LB/TH/30/2026
TH6279

EXPLAINABLE ARTIFICIAL INTELLIGENCE FOR SPEECH EMOTION RECOGNITION

S.S.J. Patabendige

239168X

Master of Science in Data Science and Artificial Intelligence

Department of Computer Science & Engineering

Faculty of Engineering

University of Moratuwa

Sri Lanka

August 2025

EXPLAINABLE ARTIFICIAL INTELLIGENCE FOR SPEECH EMOTION RECOGNITION

S.S.J. Patabendige

239168X

Thesis submitted in partial fulfillment of the requirements for the degree
Master of Science in Data Science and Artificial Intelligence

Department of Computer Science & Engineering
Faculty of Engineering

University of Moratuwa
Sri Lanka

August 2025

DECLARATION

I declare that this is my own work and this thesis/dissertation does not incorporate without acknowledgement any material previously submitted for a degree or diploma in any other University or Institute of higher learning and to the best of my knowledge and belief it does not contain any material previously published or written by another person except where the acknowledgement is made in the text. I retain the right to use this content in whole or part in future works (such as articles or books).

Signature: Date: 13 Oct 2025

The above candidate has carried out research for the Masters thesis/dissertation under my supervision. I confirm that the declaration made above by the student is true and correct.

Name of Supervisor: Dr. Uthayasanker Thayasivam

15 Oct 2025

Signature of the Supervisor: Date:

DEDICATION

This work is dedicated to the pioneers of scientific inquiry and to all those who contribute to the advancement of knowledge. May the insights shared here inspire continued exploration, fostering progress that brings us closer to a brighter, more equitable future.

ACKNOWLEDGEMENT

The successful completion of this project owes much to the invaluable guidance and assistance provided by numerous individuals, and I am deeply grateful for their unwavering support throughout this journey.

Foremost, I extend my heartfelt appreciation to my project supervisors, Dr. Uthayasanker Thayasivam, for their indispensable contribution to this work. Their constant support, patience, motivation, and vast knowledge were instrumental in shaping the outcome of this project.

I would also like to express my gratitude to all the lecturers of the Department of Computer Science and Engineering. Their insightful knowledge, shared during the Master's program, has been a source of inspiration and intellectual growth.

ABSTRACT

The rapid advancement of Artificial Intelligence (AI) has led to its widespread adoption across multiple domains; however, the inherent opacity of conventional AI models has raised concerns regarding their interpretability and trustworthiness. In the context of Speech Emotion Recognition (SER), this lack of transparency limits understanding of how specific acoustic features influence emotional predictions. This research addresses this challenge by developing an explainable and human-centered framework for SER that prioritizes interpretability over classification accuracy.

A comprehensive feature extraction pipeline was designed to derive eighteen interpretable audio features from raw speech signals, grouped into five functional categories: Acoustic, Temporal, Spectral, Harmonic/Tonal, and Spectral Energy Distribution. Traditional machine learning models Artificial Neural Network (ANN), Support Vector Machine (SVM), Random Forest, XGBoost, and k-Nearest Neighbors (k-NN) were employed to classify seven emotional states: Calm, Happy, Sad, Angry, Fearful, Surprised, and Disgust. Two model-agnostic interpretability techniques, SHapley Additive explanations (SHAP) and Local Interpretable Model-Agnostic Explanations (LIME), were used to quantify the contribution of each feature to model predictions.

A novel Aggregated Feature Importance (AFI) metric was introduced to combine interpretability outcomes by weighting each model's contribution according to its predictive performance. The analysis revealed noticeable divergence in feature importance rankings when aggregating across individual models, reflecting their inherent learning differences. However, as additional models were incorporated into the ensemble, these variations gradually diminished, showing a clear convergence and increased stability in AFI-based rankings. Key features such as mfcc_mean, mfcc_std, loudness, zero-crossing rate, and etc. consistently emerged as dominant predictors across emotional classes. Statistical validation using independent t-tests confirmed significant differences ($p < 0.05$) between emotional states, reinforcing the discriminative reliability of these features.

The resulting explainable feature–emotion map provides a transparent interpretation of how specific speech characteristics relate to distinct emotional states, supporting greater trust and usability in emotion-aware AI systems. This work establishes a reproducible framework for feature-level explainability in SER and offers a foundation for future research integrating reverse-engineered and temporally adaptive explainable models.

Keywords: Explainable Artificial Intelligence, Speech Emotion Recognition, Machine Learning, SHAP, LIME, Aggregated Feature Important

Table of Contents

1 Introduction 1
1.1 Background of Explainable AI for Speech Emotion Recognition 1
1.2 Research problem 2
1.3 Research Objectives 3
2 Literature review 4
2.1 Introduction to Artificial Intelligence (AI) and Explainable Artificial Intelligence (XAI) 4
2.1.1 Explainable Artificial Intelligence (XAI) 4
2.2 Value of XAI in Emotion Recognition 9
2.3 Methods of Speech Recognition 10
2.3.1 Feature Importance Analysis in Speech Emotion Recognition (SER) 10
2.3.2 Attention Mechanisms and Intrinsic Interpretation 11
2.3.3 Rule-Based Systems and Machine Learning Approaches 12
2.3.4 Model Visualization and Sonification 13
2.4 XAI in Speech-Emotional Recognition 14
2.4.1 Architectural Explainability in Speech Emotion Recognition 14
2.4.2 Embedding Explainability in Speech Emotion Recognition 15
2.4.3 Research Gaps of XAI for SER 16
3 Methodology 18
3.1 Methodology Overview 18
18
3.2 Step 1 Rational based feature extraction from dataset 18
Objective of step 1 is finding mp3 data set with labeled emotions and feature extraction from mp3 files 18
3.2.1 Mp3 dataset with labeled emotions 18
3.2.2 Dataset Overview 19
3.2.3 Emotional Expressions 19
3.2.4 Filename Structure 19
3.2.5 Licensing and Attribution 19
3.2.6 Feature identification and extraction from mp3 files 20
3.2.7 Rationale for Feature Selection 20

3.2.8 Why These Features Are Relevant for XAI in Voice Recognition..... 23
3.3 Step 2. Model training voice emotional recognize..... 24
3.3.1 Model Selection and Rationale..... 24
3.3.2 Data Preprocessing and Separation..... 24
3.3.3 Model Descriptions and Justification..... 24
3.3.4 Hyper parameter Tuning..... 27
3.4 Step 3. XAI for voice emotional recognize..... 27
3.4.1 Method for Selecting Feature Importance and Analyzing Value Ranges 27
3.4.2 Feature Importance Analysis..... 27
3.4.3 Instance Sampling and Explainability..... 28
3.5 Step 4. Aggregated Feature Importance (AFI) Explanation..... 28
3.6 Step 5. Value Range Analysis..... 30
3.6.1 Descriptive Statistical Analysis..... 30
3.6.2 Statistical Significance Testing with One-Sample t-Test..... 30
4 Results and Analysis..... 33
4.1 Each feature's distribution..... 33
4.2 Selected models with best parameters based on selected parameter grid ... 35
4.3 Each mode accuracy compression..... 36
4.4 Feature Importance Results for ANN, SVM, XGB, RF, KNN with LIME and SHAP..... 40
4.6 Aggregated Feature Importance (AFI) Explanation..... 45
4.7 Top five feature for each emotions..... 49
4.8 Value ranges for best features..... 50
5 Discussion..... 52
6 Conclusion and Future Work..... 55
8 References..... 57

List of tables

TABLE 2.1 Architectural explainability in speech emotion recognition..... 14

TABLE 2.2 Embedding explainability in speech emotion recognition 15

TABLE 4.4.1: Best hyperparamter and validation results for each models..... 35

TABLE 4.4.2: Precision, recall, f1, support value each class for random forest model 36

TABLE 4.4.3: Precision, recall, f1, support value each class for svm model..... 37

TABLE 4.4.4: Precision, recall, f1, support value each class for xgboost model..... 37

TABLE 4.4.5: Precision, recall, f1, support value each class for knn model 38

TABLE 4.4.6: Precision, recall, f1, support value each class for ann model..... 38

TABLE 4.4.7:Feature importance value for each emotions for m = 1..... 45

TABLE 4.4.8:Feature importance value for each emotions for m = 2..... 46

TABLE 4.4.9:Feature importance value for each emotions for m = 3..... 46

TABLE 4.4.10:Feature importance value for each emotions for m = 4..... 47

TABLE 4.4.11:Feature importance value for each emotions for m = 5..... 47

TABLE 4.4.12: Top five features for each emotion after applying afi with m=5 49

TABLE 4.4.13: Value ranges(min, max, mean, std) and t-test value for top five importance features 50

List of figures

Figure 2.2.1: Process of XAI 5
Figure 2.2.2 : XAI on different application domains with percentage.....6
Figure 2.2.3 : XAI adds new dimensions to AI by answering "wh" questions 7
Figure 3.1 :Overall workflow of Methodology.....18
Figure 4.4.1: Distribution of loudness (left) and speaking_rate (write) by Emotions 33
Figure 4.4.2: Distribution of chroma_mean (left) and mfcc_mean (write) by Emotions 33
Figure 4.4.3: Distribution of mfcc_std (left) and spectral_contrast (write) by Emotions 33
Figure 4.4.4: Distribution of zcr (left) and average_pitch (write) by Emotions 34
Figure 4.4.5: Distribution of chroma_std (left) and harmonic_mean (write) by Emotions ...34
Figure 4.4.6: Distribution of percussive_mean (left) and average_pitch (write) by Emotions 34
Figure 4.4.7: Distribution of tempo (left) and mfcc_mean (write) by Emotions 34
Figure 4.4.8: Distribution of spectral_rolloff_std (left) and spectral_rolloff_mean (write) .35
Figure 4.4.9: Distribution of tempo (left) and log_mel_spec_mean (write) by Emotions.....35
Figure 4.4.10:ANN Training & Validation Loss 36
Figure 4.4.11: ANN Training & Validation Accuracy 36
Figure 4.4.12:Precision, Recall and F1 score comparisons 39
Figure 4.4.13: Feature Importance Results for ANN with LIME 40
Figure 4.4.14: Feature Importance Results for ANN with SHAP.....40
Figure 4.4.15: Feature Importance Results for SVM with LIME 41
Figure 4.4.16: Feature Importance Results for SVM with SHAP.....41
Figure 4.4.17: Feature Importance Results for XGBoost with LIME.....42
Figure 4.4.18: Feature Importance Results for XGBoost with SHAP 42
Figure 4.4.19: Feature Importance Results for Random Forest with LIME 43
Figure 4.4.20: Feature Importance Results for Random Forest with SHAP.....43
Figure 4.4.21: Feature Importance Results for KNN with LIME 44
Figure 4.4.22: Feature Importance Results for KNN with SHAP.....44
Figure 4.4.23: Feature impotence value for each emotions after M = 5 with AFI equation...48
Figure 4.4.24: This visualization illustrates the convergence of feature importance values computed using the Aggregated Feature Importance (AFI) equation, as the number of models increases from M = 1 to M = 5. 48
Figure 4.4.25: Global feature importance for emotional recognize 49

List of abbreviations

Abbreviation	Description
XAI	Explainable Artificial Intelligence
SER	Speech Emotion Recognition
SVM	Support Vector Machine
ANN	Artificial Neural Network
RF	Random Forest
XGB	XGBoost
KNN	k-Nearest Neighbors
AFI	Aggregated Feature Importance
SHAP	SHapley Additive explanations
LIME	Local Interpretable Model-agnostic Explanations
CNN	Convolutional Neural Networks
RNN	Recurrent Neural Networks
ERNN	Enhanced Recurrent Neural Network