Emotion Detection from Speech Signals

Nijat Ahmadov

Organizational part

Personal goals

The primary target of this thesis is to acquire practical, hands-on experience in designing and evaluating deep learning models applied to real-world audio signal processing tasks. I aim to deepen my knowledge of speech emotion recognition (SER), refine my aptitudes in feature engineering, and gain proficiency in sequence modeling using hybrid architectures such as CNN-BiLSTM (Convolutional Neural Network and Bidirectional Long Short-Term Memory).

In addition to developing technical expertise, I am committed to enhancing my ability to conduct independent research, engage in critical analysis of results, and present my findings in a clear and coherent manner. This project will also serve as a stepping stone toward further exploration in the fields of affective computing and human-computer interaction.

Time

I plan to dedicate 6-8 focused hours per week to the systematic development and execution of it, structured into focused blocks for data preprocessing, model design, experimentation, result analysis, and so on. To ensure consistent progress and high-quality output, I follow a detailed weekly project plan using agile-inspired workflows, including task prioritization, milestone tracking, and iterative review cycles. This disciplined approach enables me to balance technical implementation with critical reflection, ensuring both the scientific rigor and timely completion of the project within the 4-6 week timeframe.

Compute

The development and training tasks are carried out on an “ASUS VivoBook S15 ” equipped with an Intel Core i5 Ultra processor, 16 GB RAM. This modern hybrid architecture, combining strong CPU performance with dedicated GPU acceleration, enables efficient execution of deep learning workflows using PyTorch and TensorFlow. The system handles preprocessing of audio signals and training of models effectively, especially when working with fixed-length feature sequences. For extended computational demands, I leverage cloud-based platforms, including Google Colab and Kaggle Notebooks, which provide access to high-performance GPUs like the NVIDIA T4, P100, and so forth. This hybrid local-cloud setup ensures flexibility, scalability, and uninterrupted progress throughout the model development lifecycle.

Advisor

Currently, I am actively seeking an advisor for my thesis, with a focus on identifying a researcher specializing in machine learning, speech signal processing, or affective computing within my academic institution. I am targeting professors who lead research in deep learning for audio, natural language processing, or human-computer interaction, as their expertise aligns closely with the technical and theoretical foundations of this project.

To facilitate this process, I am engaging with potential supervisors through departmental research seminars, AI/ML lab meetings, and academic networking events, while also initiating direct outreach via professional email and LinkedIn to present the scope and objectives of my work. My goal is to secure mentorship from an advisor who can provide critical feedback on model design, data interpretation, and methodological rigor, while supporting the scholarly development of this research within the broader context of artificial intelligence in speech analysis.

Once a supervisor has been formally confirmed, their academic profile, institutional affiliation, and relevant professional links such as Google Scholar, ORCID, or CV will be included in this section.

Problem statement

As we know, when we speak, our voices convey far more than just words, they carry emotions. A trembling tone might indicate fear, a loud and rapid delivery can express anger, while a bright, lively voice often reflects happiness. Our brains interpret these subtle cues effortlessly, but training a machine to do the same remains a formidable challenge. Hence, the central problem is “can we develop a system that accurately detects human emotions from speech?”. Although artificial intelligence has made significant strides in understanding spoken language, discerning the speaker’s emotional state remains an open and complex problem.

In that case, emotions are rarely straightforward. The same emotion can manifest differently depending on the individual, their cultural background, or their current state of mind. Variations in audio quality, background noise, and brief speech samples further complicate the task. Moreover, CREMA-D dataset consist of acted emotions rather than spontaneous ones, which risks the model learning performative cues rather than genuine emotional expressions. In addition, some emotions share acoustic similarities that make them difficult to distinguish. For instance, fear and surprise often present with elevated pitch and rapid speech patterns, while disgust and anger share harsh tonal qualities and volume. These overlaps challenge models to accurately differentiate between closely related emotional states. For that reason, the real challenge extends beyond merely predicting emotions. It lies in building models capable of navigating the complexities and nuances of human expression, adapting to diverse voices, and making reliable, unbiased predictions even in imperfect conditions. This project represents a crucial step toward creating machines that do more than listen; they begin to truly understand.

Relevance

Speech emotion recognition (SER) technology has significant practical applications across various domains where understanding human emotions is essential. In mental healthcare, it supports remote patient monitoring by detecting subtle vocal cues linked to depression or anxiety, providing clinicians with objective data between sessions. Customer service operations benefit from real-time emotion analysis during calls, allowing supervisors to intervene when customers become frustrated and helping companies enhance service quality. Voice assistants and chatbots become more effective when they can adjust their responses based on a user's emotional state, creating interactions that feel less robotic and more human. Educational technology platforms can utilize emotion detection to measure student engagement and tailor content delivery accordingly, while automotive systems might detect driver stress to improve safety features.

This research is particularly important now because we are at a turning point where voice interfaces are becoming ubiquitous. People interact with voice assistants daily, remote work has normalized virtual communication, and AI is increasingly expected not only to understand what we say but also how we feel. As society moves toward more natural human-computer interactions, the ability to recognize emotional states from speech transforms technology from merely functional to truly responsive. Without these capabilities, AI systems risk remaining frustratingly tone-deaf in situations where emotional intelligence makes all the difference between helpful and harmful interactions.

Input and output

The system processes speech signals to recognize emotional states through a structured pipeline of audio feature extraction and neural network classification. For input, the model utilizes mono audio recordings from the CREMA-D dataset in ".wav" format at 22,050 hz sampling rate with 16-bit depth. Each audio clip undergoes preprocessing where the signal is normalized to unit amplitude to ensure consistent dynamic range across samples. From these preprocessed audio signals, 4 key acoustic feature sets are extracted using the "librosa" library, they are:

- *Mel-Frequency Cepstral Coefficients (MFCCs)*: 13 coefficients capturing vocal tract characteristics
- *Chromagram features:* 12 pitch classes representing harmonic information
- *Mel-spectrogram*: 128 frequency bands converted to decibel scale
- *Zero-crossing rate*: 1-dimensional temporal feature indicating signal activity

These features are temporally aligned and concatenated to form a comprehensive feature matrix. To handle variable-length speech segments (ranging from 1-5 seconds), all feature sequences are standardized to a fixed length of 300 time frames through zero-padding for shorter clips and truncation for longer ones. The final input tensor has dimensions (300, 154) where 154 represents the combined feature dimension (13 + 12 + 128 + 1).

For output, the model predicts among six discrete emotional categories, (angry, disgust, fear, happy, neutral, and sad). These labels are represented as one-hot encoded vectors during training, with the final softmax layer producing a probability distribution across all six classes. The system outputs both the predicted emotion label and associated confidence score, enabling nuanced interpretation of classification certainty in real-world applications.

Metrics

To evaluate the performance of the speech emotion recognition system, I employ a comprehensive set of metrics that capture both overall effectiveness and nuanced class-specific behavior. Given the six-class classification nature of the problem, accuracy alone would be insufficient due to potential class imbalance and the varying importance of different misclassifications in emotion recognition. Metrics include:

- *Accuracy*: The proportion of correctly classified samples out of all samples. While simple, it provides an intuitive high-level performance indicator.
- *F1-score:* The harmonic mean of precision and recall. This accounts for class imbalance and provides a balanced measure of precision and recall across all emotion categories.
- *Precision, Recall, and F1-score*: These reveal which emotions the model identifies well and which are frequently confused, offering insights into specific strengths and weaknesses.
- *Confusion Matrix*: A visual representation of classification results that highlights common misclassifications.

For this project, I achieved a test accuracy of 53% with a weighted F1-score of 0.52. These values are meaningful because a plethora of approaches on CREMA-D typically achieve 50-65% accuracy depending on methodology and preprocessing. The moderate performance reflects the inherent challenges of emotion recognition such as emotional expressions are highly subjective, culturally variable, and often overlap acoustically.

Eventually, these metrics are achievable due to the architecture effectively captures both local acoustic patterns and temporal dynamics, while class weighting mitigates the dataset's natural imbalance.

Data

I utilize the CREMA-D dataset (Crowd-Sourced Emotional Multimodal Actors Dataset), which represents one of the most comprehensive and widely used resources for speech-based emotion analysis. The dataset was accessed programmatically through Kaggle using the kagglehub Python library, enabling direct integration into my data pipeline without manual downloads.

The dataset contains 7,442 audio clips recorded from 91 professional actors (48 male, 43 female) representing diverse ethnic backgrounds and age ranges (20-74 years). Each object in the dataset is a short speech utterance in ".wav" format, originally sampled at 48 kHz with 16-bit depth. In my implementation, I resampled these to 22,050 Hz to balance computational efficiency with audio quality preservation. Additionally, each audio clip represents one of 6 emotional categories, angry, disgust, fear, happy, neutral, and sad. The recordings feature standardized sentences from the IEEE Harvard sentences corpus and other neutral phrases, with each clip lasting between 1 and 5 seconds.

A notable characteristic of this dataset is its moderate class imbalance with fewer samples for "disgust" (approximately 8% of the dataset) compared to more frequently represented emotions like "happy" and "neutral" (each around 18%). Moreover, the emotional expressions are acted rather than spontaneous, which introduces a performance element that may not perfectly reflect natural emotional speech patterns. There is also inherent speaker variability, with some actors contributing significantly more samples than others, though the dataset was designed with demographic balance in mind.

Links to datasets

- CREMA-D (Crowd-Sourced Emotional Multimodal Actors Dataset) - https://www.kaggle.com/datasets/ejlok1/cremad

Validation

To ensure the reliability and generalizability of the speech emotion recognition model, I implemented a rigorous validation strategy that accounts for both dataset characteristics and the challenges inherent in emotion classification. Rather than using a simple random split, I adopted a speaker-independent validation approach where entire speakers are kept exclusive to either the training or testing set. This methodology prevents data leakage and better reflects real-world scenarios where the system must recognize emotions from voices it has never encountered during training. The dataset was partitioned into training (80%) and testing (20%) sets at the speaker level, ensuring that all audio samples from a given actor appear exclusively in one set. This speaker-stratified approach maintains the natural distribution of emotional expressions while eliminating the risk of over-optimistic performance metrics that would occur with sample-level splitting. In addition, I used “EarlyStopping” monitored validation loss with a patience of 5 epochs, halting training when no improvement was observed and restoring the best weights. Then, “ReduceLROnPlateau” reduced the learning rate by 50% when validation loss plateaued for 3 consecutive epochs, enabling finer convergence near optimal minima. It ensures that the reported performance metrics accurately reflect the model's ability to generalize to new speakers and emotional expressions, rather than merely memorizing patterns within the training data.

Solutions overview

Speech Emotion Recognition Using Convolutional Neural Networks

Makhijani, V., & Yang, Y. (2017). Emotion Recognition in Speech Using Convolutional Neural Networks.

This research demonstrated how CNNs applied to spectrogram representations could effectively capture emotional cues in speech. The authors achieved approximately 50% accuracy on the IEMOCAP dataset, establishing a baseline that many subsequent works, including my project, have sought to improve upon through architectural enhancements and feature engineering.

The CREMA-D Dataset: A Resource for Affective Computing

Cao, H., Mehta, D., Xiao, Z., & Georgiou, P. G. (2014). The CREMA-D: Crowd-Sourced Emotional Multimodal Actors Dataset. IEEE Transactions on Affective Computing.

This seminal work introduced the CREMA-D dataset, which has become a standard benchmark for speech emotion recognition research. The authors collected 7,442 audio-visual clips from 91 actors expressing six basic emotions, with labels validated through crowd-sourcing. Their methodology established best practices for emotion elicitation and labeling that my project directly builds upon

Performance Improvement of SER Systems by Combining 1D CNN and LSTM with Data Augmentation

Pan, X., & Wu, Y. (2023). Performance Improvement of SER Systems by Combining 1D CNN and LSTM with Data Augmentation. MDPI Journal of Artificial Intelligence.

This research presents a CLDNN (CNN-LSTM-DNN) model for speech emotion recognition, using MFCC features and data augmentation to improve performance on RAVDESS, EMO-DB, and IEMOCAP. By combining convolutional and recurrent layers, the approach captures both spatial and temporal audio patterns, outperforming baseline models and informing my own architecture and augmentation choices.

An Ensemble 1D-CNN-LSTM-GRU Model with Data Augmentation for Speech Emotion Recognition

Ahmed, M., Rahman, M. M., & Islam, M. R. (2022). An Ensemble 1D-CNN-LSTM-GRU Model with Data Augmentation for Speech Emotion Recognition.

This paper proposes an ensemble model combining 1D CNN for local feature extraction, LSTM for long-term dependencies, and GRU for efficient temporal modeling in speech emotion recognition. Enhanced with data augmentation to address limited data and class imbalance, the approach improves robustness and generalization across emotion categories, guiding my choices in architecture design and preprocessing strategies.

Survey of Deep Learning-Based Multimodal Emotion Recognition

Lian, S., Liu, Y., & Wang, Q. (2023). Survey of Deep Learning-Based Multimodal Emotion Recognition (Speech, Text, and Face). ACM Computing Surveys, 56(2), Article 35.

This survey reviews recent advances in multimodal emotion recognition, emphasizing deep learning methods that combine speech, text, and facial cues. It examines fusion strategies, dataset traits, and evaluation metrics. Although my work focuses on unimodal speech, the paper offers valuable context on cross-modal interactions and fusion techniques that could inspire more robust unimodal approaches, situating my research within the broader emotion recognition field.

Experiments

Baseline

To establish a baseline, I implemented a Random Forest classifier using time-averaged acoustic features. For each audio sample, I calculated the mean and standard deviation of MFCCs, chroma, mel-spectrogram, and ZCR, producing a 308-dimensional feature vector. The model, built with 100 trees in scikit-learn and class weighting to handle imbalance, offers an efficient and interpretable benchmark for speech emotion recognition.

Baseline metrics and results

The Random Forest baseline achieved 42% accuracy and a weighted F1-score of 0.40. Performance varied by emotion, with highest recall for “neutral” (47.3%) and “happy” (45.1%), and lowest for “disgust” (26.8%). The confusion matrix showed common misclassifications between acoustically similar emotions, notably “angry” mislabeled as “disgust” (32.1%) and “fear” as “sad” (29.4%). These findings highlight the limitations of ignoring temporal dynamics and set a baseline for evaluating the main model’s performance.

Main solution

The primary model uses a hybrid architecture combining convolutional and recurrent layers to capture both local acoustic features and long-term emotional dynamics. Input tensors of shape (300, 154) represent standardized audio features. It starts with a Conv1D layer (32 filters, kernel size 3) with ReLU activation and L2 regularization, followed by Batch Normalization, MaxPooling, and Dropout for stability. Furthermore, a Bidirectional LSTM layer (32 units) captures temporal context in both directions. The output passes through a dense ReLU layer (32 units) with Dropout before a 6-neuron softmax layer predicts emotions. Eventually, the model was compiled with the Nadam optimizer (learning rate 0.0005) and categorical cross-entropy loss with label smoothing (0.1). Training ran for 20 epochs with batch size 32, using class weights to address imbalance and callbacks like EarlyStopping and ReduceLROnPlateau to optimize training.

Main model metrics and results

The main model achieved a test accuracy of 53% with a weighted F1-score of 0.51, representing a 11 percentage point improvement over the Random Forest baseline. Class-specific performance revealed the model's strengths and limitations. For example, "happy" and "neutral" emotions were recognized with highest accuracy, while "disgust" remained challenging. The confusion matrix showed reduced but persistent confusion between "fear" and "sad" and between "angry" and "disgust".

Code repository

The complete implementation of this speech emotion recognition system is publicly available in a GitHub repository:

https://github.com/1453nicat/Emotion-Detection-from-Speech-Signals-with-Machine-Learning

Usage

Upon completion, the trained model is ensuring flexibility for deployment across diverse platforms. For instance,

- *Customer Service Analytics,* integration with call center systems to analyze customer emotions in real time, equipping agents with emotional insights that enhance communication and resolution effectiveness.
- *Mental Health Monitoring,* mobile applications that track vocal patterns longitudinally to identify early signs of mood changes in individuals experiencing depression, anxiety, or other affective disorders.
- *Emotionally Aware Voice Assistants,* enhancing virtual assistants like Siri or Alexa with emotional intelligence to deliver more empathetic, context-aware responses.

For production, the model can be converted to lightweight formats such as TensorFlow Lite for mobile and embedded devices or ONNX for cross-platform interoperability. The system outputs both predicted emotion labels and confidence scores, allowing applications to implement threshold-based actions. Furthermore, future developments may include streaming inference capabilities for continuous, real-time emotion tracking during conversations, offering richer, more dynamic emotional context than single-utterance classification. This would further enhance applications in human-computer interaction, mental health, and customer experience.