Lightweight Emotion Recognition from Speech using Knowledge Distillation

Shripad Pate
IIT Jodhpur
Jodhpur,India
m23mac007@iitj.ac.in

Anshav Vayeda
IIT Jodhpur
Jodhpur,India
m24csa034@iitj.ac.in

April 16, 2025

Abstract

Emotion recognition from speech plays a critical role in human-computer interaction, offering applications in areas such as mental health monitoring, customer service, and virtual assistants. However, high-performing deep learning models often demand significant computational resources, making them unsuitable for deployment in real-time and on low-power devices. This project addresses this challenge by employing knowledge distillation [2], where a powerful but computationally intensive Word2Vec-based model [1] serves as the teacher to train a lightweight student model. Using the CREMA-D dataset, which contains diverse emotional speech samples, the student model learns to classify emotions such as happiness, anger, sadness, fear, disgust, and neutrality with high efficiency. This approach aims to maintain competitive performance while significantly reducing the model size, inference time, and memory usage, making it viable for real-world, edge-based applications.

1 Introduction

Emotion recognition from vocal expressions is an increasingly vital area of research, enabling machines to interpret human affect and respond more empathetically. From mental health assessment tools to emotionally aware voice assistants, the ability to detect emotions in speech is foundational to developing responsive and human-centric AI systems. While deep learning models like Wav2Vec and HuBERT have achieved state-of-the-art performance in speech emotion recognition, their high computational requirements pose barriers to real-time and resource-constrained deployment.

To overcome this limitation, this project applies knowledge distillation (KD) as a solution to transfer the learned representations from a large, powerful teacher model to a compact student model. Specifically, we utilize a Word2Vec-based model as the teacher, leveraging its ability to capture semantic information in the speech signal. The student model, a lightweight convolutional neural network (CNN), is trained to mimic the output distributions (soft labels) of the teacher alongside the ground-truth labels. This dual supervision strategy enhances the student model's learning and generalization capabilities.

Using the CREMA-D dataset, which includes audio samples labeled with six distinct emotions, our objective is to design an efficient model that retains high classification performance while being optimized for real-time inference and low-power environments. This work contributes to the democratization of affective computing, enabling the deployment of emotion-aware systems across a wide range of accessible devices.

2 Dataset

For this study, we employ the **Crowd-sourced Emotional Multimodal Actors Dataset (CREMA-D)** [?], a high-quality and widely used benchmark dataset for emotion recognition in speech. It provides an extensive and balanced collection of emotional audio samples recorded under controlled conditions, allowing for the training and evaluation of deep learning models with strong generalization capabilities.

2.1 Overview and Composition

CREMA-D consists of a total of **7,442 audio clips** performed by **91 professional actors**, comprising **48 male** and **43 female** speakers. These actors were selected to represent a diverse population across different **ethnicities**, **age groups**, and **regional accents**, making the dataset well-suited for real-world speech modeling tasks.

Each actor was asked to read **12 semantically neutral sentences** that were carefully constructed to minimize inherent emotional bias in the verbal content. These sentences serve as a consistent linguistic framework across recordings, allowing models to focus on paralinguistic cues (e.g., intonation, pitch, and stress) to detect emotions.

2.2 Emotion Classes and Expression Levels

Each sentence was acted out with one of six core emotions:

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

Moreover, each emotional utterance is recorded in **two levels of intensity**:

- **Low/Normal intensity**
- **High/Strong intensity**

This dual-level annotation provides a nuanced understanding of expressive variability, enabling the study of both categorical and dimensional emotion recognition frameworks.

2.3 Audio Specifications

All recordings are captured in a clean, noise-controlled environment to preserve high audio fidelity. The key technical specifications are:

- **File Format:** WAV
- **Bit Depth:** 16-bit
- **Sampling Rate:** 16 kHz
- **Channels:** Mono
- **Clip Duration:** Approximately 1 to 5 seconds

These standardized formats ensure compatibility with deep learning toolkits and facilitate preprocessing steps such as framing, windowing, and spectrogram generation.

2.4 Annotation Protocol

Each audio clip is evaluated by a minimum of **six human annotators** recruited via a crowd-sourcing platform. Annotators were instructed to label the perceived emotion expressed in each utterance. The final emotion label is determined using a **majority voting mechanism**. This method has demonstrated strong **inter-rater agreement**, contributing to the overall reliability of the dataset.

In addition to categorical labels, metadata such as speaker ID, sentence ID, emotion intensity, and gender are also provided, which are valuable for speaker-level and demographic-based analysis.

2.5 Relevance to Deep Learning

CREMA-D's structured, labeled, and diverse dataset design makes it highly compatible with supervised learning pipelines in emotion recognition. Key advantages include:

- **Emotionally expressive and phonetically controlled speech samples**
- **Balanced representation across gender, emotion, and intensity**
- **High-quality labels obtained through crowd-sourced validation**
- **Robust audio specifications for deep acoustic modeling**

This dataset is particularly well-suited for training and evaluating knowledge distillation systems, where compact student models are trained to mimic the performance of larger teacher models on emotion classification tasks. The consistent and controlled setting of CREMA-D ensures that the performance of both teacher and student models can be reliably assessed without extraneous variability from environmental noise or inconsistent labeling.

3 Methodology

Knowledge distillation to create an efficient audio emotion recognition system. We transfer knowledge from a large pre-trained model to a compact neural network while preserving classification performance. The methodology encompasses model architecture design, knowledge distillation framework, dataset preparation, and training protocol.

3.1 Model Architecture

3.1.1 Teacher Model

We utilize the `Wav2Vec2ForSequenceClassification` architecture from HuggingFace Transformers as our teacher model. This model is based on the pre-trained `facebook/wav2vec2-large-960h-lv60-self` weights and comprises 94.6M parameters. The architecture includes:

- A Wav2Vec2 feature encoder that processes raw audio waveforms
- A transformer-based contextual encoder for feature representation
- A classification head consisting of a dropout layer (rate = 0.1) and a linear projection layer that maps to the emotion classes

The teacher model is fine-tuned on the CREMA-D dataset using cross-entropy loss to establish strong baseline performance for emotion recognition before serving as the knowledge source for distillation.

3.1.2 Student Model (LiteCNN)

Our proposed student model, LiteCNN, is a lightweight convolutional neural network with 1.08M parameters (approximately 87× smaller than the teacher). The architecture incorporates:

- **Initial Block:** A 1D convolution layer (kernel_size = 80, stride = 2) followed by batch normalization and ReLU activation.
- **Feature Extraction Blocks:** Three sequential blocks of depthwise separable convolutions, each comprising:
  - Depthwise convolution (groups equal to input channels)
  - Pointwise convolution (1×1 kernel) for channel mixing
  - Batch normalization and ReLU activation
  - Max pooling for downsampling (kernel_size = 3, stride = 2)
- **Classifier:** Global adaptive pooling followed by two fully connected layers with dropout (rate = 0.2) and ReLU activation, culminating in a softmax layer for class probability estimation

The use of depthwise separable convolutions significantly reduces computational complexity while maintaining representational capacity, making the model suitable for deployment in resource-constrained environments.

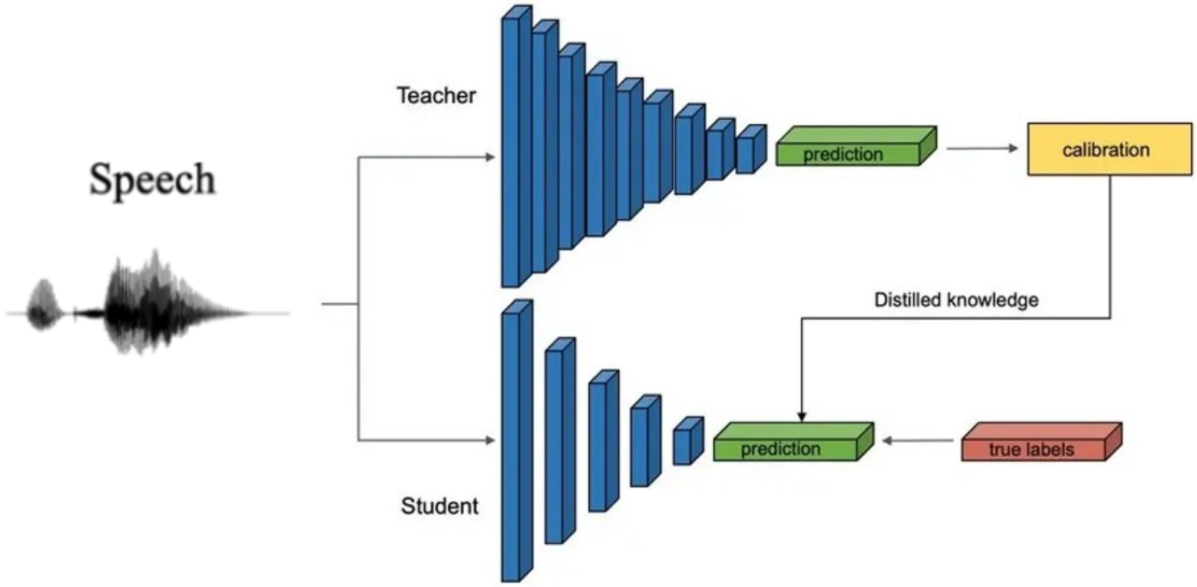

Figure 1: Overview of the knowledge distillation framework. The teacher model (Wav2Vec2) transfers knowledge to a lightweight student model (LiteCNN), enabling efficient audio emotion recognition.

3.2 Knowledge Distillation Framework

Knowledge distillation transfers the generalization capabilities of the teacher model to the student through an augmented training objective. Our implementation follows these principles:

3.2.1 Distillation Loss Function

We define a custom DistillationLoss that combines two components:

1. **Task Loss ($L_{\textbf{task}}$)**: Standard cross-entropy between student predictions and ground truth labels.
2. **Distillation Loss ($L_{\textbf{distill}}$)**: Kullback-Leibler divergence between the softened probability distributions of teacher and student models.

The total loss is computed as:

$$L_{\text{total}} = \alpha \cdot L_{\text{task}} + (1-\alpha) \cdot T^2 \cdot L_{\text{distill}}$$

Where:

- α is the balancing coefficient (set to 0.5 in our experiments),
- T is the temperature parameter (set to 2.0) that softens the probability distributions.

3.2.2 Temperature Scaling

Both teacher and student logits are divided by temperature T before applying the softmax function:

$$p_i = \text{softmax}\left(\frac{z_i}{T}\right)$$

This scaling reveals the finer structure of the teacher's knowledge by smoothing the probability distribution, allowing the student to learn subtle relationships between classes rather than just the dominant class.

3.3 Data Processing

3.3.1 Dataset

We utilize the CREMA-D (Crowd-sourced Emotional Multimodal Actors Dataset) containing audio recordings of actors expressing six distinct emotions: anger (ANG), disgust (DIS), fear (FEA), happiness (HAP), sadness (SAD), and neutral (NEU).

3.3.2 Preprocessing Pipeline

Raw audio files undergo the following preprocessing steps:

1. Resampling to 16 kHz to standardize input representation

2. Normalization to zero mean and unit variance to stabilize training

3. Fixed-length processing by either padding (with zeros) or truncating to 16,000 samples (1 second)

4. For the teacher model, audio is tokenized via the Wav2Vec2Processor

5. For the student model, raw waveforms are directly input into the convolutional layers

3.3.3 Dataset Partitioning

The dataset is split into training (80%) and validation (20%) sets using stratified sampling to maintain class distribution across both sets.

3.4 Training Protocol

3.4.1 Optimization Strategy

- **Optimizer:** Adam with an initial learning rate of 1e-4 and weight decay of 1e-5

- **Scheduler:** ReduceLROnPlateau with patience = 5 and factor = 0.5, monitoring validation loss

- **Batch Size:** 32

3.4.2 Training Procedure

1. The teacher model is fine-tuned on the CREMA-D dataset until convergence.

2. Teacher logits are computed for all training samples and cached.

3. The student model is trained using the distillation framework with both ground truth labels and teacher logits.

4. Training proceeds for 200 epochs with early stopping (patience = 20) based on validation loss.

5. Model checkpoints are saved based on best validation performance.

3.4.3 Implementation Details

- **Hardware:** NVIDIA GPUs with CUDA acceleration

- **Frameworks:** PyTorch 1.10, torchaudio, HuggingFace Transformers 4.18.0

- **Reproducibility:** All experiments use a fixed random seed (42)

This methodology enables the creation of a lightweight emotion recognition model through knowledge distillation, balancing computational efficiency with classification performance.

4 Results

We evaluate the performance of three models: the Teacher model, a lightweight Student model, and a Knowledge Distilled (KD) model trained using soft targets from the teacher network. The results are summarized in Table 1, showing training and validation accuracy, precision, and recall for each model.

The Teacher model, which is typically a larger and more powerful architecture, achieves the highest performance on both training and validation sets. It reaches over 93% training accuracy and 62% validation accuracy, along with balanced precision and recall. These results confirm the strong generalization ability of the Teacher model.

Metric	**Teacher**	**Student**	**KD**
Training			
Acc	0.937	0.494	0.568
Prec	0.938	0.519	0.552
Rec	0.937	0.496	0.560
Validation			
Acc	0.624	0.389	0.455
Prec	0.630	0.372	0.465
Rec	0.632	0.387	0.452

Table 1: Comparison of Teacher, Student, and KD models on key metrics.

In contrast, the Student model, which has significantly fewer parameters and is trained from scratch, performs poorly with less than 45% training accuracy and under 39% validation accuracy. This highlights the challenge of training small models effectively, especially when labeled data is limited or the task is complex.

To address this, we apply knowledge distillation by training the student network to match the softened outputs of the teacher. The KD model demonstrates clear improvements over the baseline student model. Training accuracy improves from 44.4% to 49.8%, while validation accuracy increases from 38.9% to 41.5%. Similar trends are observed in precision and recall.

Although the KD model does not reach the performance of the teacher, the gains over the student model are significant and demonstrate the potential of distillation to improve small model performance. These improvements are particularly valuable in real-world applications where inference time, memory, and compute are constrained (e.g., on edge devices or mobile hardware).

Moreover, the results suggest that the KD model has learned more generalized representations compared to the student, despite being trained with the same architecture. This indicates that the soft targets provided by the teacher network contain informative structure about class relationships that the student benefits from during learning.

In future work, we plan to experiment with temperature scaling, improved loss balancing between task and distillation losses, and data augmentation techniques to further improve the KD model's generalization. Exploring intermediate layer matching and attention transfer could also enhance knowledge transfer beyond final output logits.https://www.kaggle.com/datasets/ejlok1/cremad

References

[1] Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations. *Neural Information Processing Systems*, 33:12449–12460, 6 2020. 1

[2] Geoffrey E. Hinton, Oriol Vinyals, and Jeffrey Dean. Distilling the knowledge in a neural network. *arXiv (Cornell University)*, 1 2015. 1