AI Mood Detector Using Human Voice

Ghulam Ahmad (305)*, Zeehsan Anjum(299)†, Ahmed Shahfique(313)‡, Arslan Arshad (314)§
Department of Electrical Engineering
University Of Engineering and Technology Lahore
Emails: *2021ee305@uet.edu.pk, †2021ee299@uet.edu.pk, ‡2021ee313@uet.edu.pk, §2021ee314@uet.edu.pk

***Abstract*—In this study, we introduce an AI-powered system capable of detecting human emotions through voice analysis. By fine-tuning the self-supervised *wav2vec 2.0* model on established emotional speech datasets—SAVEE, RAVDESS, and TESS—we achieved a classification accuracy of 97.46%. This system identifies seven distinct emotional states: angry, disgust, fear, happy, neutral, sad, and surprise. Our approach demonstrates the effectiveness of leveraging pre-trained models for emotion recognition tasks, especially in scenarios with limited labeled data.**

***Index Terms*—Speech Emotion Recognition, wav2vec 2.0, Transfer Learning, Affective Computing, Human-Computer Interaction.**

I. Introduction

Emotions play a vital role in human communication, shaping interpersonal relationships and influencing decision-making. The automatic recognition of these emotions through speech, known as Speech Emotion Recognition (SER), has emerged as a significant research area within affective computing and human-computer interaction. SER enables various real-world applications, including personalized virtual assistants, driver fatigue monitoring systems, and mental health assessments.

Traditional SER methods heavily relied on handcrafted acoustic features, such as pitch, energy, and MFCCs, which often failed to generalize well across diverse datasets. With the rise of deep learning and the availability of large-scale datasets, end-to-end learning models have become more popular. One such model is wav2vec 2.0 [1], a self-supervised speech representation model trained on vast amounts of unlabeled audio data. Fine-tuning such models on emotional datasets allows for accurate emotion classification with minimal labeled data. This paper presents a robust pipeline for speech emotion detection using the `facebook/wav2vec2-large-xlsr-53` model, fine-tuned on a mix of gender-balanced and diverse datasets.

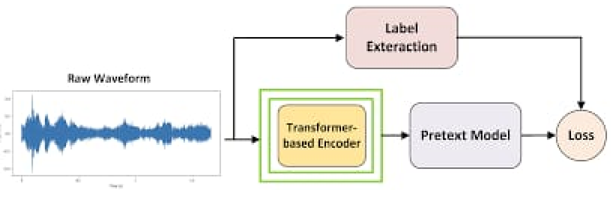


Fig. 1. System Architecture of the AI Mood Detector

II. Related Work

Early SER research focused on extracting acoustic features like MFCCs, Chroma, and spectral contrast, and feeding them into classical machine learning models, including Support Vector Machines (SVMs), Random Forests, and Hidden Markov Models (HMMs) [8]. However, these systems often lacked adaptability to new speakers or languages.

With deep learning, convolutional neural networks (CNNs) and recurrent neural networks (RNNs) brought significant improvements in capturing local and temporal speech characteristics [7]. Later, attention mechanisms and transformer-based architectures like wav2vec and HuBERT became dominant, providing superior performance through contextualized embeddings [3], [4].

Transfer learning has further enhanced SER systems by allowing models pre-trained on general audio tasks to be fine-tuned for emotion classification, requiring less labeled data and enabling domain adaptation [6].

III. Datasets and Preprocessing

Three widely used datasets were employed to train and evaluate our model:

- **SAVEE** (Surrey Audio-Visual Expressed Emotion): Contains 480 utterances from 4 male speakers across 7 emotions [9].
- **RAVDESS** (Ryerson Audio-Visual Database of Emotional Speech and Song): Comprises 1,440 audio files recorded by 24 professional actors (12 male, 12 female) [10].
- **TESS** (Toronto Emotional Speech Set): Includes 2,800 recordings from 2 female actors aged over 60 [11].

To maintain uniformity, all audio clips were resampled to 16 kHz and normalized. Noise reduction and silence trimming were applied to enhance clarity. Label encoding was used to transform categorical emotion labels into numerical format.

IV. Model Architecture

Our model is built upon the `facebook/wav2vec2-large-xlsr-53` pre-trained checkpoint. It contains the following components:

1) **Feature Extractor**: Captures speech representations from raw waveform using self-supervised pretraining. It includes convolutional layers followed by transformer encoders.

2) **Classification Head**: A task-specific fully connected neural layer added on top, projecting the 1024-dimensional embeddings into seven emotion classes.

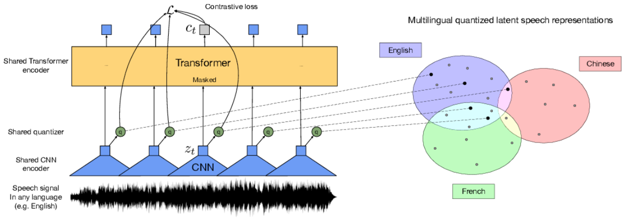

Contrastive loss
$\mathcal{L}$
c_t
Multilingual quantized latent speech representations
Shared Transformer encoder
Transformer
Masked
English
Chinese
Shared quantizer
q
z_t
Shared CNN encoder
CNN
Speech signal In any language (e.g. English)
French

Fig. 2. Model Architecture Diagram

V. Training Procedure

Fine-tuning was conducted using HuggingFace's Transformers and Datasets libraries. We adopted the Adam optimizer with a low learning rate to preserve pre-trained knowledge. A small batch size was used due to GPU memory constraints.

A. Training Hyperparameters

- **learning_rate**: 0.0001
- **train_batch_size**: 4
- **eval_batch_size**: 4
- **eval_steps**: 500
- **seed**: 42
- **gradient_accumulation_steps**: 2
- **optimizer**: Adam (betas=(0.9, 0.999), epsilon=1e-08)
- **num_epochs**: 4
- **max_steps**: 7500
- **save_steps**: 1500

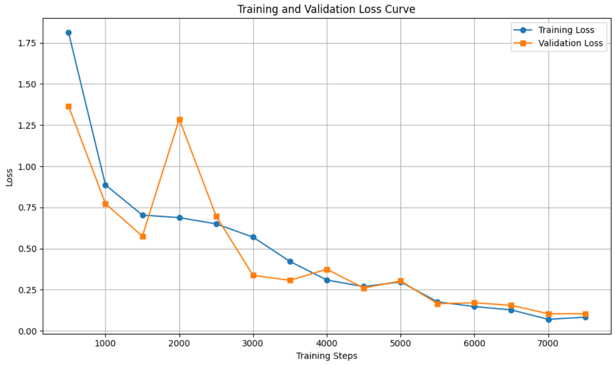


Fig. 3. Training and Validation Loss Over Epochs

VI. Experimental Setup

The full dataset was divided into 80% training, 10% validation, and 10% testing sets. Cross-validation was employed to reduce overfitting risks. Model performance was evaluated using accuracy, precision, recall, and F1-score.

Data augmentation techniques, such as pitch shifting and time stretching, were selectively applied to improve robustness without distorting emotional content.

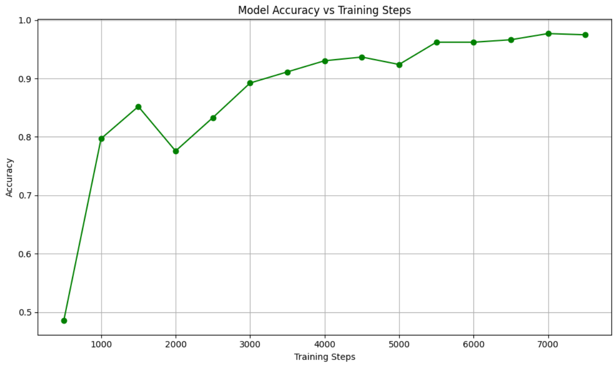
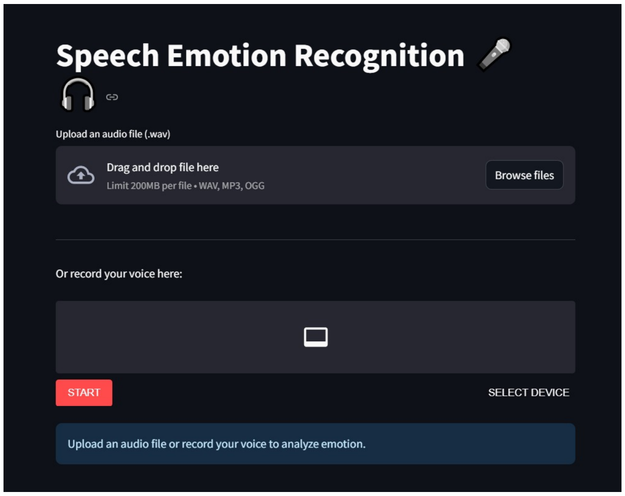


Fig. 4. Training and Validation Accuracy Over Epochs

VII. Results and Discussion

The fine-tuned wav2vec 2.0 model achieved a test accuracy of **97.46%**. The classification report indicated high precision and recall across most classes. Confusion between 'fear' and 'surprise' was observed, likely due to acoustic similarities.



Speech Emotion Recognition 🎤 🎧

Upload an audio file (.wav)

Drag and drop file here
Limit 200MB per file • WAV, MP3, OGG

Or record your voice here:

Upload an audio file or record your voice to analyze emotion.

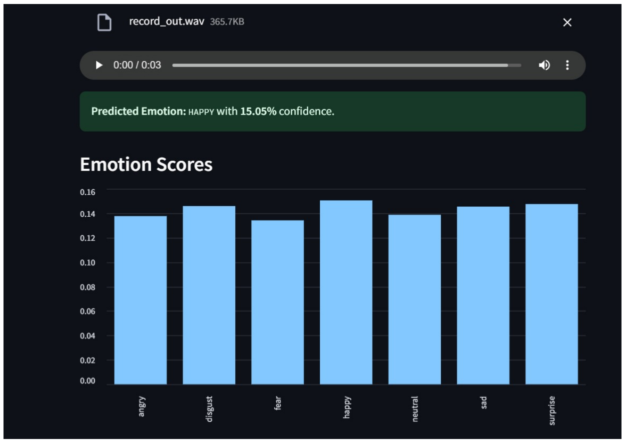
Fig. 5. Streamlit Application Frontend for Real-Time Emotion Detection

Figure 6 illustrates the system's real-time inference result for an uploaded audio file. This visual feedback is crucial for understanding the confidence levels of each predicted emotion class.

VIII. Conclusion and Future Work

This study demonstrates the strength of transfer learning in SER. By leveraging a pre-trained wav2vec 2.0 model, we achieved high performance with limited labeled emotional data. The Streamlit interface bridges the gap between research and end-user deployment.

Future work includes exploring multilingual datasets, incorporating multimodal features (e.g., facial expressions), and optimizing latency for real-time applications on mobile or embedded systems.



record_out.wav 365.7KB

0:00 / 0:03

Predicted Emotion: HAPPY with **15.05%** confidence.

Emotion Scores

Fig. 6. Model Output: Emotion Classification Scores for an Uploaded Audio File

Acknowledgment

We thank the authors and contributors of the SAVEE, RAVDESS, and TESS datasets. Their work was instrumental in training and evaluating our model.

References

[1] A. Baevski et al., “wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations,” *NeurIPS*, 2020.

[2] S. Schneider et al., “wav2vec: Unsupervised Pre-training for Speech Recognition,” *arXiv:1904.05862*, 2019.

[3] L.-W. Chen and A. Rudnicky, “Exploring Wav2vec 2.0 Fine-Tuning for Improved Speech Emotion Recognition,” *arXiv:2110.06309*, 2021.

[4] L. Pepino et al., “Emotion Recognition from Speech Using Wav2vec 2.0 Embeddings,” *arXiv:2104.03502*, 2021.

[5] P. Jafarzadeh et al., “Speaker Emotion Recognition with Wav2Vec2 and HuBERT,” *arXiv:2411.02964*, 2024.

[6] S. Latif et al., “Speech Technology for Health: A Taxonomy and Review,” *ACM Computing Surveys*, 2020.

[7] A. Satt et al., “Efficient Emotion Recognition from Speech Using Deep Learning on Spectrograms,” *Interspeech*, 2017.

[8] M. El Ayadi et al., “Survey on Speech Emotion Recognition: Features, Classification Schemes, and Databases,” *Pattern Recognition*, 2011.

[9] SAVEE Dataset, University of Surrey. [Online]. Available: http://kahlan.eps.surrey.ac.uk/savee/

[10] S. Livingstone and F. Russo, “The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS),” *PLOS ONE*, 2018.

[11] S. Dupuis and P. Kastner, “Toronto Emotional Speech Set (TESS),” 2010. [Online]. Available: https://tspace.library.utoronto.ca/handle/1807/24487