

Automatic Recognition of Emotions in Speech With Large Self-Supervised Learning Transformer Models

Mrunal Prakash Gavali

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

Abhishek Verma

Department of Computer Science
California State University, Northridge
Northridge, CA, USA
abhishek.verma@csun.edu

Abstract—Speech Emotion Recognition (SER) is an important area of research in the realm of collaborative and social robotics, which aims to enhance human-robot interaction (HRI) and serves as a feedback mechanism for affective computing. Despite the recent progress in SER research area, it remains a challenging research problem due to the profound variations in the complexity, subjectivity, and contextual heterogeneity of human emotional expressions. Consequently, the inherent difficulties of modeling paralinguistic emotional information embedded in speech signals are further compounded when employing supervised learning, as it necessitates annotated labels for a large-scale dataset for satisfactory model performance. To this end, self-supervised learning (SSL) approach is widely adopted in the speech domain to address this problem of limited availability of annotated data.

Therefore, the focus of our research is to investigate and evaluate several state-of-the-art large attention-based self-supervised learning (SSL) models for the task of automatic speech emotion recognition (SER) on the challenging RAVDESS dataset. Results of the four large SSL models on the RAVDESS dataset are promising. In particular HuBERT large model achieved highest accuracy of 88% with a much lower training time and lower model size on disk compared to the rest of the models.

Index Terms—speech emotion recognition; self-supervised learning; emotion AI; transformers; speech processing; acoustic features

I. INTRODUCTION

Inspired by the challenge of endowing collaborative and social robots with emotional intelligence to make human-robot interaction more intuitive, this research investigates the effect of using different Self-Supervised Learning (SSL) models for emotion recognition using the speech modality. Speech emotion recognition (SER) research area is a challenging research problem as its difficult to model the inherently complex, extremely ambiguous and heterogeneous emotional representations in speech signals. In addition to the complicated speech representation learning, large-scale annotated and publicly available datasets for supervised learning often do not have enough speakers or lexical variations to adequately cover the highly personal variations in the emotion expressions, which makes self-supervised learning an ideal approach to address this research problem. Our research seeks to provide a comparison

for Self-Supervised Learning (SSL) algorithms for speech emotion recognition on the RAVDESS dataset.

Atmaja et al. [1] states that despite the self-supervised learning approach is claimed to be universal for speech representation learning in different speech processing tasks in the SUPERB (Speech Processing Universal PERFORMANCE) Benchmark list proposed by Yang et al. [2], it is worth investigating the different SSL models for specific speech processing tasks like SER instead of generic tasks. Speech emotion recognition (SER) task, which is a subset of non-semantic speech processing task, unlike Automatic Speech Recognition (ASR), does not necessitate the input granularity to be at the level of words or phonemes. However, most of the speech representations and available speech datasets for SSL are based on ASR, which seeks to recognize the linguistic speech content instead of the paralinguistic speech content as in the case of SER task. Hence, the authors in [1] suggest the potential for exploring the use of SSL models pretrained for Automatic Speech Recognition (ASR) as a basis for fine-tuning for Speech Emotion Recognition (SER) downstream task, since both the linguistic and paralinguistic information may be contained in the same extracted acoustic features from SSL models.

Our research paper investigates the different attention-based SSL models for speech emotion recognition (SER) task on the RAVDESS dataset [3] using upstream-downstream paradigm and merge mean pooling strategy by transfer learning from the ASR based pretrained SSL models from the Hugging Face repositories. It is important to note that no direct comparison with the state-of-the-art methods for speaker-independent evaluation on the RAVDESS dataset has been previously conducted. The speakers are different from the actors in the RAVDESS dataset used for training, making speaker-independent evaluation essential for developing a more universal and realistic model. Our research holds value in exploring and evaluating experiments under the more challenging text and speaker-independent conditions when compared to other researches performed on the speaker-dependent settings on the RAVDESS dataset.

This research paper is organized as follows. Section II 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 followed by section III that covers the description of the RAVDESS dataset used in the experiments. Section IV introduces different Self-Supervised Learning (SSL) models used in this research. Section V presents the experimental setup used for performing the experiments including development tools, model configuration and evaluation framework. Section VI presents the experimental results such as confusion matrix, classification reports and training graphs for model performance comparison and discuss findings from the experiment by summarizing the model performance of several models. Finally, section VII closes with the conclusion and future directions for speech emotion recognition research.

II. RELATED WORK

Han et al. (2021) proposed a parallel network of Resnet-CNN-Transformer Encoder for SER on RAVDESS dataset [4] and achieved average test accuracy of 80.89% after training the model for 500 epochs by repeating the experiment 5 times. The human accuracy for speech emotion recognition for RAVDESS dataset is 67% which indicates that SER for RAVDESS is complex even for human evaluators [3].

Luna-Jiménez et al. (2021) [5] used transfer learning by utilizing pre-existing knowledge captured by a supervised pre-trained models like the CNN-14 of the PANNs [6] framework for SER task on RAVDESS to improve performance through fine-tuning. The authors reported the SER accuracy (using speech modality) of only 76.58% on RAVDESS.

More recently, 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 results, especially in a low-data setting as demonstrated in [7] where Luna-Jiménez et al. 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 in their previous work, which has 76.58% accuracy.

IBM AI research (2022) presents downstream-upstream paradigm on IEMOCAP dataset for End-to-End (E2E) SER downstream task [8]. Moreover, [1] presents comprehensive research on five emotion datasets in different languages with 20 different deep learning models. These researches provide a benchmark for our research expectations regarding the performance of Self-Supervised Learning pretrained models on the SER task.

Wang et al. (2021) 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 the

Table I: Summary of self-supervised learning (SSL) models used in our research based on pretraining and parameter size.

Model	Pretraining dataset	Parameter Size
Wav2vec2.0 base model [9]	53k hours of raw English speech data sampled from audiobooks	95M
Wav2vec2.0-XLS-R-300M [10]	436k hours of unlabeled speech, including VoxPopuli, MLS, Common-Voice, BABEL, and VoxLingua107 in 128 languages.	300M
Wav2vec2.0-XLS-R-1B [11], [12]	436k hours of unlabeled speech, including VoxPopuli, MLS, Common-Voice, BABEL, and VoxLingua107 in 128 languages.	1B
HuBERT large model [13]	60k hours of Libri-light audio	317M

Table II: Configuration of self-supervised learning (SSL) models used in our research: Wav2vec2.0 base, Wav2vec2.0-XLS-R-300M, Wav2vec2.0-XLS-R-1B, and HuBERT large.

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

	precision	recall	f1-score	support
angry	0.94	0.84	0.89	38
calm	0.85	1.00	0.92	39
disgust	0.90	0.97	0.94	38
fearful	0.94	0.82	0.88	39
happy	0.73	0.84	0.78	38
neutral	0.64	0.47	0.55	19
sad	0.66	0.66	0.66	38
surprised	0.95	0.90	0.92	39
accuracy			0.84	288
macro avg	0.83	0.81	0.82	288
weighted avg	0.84	0.84	0.83	288

Fig. 1: Classification report for Wav2vec2.0 base model on RAVDESS dataset (at 1300 steps).

SER task on IEMOCAP dataset. This 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.

III. DESCRIPTION OF THE RAVDESS DATASET

The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS) dataset [3] is an audio-visual dataset with collection of speech as well as song audio files of .wav

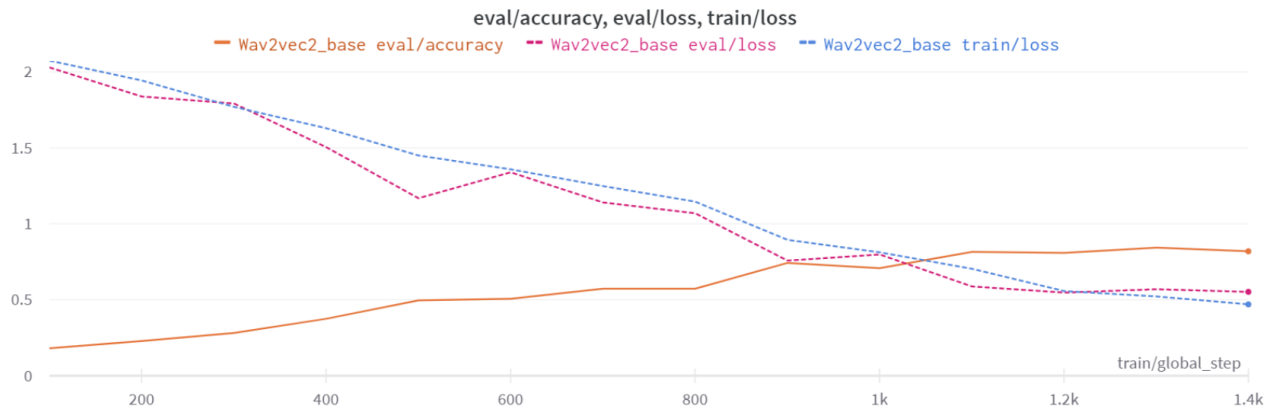


Fig. 2: Training graph for Wav2vec2.0 Base model on RAVDESS dataset. X-axis shows train/global_step and Y-axis shows loss and accuracy.

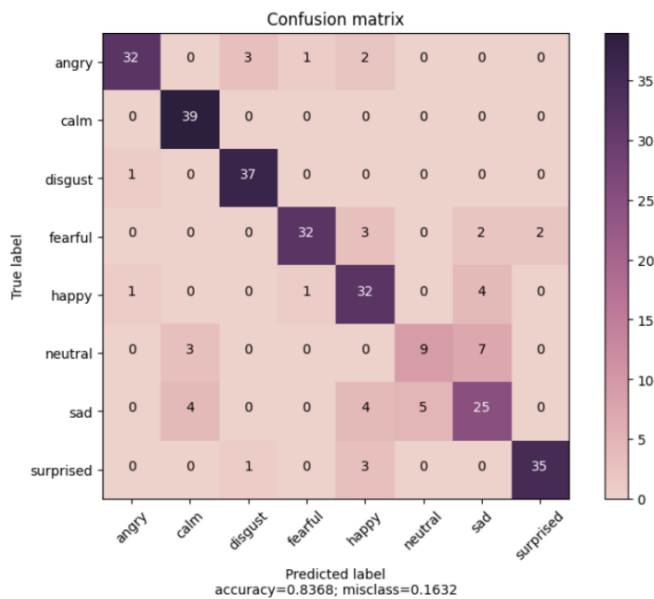


Fig. 3: Numerical confusion matrix for Wav2vec2.0 base model on RAVDESS dataset (at 1300 steps).

format. We use the speech data of 1440 audio files (excluding song files), grouped in different folders based on speaker instead of emotions, from the original RAVDESS dataset in our research. The speech audio in RAVDESS dataset consists of 24 professional actors (with the same ratio of male-female actors) speaking in neutral North American accent, which is recorded at 16-bit and 48kHz to vocalize two lexically similar statements in 8 different emotions (neutral, calm, happy, sad, angry, fear, surprise, and disgust) with normal and strong intensity - “Kids are talking by the door” and “Dogs are sitting by the door”. The ‘neutral’ emotion category is not recorded with strong intensity level.

The 1440 speech audio files in RAVDESS dataset have

a unique filename consisting of seven numerical identifiers. Therefore, the audio files are organized based on the RAVDESS dataset’s naming convention grouped by different emotion labels. Subsequently, a CSV file is generated to indicate the file path of each audio file and its corresponding emotion label. This allows for the data to be easily loaded for training using a data loader without having to repeat the prior steps for each training experiment. Once the data is loaded using the data loader, the SER dataset is divided into train and test CSV files, maintaining an 80:20 ratio with 1152 and 288 speech audio files for training and validation purposes respectively.

IV. RESEARCH METHODOLOGY

In this research, the SSL models employed exhibit a shared basis in terms of their model architecture and functionality. These models take raw audio input and generate vector representations as output. The distinction among these SSL frameworks primarily lies in their pre-training stage. This characteristic makes the exploration of various SSL models for the speech emotion recognition (SER) task particularly suitable for comparative analysis, as further elucidated in our research.

A. Wav2vec2.0

The Wav2vec 2.0 base model has parameter size of 95 million as presented in Table I. This model variant, based on self-training objective [14], is pretrained and fine-tuned on 960 hours of Librispeech [15] containing only English language speech data. The speech audio samples are sampled at 16kHz. This specific variant of the Wav2vec2.0 base model used in the research can be found in the Hugging Face repository [9].

B. Wav2vec2-large-XLS-R

XLS-R is a single large-scale cross-lingual model released by Meta AI for speech representations, pretrained on 436k hours of raw unlabeled acoustic speech in 128 languages using following datasets - VoxPopuli, MLS, CommonVoice, BABEL, and VoxLingua107. This multilingual pretrained model uses

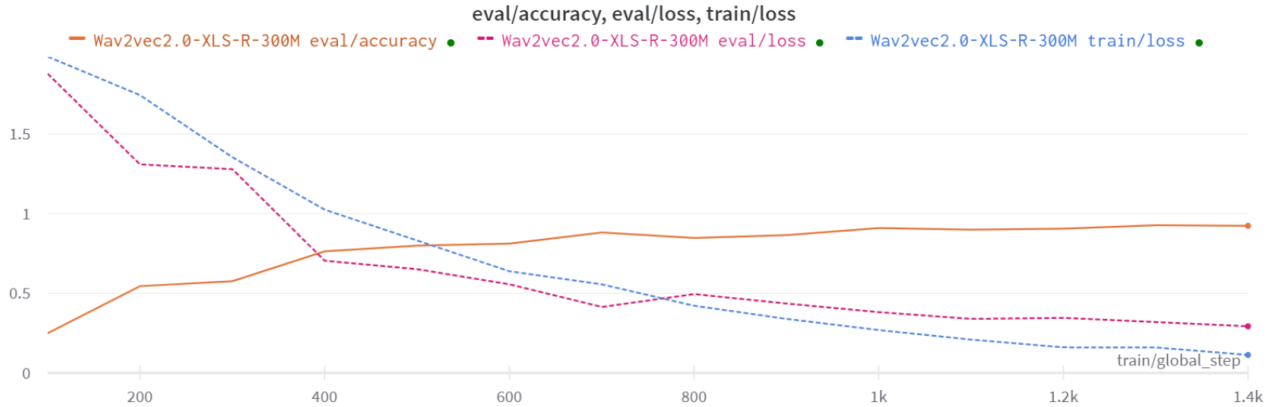


Fig. 4: Training graph for Wav2vec2.0-XLS-R-300M model on RAVDESS dataset. X-axis shows train/global_step and Y-axis shows loss and accuracy.

	precision	recall	f1-score	support
angry	0.91	0.84	0.88	38
calm	0.78	0.97	0.86	39
disgust	0.90	0.95	0.92	38
fearful	0.92	0.85	0.88	39
happy	0.79	0.82	0.81	38
neutral	0.72	0.68	0.70	19
sad	0.78	0.74	0.76	38
surprised	0.94	0.85	0.89	39
accuracy			0.85	288
macro avg	0.84	0.84	0.84	288
weighted avg	0.85	0.85	0.85	288

Fig. 5: Classification report for Wav2vec2-XLS-R-300M model on RAVDESS dataset (at 800 steps).

Wav2vec2.0 as its foundation with shared quantization module for outputting multilingual quantized speech units to be fed into the transformer for contrastive learning. The two model versions of this pretrained model from the Hugging Face repository used for the experiments in this research are Wav2vec2.0-XLS-R-300M [10] and Wav2vec2.0-XLS-R-1B [12], which are both larger than wav2vec2.0-base model in terms of languages and model size with 300M and 1B parameter size respectively. For finetuning on labeled RAVDESS dataset, the input speech audio is sampled at 16kHz. The Wav2vec2-xls-r-1b-common_voice-tr-ft model [12] version used in this experiment is fine-tuned on the COMMON_VOICE - TR dataset using the original Wav2vec2-xls-r-1b [11] released by Meta AI as the pretrained model.

C. HuBERT

For the experiments in this research, we use the HuBERT large model version with 317 million parameter size as shown in Table I. This model is pre-trained on 60k hours of English audiobook speech data and fine-tuned on 960h of Librispeech on 16kHz sampled speech audio, available in the following

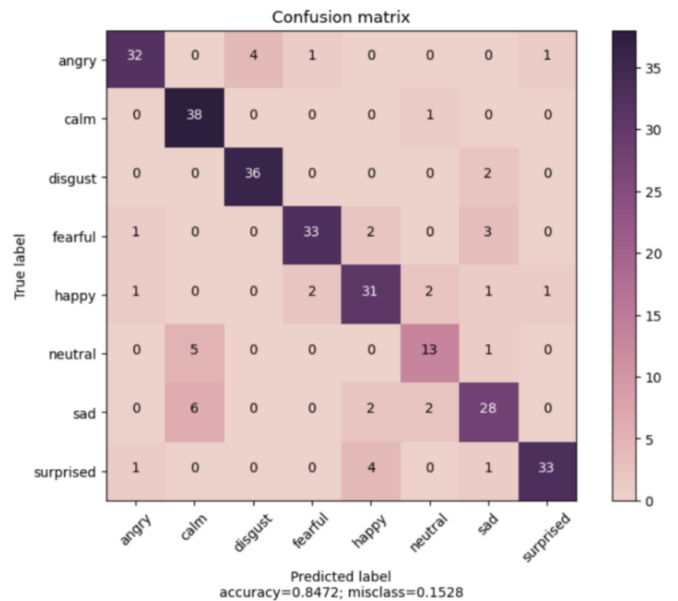


Fig. 6: Numerical confusion matrix for Wav2vec2-XLS-R-300M model on RAVDESS dataset (at 800 steps).

Hugging Face repository [13]. The HuBERT self-supervised learning (SSL) framework utilizes an offline clustering step to generate aligned target labels for a prediction loss similar to that in the BERT model [16].

V. EXPERIMENTAL SETUP

The implementation of SSL deep learning experiments is done using the Huggingface Transformers library [17], Huggingface Datasets library [18], as well as the Pytorch [19] and Torchaudio [20] library. Weights and Biases (wandb) [21] is used for tracking experiments and plotting training graphs. The final fine-tuned model checkpoint is uploaded directly to the Hugging Face Hub. The Google Colab Pro Plus

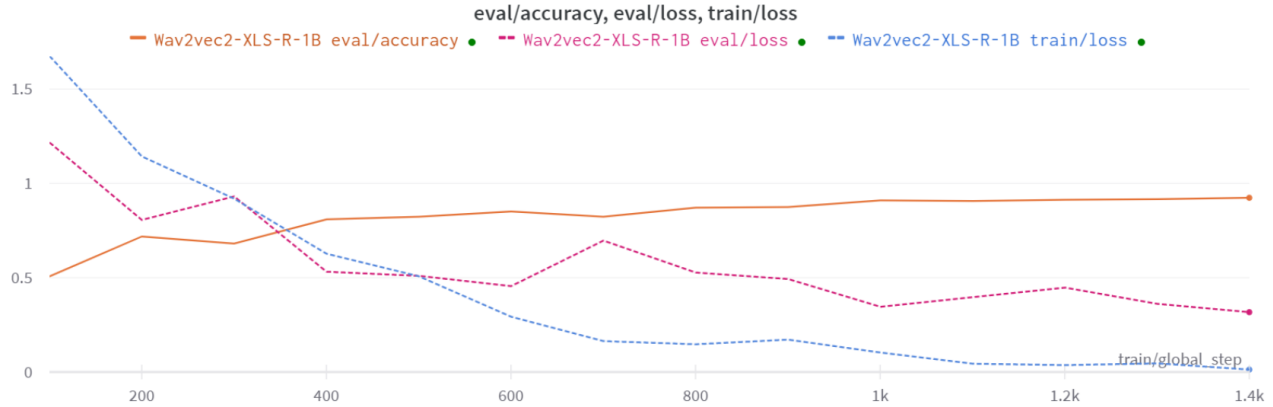


Fig. 7: Training graph for Wav2vec2.0-XLS-R-1B model on RAVDESS dataset. X-axis shows train/global_step and Y-axis shows loss and accuracy.

	precision	recall	f1-score	support
angry	0.88	0.92	0.90	38
calm	1.00	0.64	0.78	39
disgust	0.95	0.95	0.95	38
fearful	0.78	0.97	0.86	39
happy	0.95	0.50	0.66	38
neutral	0.46	1.00	0.63	19
sad	0.82	0.74	0.78	38
surprised	0.88	0.92	0.90	39
accuracy			0.82	288
macro avg	0.84	0.83	0.81	288
weighted avg	0.86	0.82	0.82	288

Fig. 8: Classification report for Wav2vec2-XLS-R-1B model on RAVDESS dataset (at 500 steps).

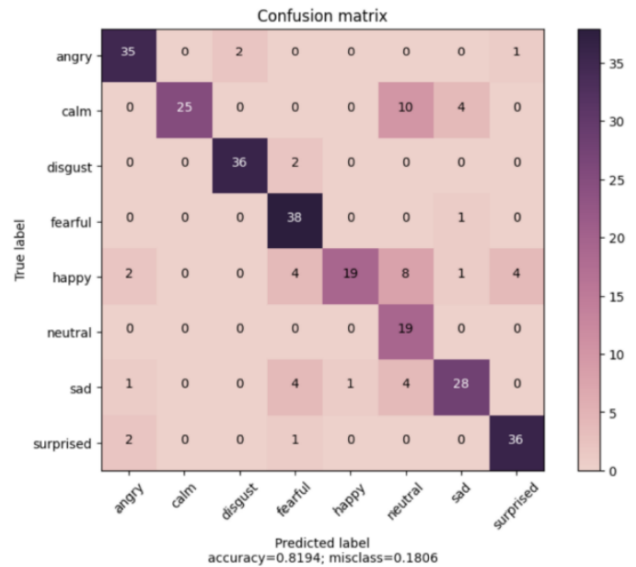


Fig. 9: Numerical confusion matrix for Wav2vec2-XLS-R-1B model on RAVDESS dataset (at 500 steps).

subscription is used which enables the execution of research experiments and fulfillment of demanding computational needs for RAM and GPU by utilizing the power of NVIDIA Tesla T4 and NVIDIA A100-SXM4-40GB. The model checkpoint is saved in Google Drive after every 100 steps.

The model configuration for the research experiments is described in Table II. A batch size of 4 and a learning rate of 0.0001 is used for the all SSL model experiments to train them over 1400 steps. These hyperparameters are passed as training arguments to the Hugging Face trainer API for fine-tuning all SSL models used in this research. All the training samples are padded to the length of the longest sample in their batch, rather than the overall longest sample using a special padding data collator.

The evaluation framework for all the different SSL models involves constructing the classification reports and numerical confusion matrix using Scikit-learn library [22] to record the different evaluation metrics like accuracy along with the F1-score, recall, and precision on a label-basis for the 8 emotion categories in the RAVDESS dataset.

VI. EXPERIMENTAL RESULTS AND DISCUSSION

Each of the SSL models along with its variations is trained and evaluated to determine the best candidates for the SER downstream task on the RAVDESS dataset.

1) *Wav2vec2.0 Base Model* : The classification report depicts that the accuracy of the Wav2vec2.0 Base model on the RAVDESS dataset is 84%, as shown in Figure 1. The confusion matrix for Wav2vec2.0 base model on test data is presented in Figure 3. The training graph in Figure 2, indicates that the Wav2vec2.0-base model has converged without overfitting.

2) *Wav2vec2.0-XLS-R-300M Model*: The classification report depicts that the accuracy of the Wav2vec2.0-XLS-R-300M model on the RAVDESS dataset is 85%, as shown in Figure 5. The confusion matrix for Wav2vec2.0-XLS-R-300M model

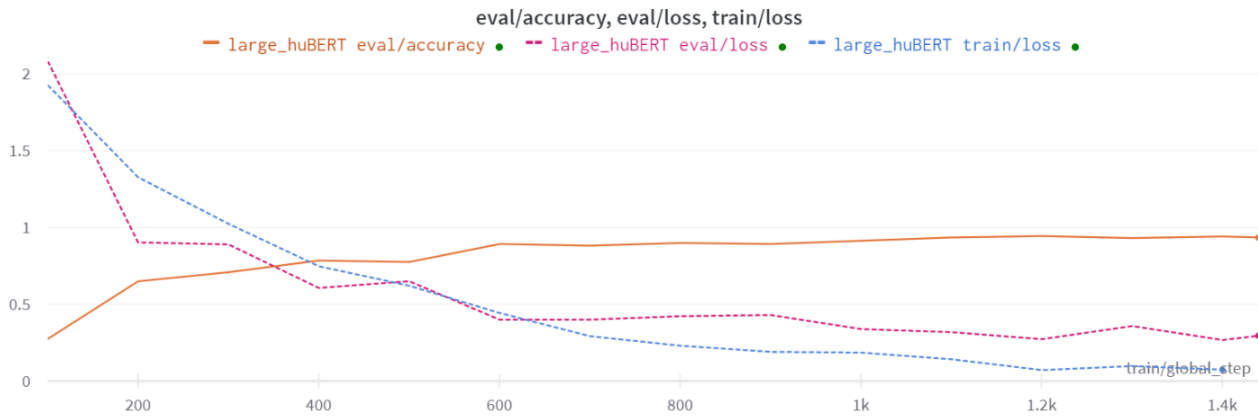


Fig. 10: Training graph for HuBERT large model on RAVDESS dataset. X-axis shows train/global_step and Y-axis shows loss and accuracy.

	precision	recall	f1-score	support
angry	0.94	0.82	0.87	38
calm	0.90	0.95	0.92	39
disgust	0.79	1.00	0.88	38
fearful	1.00	0.90	0.95	39
happy	0.90	0.68	0.78	38
neutral	0.76	0.84	0.80	19
sad	0.92	0.89	0.91	38
surprised	0.82	0.92	0.87	39
accuracy			0.88	288
macro avg	0.88	0.88	0.87	288
weighted avg	0.89	0.88	0.88	288

Fig. 11: Classification report for HuBERT large model on RAVDESS dataset (at 600 steps).

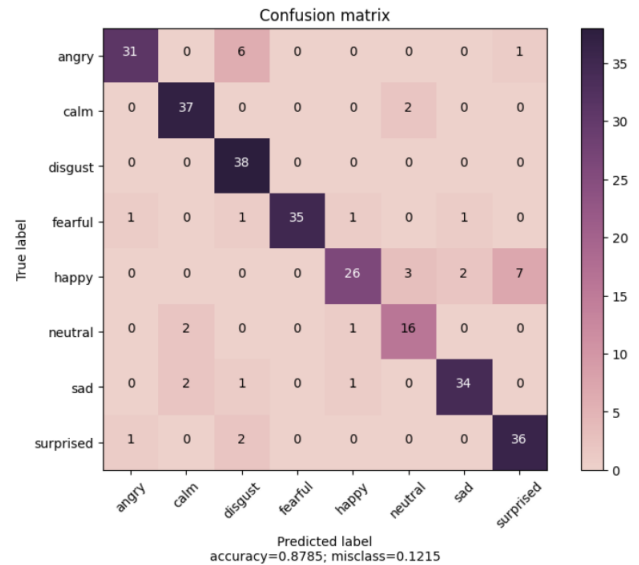


Fig. 12: Numerical confusion matrix for HuBERT large model on RAVDESS dataset (at 600 steps).

on test data is presented in Figure 6. We use early stopping prior to overfitting at 800 steps to build the model. After 800 steps, the training graph displayed in Figure 4 indicates a divergence between the training and validation losses. This divergence suggests the occurrence of overfitting, where the model becomes too specialized to the training data and may not generalize well on unseen real world data after 800 steps.

3) *Wav2vec2.0-XLS-R-1B Model*: The classification report depicts that the accuracy of the the Wav2vec2.0-XLS-R-1B model on the RAVDESS dataset is 82%, as shown in Figure 8. The confusion matrix for Wav2vec2.0-XLS-R-1B model on test data is presented in Figure 9. We use early stopping prior to overfitting at 500 steps to build the model. The training graph in Figure 7, indicates a divergence between the training and validation losses after 500 steps. This divergence suggests the occurrence of overfitting after 500 steps.

4) *HuBERT Large Model*: The classification report depicts that the accuracy of the the HuBERT Large model on the RAVDESS dataset is 88% on the RAVDESS dataset as shown in Figure 11. We use early stopping prior to overfitting at 600

steps to build the model. The confusion matrix for HuBERT Large model on test data is presented in Figure 12. The training graph in Figure 10 indicates divergence between training and validation losses after 600 steps which suggests the occurrence of overfitting.

5) *Overall Comparison Across Four Models*: The results presented in Table III demonstrate that SSL model can achieve satisfactory performance when applied to a larger set of emotion categories. From the Table III, it is clear that the HuBERT large model achieved the highest accuracy on the RAVDESS dataset with 88% at 600 steps. It takes less time to train HuBERT large model than the Wav2vec2.0 family variants with only 16 minutes and 34 seconds of time on the RAVDESS dataset.

Table III: Summary of performance of different self-supervised learning (SSL) models in our research on the RAVDESS dataset.

Models	Accuracy	Model Disk Size	Checkpoint runtime as per number of steps (approx.)
Wav2vec2.0 Base	84% (at 1300 steps)	380MB	39 minutes (at 1300 steps)
Wav2vec2.0-XLS-R-300M	85% (at 800 steps)	1.27GB	21 minutes and 40 seconds (at 800 steps)
Wav2vec2.0-XLS-R-1B	82% (at 500 steps)	3.86GB	25 minutes and 40 seconds (at 500 steps)
HuBERT large	88% (at 600 steps)	1.27GB	16 minutes and 34 seconds (at 600 steps)

VII. CONCLUSION AND FUTURE WORK

In this research, we performed evaluation of four state-of-the-art large self-supervised learning models for the complex speech emotion recognition downstream task with an Upstream + Downstream paradigm, independent of speaker, and text on the challenging RAVDESS emotion dataset.

It is important to note that no direct comparison with the state-of-the-art methods for speaker-independent evaluation on the RAVDESS dataset has been previously conducted. The speakers are different from the actors in the RAVDESS dataset used for training, making speaker-independent evaluation essential for developing a more universal and realistic model. Our research holds value in exploring and evaluating experiments under the more challenging text and speaker-independent conditions when compared to other researches performed on the speaker-dependent settings on the RAVDESS dataset.

Results of the four large SSL models on the RAVDESS dataset are promising considering the challenging nature of emotion recognition from speech. HuBERT large model achieved highest accuracy with a much lower training time and lower model size on disk compared to the rest of the models.

For future work, our plan is to assess a wider range of SSL models on diverse SER benchmarks in different spoken languages. This will involve incorporating various ensemble strategies with myriad data augmentation and cross-corpus techniques to handle highly unbalanced datasets, including those with a large number of emotion categories and acoustic emotion burst sounds like laughter and gasps.

REFERENCES

- [1] B. T. Atmaja and A. Sasou, "Evaluating self-supervised speech representations for speech emotion recognition," *IEEE Access*, vol. 10, pp. 124 396–124 407, 2022.
- [2] S. wen Yang, P.-H. Chi, Y.-S. Chuang, C.-I. J. Lai, K. Lakhotia, Y. Y. Lin, and et al., "SUPERB: Speech Processing Universal PERFORMANCE Benchmark," in *Proc. Interspeech 2021*, 2021, pp. 1194–1198.
- [3] S. R. Livingstone and F. A. Russo, "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*, vol. 13, no. 5, pp. e0196391–e0196391, 2018.
- [4] S. Han, F. Leng, and Z. Jin, "Speech emotion recognition with a resnet-cnn-transformer parallel neural network," in *2021 International Conference on Communications, Information System and Computer Engineering (CISCE)*, 2021, pp. 803–807.
- [5] C. Luna-Jiménez, D. Griol, Z. Callejas, R. Kleinlein, J. M. Montero, and F. Fernández-Martínez, "Multimodal emotion recognition on ravdess dataset using transfer learning," *Sensors*, vol. 21, no. 22, 2021. [Online]. Available: <https://www.mdpi.com/1424-8220/21/22/7665>
- [6] Q. Kong, Y. Cao, T. Iqbal, Y. Wang, W. Wang, and M. D. Plumbley, "Panns: Large-scale pretrained audio neural networks for audio pattern recognition," *IEEE/ACM Trans. Audio, Speech and Lang. Proc.*, vol. 28, p. 2880–2894, nov 2020. [Online]. Available: <https://doi.org/10.1109/TASLP.2020.3030497>
- [7] C. Luna-Jiménez, R. Kleinlein, D. Griol, Z. Callejas, J. M. Montero, and F. Fernández-Martínez, "A proposal for multimodal emotion recognition using aural transformers and action units on ravdess dataset," *Applied Sciences*, vol. 12, no. 1, 2022. [Online]. Available: <https://www.mdpi.com/2076-3417/12/1/327>
- [8] E. Morais, R. Hoory, W. Zhu, I. Gat, M. Damasceno, and H. Aronowitz, "Speech emotion recognition using self-supervised features," in *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022, pp. 6922–6926.
- [9] Hugging Face, "Facebook wav2vec2-base-960h," <https://huggingface.co/facebook/wav2vec2-base-960h>, 2021, accessed: May 11, 2023.
- [10] —, "Facebook wav2vec2-xls-r-300m," <https://huggingface.co/facebook/wav2vec2-xls-r-300m>, 2021, accessed: May 11, 2023.
- [11] —, "Facebook wav2vec2-xls-r-1b," <https://huggingface.co/facebook/wav2vec2-xls-r-1b>, 2021, accessed: May 11, 2023.
- [12] —, "P. von Platen wav2vec2-xls-r-1b-common_voice-tr-ft," https://huggingface.co/patrickvonplaten/wav2vec2-xls-r-1b-common_voice-tr-ft, 2021, accessed: May 11, 2023.
- [13] —, "Facebook hubert-large-ls960-ft," <https://huggingface.co/facebook/hubert-large-ls960-ft>, 2021, accessed: May 11, 2023.
- [14] Q. Xu, A. Baevski, T. Likhomanenko, P. Tomaso, A. Conneau, R. Collobert, and et al., "Self-training and pre-training are complementary for speech recognition," in *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2021, pp. 3030–3034.
- [15] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: An asr corpus based on public domain audio books," in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2015, pp. 5206–5210.
- [16] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," 2019.
- [17] T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, and et al., "Transformers: State-of-the-art natural language processing," in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*. Online: Association for Computational Linguistics, Oct. 2020, pp. 38–45. [Online]. Available: <https://aclanthology.org/2020.emnlp-demos.6>
- [18] Q. Lhoest, A. Villanova del Moral, Y. Jernite, A. Thakur, P. von Platen, S. Patil, and et al., "Datasets: A community library for natural language processing," in *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*. Online and Punta Cana, Dominican Republic: Association for Computational Linguistics, Nov. 2021, pp. 175–184. [Online]. Available: <https://aclanthology.org/2021.emnlp-demo.21>
- [19] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, and et al., "Pytorch: An imperative style, high-performance deep learning library," in *Advances in Neural Information Processing Systems*, H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, Eds., vol. 32. Curran Associates, Inc., 2019.
- [20] Y.-Y. Yang, M. Hira, Z. Ni, A. Chourdia, A. Astafurov, C. Chen, and et al., "Torchaudio: Building blocks for audio and speech processing," *arXiv preprint arXiv:2110.15018*, 2021.
- [21] L. Biewald, "Experiment tracking with weights and biases," 2020, software available from wandb.com. [Online]. Available: <https://www.wandb.com/>
- [22] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, and et al., "Scikit-learn: Machine learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.