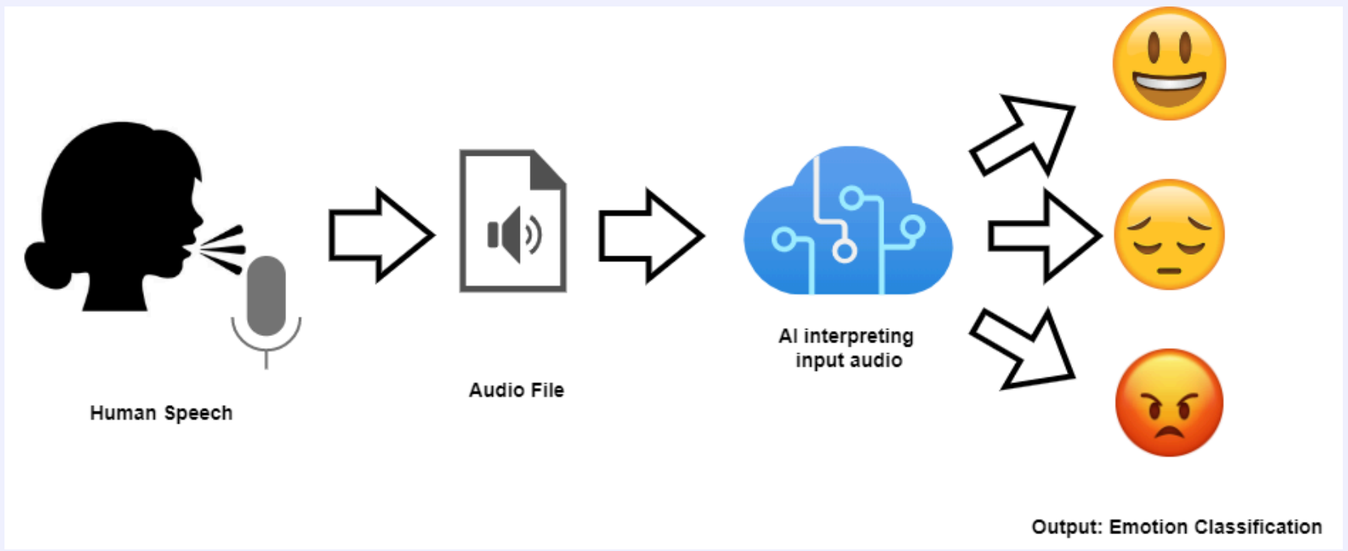


Speech Emotion Recognition Using Ensemble Deep Learning and Multimodal Features: An Improvement Over Gender-Dependent CNN Models

Introduction

Speech Emotion Recognition (SER) is a rapidly advancing field aimed at enhancing human-computer interaction by detecting emotions from vocal patterns.

Traditional models often fall short due to emotional intensity variation, background noise, and speaker diversity.
Our work addresses these limitations by combining deep learning (CNN+LSTM) with traditional classifiers (SVM, Random Forest), using fused audio features (MFCC + spectrogram).

This hybrid approach improves classification accuracy and generalizability, offering potential use in virtual assistants, mental health monitoring, and customer service.

Research Objectives

- To investigate whether an ensemble of CNN+LSTM, SVM, and Random Forest—trained using a combination of MFCC and spectrogram features—can outperform standalone models in classifying 8 emotional states from the RAVDESS dataset.

  We aim to enhance performance through model diversity and multi-feature representation.

Research Question

What factors contribute to the quality and successful completion of EFL learners' undergraduate theses?

Literature Review

📌 Baseline Study: Used CNNs with gender-based model training and only MFCC features.
📌 Gaps: No temporal modeling (no LSTM), no traditional model comparison, no feature fusion, no ensemble strategies.

📌 Our Approach:
- Adds LSTM to handle temporal audio patterns.
- Compares deep learning and traditional models directly.
- Uses an ensemble voting strategy to combine strengths.
- Enhances generalization using fused MFCC + spectrogram features.

Dataset Used

🎧 **RAVDESS Dataset**

- 1440 audio files from 24 actors (balanced male/female)
- 8 emotions: neutral, calm, happy, sad, angry, fearful, disgust, surprise
- Clean, labeled, and preprocessed using file naming conventions
- Dataset partitioned into 80% training, 20% test

Technologies Used

Programming: Python, Jupyter Notebook

🔊 Audio Processing: Librosa

📈 Visualization: Matplotlib, Seaborn

📦 ML Libraries:

TensorFlow/Keras, Scikit-learn
🧪 Deployment Ready: Scaler, Encoder, and Models saved via Pickle

Methodology

Preprocess and extract features:

- MFCCs (n=40)
- Spectrograms (top 40 frequency bins)
- Combined shape: (130, 80)

Train Models Individually:

- CNN+LSTM for sequence learning
- SVM for decision boundaries
- Random Forest for non-linear patterns

Apply Ensemble Voting:

- Predict emotion from majority output of three models

Evaluate using:
- Accuracy
- Confusion Matrix
- Classification Report

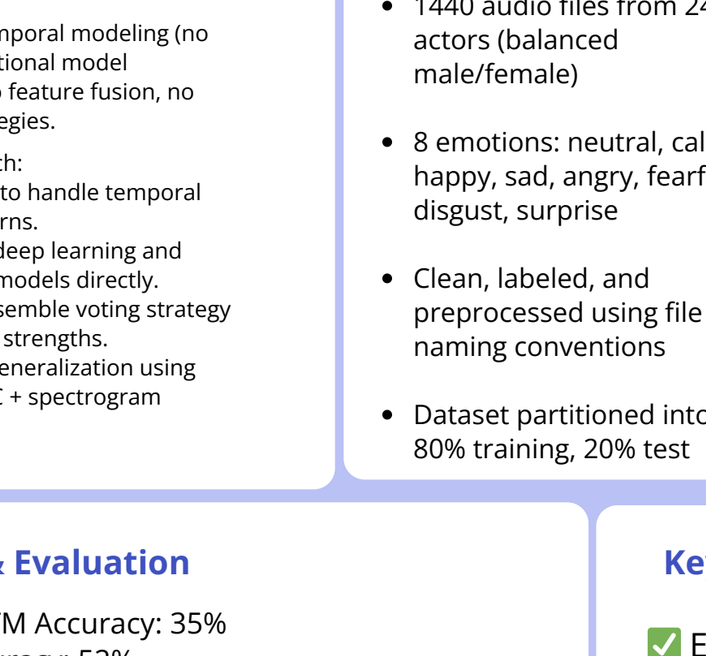
Results & Evaluation

📊 CNN+LSTM Accuracy: 35%
📊 SVM Accuracy: 52%
📊 Random Forest Accuracy: 44%
📊 Ensemble Accuracy: 45%

Confusion matrices show better performance by SVM and ensemble on difficult classes like surprise and calm.
CNN+LSTM struggled due to overfitting but added temporal context that helped ensemble predictions.

Key Insights

✅ Ensemble learning improves robustness
✅ Feature fusion improves emotion separability
✅ Traditional models outperform deep learning in smaller audio datasets
✅ CNN+LSTM is valuable when combined with other models



Conclusion

Our research bridges gaps in existing SER studies by integrating deep and traditional models, using diverse feature sets, and combining them in a voting ensemble.

Future work includes:
- Data augmentation
- Use of larger multimodal datasets (TESS, CREMA-D)
- Incorporating attention/transformer mechanisms
- Live inference and multilingual support

References

- Singh, V., & Prasad, S. (2023). Speech emotion recognition system using gender-dependent convolution neural network. Procedia Computer Science, 218, 2533–2540. Rimberio. (2022). Principles of Language Learning and Teaching Journal
- 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. Rimberio. (2022). Principles of Language Learning and Teaching Journal
- Issa, D., Demirci, M. F., & Yazici, A. (2020). Speech emotion recognition with deep convolutional neural networks. Biomedical Signal Processing and Control, 59, 101894. Rimberio. (2022). Principles of Language Learning and Teaching Journal
- Tzirakis, P., Trigeorgis, G., Nicolaou, M. A., Schuller, B., & Zafeiriou, S. (2017). End-to-end multimodal emotion recognition using deep neural networks. IEEE Journal of Selected Topics in Signal Processing, 11(8), 1301–1309. Rimberio. (2022). Principles of Language Learning and Teaching Journal