-Efficient Fine-Tuning (PEFT) with LoRA on (K, Q, V) Projection Layers abbreviated as Hubert(QKV) , and Parameter-Efficient Fine-Tuning (PEFT) with LoRA on classifier Layers abbreviated as Hubert (classifier). The evaluation metrics used are F1 score, Equal Error Rate (EER), and Accuracy. The results are summarized in Table 3. These results indicate that the fully fine tuned Hubert model outperforms the modified versions in all evaluated metrics.

Models / Metrics	F1_score	EER	Accuracy
Hubert (FT)	0.8610	0.0713	86.10
Hubert (QKV)	0.6715	0.164	67.146
Hubert (classifier)	0.4461	0.2870	44.61

Table 3: Performance metrics for different Hubert models.

Confusion Matrices The confusion matrices provide a detailed breakdown of the model's performance across different emotion categories.(Liang, 2022) Each matrix shows the percentage of correct and incorrect predictions for each emotion, allowing us to analyze the strengths and weaknesses of each model. The confusion matrices for the full fine tuned Hubert Model, PEFT on Hubert (QKV), and PEFT on Hubert (Classifier) models are presented in Tables 4 , 5 , and 6 respectively. These matrices reveal that the full fine tuning of the Hubert model yields the highest accuracy across most emotion categories, while the modified versions show varying degrees of misclassification.

label	ang	cal	dis	fea	hap	neu	sad	sur
ang	**96.99**	0.00	1.10	1.10	0.55	0.27	0.00	0.00
cal	0.00	**89.74**	0.00	0.00	0.00	10.26	0.00	0.00
dis	4.77	0.00	**82.16**	2.51	3.77	2.01	4.52	0.25
fea	2.01	0.00	2.76	**76.94**	6.52	2.76	8.02	1.00
hap	2.65	0.27	0.53	1.86	**90.45**	3.71	0.27	0.27
neu	0.31	0.00	0.00	0.00	1.57	**98.11**	0.00	0.00
sad	0.76	0.00	3.82	9.41	1.53	13.23	**71.25**	0.00
sur	0.70	0.00	0.00	0.00	2.80	0.00	0.00	**96.50**

Table 4: Confusion Matrix for Hubert full finetuning (in percentage)

label	ang	cal	dis	fea	hap	neu	sad	sur
ang	**92.88**	0.00	1.64	0.27	1.92	1.37	0.00	1.92
cal	0.00	**84.62**	0.00	0.00	0.00	0.00	15.38	0.00
dis	9.30	1.26	**64.57**	0.75	2.51	16.33	3.27	2.01
fea	5.01	0.75	2.01	**42.86**	17.54	15.54	13.78	2.51
hap	11.67	2.92	2.12	2.65	**59.42**	13.26	1.59	6.37
neu	1.57	7.55	0.00	0.00	0.94	**88.99**	0.94	0.00
sad	1.02	4.83	3.05	2.54	1.78	35.62	**49.87**	1.27
sur	1.40	0.70	0.70	0.00	4.90	1.40	0.00	**90.91**

Table 5: Confusion Matrix for Hubert-PEFT-KQV (in percentage)

label	ang	cal	dis	fea	hap	neu	sad	sur
ang	**84.38**	0.00	3.84	0.82	1.64	8.77	0.27	0.27
cal	0.00	**92.31**	0.00	0.00	0.00	2.56	5.13	0.00
dis	14.82	2.01	**38.44**	1.26	0.75	23.87	18.09	0.75
fea	15.29	1.25	1.75	20.05	8.52	24.56	**27.82**	0.75
hap	29.71	4.77	11.94	3.98	4.51	**33.69**	6.37	5.04
neu	1.26	7.23	0.31	0.31	0.00	**87.11**	3.77	0.00
sad	0.51	5.34	2.80	3.31	0.76	**46.31**	40.97	0.00
sur	24.48	8.39	10.49	2.80	0.70	16.08	0.00	**37.06**

Table 6: Confusion Matrix for Hubert-PEFT-Classifier (in percentage)

In conclusion, the fully fine-tuned Hubert model outperforms its PEFT counterparts in all metrics, highlighting the trade-off between computational efficiency and model accuracy in emotion classification.

5 Limitations

While our experimental setup has demonstrated the efficiency of the HuBERT model and its variations in speech emotion recognition tasks through full fine-tuning, fine-tuning of QKV layers, and fine-tuning of the classifier, there are several limitations to consider. First, the dataset composition, though diverse, may still not capture the full variability of real-world speech emotions, potentially limiting the generalizability of our findings. The reliance on publicly available datasets may introduce biases inherent to these datasets. Additionally, the pre-trained models used in this study are initially trained on general speech data and might not be optimized for emotion-specific nuances, even after fine-tuning, which could affect performance. The feature extraction and classification processes are also computationally intensive, requiring significant processing power and memory, which could be a constraint for deployment in resource-limited environments. Furthermore, our evaluation focuses primarily on accuracy, F1 score, and EER; other important metrics like latency and robustness to noise were not explored. While we explored different fine-tuning strategies, the potential benefits of combining these strategies or exploring alternative fine-tuning approaches represent areas for further research.

6 Ethical Considerations

Some part of sentences were rephrased using chatGPT. Since we used publicly available datasets no other considerations were required.

References

Junyi Ao, Ziqiang Zhang, Long Zhou, Shujie Liu, Haizhou Li, Tom Ko, Lirong Dai, Jinyu Li, Yao Qian, and Furu Wei. 2022. Pre-training transformer decoder for end-to-end asr model with unpaired speech data. *arXiv preprint arXiv:2203.17113*.

Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. 2020. wav2vec 2.0: A framework for self-supervised learning of speech representations. *Advances in neural information processing systems*, 33:12449–12460.

Carlos Busso, Zhigang Deng, Serdar Yildirim, Murtaza Bulut, Chul Min Lee, Abe Kazemzadeh, Sungbok Lee, Ulrich Neumann, and Shrikanth Narayanan. 2004. Analysis of emotion recognition using facial expressions, speech and multimodal information. In *Proceedings of the 6th international conference on Multimodal interfaces*, pages 205–211.

Houwei Cao, David G Cooper, Michael K Keutmann, Ruben C Gur, Ani Nenkova, and Ragini Verma. 2014. Crema-d: Crowd-sourced emotional multimodal actors dataset. *IEEE transactions on affective computing*, 5(4):377–390.

Adil CHAKHTOUNA, Sara SEKKATE, and ADIB Abdellah. 2024. Unveiling embedded features in wav2vec2 and hubert msodels for speech emotion recognition. *Procedia Computer Science*, 232:2560–2569.

Li-Wei Chen and Alexander Rudnicky. 2023. Exploring wav2vec 2.0 fine tuning for improved speech emotion recognition. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE.

Karine Durand, Mathieu Gallay, Alix Seigneuric, Fabrice Robichon, and Jean-Yves Baudouin. 2007. The development of facial emotion recognition: The role of configural information. *Journal of experimental child psychology*, 97(1):14–27.

Tiantian Feng and Shrikanth Narayanan. 2023. Peft-ser: On the use of parameter efficient transfer learning approaches for speech emotion recognition using pre-trained speech models. In *2023 11th International Conference on Affective Computing and Intelligent Interaction (ACII)*, pages 1–8. IEEE.

Yuan Gao, Chenhui Chu, and Tatsuya Kawahara. 2023. Two-stage finetuning of wav2vec 2.0 for speech emotion recognition with asr and gender pretraining. In *Proc. Interspeech*.

Yuan Gao, Hao Shi, Chenhui Chu, and Tatsuya Kawahara. 2024. Enhancing two-stage finetuning for speech emotion recognition using adapters. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 11316–11320. IEEE.

Ioulia Grichkovtsova, Michel Morel, and Anne Lacheret. 2012. The role of voice quality and prosodic contour in affective speech perception. *Speech Communication*, 54(3):414–429.

Wei-Ning Hsu, Benjamin Bolte, Yao-Hung Hubert Tsai, Kushal Lakhotia, Ruslan Salakhutdinov, and Abdelrahman Mohamed. 2021. Hubert: Self-supervised speech representation learning by masked prediction of hidden units. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29:3451–3460.

Nineli Lashkarashvili, Wen Wu, Guangzhi Sun, and Philip Woodland. 2024. Parameter efficient finetuning for speech emotion recognition and domain adaptation. pages 10986–10990.

Yingting Li, Ambuj Mehrish, Rishabh Bhardwaj, Navonil Majumder, Bo Cheng, Shuai Zhao, Amir Zadeh, Rada Mihalcea, and Soujanya Poria. 2023.

Evaluating parameter-efficient transfer learning approaches on sure benchmark for speech understanding. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE.

Jingsai Liang. 2022. Confusion matrix: Machine learning. *POGIL Activity Clearinghouse*, 3(4).

Steven R Livingstone and Frank A Russo. 2018. The ryerson audio-visual database of emotional speech and song (ravdess): A dynamic, multimodal set of facial and vocal expressions in north american english. *PloS one*, 13(5):e0196391.

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

M. Kathleen Pichora-Fuller and Kate Dupuis. 2020. Toronto emotional speech set (TESS).

Srinivasan Ramakrishnan and Ibrahiem MM El Emary. 2013. Speech emotion recognition approaches in human computer interaction. *Telecommunication Systems*, 52:1467–1478.

David Rybach, Christian Gollan, Ralf Schluter, and Hermann Ney. 2009. Audio segmentation for speech recognition using segment features. In *2009 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 4197–4200. IEEE.

Shamane Siriwardhana, Andrew Reis, Rivindu Weerasekera, and Suranga Nanayakkara. 2020. Jointly fine-tuning" bert-like" self supervised models to improve multimodal speech emotion recognition. *arXiv preprint arXiv:2008.06682*.

Wenxuan Wu, Xueyuan Chen, Xixin Wu, Haizhou Li, and Helen Meng. 2024. Target speech extraction with pre-trained av-hubert and mask-and-recover strategy. *arXiv preprint arXiv:2403.16078*.

Shu-wen Yang, Po-Han Chi, Yung-Sung Chuang, Cheng-I Jeff Lai, Kushal Lakhotia, Yist Y Lin, Andy T Liu, Jiatong Shi, Xuankai Chang, Guan-Ting Lin, et al. 2021. Superb: Speech processing universal performance benchmark. *arXiv preprint arXiv:2105.01051*.