FPT Education

FPT UNIVERSITY

July 19th 2025

DEEP LEARNING PROJECT

Development and Evaluation of a Speech Recognition Model

Presented by Group 1

AI1914

MEMBERS

Tran Hoang Tuan Hung
Model Engineer
mardeusvn@gmail.com

Tran Quoc Huan
Report Author
thung2735@gmail.com

Ho Le Minh Toan
Lecturer
toanhlm@fe.edu.vn

Phan Quoc Anh
Model Engineer
pqa1085@gmail.com

Nguyen Huu Gia Bao
Report Author
blank@email.com

Nguyen Van Anh Duy
Leader
duynvase181823@fpt.edu.vn

Nguyen Truong Phuc Thinh
Presentation Designer
nguyentruongphucthinh@gmail.com

Huynh Han Dong
Report Author
huynhhandong@gmail.com

TABLE
of contents

Abstract

- This project presents the comprehensive design and architecture of a Speech Emotion Recognition (SER) model trained on multiple datasets of emotional speech. It stands as a major milestone for our team, showcasing the skills and knowledge we gained in the DPL302m course.
- The model structure we’re using is a 3D convolutional neural network.
- The primary objective is to achieve high classification accuracy, targeting 80–90% or higher.

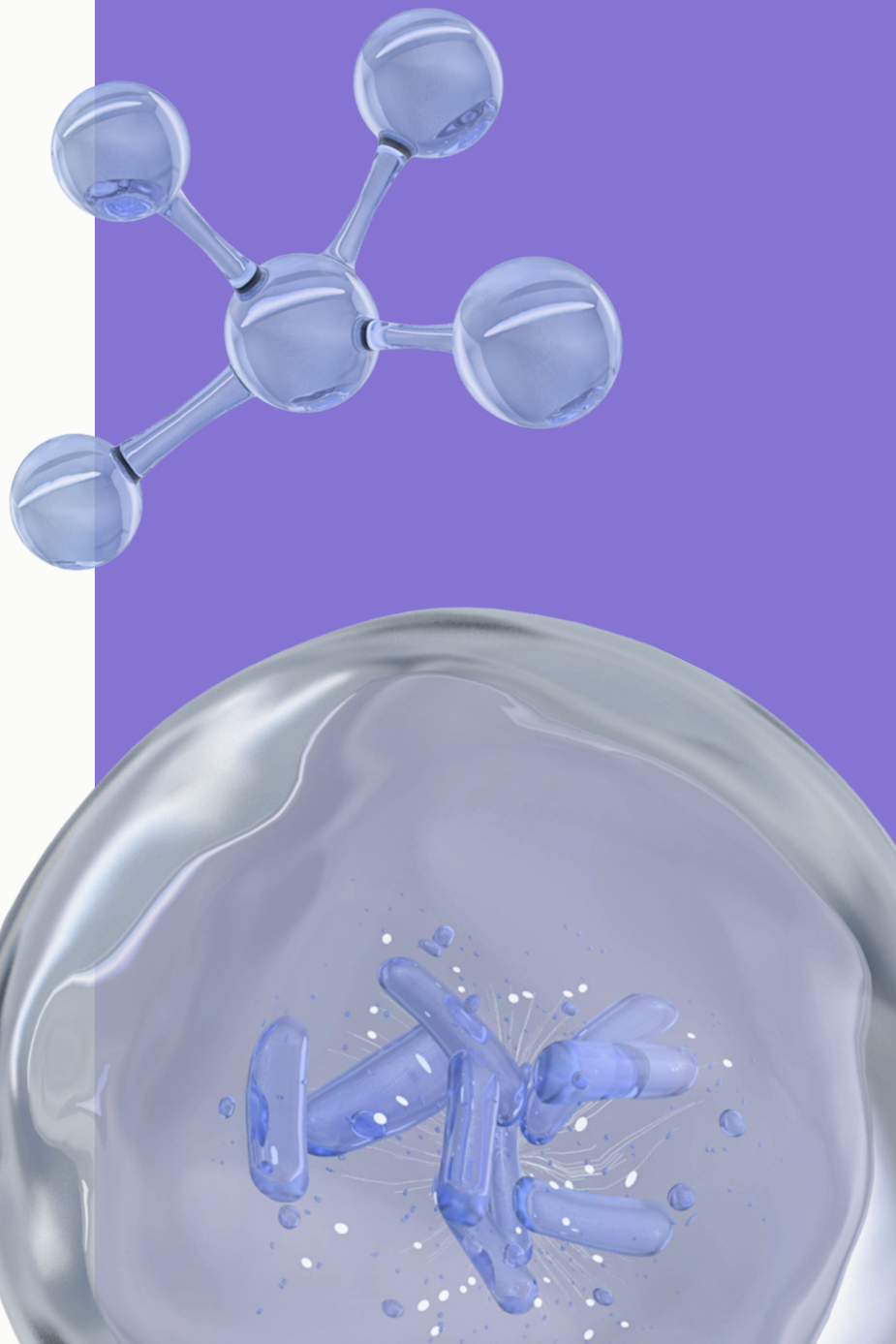
Introduction

Why SER matters?

Speech Emotion Recognition (SER) serves as the bridge between functional AI and emotionally aware AI. By enabling machines to perceive and respond to human emotions, we can expand the role of AI into traditionally human-centric fields such as therapy, education, entertainment, and even conflict resolution. This creates a deeper, more meaningful synergy between humans and machines.

Motivation

Speech emotion is complex and varies by individual. The same sentence can express joy, sarcasm, or sadness depending on tone, pitch, and tempo. We want to solve that challenge.

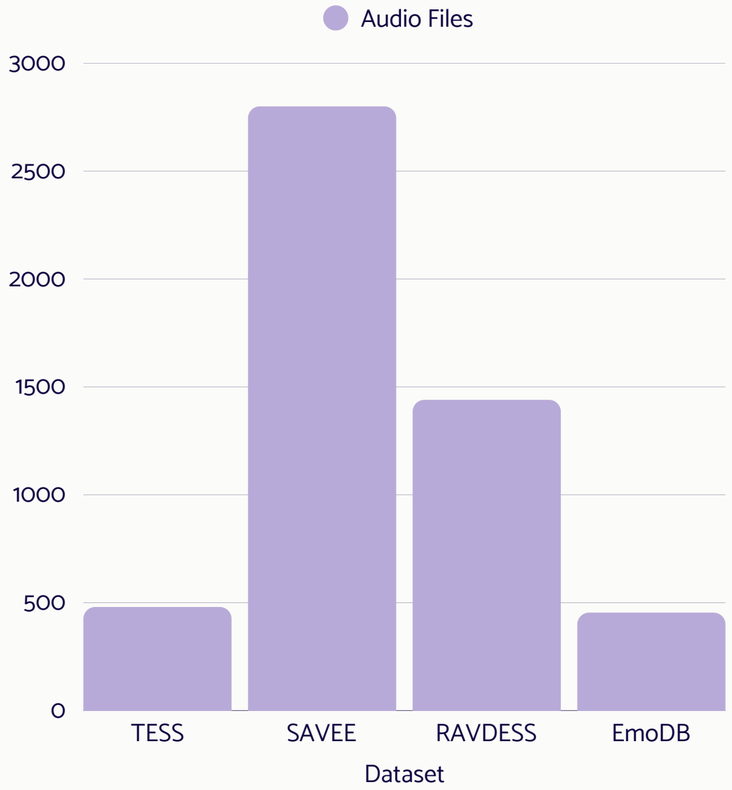


Methodology

This section outlines the development process of our Speech Emotion Recognition model, targeting high accuracy in detecting emotions from voice data.

- **Key Steps**
  - Collect and combine RAVDESS, TESS, SAVEE datasets
  - Preprocess audio into mel-spectrograms
  - Train using 3D CNN and Transformer Encoder
  - Evaluate with accuracy, precision, recall, and confusion matrix

Datasets & Audiomentations



The speech emotion recognition model integrates RAVDESS, TESS, SAVEE and EmoDB datasets. Together, these datasets provide a comprehensive collection of 5147 audio samples spanning seven emotion categories: anger, happiness, sadness, fear, disgust, surprise, neutral.

To boost generalization and mimic real-world audio, we used audiomentations for:

- Shifting: Applies temporal offsets.
- Gaussian noise: Adds noise for low-quality simulation.
- Pitch shifting: Adjusts tone for speaker variation.
- Time stretching: Alters tempo, preserving pitch.

Preprocessing & Feature Extraction

NORMALIZATION

Scale audio amplitudes to a uniform range for consistent analysis.

FEATURE EXTRACTION

Convert waveforms into mel-spectrograms, a 3D representation (fixed_n_frames, n_mels, frame_per_timesteps) capturing frequency and temporal dynamics.

FEATURE-LABEL MAPPING

Link each mel-spectrogram to its emotion label for effective model training.

DATA AUGMENTATION

Enrich the dataset with noise, pitch shifts, time stretching, and audio shifts to boost model resilience and real-world adaptability.

Model Architecture

The model comprises 9 blocks, each designed to process audio features at increasing levels of abstraction, culminating in emotion classification. The architecture leverages 3D convolutions to capture spectro-temporal patterns, with batch normalization and dropout ensuring stable training and preventing overfitting.

3D Convolution Block 1

Low-Level Feature Extraction

- Conv3D (64, 1×5×5, ReLU)
- BatchNormalization
- MaxPooling3D (1×2×2)
- Dropout (20%)

3D Convolution Block 2

Mid-Level Pattern Learning

- Conv3D (128, 1×3×3, ReLU)
- BatchNormalization
- MaxPooling3D (1×2×2)
- Dropout (30%)

3D Convolution Block 3

High-Level Abstraction

- Conv3D (256, 1×3×3, ReLU)
- BatchNormalization
- MaxPooling3D (1×2×2)
- Dropout (30%)

Model Architecture

3D Convolution Block 4
Advanced Pattern Learning

- Conv3D (512, 1×3×3, ReLU)
- BatchNormalization
- MaxPooling3D (1×2×2)
- Dropout (30%)

3D Convolution Block 5
Bottleneck–Feature Expansion

- Conv3D (512, 1×1×1, ReLU)
- BatchNormalization

Global Average Pooling
Feature Consolidation

- GlobalAveragePooling3D

Fully Connected Block 1
High-Capacity Interpretation

- Dense (512, ReLU)
- BatchNormalization
- Dropout (40%)

Fully Connected Block 2
Refinement

- Dense (256, ReLU)
- BatchNormalization
- Dropout (30%)

Fully Connected Block 3
Output

- Dense (7, SoftMax)

Training Strategy

The dataset was split into training, validation and test subsets with a 70/15/15 ratio to ensure we could properly evaluate the model's performance. The model was trained using the categorical cross-entropy loss function to measure how far off the model's predictions are from the actual classes, improving its accuracy for problems with multiple categories.

Additionally, we integrated several callbacks to further improve training stability and convergence:

Early Stopping

Stops training if validation loss doesn’t improve for a set number of epochs, preventing overfitting.

Model Checkpoint

Saves the best-performing model weights based on validation accuracy.

Reduce LR on Plateau

Reduces learning rate when validation loss stops improving, aiding convergence.

Hyperparameter Tuning

- Architecture
- Activation function
- Callbacks
- Regularization techniques
- Loss function

Model Evaluation

Metrics

- Accuracy
- Precision
- Recall
- F1-Score

Confusion Matrix

- Prediction distribution
- High confusion
- Good discrimination

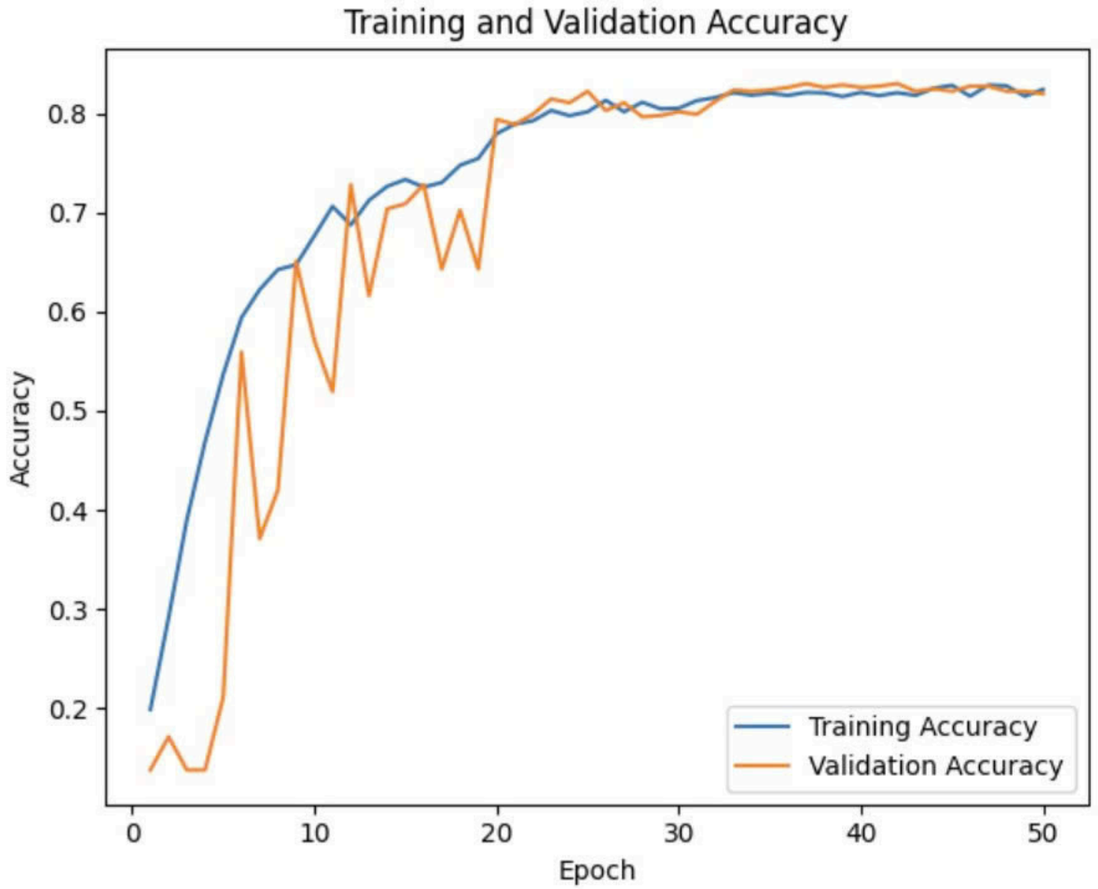
Visualization

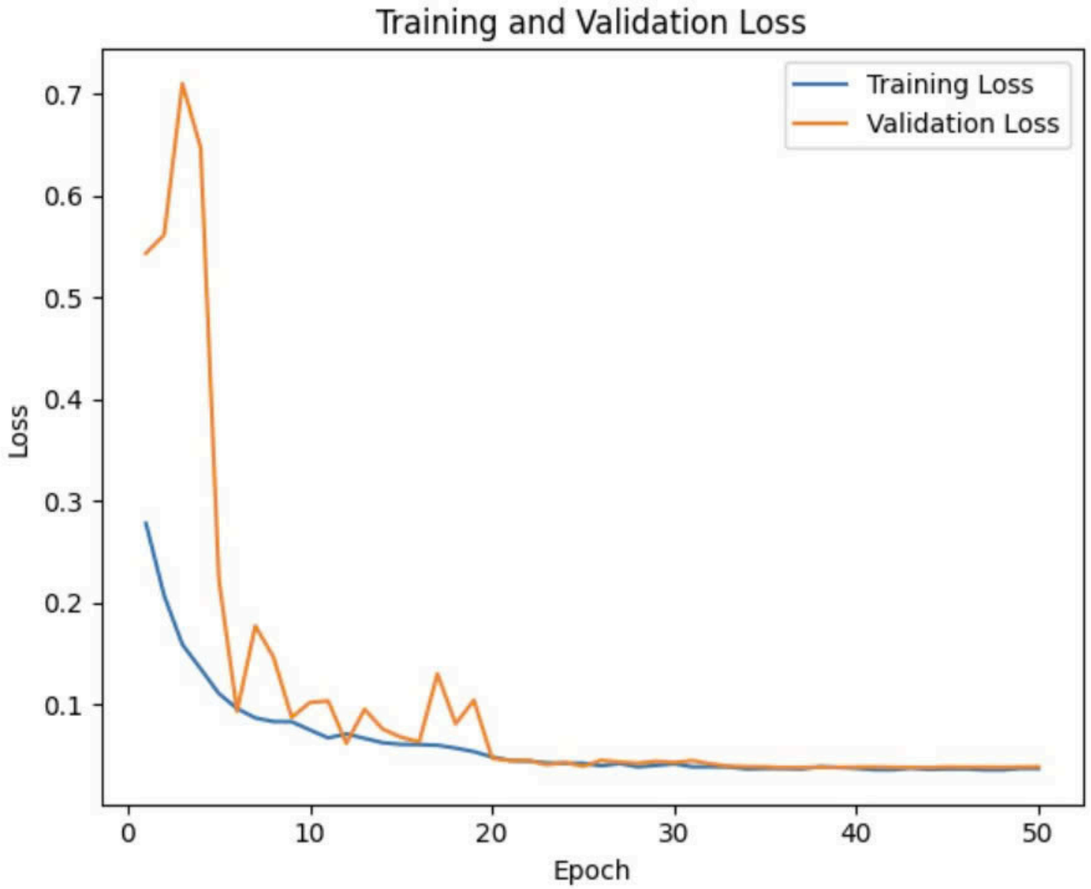
- Convergence of curves
- Metrics divergence
- Validation fluctuations

Error Analysis

- Prosodic features
- Signal-to-noise ratio
- Audio segments length

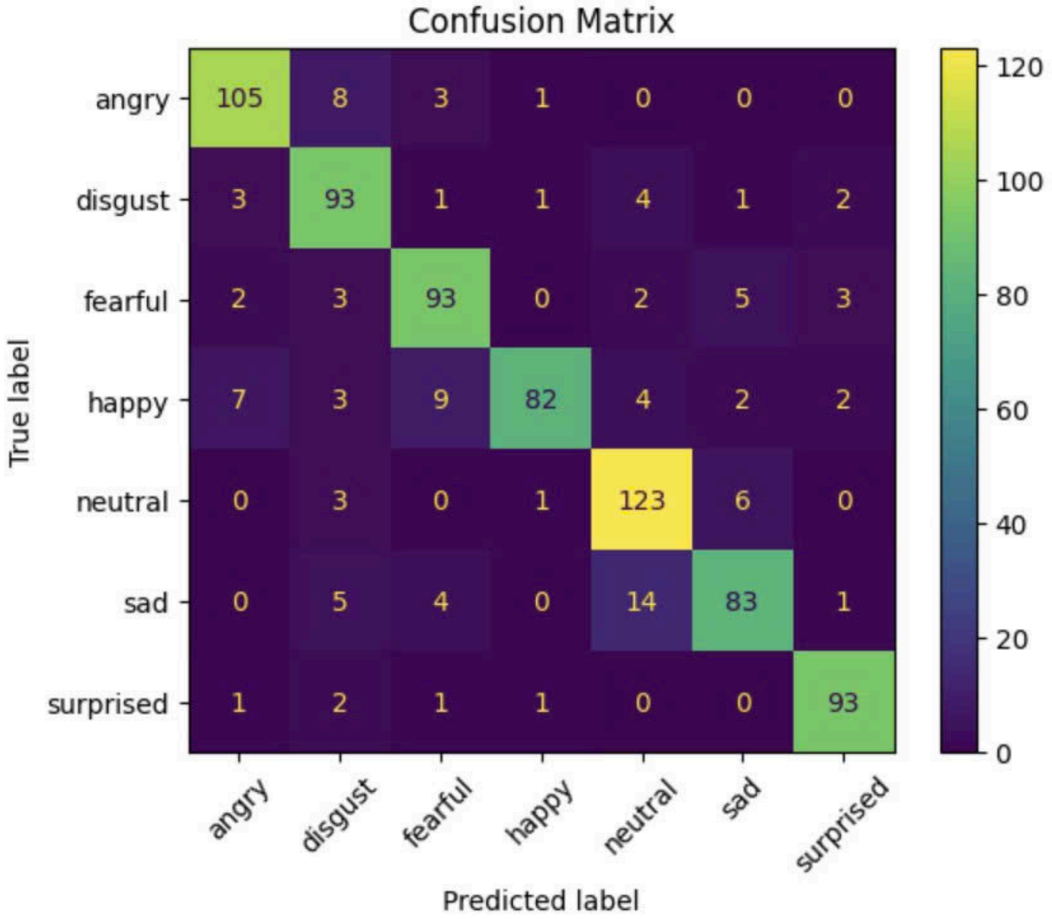
Results


Training and Validation Accuracy
Accuracy
Epoch
Training Accuracy
Validation Accuracy
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0
10
20
30
40
50


Training and Validation Loss
Loss
Epoch
Training Loss
Validation Loss
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0
10
20
30
40
50

Results

	precision	recall	f1-score	support
angry	0.89	0.90	0.89	117
disgust	0.79	0.89	0.84	105
fearful	0.84	0.86	0.85	108
happy	0.95	0.75	0.84	109
neutral	0.84	0.92	0.88	133
sad	0.86	0.78	0.81	107
surprised	0.92	0.95	0.93	98
accuracy			0.86	777
macro avg	0.87	0.86	0.86	777
weighted avg	0.87	0.86	0.86	777



Usage

```python
model = tf.keras.models.load_model(
    'MardeusNet.keras',
    custom_objects={'FocalLoss': FocalLoss}
)
```

[51] Python

```python
testing = []
folder = 'EmoDB\wav'
for audio in os.listdir(folder)[:5]:
    testing.append(os.path.join(folder, audio))
```

[60] Python

```python
for path in testing:
    pred_class, conf, prob_vec = predict_file(path, model, classes)
    print(f'Predicted: {pred_class} (Confidence: {conf:.2f})')
```

[64] Python

```
Predicted: happy (Confidence: 0.44)
Predicted: neutral (Confidence: 0.68)
Predicted: angry (Confidence: 0.90)
Predicted: happy (Confidence: 0.36)
Predicted: neutral (Confidence: 0.43)
```

Conclusion

- The system successfully processes an audio dataset, trains a deep learning model, and predicts emotions with reasonable confidence on an EmoDB test file.
- The model’s performance (based on plots and predictions) suggests it generalizes well, though the slight accuracy drop for some predictions indicates potential for further tuning or data balancing.
- The use of data augmentation and focal loss shows attention to robustness and handling imbalanced classes.

References

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.
Dataset available at https://doi.org/10.1371/journal.pone.0196391

Sahar, M., & Dupuis, K. (2010)
Toronto emotional speech set (TESS). Department of Psychology, Toronto Metropolitan University.
Dataset available at https://tspace.library.utoronto.ca/handle/1807/24487

Haq, S., & Jackson, P. J. B. (2009)
Speaker-dependent audio-visual emotion recognition. In Proceedings of the International Conference on Auditory-Visual Speech Processing (AVSP).
Dataset available at http://kahlan.eps.surrey.ac.uk/savee/Download.html

Burkhardt, F., Paeschke, A., Rolfes, M., Sendlmeier, W. F., & Weiss, B. (2005)
A database of German emotional speech. In Interspeech 2005 (pp. 1517–1520).
Dataset available at https://doi.org/10.21437/Interspeech.2005-482