Ensemble of Large Self-Supervised Transformers for Improving Speech Emotion Recognition

Mrunal Prakash Gavali

Department of Computer Science,
California State University,
Northridge, CA, USA
E-mail: mrunal-prakash.gavali.927@csun.edu

Abhishek Verma*

Department of Computer Science,
California State University,
Northridge, CA, USA
E-mail: abhishek.verma@csun.edu
* corresponding author

Abstract:
Speech emotion recognition (SER) is a challenging and active field of collaborative, social robotics to improve human-robot interaction (HRI) and affective computing as a feedback mechanism. More recently self-supervised learning (SSL) approaches have become an important method for learning speech representations. We present results of experiments on the challenging large-scale speech emotion RAVDESS dataset. Six very large state-of-the-art self-supervised learning transformer models were trained on the speech emotion dataset. Wav2vec2.0-XLSR-53 was the most successful of the six level-0 models and achieved classification accuracy of 93%. We propose majority voting ensemble models that combined three and five level-0 models. The five-model and three-model majority voting ensemble models achieved 96.88% and 96.53% accuracy respectively and thereby significantly outperformed the best level-0 model and surpassed the state-of-the-art.

Keywords: Speech Emotion Recognition; self-supervised learning; Emotion AI; transformers; speech processing; acoustic features

Reference to this paper should be made as follows: Gavali, M. P. and Verma, A. (202X) ‘Ensemble of Large Self-Supervised Transformers for Improving Speech Emotion Recognition,’ *International Journal of Data Mining, Modeling, and Management*, Vol. x, No. x, pp.xxx–xxx.

Biographical notes: Mrunal Prakash Gavali received her MS degree in Software Engineering from California State University, Northridge. Her research interests are in deep learning, machine learning, data science, natural language processing (NLP), information retrieval, prompt engineering, ensemble learning, and speech emotion recognition.

Abhishek Verma received his Ph.D. in Computer Science from New Jersey Institute of Technology, NJ, USA. His research interests are in data science, big data analytics, machine learning, and deep learning on big datasets.

1 Introduction

Emotion recognition is a challenging and popular field of collaborative, social robotics to improve human-robot interaction (HRI) and affective computing as a feedback mechanism. However, human emotional expression and experience is complex, subjective and contextually heterogenous with endless variations. Thus, the innate challenges of modeling emotional representations in speech signals are augmented in supervised learning where annotated labels for a large dataset are required to improve performance. Self-supervised learning approach, specifically contrastive learning resolves this issue of dealing with the lack of available annotated data, either for low-resource language or an overall lack of available data for the downstream task itself by learning the generic representation from large-scale data without any manual external supplementary labeling. Although Convolutional Neural Networks (CNN) learn representations with fewer parameters than attention-based transformer models, CNNs are spatially local. Therefore, attention-based Transformer models are adopted to learn global representations.

Hence, this research paper investigates the different model architectures for downstream task of speech emotion recognition (SER) using self-supervised learning (SSL), and attention-based transformer modules by fine tuning of automatic speech recognition (ASR) models. Moreover, the best performing SSL models would be further chosen to implement majority voting ensemble to investigate improvement in performance of emotion prediction.

This research paper is organized as follows. Section 2 provides a thorough overview, beginning with an introduction to speech emotion recognition (SER) and the relevant deep learning concepts and techniques in this research area. Section 3 introduces different self-supervised learning (SSL) models used in this research. This section also presents description of majority voting ensemble using the SSL models. Section 4 covers the description of the RAVDESS dataset used in the experiments. Section 5 presents the experiment setup used for performing the experiments including development tools and data pre-processing steps. Section 6 presents the experiment results such as confusion matrices, classification reports, tables for model performance comparison, evaluation metrics like weighted accuracy and unweighted accuracy with additional context describing their interpretation. This section also includes information about model configurations used in the experiments and then discusses findings from the experiments by summarizing the model performance of several models. Finally, Section 7 presents the conclusion on speech emotion recognition using self-supervised learning (SSL) approach and the majority voting ensembles along with possible future work.

2 Related Work

This section provides a summary of recent research conducted on speech emotion recognition (SER) with various deep learning algorithms. Ultimately, this paper seeks to provide a generalizable comparison for self-supervised learning (SSL) algorithms for speech emotion recognition on RAVDESS dataset specifically using majority voting ensemble.

Han et al. (2021) proposed a parallel network of Resnet-CNN-Transformer Encoder for SER on RAVDESS dataset and achieved average test accuracy of 80.89% after training the model for 500 epochs by repeating the experiments five times. The human accuracy for speech emotion recognition on RAVDESS dataset is 67%, which indicates that SER for RAVDESS is complex even for human evaluators Livingstone & Russo (2018).

In another research, Bautista et al. (2022) used combination of a CNN and attention based networks, running in parallel to classify emotions using speech modality on the RAVDESS dataset. The parallel hybrid model is used to model both the spatial as well as temporal features. The authors transformed the raw acoustic speech data from RAVDESS dataset into Mel-Spectrograms and applied different acoustic data augmentation techniques like Additive White Gaussian Noise (AWGN), Room Impulse Response (RIR), SpecAugment, and Tanh distortion techniques to generalize model representation. They conclude that supervised parallel hybrid model architectures perform better with a substantially lower number of training parameters in comparison to the standalone CNN or attention based models as well as hybrid architectures that combine CNN layers within time-distributed wrappers stacked on attention-based modules Zenkov (2020).

Luna-Jiménez et al. (2021) presents multimodal emotion recognition by combining two modalities, speech and facial expressions with a late fusion strategy on the RAVDESS dataset using transfer learning. Transfer learning uses pre-existing knowledge captured by a supervised pre-trained models like the CNN-14 of the PANNs Kong et al. (2020) framework for SER task to improve performance through fine-tuning. Whereas the facial expressions were classified using transfer learning approach with a pre-trained spatial transformer model on both saliency maps and facial images, which is then followed by a Bi-LSTM that incorporates an attention mechanism. By combining these models trained on different modalities using a late fusion strategy, the authors were able to produce higher accuracy of 80.08% on RAVDESS dataset for subject-wise 5-fold cross-validation when compared to if the models were executed separately. For instance, the authors reported the SER accuracy using speech modality of only 76.58% on RAVDESS, without multimodal fusion strategy. The deep learning research in the speech domain has majorly adopted a pre-training approach with self-supervised learning of speech representations from raw speech acoustic data over using supervised learning architectures. When fine-tuned on standard benchmarks, the self-supervised pretraining approach with wav2vec2 as feature extractor has simplified and improved performance, especially in a low-data setting as demonstrated in Luna-Jiménez et al. (2022) where it achieved 81.82% accuracy on RAVDESS for speech emotion recognition task by incorporating self-supervised model like wav2vec2-xlsr + multilayer perceptron (MLP) instead of supervised models like CNNs or PANNs used in their previous work that gave 76.58% accuracy.

However, due to the nature of self-supervised pretraining, the audio encoders like Wav2Vec2.0 and HuBERT lack suitable decoding to transform speech representations into functional outputs, which necessitates fine-tuning for ASR and acoustic classification tasks as shown in IBM AI research Morais et al. (2022) with downstream-upstream paradigm on IEMOCAP dataset for End-to-End (E2E) SER downstream task. Moreover, Atmaja & Sasou (2022) presents comprehensive research on five emotion datasets in different languages with 20 different deep learning models. The prior research provide a benchmark for our research expectations regarding the performance of self-supervised learning pretrained models for SER task.

Besides, Wang et al. (2021) also presents a comprehensive fine-tuning of Wav2vec2.0 and HuBERT pretrained ASR models for other downstream tasks like speech emotion recognition (SER), spoken language understanding (SLU) and speaker verification (SV). The authors achieved competitive weighted accuracy (WA) results of 79.58% and 73.01% on speaker-dependent setting and on speaker-independent setting respectively for SER task on IEMOCAP dataset. Their research illustrates the strength of the fine-tuned SSL models for learning speech representations like audio prosody, voice prints and semantics effectively on a large dataset.

Since fine-tuning self-supervised models is extremely complex requiring skilled practitioner, OpenAI released Whisper architecture model family in 2022 based on large-scale weak supervised training on a large and diverse labelled dataset of 680,000 hours of multilingual and multitask supervision with concentration on zero-shot transfer to demonstrate improvement in the robustness and generalization of speech recognition to predict massive amounts of audio transcriptions on the internet without any finetuning Radford et al. (2022). Whisper showed competitive results on standard benchmarks with previous fully supervised research results in a zero-shot transfer setting, even outperforming XLS-R in Multilingual speech recognition performance on Multilingual LibriSpeech (MLS) dataset but is still substantially behind XLS-R on Multilingual VoxPopuli datasest.

The self-supervised fine-tuned models are adroit at finding patterns that are stiff and forged and do not generalize well to other datasets or distributions. This can be avoided by using large-scale weakly supervised Whisper for speech recognition. Vásquez-Correa (2023) presents a comprehensive comparison of wav2vec2.0 and Whisper for speech recognition in a privacy preserving federated learning setting.

Note that model fine-tuning is focused on adapting a pre-trained model to a new dataset or a task by fine-tuning the model on a new labeled dataset. This would involve adjusting most of the model parameters. Compared to model-tuning, transfer learning freezes certain layers of the pre-trained model and only modifies some model parameters. In another research by Shirian & Guha (2020), a Graph Convolution Network (GCN)-based architecture is proposed to model speech signal as a cycle graph or a line graph to solve SER task using graph classification. Their research uses IEMOCAP and MSP-IMPROV databases to take advantage of graph signal processing results. This approach achieves state-of-the-art results with substantially less parameters of only 30K, which is ideal for deploying this model on resource-constrained devices.

Besides, using only speech modality from the audio-visual datasets, some recent research works introduce a generalized modality-agnostic approach to emotion recognition that can adapt across different modalities like text, speech in audio data, facial expressions in video data and body gestures using motion sensors data by modeling dynamic data as structured graphs. An example of such research proposed by Shirian et al. (2022) is the compact Learnable Graph Inception Network (L-GrIN) for joint graph learning and classification to cooperatively learn emotion recognition and the underlying graph structure identification in the dynamic data of five benchmark emotion recognition datasets.Vaiani et al. (2022) presents another recent model-agnostic example called ViPER with a multimodal architecture to leverage a modality-agnostic late fusion transformer-based network to merge video, textual and audio labels for human emotion recognition.

Most SER approaches are trained on centralized architecture raising ethical concerns about the privacy of the personal data. To address this concern for persevering privacy, a distributed machine learning approach like federated learning could be useful through decentralization of personal data. Tsouvalas et al. (2022) employs SSL models in federated learning (FL) paradigm to utilize both labeled and unlabeled on-device data. The authors evaluated their experiments on IEMOCAP benchmark to demonstrate that their approach

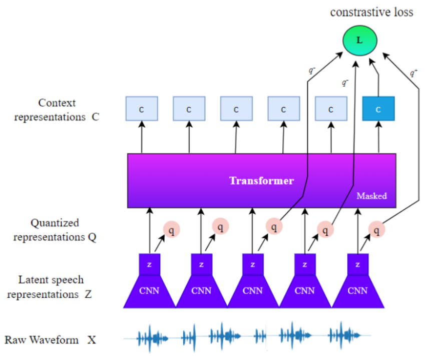


Figure 1 Wav2vec2.0 model architecture. The raw acoustic speech signals are mapped to latent speech representations and sent as input to a transformer module to produce context representations Morais et al. (2022).

of using SSL in semi-supervised federated learning setting generalizes well even under sparse labelled data and highly non-independent and non-identically distributed setting and achieves better accuracy than fully supervised federated settings under the similar availability of labeled data.

3 Research Methodology

The SSL models used in this research share a common foundation in terms of their model architecture structure and function, which involves taking input of raw audio and outputs a vector representation. The variations amongst these SSL frameworks are found in the pre-training stage. This makes the study of different SSL models for SER task ideal for comparison as explored further in our research.

3.1 Wav2vec2.0

The wav2vec 2.0 large variant with 317 million parameter size as shown in Table 1 is chosen for our research. For the experiments in this research, the version of the Wav2vec2.0 large model, pretrained and fine-tuned on 960 hours of only English language speech datasets - Libri-Light Kahn et al. (2020) and Librispeech Panayotov et al. (2015) with self-training objective Xu et al. (2021) on 16kHz sampled speech audio were obtained from the Hugging Face repository Hugging Face (2021c). To prepare for the pretext task, a segment of the latent speech representation obtained from the output of the feature extractor is concealed as well as substituted with a trained feature vector that is common to all masked components. This modified representation is then used as input to the transformer. Following that, a quantization module is used that uses product quantization to convert the unmasked, unmodified latent output of the feature encoder into discrete values $q1, q2, q3..., qt$ as shown in Figure 1. The goal of the contrastive task is to identify the right quantized latent speech representation (qt) for a masked segment of the input from a pool of $K+1$ candidates, where one of them is the true qt, and the remaining K are distractors that are uniformly sampled from the quantization of the other masked outputs in the same utterance.

Table 1 Summary of SSL model variants according to their pretraining dataset and parameter size.

Model	**Pretraining dataset**	**Parameter Size**
Wav2vec2.0 large model Hugging Face (2021*c*)	53k hours of raw English speech data sampled from audiobooks	317M
Wav2vec2-large-XLSR-53 Hugging Face (2021*d*)	56k hours of unlabeled multilingual speech datasets, sourced from Multilingual LibriSpeech, CommonVoice and BABEL.	317M
HuBERT base model Hugging Face (2021*a*)	960 hours of LibriSpeech audio	95M
HuBERT x-large model Hugging Face (2021*b*)	60k hours of Libri-light audio	1B
UniSpeech large model Hugging Face (2021*e*)	Labeled: 1350 hours English (en)	317M
Unispeech large multilingual model Hugging Face (2021*f*)	Labeled: 1350 hours English (en) + 353 hours French (fr) + 168 hours Spanish (es) + 90 hours Italian (it) = 1961 hours	317M

3.2 Wav2vec2-large-XLSR-53

XLSR is an acronym for cross-lingual speech representations. As the name suggests, the XLSR model is designed for multilingual speech recognition and utilizes wav2vec 2.0 as its foundation but it is larger in terms of both languages and model size. This model is pretrained from the raw speech waveform in numerous languages as shown in Figure 2. It trains by solving a contrastive task on masked latent speech representations, and it earns a shared quantization of the latents for different languages. After fine-tuning on labeled data, experiments have demonstrated that cross-lingual pretraining is more effective than monolingual pretraining. To achieve this, the model employs a shared quantization module over feature encoder representations to produce multilingual quantized speech units. Those units are then used as targets for a transformer that is trained via contrastive learning. Through this approach, the model can share discrete tokens across multiple languages, effectively creating connections between them. The XLSR approach needs only raw unlabeled acoustic speech data in different languages.

3.3 HuBERT

Self-supervised speech representation learning using HuBERT framework incorporates an offline clustering step to generate aligned target labels for a prediction loss similar to BERT. The primary method used by Hubert to learn speech sequences involves partially masking speech frame features. This is achieved through a series of steps. Initially, feature sequences and pseudo tags are generated for a given audio content using a CNN and K-means clustering. Following this, the feature sequence is masked by randomly selecting the starting index as a time step of K, masking the feature sequence with length of S steps, and inputting the resulting masked feature sequence into the transformers to generate contextual feature representations and to compare with the generated pseudo-labels. This allows for the prediction of the masked audio features.

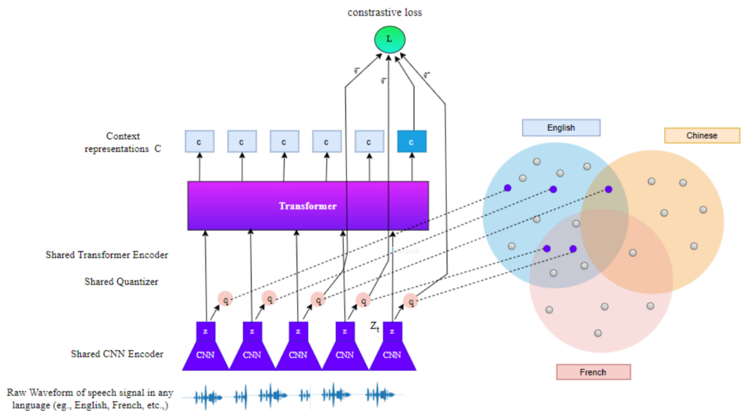

constrastive loss
L
q^-
q^-
q^+
Context representations C
c
c
c
c
c
c
Transformer
Shared Transformer Encoder
Shared Quantizer
q
q
q
q
q
Z_t
z
CNN
Shared CNN Encoder
Raw Waveform of speech signal in any language (eg., English, French, etc.,)
English
Chinese
French

Figure 2 Wav2vec2.0-XLSR model architecture Conneau et al. (2021).

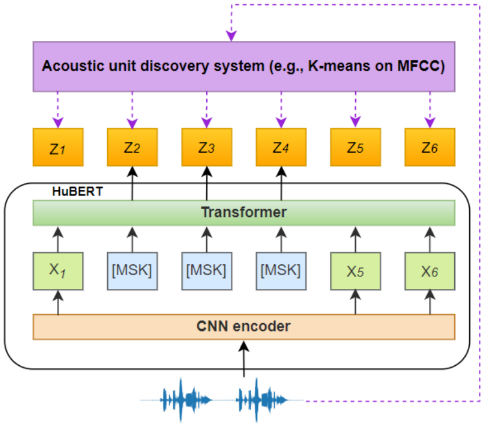


Figure 3 HuBERT model architecture Hsu et al. (2021).

In wav2vec2 framework, a portion of the CNN feature extractor's output is masked prior to inputting it into the transformer module for pre-training. However, as a replacement for the quantization module utilized in wav2vec2.0 framework, HuBERT predicts hidden cluster assignments of the masked timesteps, which are presented with Zi as shown in Figure 3. These pseudo-labels are based on following two approaches. During the first step of pre-training objective, the KNN clusters are constructed using the Mel-frequency cepstral coefficients (MFFCs) of the training data. However, for consequent clustering steps, latent representations from an intermediate layer of the HuBERT transformer encoder, from the previous iteration, are re-used. The training process alternates between two steps: a clustering step to create pseudo-targets, and a prediction step where the model tries to guess these targets at masked positions. HuBERT relies primarily on the consistency of the unsupervised clustering step rather than the intrinsic quality of the assigned cluster labels.

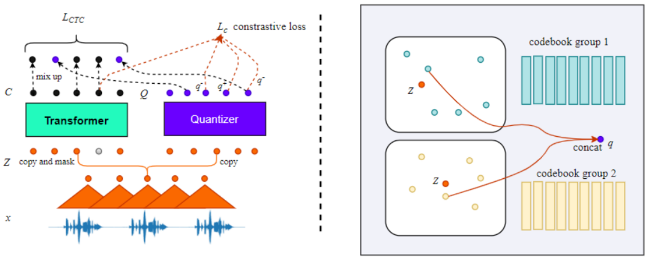


Figure 4 UniSpeech model architecture Wang et al. (2021).

3.4 UniSpeech

Wang et al. (2021) proposed UniSpeech framework as a unified approach to learn phonetically aware contextual representations. Given a set of datasets:

- X set of labeled datasets from a high-resource domains at a large scale.
- Another set of unlabeled datasets, Y from low-resource domains.
- A final set of labeled datasets Z from the same low-resource setting.

UniSpeech framework follows the principles of wav2vec2.0. The model structure comprises of the following components - a feature encoder to extract latent speech representations, a transformer network to learn contextual representations and finally, latent representations are discretized with the help of a quantizer as shown in Figure 4.

The objective is to leverage both X and Y data to learn robust representations by pre-training the model. Subsequently, the feature extractor is frozen and the transformer part is fine-tuned on a small portion of the labeled low-resource data (Z) using the multitask learning (MLT) method, which consists of three main components. Firstly, a phonetic Connectionist Temporal Classification (CTC) loss Graves et al. (2006) is applied on labeled high-resource data (X) to fulfill the first learning objective. The remaining two objectives are achieved by adopting the same technique as in wav2vec2.0, which involve recognizing the right quantized latent speech representation from a set of distractors, on both X and Y data. This process involves masking a specific portion of the feature extractor's output.

3.5 Proposed Majority Voting Ensemble Models

We propose several voting ensembles of different SSL models created using the best models, which gave high accuracy of 90% and above without overfitting in combinations of three and five as shown in Figure 5 and Figure 6 respectively. Each of the individual SSL models in the ensemble was trained individually with the same data, and evaluated on the same test data. Then, the predictions from the base models were combined using majority voting. An odd number of member SSL models in the ensemble were chosen to avoid ties across predictions.

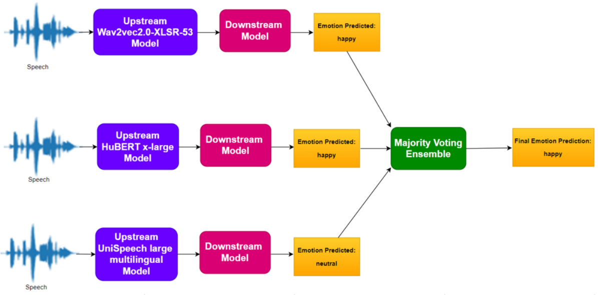

Upstream Wav2vec2.0-XLSR-53 Model
Downstream Model
Emotion Predicted: happy
Speech
Upstream HuBERT x-large Model
Downstream Model
Emotion Predicted: happy
Majority Voting Ensemble
Final Emotion Prediction: happy
Speech
Upstream UniSpeech large multilingual Model
Downstream Model
Emotion Predicted: neutral
Speech

Figure 5 Architecture of the proposed majority voting ensemble of 3 models - Wav2vec2.0-XLSR-53, HuBERT x-large and UniSpeech large multi-lingual Model.

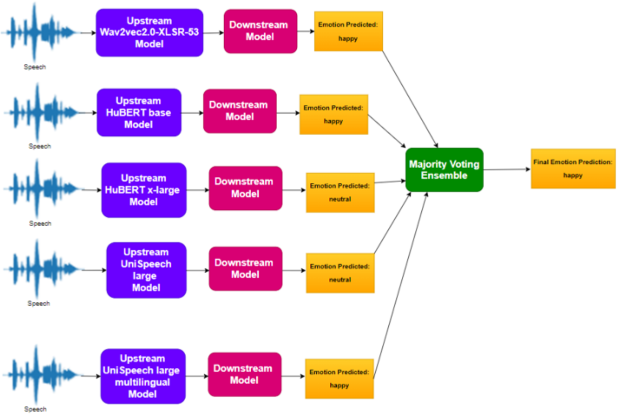

Upstream Wav2vec2.0-XLSR-53 Model
Downstream Model
Emotion Predicted: happy
Speech
Upstream HuBERT base Model
Downstream Model
Emotion Predicted: happy
Speech
Upstream HuBERT x-large Model
Downstream Model
Emotion Predicted: neutral
Majority Voting Ensemble
Final Emotion Prediction: happy
Speech
Upstream UniSpeech large Model
Downstream Model
Emotion Predicted: neutral
Speech
Upstream UniSpeech large multilingual Model
Downstream Model
Emotion Predicted: happy
Speech

Figure 6 Architecture of the proposed majority voting ensemble of 5 models - Wav2vec2.0-XLSR-53, HuBERT base, HuBERT x-large, UniSpeech large and UniSpeech large multi-lingual.

4 Description of RAVDESS Dataset

The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS) dataset Livingstone & Russo (2018) is an audio-visual dataset with speech as well as song files in a neutral North American accent. From the original dataset, only speech audio data consists of 1,440 audio files recorded in 16-bit, 48kHz .wav format. 24 professional actors, equally divided into 12 males and 12 females, vocalized two lexically matching statements, where each actor recorded 60 trials, resulting in 1,440 files in total. The speech emotions in the dataset are classified into eight categories, including neutral, calm, happy, sad, angry, fear, surprise, and disgust. In the dataset, each emotion is expressed at two levels of intensity, normal and strong, with the exception of 'neutral' emotion, which does not have strong intensity. The 1,440 speech data is split into train and validation sets in 80:20 ratio with 1,152 and 288 speech audio files in the train and validation sets respectively.

5 Experiment Setup

5.1 Hardware and Software Setup

- Python programming language is chosen as it's the language of choice for almost all the audio processing and speech deep learning packages with access to well documented libraries.

- Huggingface transformers library Wolf et al. (2020) is a cutting-edge package in Python containing open-source implementations of the state-of-the-art transformer models for different modalities in PyTorch and other deep learning libraries.

- Huggingface datasets library Lhoest et al. (2021) is simple, fast and reproducible data pre-processing library for public and custom datasets in different formats to efficiently prepare the dataset for inspection, model evaluation, and training.

- PyTorch is an open-source machine learning (ML) framework built as Python wrapper for Torch library. PyTorch was originally developed by Facebook's artificial intelligence research group (FAIR) and is now part of the Linux foundation. It provides a high-level feature of GPU accelerated tensor computation for computer vision, natural language processing, acoustic deep learning and graph learning applications. Our research specifically used the Torchaudio library, which is a powerful library for audio and signal processing in PyTorch framework Paszke et al. (2019).

- Weights & Biases for MLOps (WandB) is a real-time automated deep learning experiment tracking tool and can be used seamlessly with popular deep learning frameworks such as Pytorch and other deep learning frameworks Biewald (2020).

- Librosa is an opensource Python library for audio and music processing McFee et al. (2015).

- Scikit-learn is an opensource, commercially usable - BSD license, Python library, which includes numerous efficient tools for machine learning and predictive data analysis Pedregosa et al. (2011).

- Pandas is a data frame-based library used for performing disk read and write operations, and for importing the train and test dataset from a csv file as well as saving and loading predictions of different SSL models for majority voting ensemble Wes McKinney (2010), pandas development team (2020).

- NumPy library is used for performing various matrix operations in the speech pre-processing step Harris et al. (2020).

- Matplotlib for plotting the confusion matrices to compare the prediction of SSL models and majority voting ensemble Hunter (2007).

- Google Colab is a hosted Jupyter notebook environment, which includes numerous built-in deep learning libraries as well as GPU and TPU support. To conduct experiments in our research, Google Colab Pro Plus subscription is used to accelerate training and satisfy the demands of high computational requirements for RAM and GPU, by leveraging NVIDIA Tesla T4 and NVIDIA A100-SXM4-40GB.

5.2 Data Preprocessing, Parameter Setup, and Implementation

The speech data in the SER RAVDESS dataset is a collection of the .wav audio files grouped in different folders based on speakers instead of emotions. So, the files are rearranged according to the naming convention used for the RAVDESS dataset to group the different audio files as per different emotions labels. After this step, a csv file was created to specify the audio file path and its corresponding emotion label so that the data could be loaded for training by using a data loader without repeating the above steps each time before training. After the data is loaded using a data loader, the SER dataset is split into train and test csv files in 80:20 ratio.

To preprocess the speech audio data into our emotion classification model, the relevant Wav2vec2 assets pertaining to the language of the dataset need to be set up from the Hugging Face model cards for fine-tuning. A merge strategy, mean pooling mode, is utilized to handle the context representations in any audio length for concatenating those 3D representations into 2D representations. Note that three merge strategies namely, mean, sum, and max could be employed in pooling mode. The mean merge strategy is employed during the deep learning experiments in this paper to achieve better results for SER. Consequently, the configurator and the feature extractor from Hugging Face transformers need to be initiated for the same. The .wav or .mp3 audio files are read using Torchaudio library to resample the audio files to 16kHz, and to map each audio to the corresponding emotion label. This preprocessed data is then used for the training by emotion classification model based on the merge strategy by specifying the pooling mode type as ‘mean’.

The SSL models are trained using Hugging Face library, in particular its Trainer class API over writing a long boilerplate training loop in PyTorch to simplify the intricacies involved in writing a training loop workflow in a single line after specifying the hyperparameters used as training arguments when initializing the Trainer class. Thus, the Trainer API helps in developing an organized and clean codebase. Before proceeding to use the Hugging Face Trainer class API, the data collator is defined in the implementation as described below.

SSL models like XLSR-Wav2Vec2.0 are different to most NLP models in that the SSL models have a much larger input length than output length. For example, a sample of input length 50,000 has an output length of no more than 100. Given the large input sizes of SSL models, it is substantially more effective to dynamically pad the training batches. This approach involves padding all the training samples to the length of the longest sample in their batch, rather than the overall longest sample. Therefore, a special padding data collator is required when fine-tuning an SSL model.

To put it simply, the special padding data collator is different from the common data collators in that it applies separate padding functions to the input values and labels, taking advantage of XLSR-Wav2Vec2’s context manager. This is crucial because input and output are of different modalities in speech, and therefore require different padding functions. Similar to the common data collators, the padding tokens in the labels are set to 100 so that they do not affect the loss computation.

Next, the SSL model is imported from the pretrained checkpoint in the model cards in the Hugging Face hub using Hugging Face API. Subsequently, training arguments and the evaluation metrics are defined for the training to record the evaluation metrics between the training and validation dataset. Finally, the Hugging Face Trainer class API is used for training and the trained model checkpoint is saved after each 100 steps. The fine-tuned model is then evaluated on the test data to confirm that the model has learned the

Table 2 Model Configuration

Training Hyperparameter	**Value**
Train batch size	4
Evaluation batch size	4
Gradient accumulation steps	2
Evaluation strategy	steps
Number of training epochs	10
Save steps	100
Evaluation steps	100
Logging steps	100
Learning rate	1e-4
save_total_limit (number of checkpoints)	2
do_train	True
do_eval	True
do_predict	True

labels for the speech emotion recognition task by constructing the confusion matrices, classification reports and training graphs in weights and biases along with computing the weighted accuracy (WA) and unweighted accuracy (UA) across different SSL experiments. The fine-tuned model checkpoint with satisfactory performance on the dataset is also pushed to Hugging Face Hub private personal repository to inference later. Moreover, the predictions of the different SSL models on the test dataset are saved in the csv file. These best performing SSL model predictions are further used for majority voting ensemble in combinations of three or five models based on hard voting by taking mode of the emotion label predictions to further investigate improvement in performance.

5.3 Model Configuration

The training hyperparameters included a batch size of four, a learning rate of 0.0001 and a gradient accumulation of two steps as shown in Table 2. The training was performed for 10 epochs (1,400 steps) for all the SSL models. The parameters presented in Table 2 are passed to the Hugging Face trainer API as training arguments for training the SSL models from the Hugging Face Transformers library. The checkpoint is saved after every 100 steps.

5.4 Model Evaluation Metrics

This section describes the evaluation framework for all the different deep learning experiments. First, numerical confusion matrices along with normalized confusion matrices are computed using Scikit-learn library to illustrate the proportions of confusion rather than the exact numerical values and to provide a clear summary of where confusions among emotion labels arise. Secondly, the classification reports are constructed using Scikit-learn library to record the different evaluation metrics like recall, precision, F1-score and accuracy on a label-basis for the different SSL models. Thirdly, the overall accuracy metrics of the experiments is computed using two evaluation metrics - weighted accuracy (WA) and unweighted accuracy (UA), which are commonly used in the deep learning literature for SER and fit well on the dataset used for the experiments because of their skewness in the emotion label distributions. For eight emotion categories classification problem in the RAVDESS dataset, the precision score, recall score and F1 score is computed separately for each of the eight categories. This results in eight precision scores, one for each emotion category. Similarly, there are eight recall scores and eight F1 scores, for each emotion category in the RAVDESS dataset.

5.4.1 Unweighted accuracy (UA)

Unweighted accuracy (UA) represents the proportion of correctly predicted instances, calculated by dividing the total number of correct predictions by the total number of instances. Unweighted accuracy assigns equal importance to all classes, regardless of the dataset's class distribution.

5.4.2 Weighted accuracy (WA)

The computation of weighted accuracy involves taking the average of the fraction of correct predictions within each emotion category, which is determined by dividing the number of correctly predicted instances in an emotion category by the total number of instances in that category. "weighted averaged recall", i.e., weighted accuracy (WA) is shown for different emotion categories of the dataset in the classification report table.

6 Experiment Results and Discussion

6.1 Experiment Results of SSL Models

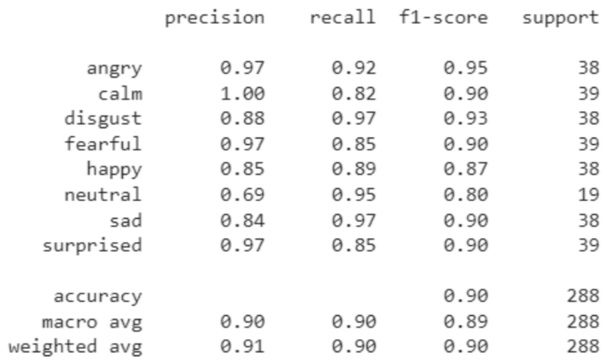
Each of the SSL-model along with its variants have been trained and evaluated to determine the best candidates for the SER downstream task on the RAVDESS dataset. These experiment results decide which models are further selected for majority voting ensemble.

6.1.1 Wav2vec2.0 Large Model

The classification report computed for Wav2vec2.0 large model (at 900 steps) is shown in Figure 7. From the classification report for test data, the weighted accuracy (WA) is 90% and the unweighted accuracy (UA) is 90% on the RAVDESS dataset. The numerical and normalized confusion matrices based on the classification report for test data using Wav2vec2.0 large model are presented in Figure 8. From the training graph in Figure 9, it can be inferred that the Wav2vec2.0-large model has converged and that there is no overfitting..

6.1.2 Wav2vec2.0-XLSR-53 Model

The classification report computed for Wav2vec2-XLSR-53 model is shown in Figure 10. From the classification report for test data, the weighted accuracy (WA) is 93% and the unweighted accuracy (UA) is 93% on the RAVDESS dataset. The numerical as well as normalized confusion matrices based on the classification report for test data using Wav2vec2-XLSR-53 model are presented in Figure 11. From the training graph in Figure 12, it can be inferred that the Wav2vec2.0-XLSR-53 model has converged and that there is no overfitting.

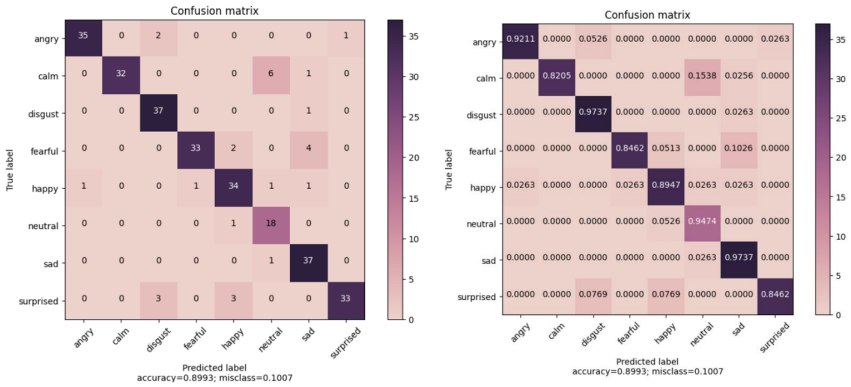


```
              precision    recall  f1-score   support

       angry       0.97      0.92      0.95        38
        calm       1.00      0.82      0.90        39
     disgust       0.88      0.97      0.93        38
     fearful       0.97      0.85      0.90        39
       happy       0.85      0.89      0.87        38
     neutral       0.69      0.95      0.80        19
         sad       0.84      0.97      0.90        38
   surprised       0.97      0.85      0.90        39

    accuracy                           0.90       288
   macro avg       0.90      0.90      0.89       288
weighted avg       0.91      0.90      0.90       288
```

Figure 7 Classification report for Wav2vec-2.0 large model on RAVDESS dataset (at 900 steps).



Confusion matrix

True label \ Predicted label	angry	calm	disgust	fearful	happy	neutral	sad	surprised
angry	35	0	2	0	0	0	0	1
calm	0	32	0	0	0	6	1	0
disgust	0	0	37	0	0	0	1	0
fearful	0	0	0	33	2	0	4	0
happy	1	0	0	1	34	1	1	0
neutral	0	0	0	0	1	18	0	0
sad	0	0	0	0	0	1	37	0
surprised	0	0	3	0	3	0	0	33

accuracy=0.8993; misclass=0.1007

(a)

Confusion matrix

True label \ Predicted label	angry	calm	disgust	fearful	happy	neutral	sad	surprised
angry	0.9211	0.0000	0.0526	0.0000	0.0000	0.0000	0.0000	0.0263
calm	0.0000	0.8205	0.0000	0.0000	0.0000	0.1538	0.0256	0.0000
disgust	0.0000	0.0000	0.9737	0.0000	0.0000	0.0000	0.0263	0.0000
fearful	0.0000	0.0000	0.0000	0.8462	0.0513	0.0000	0.1026	0.0000
happy	0.0263	0.0000	0.0000	0.0263	0.8947	0.0263	0.0263	0.0000
neutral	0.0000	0.0000	0.0000	0.0000	0.0526	0.9474	0.0000	0.0000
sad	0.0000	0.0000	0.0000	0.0000	0.0000	0.0263	0.9737	0.0000
surprised	0.0000	0.0000	0.0769	0.0000	0.0769	0.0000	0.0000	0.8462

accuracy=0.8993; misclass=0.1007

(b)

Figure 8 (a) Numerical confusion matrix for Wav2vec-2.0 large model on RAVDESS dataset (at 900 steps). (b) Normalized confusion matrix for Wav2vec-2.0 large model on RAVDESS dataset (at 900 steps).

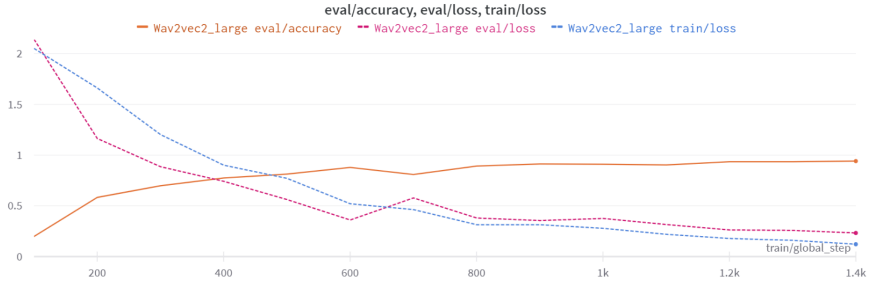
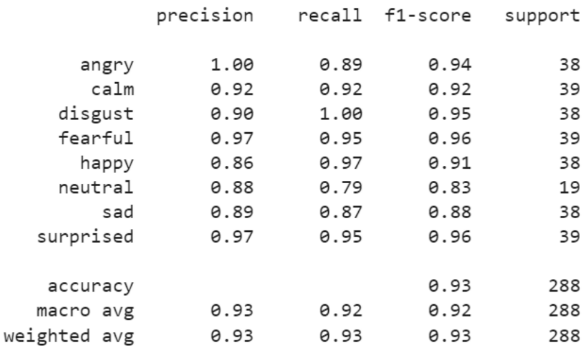


Figure 9 Training graph for Wav2vec-2.0 large model. X-axis shows train/global_step and Y-axis shows loss and accuracy.

6.1.3 HuBERT Base Model

The classification report computed for HuBERT base model is shown in figure 13. From the classification report for test data, the weighted accuracy (WA) is 92% and the unweighted

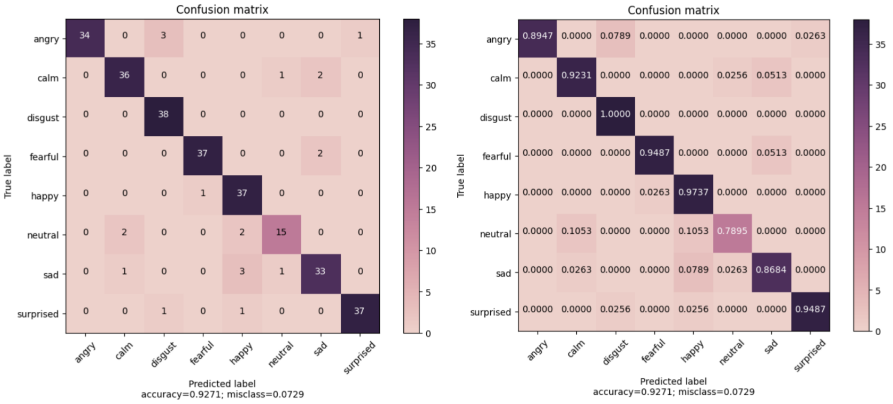


```
              precision    recall  f1-score   support

       angry       1.00      0.89      0.94        38
        calm       0.92      0.92      0.92        39
     disgust       0.90      1.00      0.95        38
     fearful       0.97      0.95      0.96        39
       happy       0.86      0.97      0.91        38
     neutral       0.88      0.79      0.83        19
         sad       0.89      0.87      0.88        38
   surprised       0.97      0.95      0.96        39

    accuracy                           0.93       288
   macro avg       0.93      0.92      0.92       288
weighted avg       0.93      0.93      0.93       288
```

Figure 10 Classification report for Wav2vec2-XLSR-53 model on RAVDESS dataset (at 1,100 steps).



(a) **(b)**

Figure 11 (a) Numerical confusion matrix for Wav2vec2-XLSR-53 model on RAVDESS dataset (at 1,100 steps). (b) Normalized confusion matrix for Wav2vec2-XLSR-53 model on RAVDESS dataset (at 1,100 steps).

accuracy (UA) is 92% on the RAVDESS dataset. The numerical as well as normalized confusion matrices based on the classification report for test data using HuBERT base model is presented in Figure 14. From the training graph in Figure 15, it can be inferred that there is divergence between training and validation losses after 900 steps, which is indicative of overfitting. Hence, we should use the trained model at 900 steps.

6.1.4 HuBERT X-large model

The classification report computed for HuBERT x-large model is shown in Figure 16. From the classification report for test data, the weighted accuracy (WA) is 90% and the unweighted accuracy (UA) is 90% on the RAVDESS dataset. The numerical and normalized confusion matrices based on the classification report for test data using HuBERT x-large model are presented in Figure 17. The confusion matrices show that HuBERT x-large model predicts sad and neutral emotion categories poorly in comparison to other emotions. From

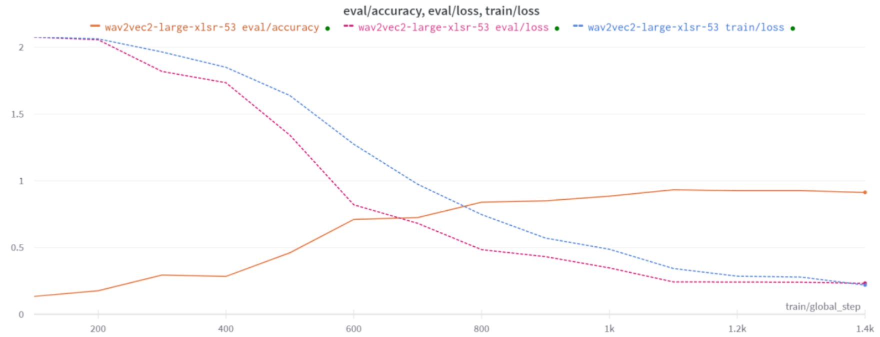
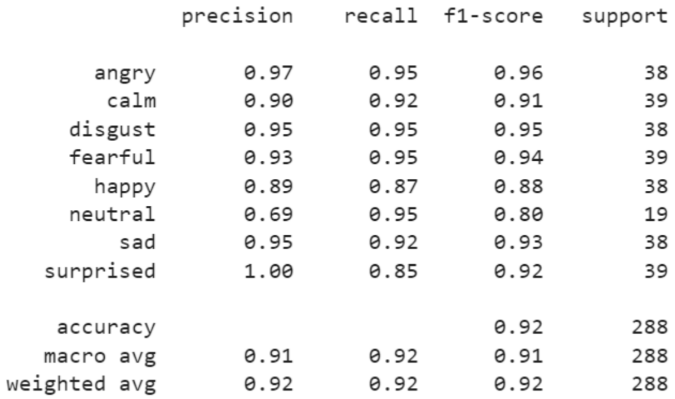


Figure 12 Training graph for Wav2vec2-XLSR-53 model. X-axis shows train/global_step and Y-axis shows loss and accuracy.



```
              precision    recall  f1-score   support

       angry       0.97      0.95      0.96        38
        calm       0.90      0.92      0.91        39
     disgust       0.95      0.95      0.95        38
     fearful       0.93      0.95      0.94        39
       happy       0.89      0.87      0.88        38
     neutral       0.69      0.95      0.80        19
         sad       0.95      0.92      0.93        38
   surprised       1.00      0.85      0.92        39

    accuracy                           0.92       288
   macro avg       0.91      0.92      0.91       288
weighted avg       0.92      0.92      0.92       288
```

Figure 13 Classification report for HuBERT base model on RAVDESS dataset (at 900 steps).

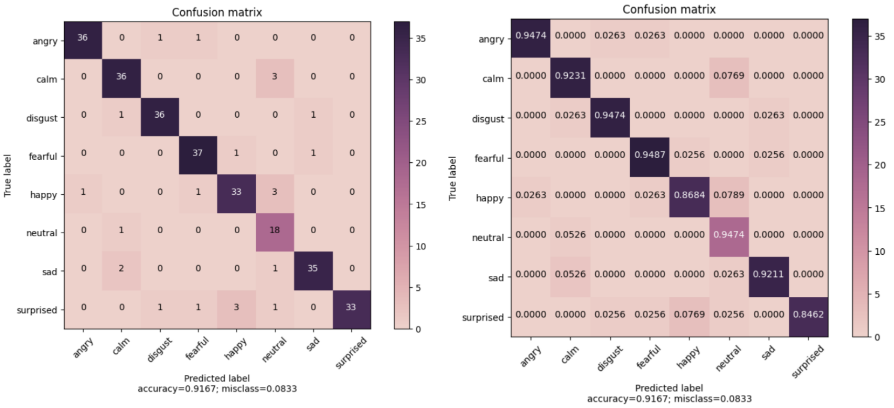
the training graph in Figure 18, it can be inferred that there is divergence between training and validation losses after 600 steps. Hence, we should use the trained model at 600 steps.

6.1.5 UniSpeech Large Model

The classification report computed for UniSpeech-large model is shown in Figure 19. From the classification report for test data, the weighted accuracy (WA) is 91% and the unweighted accuracy (UA) is 91% on the RAVDESS dataset. The numerical and normalized confusion matrices based on the classification report for test data using UniSpeech-large model are presented in Figure 20. The confusion matrices show that UniSpeech-large model predicts sad emotion category poorly in comparison to other emotions. From the training graph in Figure 21, it can be inferred that there is divergence between training and validation losses after 700 steps. Hence, we should use the trained model at 700 steps.

6.1.6 UniSpeech Large Multi-Lingual Model

The classification report computed for UniSpeech large multilingual model is shown in Figure 22. From the classification report for test data, the weighted accuracy (WA) is 92%



True label \ Predicted label	angry	calm	disgust	fearful	happy	neutral	sad	surprised
angry	36	0	1	1	0	0	0	0
calm	0	36	0	0	0	3	0	0
disgust	0	1	36	0	0	0	1	0
fearful	0	0	0	37	1	0	1	0
happy	1	0	0	1	33	3	0	0
neutral	0	1	0	0	0	18	0	0
sad	0	2	0	0	0	1	35	0
surprised	0	0	1	1	3	1	0	33

True label \ Predicted label	angry	calm	disgust	fearful	happy	neutral	sad	surprised
angry	0.9474	0.0000	0.0263	0.0263	0.0000	0.0000	0.0000	0.0000
calm	0.0000	0.9231	0.0000	0.0000	0.0000	0.0769	0.0000	0.0000
disgust	0.0000	0.0263	0.9474	0.0000	0.0000	0.0000	0.0263	0.0000
fearful	0.0000	0.0000	0.0000	0.9487	0.0256	0.0000	0.0256	0.0000
happy	0.0263	0.0000	0.0000	0.0263	0.8684	0.0789	0.0000	0.0000
neutral	0.0000	0.0526	0.0000	0.0000	0.0000	0.9474	0.0000	0.0000
sad	0.0000	0.0526	0.0000	0.0000	0.0000	0.0263	0.9211	0.0000
surprised	0.0000	0.0000	0.0256	0.0256	0.0769	0.0256	0.0000	0.8462

(a) **(b)**

Figure 14 (a) Numerical confusion matrix for HuBERT base model on RAVDESS dataset (at 900 steps). (b) Normalized confusion matrix for HuBERT base model on RAVDESS dataset (at 900 steps).

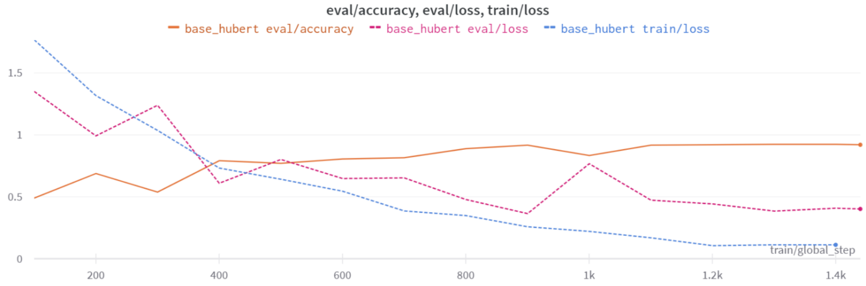
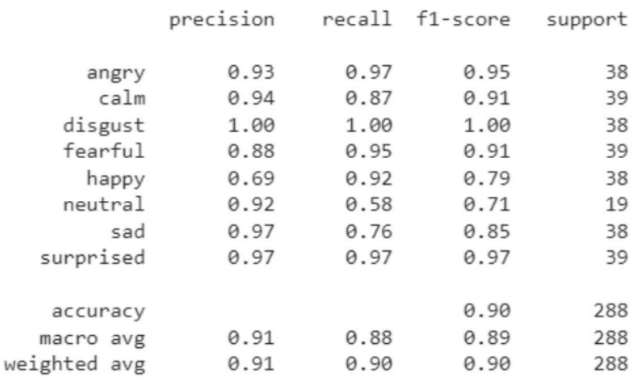


Figure 15 Training graph for HuBERT base model. X-axis shows train/global_step and Y-axis shows loss and accuracy.

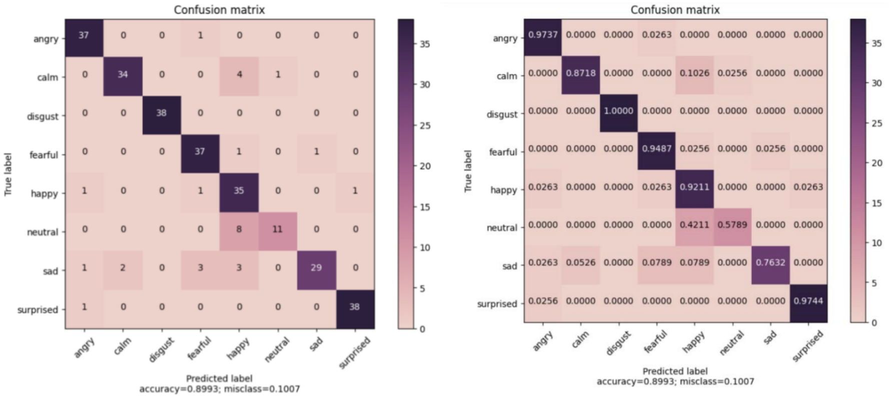


```
              precision    recall  f1-score   support

       angry       0.93      0.97      0.95        38
        calm       0.94      0.87      0.91        39
     disgust       1.00      1.00      1.00        38
     fearful       0.88      0.95      0.91        39
       happy       0.69      0.92      0.79        38
     neutral       0.92      0.58      0.71        19
         sad       0.97      0.76      0.85        38
   surprised       0.97      0.97      0.97        39

    accuracy                           0.90       288
   macro avg       0.91      0.88      0.89       288
weighted avg       0.91      0.90      0.90       288
```

Figure 16 Classification report for HuBERT x-large model on RAVDESS dataset (at 600 steps).



(a) **(b)**

Figure 17 (a) Numerical confusion matrix for HuBERT x-large model on RAVDESS dataset (at 600 steps). (b) Normalized confusion matrix for HuBERT x-large model on RAVDESS dataset (at 600 steps).

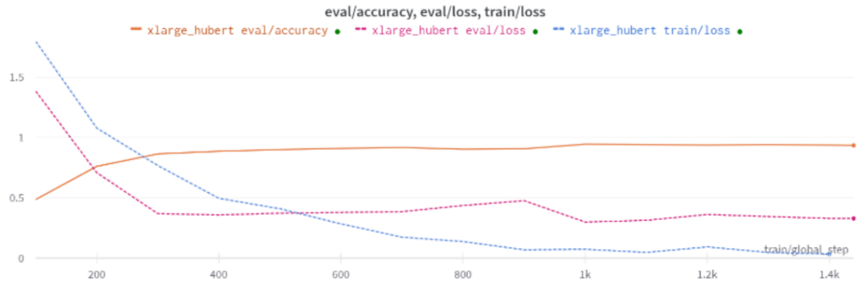
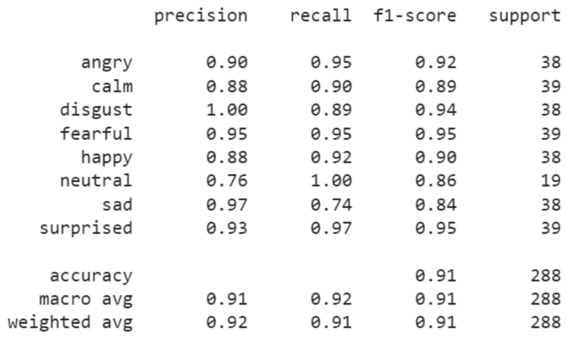


Figure 18 Training graph for HuBERT x-large model. X-axis shows train/global_step and Y-axis shows loss and accuracy.



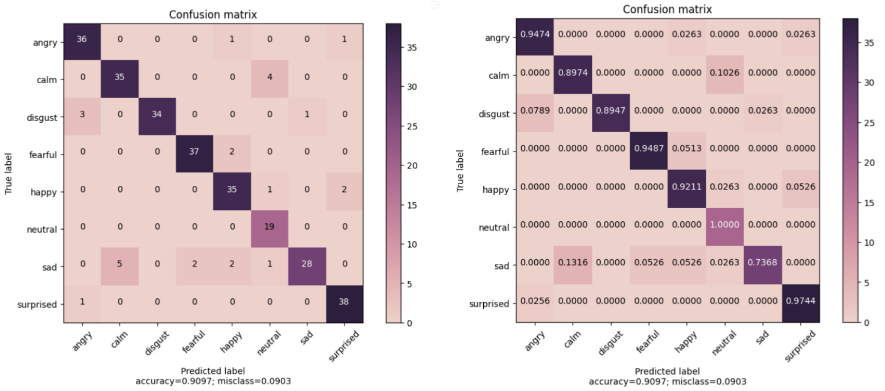
```
              precision    recall  f1-score   support

       angry       0.90      0.95      0.92        38
        calm       0.88      0.90      0.89        39
     disgust       1.00      0.89      0.94        38
     fearful       0.95      0.95      0.95        39
       happy       0.88      0.92      0.90        38
     neutral       0.76      1.00      0.86        19
         sad       0.97      0.74      0.84        38
   surprised       0.93      0.97      0.95        39

    accuracy                           0.91       288
   macro avg       0.91      0.92      0.91       288
weighted avg       0.92      0.91      0.91       288
```

Figure 19 Classification report for UniSpeech large model on RAVDESS dataset (at 700 steps).

and the unweighted accuracy (UA) is 92% on the RAVDESS dataset. The numerical as well as normalized confusion matrices based on the classification report for test data using UniSpeech large multilingual model are presented in Figure 23. From the training graph in



(a) **(b)**

Figure 20 (a) Numerical confusion matrix for UniSpeech large model on RAVDESS dataset (at 700 steps). (b) Normalized confusion matrix for UniSpeech large model on RAVDESS dataset (at 700 steps).

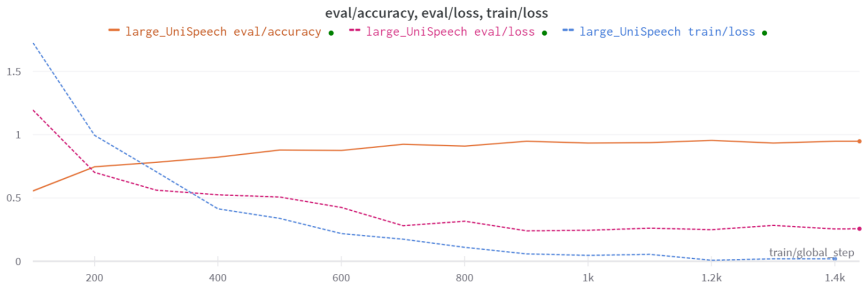


Figure 21 Training graph for UniSpeech large model. X-axis shows train/global_step and Y-axis shows loss and accuracy.

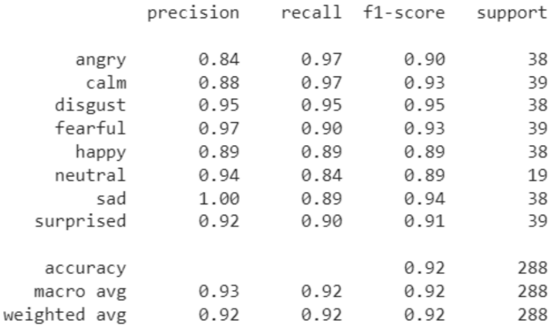
Figure 24, it can be inferred that there is divergence between training and validation losses after 600 steps. Hence, we should use the trained model at 600 steps.

6.2 Experiment Results of Proposed Majority Voting Ensemble Models

Table 3 shows the results of majority voting ensemble of the SSL models across five and three combinations based on hard voting. The sub sections below present the classification reports and the confusion matrices of the majority voting ensemble models.

6.2.1 Ensemble combination of five SSL models

The classification report computed for the ensemble model of Wav2vec2.0-XLSR-53, HuBERT-Base, HuBERT x-large, UniSpeech-large, UniSpeech-large-multilingual based on majority hard voting is shown in Figure 25. From the classification report, the weighted accuracy (WA) is 97% and unweighted accuracy (UA) is 97% on the RAVDESS dataset. The numerical as well as normalized confusion matrices based on the classification report for the test data using majority ensemble model with five SSL models are presented in

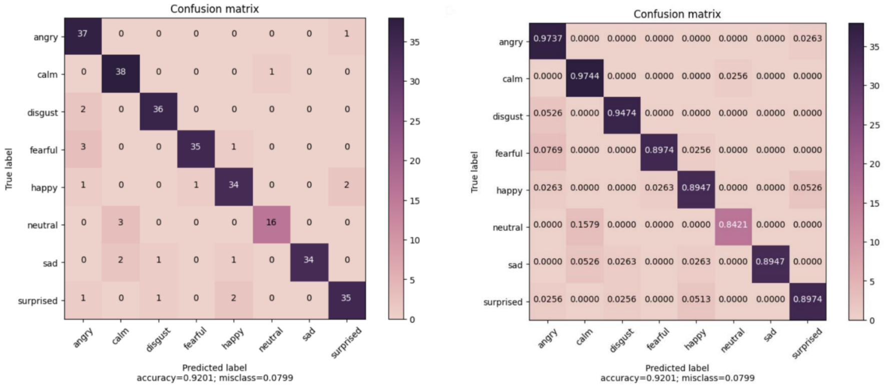


```
              precision    recall  f1-score   support

       angry       0.84      0.97      0.90        38
        calm       0.88      0.97      0.93        39
     disgust       0.95      0.95      0.95        38
     fearful       0.97      0.90      0.93        39
       happy       0.89      0.89      0.89        38
     neutral       0.94      0.84      0.89        19
         sad       1.00      0.89      0.94        38
   surprised       0.92      0.90      0.91        39

    accuracy                           0.92       288
   macro avg       0.93      0.92      0.92       288
weighted avg       0.92      0.92      0.92       288
```

Figure 22 Classification report for UniSpeech large Multi-lingual model on RAVDESS dataset (at 600 steps).



	angry	calm	disgust	fearful	happy	neutral	sad	surprised
angry	37	0	0	0	0	0	0	1
calm	0	38	0	0	0	1	0	0
disgust	2	0	36	0	0	0	0	0
fearful	3	0	0	35	1	0	0	0
happy	1	0	0	1	34	0	0	2
neutral	0	3	0	0	0	16	0	0
sad	0	2	1	0	1	0	34	0
surprised	1	0	1	0	2	0	0	35

	angry	calm	disgust	fearful	happy	neutral	sad	surprised
angry	0.9737	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0263
calm	0.0000	0.9744	0.0000	0.0000	0.0000	0.0256	0.0000	0.0000
disgust	0.0526	0.0000	0.9474	0.0000	0.0000	0.0000	0.0000	0.0000
fearful	0.0769	0.0000	0.0000	0.8974	0.0256	0.0000	0.0000	0.0000
happy	0.0263	0.0000	0.0000	0.0263	0.8947	0.0000	0.0000	0.0526
neutral	0.0000	0.1579	0.0000	0.0000	0.0000	0.8421	0.0000	0.0000
sad	0.0000	0.0526	0.0263	0.0000	0.0263	0.0000	0.8947	0.0000
surprised	0.0256	0.0000	0.0256	0.0000	0.0513	0.0000	0.0000	0.8974

(a) **(b)**

Figure 23 (a) Numerical confusion matrix for UniSpeech large Multi-lingual model on RAVDESS dataset (at 600 steps). (b) Normalized confusion matrix for UniSpeech large Multi-lingual model on RAVDESS dataset (at 600 steps).

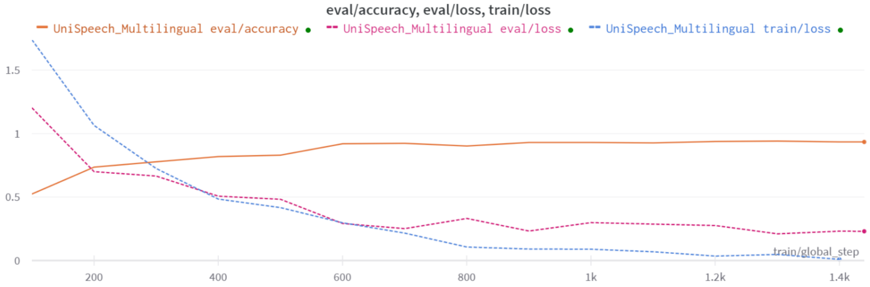
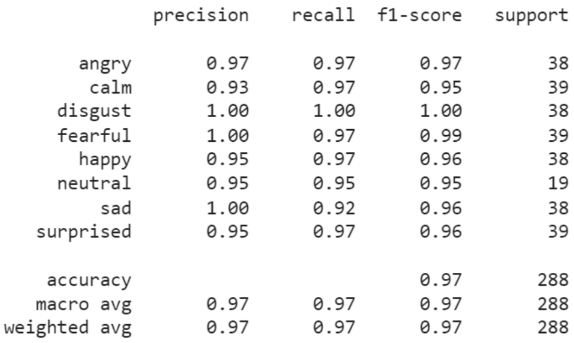


Figure 24 Training graph for UniSpeech large Multi-lingual model. X-axis shows train/global_step and Y-axis shows loss and accuracy.

Figure 26. We tested all possible five model voting ensembles and report in this paper the results of the most promising five model voting ensemble.

Table 3 Results of Proposed Majority Voting Ensemble Models

Ensemble model combinations of 5 and 3 models	**Accuracy**
Wav2vec2.0-XLSR-53, HuBERT-base, HuBERT x-large, UniSpeech-large, UniSpeech-large-multilingual	96.88%
Wav2vec2.0-XLSR-53, HuBERT x-large, UniSpeech-large-multilingual	96.53%

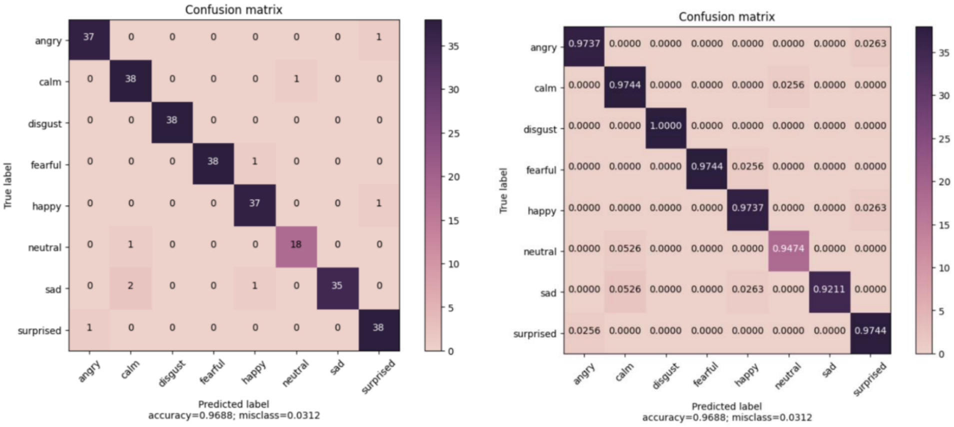


```
              precision    recall  f1-score   support

       angry       0.97      0.97      0.97        38
        calm       0.93      0.97      0.95        39
     disgust       1.00      1.00      1.00        38
     fearful       1.00      0.97      0.99        39
       happy       0.95      0.97      0.96        38
     neutral       0.95      0.95      0.95        19
         sad       1.00      0.92      0.96        38
   surprised       0.95      0.97      0.96        39

    accuracy                           0.97       288
   macro avg       0.97      0.97      0.97       288
weighted avg       0.97      0.97      0.97       288
```

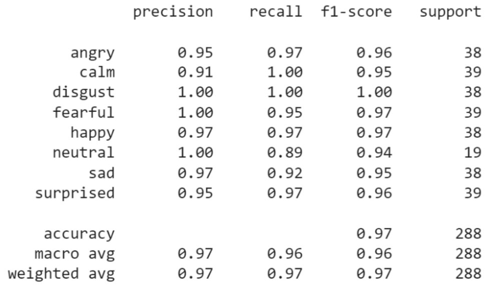
Figure 25 Classification report for majority voting ensemble model of five models combination (Wav2vec2.0-XLSR-53, HuBERT-Base, HuBERT x-large, UniSpeech-large, UniSpeech-large-multilingual) using hard voting on RAVDESS dataset.



	angry	calm	disgust	fearful	happy	neutral	sad	surprised
angry	37	0	0	0	0	0	0	1
calm	0	38	0	0	0	1	0	0
disgust	0	0	38	0	0	0	0	0
fearful	0	0	0	38	1	0	0	0
happy	0	0	0	0	37	0	0	1
neutral	0	1	0	0	0	18	0	0
sad	0	2	0	0	1	0	35	0
surprised	1	0	0	0	0	0	0	38

	angry	calm	disgust	fearful	happy	neutral	sad	surprised
angry	0.9737	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0263
calm	0.0000	0.9744	0.0000	0.0000	0.0000	0.0256	0.0000	0.0000
disgust	0.0000	0.0000	1.0000	0.0000	0.0000	0.0000	0.0000	0.0000
fearful	0.0000	0.0000	0.0000	0.9744	0.0256	0.0000	0.0000	0.0000
happy	0.0000	0.0000	0.0000	0.0000	0.9737	0.0000	0.0000	0.0263
neutral	0.0000	0.0526	0.0000	0.0000	0.0000	0.9474	0.0000	0.0000
sad	0.0000	0.0526	0.0000	0.0000	0.0263	0.0000	0.9211	0.0000
surprised	0.0256	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.9744

Figure 26 On the left numerical confusion matrix for majority voting ensemble model with five SSL models (Wav2vec2.0-XLSR-53, HuBERT-base, HuBERT x-large, UniSpeech-large, UniSpeech-large-multilingual) using hard voting on RAVDESS dataset. On the right normalized confusion matrix for majority voting ensemble model with five SSL models (Wav2vec2.0-XLSR-53, HuBERT-base, HuBERT x-large, UniSpeech-large, UniSpeech-large-multilingual) using hard voting on RAVDESS dataset.

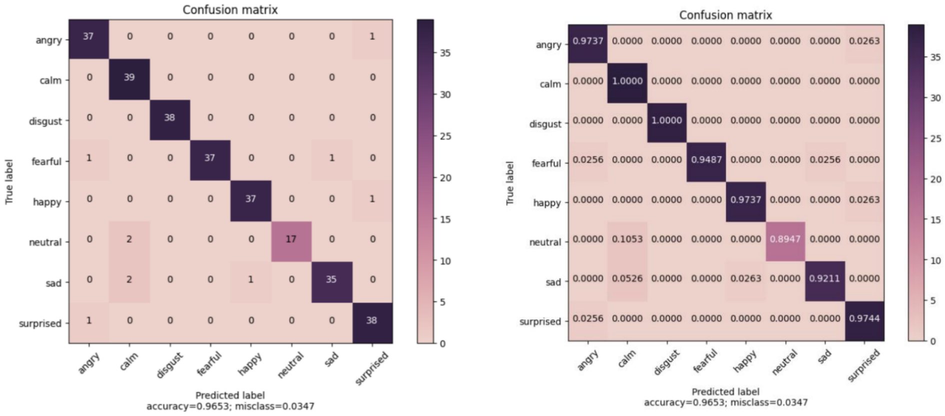


```
              precision    recall  f1-score   support

       angry       0.95      0.97      0.96        38
        calm       0.91      1.00      0.95        39
     disgust       1.00      1.00      1.00        38
     fearful       1.00      0.95      0.97        39
       happy       0.97      0.97      0.97        38
     neutral       1.00      0.89      0.94        19
         sad       0.97      0.92      0.95        38
   surprised       0.95      0.97      0.96        39

    accuracy                           0.97       288
   macro avg       0.97      0.96      0.96       288
weighted avg       0.97      0.97      0.97       288
```

Figure 27 Classification report for majority voting ensemble model with three SSL models (Wav2vec2.0-XLSR-53, HuBERT x-large, UniSpeech-large-multilingual) using hard voting on RAVDESS dataset.



	angry	calm	disgust	fearful	happy	neutral	sad	surprised
angry	37	0	0	0	0	0	0	1
calm	0	39	0	0	0	0	0	0
disgust	0	0	38	0	0	0	0	0
fearful	1	0	0	37	0	0	1	0
happy	0	0	0	0	37	0	0	1
neutral	0	2	0	0	0	17	0	0
sad	0	2	0	0	1	0	35	0
surprised	1	0	0	0	0	0	0	38

	angry	calm	disgust	fearful	happy	neutral	sad	surprised
angry	0.9737	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0263
calm	0.0000	1.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
disgust	0.0000	0.0000	1.0000	0.0000	0.0000	0.0000	0.0000	0.0000
fearful	0.0256	0.0000	0.0000	0.9487	0.0000	0.0000	0.0256	0.0000
happy	0.0000	0.0000	0.0000	0.0000	0.9737	0.0000	0.0000	0.0263
neutral	0.0000	0.1053	0.0000	0.0000	0.0000	0.8947	0.0000	0.0000
sad	0.0000	0.0526	0.0000	0.0000	0.0263	0.0000	0.9211	0.0000
surprised	0.0256	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.9744

(a) **(b)**

Figure 28 (a) Numerical confusion matrix for majority voting ensemble model with three SSL models (Wav2vec2.0-XLSR-53, HuBERT x-large, UniSpeech-large-multilingual) using hard voting on RAVDESS dataset. (b) Normalized confusion matrix for majority voting ensemble model with three SSL models (Wav2vec2.0-XLSR-53, HuBERT x-large, UniSpeech-large-multilingual) using hard voting on RAVDESS dataset.

6.2.2 Ensemble combination of three SSL models

The classification report computed for the ensemble model of Wav2vec2.0-XLSR-53, HuBERT x-large, UniSpeech-large-multilingual based on majority hard voting is shown in Figure 27. From the classification report, the weighted accuracy (WA) is 97% and unweighted accuracy (UA) is 97% on the RAVDESS dataset. The numerical as well as normalized confusion matrices based on the classification report for the test data using majority ensemble model with three SSL models are presented in Figure 28. We tested all possible three model voting ensembles and report in this paper the results of the most promising three model voting ensemble.

Table 4 Summary of Performance of the SSL Models and the Proposed Majority Voting Ensemble Models on the RAVDESS Dataset

Models	**Accuracy**
Wav2vec2.0-large	90% (at 900 steps)
Wav2vec2.0-large-XLSR-53	93% (at 1,100 steps)
HuBERT base	92% (at 900 steps)
HuBERT x-large	90% (at 600 steps)
UniSpeech large	91% (at 700 steps)
UniSpeech large multilingual	92% (at 600 steps)
Wav2vec2.0-XLSR-53, HuBERT-base, HuBERT x-large, UniSpeech-large, UniSpeech-large-multilingual	96.88%
Wav2vec2.0-XLSR-53, HuBERT x-large, UniSpeech-large-multilingual	96.53%

Table 5 Summary of Runtimes of the SSL Models on the RAVDESS Dataset

No.	**SSL Models**	**Total Runtime (approx.)**	**Average runtime for 100 steps (approx.)**	**Checkpoint runtime as per number of steps (approx.)**
1.	Wav2vec2.0-large	1 hour 7 minutes and 5 seconds	4 minutes and 40 seconds	43 minutes and 6 seconds (at 900 steps)
2.	Wav2vec2.0-large-XLSR-53	1 hour 11 minutes and 17 seconds	5 minutes and 4 seconds	56 minutes (at 1,100 steps)
3.	HuBERT base	21 minutes and 5 seconds	1 minute and 25 seconds	13 minutes and 26 seconds (at 900 steps)
4.	HuBERT x-large	1 hour 13 minutes and 27 seconds	5 minutes and 14 seconds	31 minutes and 24 seconds (at 600 steps)
5.	UniSpeech large	35 minutes and 6 seconds	2 minutes and 27 seconds	18 minutes (at 700 steps)
6.	UniSpeech large multilingual	46 minutes and 42 seconds	3 minutes and 15 seconds	20 minutes (at 600 steps)

6.2.3 Comparison of results

Through the implementation of popular SSL models for speech emotion recognition (SER) downstream task, an AI pipeline workflow is created using Upstream + Downstream model paradigm with merge pooling strategy for speaker independent setting. The results achieved on the six self-supervised Learning base models are state-of-the-art with significantly easy workflow and less training time for complex task like speech emotion recognition as shown in Table 4. Wav2vec2.0-XLSR-53 from wav2vec2.0 family has the highest WA as well as UA accuracy of 93% at 1,100 steps with training time of approximately 56 minutes. Similarly, HuBERT-base model achieved 92% accuracy (both WA and UA) at 900 steps

Table 6 Summary of Model Size on Disk of the Different SSL Models Fine-Tuned on the RAVDESS Dataset

Models	**Model Size on Disk**
Wav2vec2.0-large	1.27GB
Wav2vec2.0-large-XLSR-53	1.27GB
HuBERT base	380MB
HuBERT x-large	3.86GB
UniSpeech large	1.27GB
UniSpeech large multilingual	1.27GB

with 13 minutes of training time. Whereas, UniSpeech-large-multilingual model achieved 92% accuracy at 600 steps in 20 minutes of training time as shown in Table 5. Thus, based on the time taken to train the model to achieve high accuracy HuBERT-base could be an ideal candidate. Besides, HuBERT-base model disk size is also significantly small of 380MB in contrast to other high performing SSL models like Wav2vec2.0-XLSR-53 and UniSpeech-large-multilingual, both of which have comparatively high model disk size of 1.27GB as shown in Table 6.

The experiments in this research also demonstrated that proposed majority voting ensemble models of different combinations of three or five of the top performing SSL models significantly improved in terms of the overall accuracy of the model predictions across different emotion classes of the dataset as shown in Table 3. The majority voting ensemble of five SSL models - Wav2vec2.0-XLSR-53, HuBERT-base, HuBERT x-large, UniSpeech-large, UniSpeech-large-multilingual based on hard voting provides more confidence in prediction of emotion categories with a high weighted and unweighted accuracy of 96.88% on the RAVDESS dataset, which is 3.88% higher than the top performing single SSL model, Wav2vec2.0-XLSR-53. As for the majority ensemble with combination of three SSL models - Wav2vec2.0-XLSR-53, HuBERT x-large, UniSpeech-large-multilingual achieved a high weighted and unweighted accuracy of 96.53% on the RAVDESS dataset, which is 3.53% higher than the single top performing SSL model, Wav2vec2.0-XLSR-53. This proves that majority voting ensemble models help to increase confidence in the emotion predictions.

7 Conclusion and Future Work

We conducted experiments on the challenging large-scale speech emotion RAVDESS dataset. Six very large state-of-the-art self-supervised transformers were trained on the speech emotion dataset. Wav2vec2.0-XLSR-53 was the most successful of the six level-0 models. We proposed majority voting ensemble models that combined 3 and 5 level-0 models, both the voting models significantly outperformed the level-0 models.

As future work different data augmentation techniques could be explored to improve the performance for the SER task. Moreover, speech emotion recognition could be further explored using multilingual settings. Besides, ethical considerations in developing SER models for privacy perseveration suggests usage of federated learning to be the next suitable step in this domain. As such investigations pursuing federated learning paradigm using SSL models for emotion recognition are possible extensions of this research work.

References

Atmaja, B. T. & Sasou, A. (2022), 'Evaluating self-supervised speech representations for speech emotion recognition', *IEEE Access* **10**, 124396–124407.

Bautista, J. L., Lee, Y. K. & Shin, H. S. (2022), 'Speech emotion recognition based on parallel cnn-attention networks with multi-fold data augmentation', *Electronics*.

Biewald, L. (2020), 'Experiment tracking with weights and biases'. Software available from wandb.com.

Conneau, A., Baevski, A., Collobert, R., Mohamed, A. & Auli, M. (2021), Unsupervised Cross-Lingual Representation Learning for Speech Recognition, *in* 'Proc. Interspeech 2021', pp. 2426–2430.

Graves, A., Fernández, S., Gomez, F. & Schmidhuber, J. (2006), Connectionist temporal classification: Labelling unsegmented sequence data with recurrent neural networks, *in* 'Proceedings of the 23rd International Conference on Machine Learning', ICML '06, Association for Computing Machinery, New York, NY, USA, p. 369–376.

Han, S., Leng, F. & Jin, Z. (2021), Speech emotion recognition with a resnet-cnn-transformer parallel neural network, *in* '2021 International Conference on Communications, Information System and Computer Engineering (CISCE)', pp. 803–807.

Harris, C. R., Millman, K. J., van der Walt, S. J., Gommers, R., Virtanen, P., Cournapeau, D., Wieser, E., Taylor, J., Berg, S., Smith, N. J., Kern, R., Picus, M., Hoyer, S., van Kerkwijk, M. H., Brett, M., Haldane, A., del Río, J. F., Wiebe, M., Peterson, P., Gérard-Marchant, P., Sheppard, K., Reddy, T., Weckesser, W., Abbasi, H., Gohlke, C. & Oliphant, T. E. (2020), 'Array programming with NumPy', *Nature* **585**(7825), 357–362.

Hsu, W.-N., Bolte, B., Tsai, Y.-H. H., Lakhotia, K., Salakhutdinov, R. & Mohamed, A. (2021), 'Hubert: Self-supervised speech representation learning by masked prediction of hidden units', *IEEE/ACM Trans. Audio, Speech and Lang. Proc.* **29**, 3451–3460.

Hugging Face (2021*a*), 'Facebook Hubert-Base LS960', https://huggingface.co/facebook/hubert-base-ls960. Accessed: May 11, 2023.

Hugging Face (2021*b*), 'Facebook Hubert-XLarge LS960-FT', https://huggingface.co/facebook/hubert-xlarge-ls960-ft. Accessed: May 11, 2023.

Hugging Face (2021*c*), 'Facebook Wav2Vec2-Large 960h LV60-Self', https://huggingface.co/facebook/wav2vec2-large-960h-lv60-self. Accessed: May 11, 2023.

Hugging Face (2021*d*), 'Facebook Wav2Vec2-Large XLSR-53', https://huggingface.co/facebook/wav2vec2-large-xlsr-53. Accessed: May 11, 2023.

Hugging Face (2021*e*), 'Microsoft UniSpeech-Large 1500h CV', https://huggingface.co/microsoft/unispeech-large-1500h-cv. Accessed: May 11, 2023.

Hugging Face (2021*f*), 'Microsoft UniSpeech-Large Multi-Lingual 1500h CV', https://huggingface.co/microsoft/unispeech-large-multi-lingual-1500h-cv. Accessed: May 11, 2023.

Hunter, J. D. (2007), ‘Matplotlib: A 2d graphics environment’, *Computing in Science & Engineering* **9**(3), 90–95.

Kahn, J., Rivière, M., Zheng, W., Kharitonov, E., Xu, Q., Mazaré, P., Karadayi, J., Liptchinsky, V., Collobert, R., Fuegen, C., Likhomanenko, T., Synnaeve, G., Joulin, A., Mohamed, A. & Dupoux, E. (2020), Libri-light: A benchmark for asr with limited or no supervision, *in* ‘ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)’, pp. 7669–7673.

Kong, Q., Cao, Y., Iqbal, T., Wang, Y., Wang, W. & Plumbley, M. D. (2020), ‘Panns: Large-scale pretrained audio neural networks for audio pattern recognition’, *IEEE/ACM Trans. Audio, Speech and Lang. Proc.* **28**, 2880–2894.

Lhoest, Q., Villanova del Moral, A., Jernite, Y., Thakur, A., von Platen, P., Patil, S., Chaumond, J., Drame, M., Plu, J., Tunstall, L., Davison, J., Šaško, M., Chhablani, G., Malik, B., Brandeis, S., Le Scao, T., Sanh, V., Xu, C., Patry, N., McMillan-Major, A., Schmid, P., Gugger, S., Delangue, C., Matussière, T., Debut, L., Bekman, S., Cistac, P., Goehringer, T., Mustar, V., Lagunas, F., Rush, A. & Wolf, T. (2021), Datasets: A community library for natural language processing, *in* ‘Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing: System Demonstrations’, Association for Computational Linguistics, Online and Punta Cana, Dominican Republic, pp. 175–184.

Livingstone, S. R. & Russo, F. A. (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–e0196391.

Luna-Jiménez, C., Griol, D., Callejas, Z., Kleinlein, R., Montero, J. M. & Fernández-Martínez, F. (2021), ‘Multimodal emotion recognition on ravdess dataset using transfer learning’, *Sensors*.

Luna-Jiménez, C., Kleinlein, R., Griol, D., Callejas, Z., Montero, J. M. & Fernández-Martínez, F. (2022), ‘A proposal for multimodal emotion recognition using aural transformers and action units on ravdess dataset’, *Applied Sciences*.

McFee, B., Raffel, C., Liang, D., Ellis, D. P., McVicar, M., Battenberg, E. & Nieto, O. (2015), librosa: Audio and music signal analysis in python, *in* ‘Proceedings of the 14th python in science conference’, Vol. 8.

Morais, E., Hoory, R., Zhu, W., Gat, I., Damasceno, M. & Aronowitz, H. (2022), Speech emotion recognition using self-supervised features, *in* ‘ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)’, IEEE, pp. 6922–6926.

Panayotov, V., Chen, G., Povey, D. & Khudanpur, S. (2015), Librispeech: An asr corpus based on public domain audio books, *in* ‘2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)’, pp. 5206–5210.

pandas development team, T. (2020), ‘pandas-dev/pandas: Pandas’.

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., Desmaison, A., Kopf, A., Yang, E., DeVito, Z., Raison, M., Tejani, A., Chilamkurthy, S., Steiner, B., Fang, L., Bai, J. & Chintala, S. (2019), Pytorch: An imperative style, high-performance deep learning library, *in* H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox & R. Garnett, eds, 'Advances in Neural Information Processing Systems', Vol. 32, Curran Associates, Inc.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M. & Duchesnay, E. (2011), 'Scikit-learn: Machine learning in Python', *Journal of Machine Learning Research* **12**, 2825–2830.

Radford, A., Kim, J. W., Xu, T., Brockman, G., McLeavey, C. & Sutskever, I. (2022), 'Robust speech recognition via large-scale weak supervision', *arXiv preprint arXiv:2212.04356*.

Shirian, A. & Guha, T. (2020), 'Compact graph architecture for speech emotion recognition', *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* pp. 6284–6288.

Shirian, A., Tripathi, S. & Guha, T. (2022), 'Dynamic emotion modeling with learnable graphs and graph inception network', *IEEE transactions on multimedia* **24**, 780–790.

Tsouvalas, V., Ozcelebi, T. & Meratnia, N. (2022), 'Privacy-preserving speech emotion recognition through semi-supervised federated learning'.

Vaiani, L., La Quatra, M., Cagliero, L. & Garza, P. (2022), Viper: Video-based perceiver for emotion recognition, *in* 'Proceedings of the 3rd International on Multimodal Sentiment Analysis Workshop and Challenge', MuSe' 22, Association for Computing Machinery, New York, NY, USA, p. 67–73.

Vásquez-Correa, Juan Camilo; Álvarez Muniain, A. (2023), 'Novel speech recognition systems applied to forensics within child exploitation: Wav2vec2.0 vs. whisper', *Sensors (Basel)* **23**(4), 1843.

Wang, C., Wu, Y., Qian, Y., Kumatani, K., Liu, S., Wei, F., Zeng, M. & Huang, X. (2021), Unispeech: Unified speech representation learning with labeled and unlabeled data, *in* M. Meila & T. Zhang, eds, 'Proceedings of the 38th International Conference on Machine Learning', Vol. 139 of *Proceedings of Machine Learning Research*, PMLR, pp. 10937–10947.

Wes McKinney (2010), Data Structures for Statistical Computing in Python, *in* Stéfan van der Walt & Jarrod Millman, eds, 'Proceedings of the 9th Python in Science Conference', pp. 56 – 61.

Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., Davison, J., Shleifer, S., von Platen, P., Ma, C., Jernite, Y., Plu, J., Xu, C., Le Scao, T., Gugger, S., Drame, M., Lhoest, Q. & Rush, A. (2020), Transformers: State-of-the-art natural language processing, *in* 'Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations', Association for Computational Linguistics, Online, pp. 38–45.

Xu, Q., Baevski, A., Likhomanenko, T., Tomasello, P., Conneau, A., Collobert, R., Synnaeve, G. & Auli, M. (2021), Self-training and pre-training are complementary for speech recognition, *in* 'ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)', pp. 3030–3034.

Zenkov, I. (2020), 'transformer-cnn-emotion-recognition', https://github.com/IliaZenkov/transformer-cnn-emotion-recognition. Accessed: May 11, 2023.