Project Proposal

I. Introduction

Emotion recognition from speech is a crucial technology with broad applications, including mental health diagnostics, human-computer interaction, and customer service automation. Understanding emotions in speech enables AI systems to provide more natural and empathetic responses, enhancing user experience and decision-making processes.

State-of-the-art (SOTA) speech models like Wav2Vec 2.0 and HuBERT have demonstrated remarkable accuracy in speech-based emotion recognition. However, these models require substantial computational resources, making them impractical for deployment on edge devices and real-time applications with limited processing power. To address this challenge, knowledge distillation (KD) provides an effective approach by compressing the knowledge from a high-performing "teacher" model into a more compact and efficient "student" model without significant loss in accuracy.

This project proposes training a lightweight emotion recognition model using KD, leveraging the CREMA (Crowd-Sourced Emotional Multimodal Actors) dataset. The dataset consists of audio recordings labeled with six primary emotions: Sadness (SAD), Anger (ANG), Disgust (DIS), Fear (FEA), Happiness (HAP), and Neutral (NEU). By distilling knowledge from a large-scale, computationally expensive model into a smaller, optimized model, we aim to develop an efficient emotion recognition system capable of running on resource-constrained devices while maintaining high classification performance.

II. Objectives

A. Primary Objective

Develop a computationally efficient student model using knowledge distillation to replicate the performance of a state-of-the-art (SOTA) teacher model on the CREMA dataset for emotion recognition from vocal expressions.

B. Secondary Objectives

Train a compact deep learning model that minimizes computational resources while maintaining high performance. Improve accuracy and generalization in recognizing emotions from vocal expressions. Optimize for real-time inference on resource-constrained devices (e.g., smartphones). Evaluate and compare the student model against the teacher model on multiple metrics:

- Classification accuracy
- F1-score
- Inference speed
- Model size
- Memory usage

III. Methodology

A. Dataset

The Crowd-sourced Emotional Multimodal Actors Dataset (CREMA) is a well-established resource for speech emotion recognition research. This dataset consists of 7,442 original audio clips from 91 actors (48 male and 43 female) of diverse ethnic backgrounds. Each actor recorded a set of 12 sentences with varying emotional intensities across six distinct emotional states.

Dataset Characteristics

1)Size: 7,442 audio samples

2)Participants: 91 actors (48 male, 43 female)

3)Demographics: Diverse ethnic backgrounds and age ranges

4)Audio Format: 16-bit WAV files sampled at 16kHz

5)Duration: Audio clips range from approximately 1 to 5 seconds

The dataset contains recordings expressing six emotional states:

- SAD (Sadness)
- ANG (Anger)
- DIS (Disgust)
- FEA (Fear)
- HAP (Happiness)
- NEU (Neutral)

B. Teacher and Student Model Selection

SOTA Models: Utilize pretrained models such as Wav2Vec as the teacher model.

Student Model Architecture: Implement a lightweight model, such as a CNN, LSTM, or MobileNet-style network, to serve as the student model.

Knowledge Distillation Strategy:

Utilize soft labels (teacher's probabilistic outputs) and hard labels (ground truth).

Apply temperature scaling to soften teacher logits for better knowledge transfer.

3.4 Training Pipeline

Train the teacher model on the CREMA dataset to ensure high performance.

Extract teacher predictions (logits) for the training set to be used in distillation.

Train the student model using both ground truth labels and the teacher's soft targets to achieve optimal performance with reduced computational complexity.

IV. Timeline

- Week 1-2: Dataset preprocessing and selection of the SOTA model.
- Week 3-4: Fine-tuning the teacher model.
- Week 5-6: Implementing knowledge distillation techniques.
- Week 7-8: Training, evaluation, and optimization of the student model.
- Week 9: Final testing and documentation.

V. Team

Name : Shripad Pate
Roll no :M23MAC007

VI. Conclusion

This project aims to democratize emotion recognition by creating a compact, efficient model using KD. By leveraging CREMA and SOTA teacher models, we expect to achieve near-teacher accuracy with significantly lower computational costs, enabling real-time applications in healthcare, education.