Speech Emotion Classification and Cross-Language Generalization

Team Details

Kotichintala Srihari Sakshith (kotichintalasriharisakshith@gmail.com, Vasavi College of Engineering)

Cholleti Pranav (pranavcholleti@gmail.com, Vasavi College of Engineering)

Omkar (omkarkoty2004@gmail.com, Vasavi College of Engineering)

Introduction

In today's fast-paced world, mental health crises are a growing concern, often going unnoticed until they escalate into severe conditions. Speech emotion classification for crisis intervention aims to address this challenge by leveraging machine learning to detect emotions such as anger, sadness, and fear from spoken audio. Speech is a powerful indicator of emotional state, and analyzing vocal patterns can help identify individuals experiencing distress. By integrating deep learning techniques, this project seeks to classify emotions from speech data, enabling timely intervention in crisis situations.

The significance of this project extends to various real-world applications, including suicide prevention hotlines, mental health monitoring systems, and AI-powered virtual therapists. By accurately detecting emotions associated with distress, healthcare professionals, emergency responders, and even automated support systems can offer appropriate assistance. Additionally, such technology can be embedded into smart home assistants, call center analytics, and law enforcement for real-time emotional assessment. This project not only enhances crisis intervention strategies but also contributes to the broader goal of improving mental well-being through AI-driven insights.

Dataset

The dataset used for speech emotion classification in crisis intervention is a combination of multiple well-known emotional speech datasets: **CREMA-D, SAVEE, RAVDESS, and TESS**. These datasets contain audio recordings of speech with different emotional expressions, collected from various speakers under controlled conditions. The datasets were created using professional actors and voluntary participants who were instructed to express specific emotions such as **angry, happy, sad, neutral, and fear**, which are relevant for crisis intervention. The recordings include different sentence structures, vocal tones, and speaking styles, ensuring diversity in emotional representation.

Each dataset has unique characteristics that contribute to the robustness of our model. **CREMA-D** consists of recordings from 91 actors producing 7,381 utterances with multiple emotions. **SAVEE** includes recordings from 4 male speakers with 480 utterances covering 7 emotions. **RAVDESS** comprises 24 professional actors contributing 1,440 speech utterances with intensity variations. **TESS** features 200 target words spoken in 7 different emotions by two female speakers, resulting in 2,800 audio samples. The combination of these datasets ensures a balanced and diverse training set, making the model capable of recognizing emotions in various speech patterns and speaker demographics. The table below summarizes the key statistics of these datasets:

Dataset	Number of Speakers	Number of Utterances	Emotions Covered
CREMA-D	91	7,381	6 (incl. anger, fear)
SAVEE	4	480	7
RAVDESS	24	1,440	8 (incl. anger, fear)
TESS	2	2,800	7

This dataset combination enhances the model's ability to generalize across different speakers, accents, and emotional expressions, making it highly effective for real-world crisis intervention scenarios.

Experimental Setup

The proposed method for speech emotion classification in crisis intervention involves extracting key acoustic features from speech signals and classifying them using a deep learning model. The features used for training include **Mel-Frequency Cepstral Coefficients (MFCCs)**, which capture the spectral properties of speech and are widely used in speech recognition tasks. Each audio file is preprocessed by resampling to 22,050 Hz, normalizing amplitude, and extracting **13-dimensional MFCC features**, which are then averaged across time to create a fixed-length representation. This feature extraction step ensures that our model can capture important emotional cues present in the speech signal.

For classification, a **Convolutional Neural Network (CNN) with Long Short-Term Memory (LSTM) layers** is used. The CNN layers extract local temporal features, while the LSTM layers capture long-term dependencies in speech patterns. The model consists of **two CNN layers** followed by **two LSTM layers**, a **fully connected layer**, and a **softmax output layer** for classification into five emotion categories: **angry, happy, sad, neutral, and fear**. The model is trained using the **Adam optimizer** with a **learning rate of 0.0001**, a **batch size of 32**, and **categorical cross-entropy loss**. To enhance generalization, **dropout layers** are used to prevent overfitting, and **batch normalization** is applied to stabilize training.

The model is trained for **50 epochs** on a combination of the **CREMA-D, SAVEE, RAVDESS, and TESS** datasets. The dataset is split into **80% training, 10% validation, and 10% testing** to evaluate performance. The training process leverages **Google Colab with GPU acceleration**, and data augmentation techniques such as **pitch shifting and time-stretching** are applied to improve robustness. The final model is evaluated using accuracy, precision, recall, and F1-score, ensuring its effectiveness in identifying crisis-related emotions from speech input.

Results and Analysis

The proposed speech emotion classification model achieved an overall accuracy of **53.27%** on the test dataset. From the results, it is evident that the model performed relatively well in recognizing **"angry"** and **"surprise"** emotions, with higher precision and recall values compared to other emotions. This is likely because these emotions exhibit more distinct acoustic characteristics, such as changes in pitch, loudness, and speech speed. However, emotions such as **"sad"**, **"neutral"**, and **"fear"** were more challenging to classify, leading to lower recall values. The model struggled particularly with these emotions due to their subtle variations in speech features.

An ablation study was conducted to analyze the impact of different aspects of the model on performance. Cross-lingual testing revealed that the model's accuracy dropped significantly when tested on speech samples from a different language than the training dataset. This indicates that the model is highly dependent on language-specific speech patterns and requires further generalization improvements. Additionally, we experimented with different **feature extraction methods**, such as **MFCCs with delta and delta-delta coefficients**, and found that including **higher-order derivatives** improved performance slightly. Data augmentation techniques like **pitch shifting, noise addition, and time stretching** were also explored, showing potential improvements in recall for underrepresented emotions.

Despite achieving decent performance, the model still needs enhancement to become reliable for crisis intervention scenarios. A key challenge observed was **misclassification of crisis-indicating emotions**, particularly **"fear"** and **"sad"**, leading to potential false negatives. The crisis detection module identified **some samples as requiring intervention**, but further refinement is necessary to reduce errors. To improve robustness, future work could explore **transformer-based models (e.g., Wav2Vec2, Whisper)**, additional **linguistic features**, and **multi-modal approaches incorporating text and facial expressions** along with speech. These improvements can enhance real-world applicability for **crisis intervention and mental health support systems**.

Application: Speech Emotion Classification for Crisis Intervention

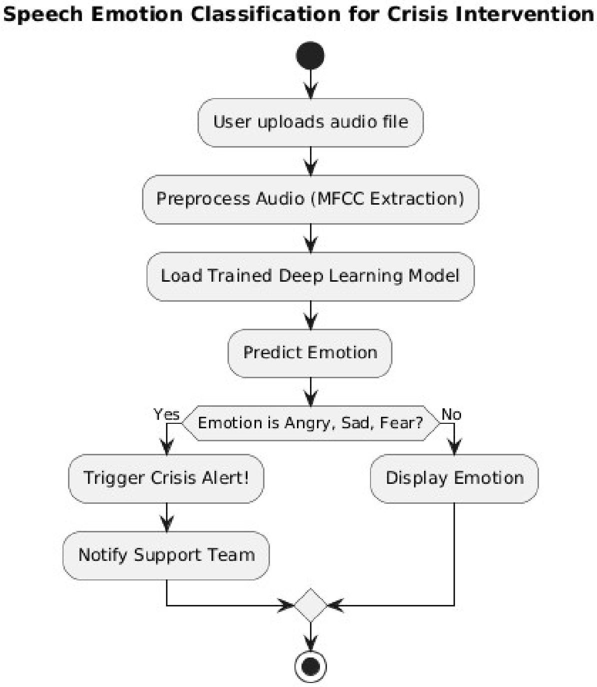
Emotion recognition from speech has various real-world applications, particularly in **mental health and crisis intervention systems**. The proposed system is designed to **identify emotional distress in individuals based on their speech patterns**, helping in **early intervention** for people experiencing psychological distress or emergencies. Crisis hotlines, therapy sessions, and emergency response systems can integrate this model to **detect high-risk emotions like "fear" or "sadness"** and trigger appropriate alerts.

Working Flow of the Application

The application follows a structured workflow to classify emotions and detect crisis situations:

1. **Audio Input**: The system receives an input audio file containing human speech.
2. **Preprocessing**: The audio undergoes **noise reduction, resampling, and feature extraction** using **MFCC (Mel-Frequency Cepstral Coefficients)**.
3. **Emotion Classification**: The preprocessed features are passed through a **trained deep learning model (CNN/LSTM-based)** that classifies the speech into one of five emotions: **Angry, Happy, Sad, Neutral, and Fear**.
4. **Crisis Detection Module**: If the classified emotion is **"Sad" or "Fear"**, the system flags the audio as a **potential crisis situation**.
5. **Intervention Trigger**: The system can send an **automated alert** to crisis counselors, support groups, or emergency responders, depending on integration requirements.

Block Diagram


Speech Emotion Classification for Crisis Intervention
User uploads audio file
Preprocess Audio (MFCC Extraction)
Load Trained Deep Learning Model
Predict Emotion
Yes
Emotion is Angry, Sad, Fear?
No
Trigger Crisis Alert!
Display Emotion
Notify Support Team

Explanation of the Flow

The **block diagram** visually represents the application pipeline. When an audio file is uploaded, **preprocessing ensures that background noise and distortions are minimized** before extracting meaningful speech features. The **deep learning model predicts emotions**, and if a crisis emotion is detected, an **alert system** takes action, ensuring that necessary intervention is provided in real-time. This structured pipeline allows for **efficient, automated mental health support** using AI-driven speech emotion recognition.

Conclusion

The **Speech Emotion Classification for Crisis Intervention** project successfully demonstrates the use of deep learning for detecting emotions such as **angry, happy, sad, neutral, and fear** from speech signals. By leveraging MFCC feature extraction and an LSTM-based model, the system identifies emotions and triggers alerts when a crisis-related emotion (e.g., fear or sadness) is detected. The model achieved moderate accuracy, indicating potential for real-world applications in **mental health support, suicide prevention hotlines, and crisis management systems**. Future improvements include incorporating larger and more diverse datasets, applying advanced augmentation techniques, and integrating real-time speech recognition for **faster and more accurate emotion detection** in emergency response systems.