Speech Emotion Recognition

Archita Goyal
Dept. of CSE, IITK
architag23@iitk.ac.in

Aritra Ambudh Dutta
Dept. of CSE, IITK
aritraad23@iitk.ac.in

Harshpreet Kaur
Dept. of CSE, IITK
harshpreet23@iitk.ac.in

Saksham Verma
Dept. of CSE, IITK
sakshamv23@iitk.ac.in

Suyash Kapoor
Dept. of CSE, IITK
suyashk23@iitk.ac.in

***Abstract*—Speech audio models are crucial for advancing research in speech-related technologies and are crucial in applications like automatic speech recognition and emotion recognition. In this paper, we build several machine learning models to classify speech sounds based on eight emotion categories: neutral, calm, happy, sad, angry, fearful, disgust, and surprised. Our system derives 64 audio features, namely Mel frequency spectral coefficients (MFCCs) and Mel spectrograms, from the original speech signals. Initial model exploration with Decision Trees, Random Forests, and Multi-Layer Perceptrons yielded moderate performance (42%, 64%, and 50% accuracy, respectively). Hence, we introduced a deep learning model based on the Convolutional Neural Network (CNN), which consists of 3 hidden layers and efficiently classifies seven different emotional states on the given data, indicating its feasibility for real-time human–machine interaction.**

***Index Terms*—Mel frequency spectral coefficients, Decision Trees, Random Forests, Multi layer Perceptrons, Convolutional Neural Network, Deep Learning**

I. Objective

By examining vocal traits and patterns, this project aims to create a model for identifying emotions in speech audio files. By giving systems the ability to recognize and process emotions, this model seeks to improve human-computer interaction. The suggested system has useful applications in emotion-aware systems for sentiment analysis, virtual assistants for enhancing user engagement, and therapeutic applications for mental health evaluation and assistance.

II. Data Acquisition and Pre-processing

A. Dataset

We utilized a custom emotional speech dataset comprising recordings from 19 speakers for training and evaluation. Each speaker contributed approximately 60 audio files, resulting in a total of 1140 samples. The recordings capture various emotions expressed in controlled settings, categorized into eight distinct classes: neutral (01), calm (02), happy (03), sad (04), angry (05), fearful (06), disgust (07), and surprised (08). The dataset was processed using a custom AudioDataset class that handles the loading and transformation of audio files.

B. Pre-processing Pipeline

Our preprocessing pipeline consists of the following steps:

- Dataset Organization: Audio files are organized in folders by actor ("Actor_*") and processed to extract emotion labels from filenames, mapping emotion codes to numerical classes (0-7).
- Audio Loading and Conversion: Raw audio files are loaded using torchaudio and stereo files are converted to mono by averaging channels.
- Spectrogram Generation: Audio signals are transformed into mel spectrograms using a window size of 1024, hop length of 512, and 64 mel bands, then converted to decibel scale for better representation.
- Input Preparation: Single-channel spectrograms are expanded to three channels using repeat(3, 1, 1) to match CNN requirements, and a custom pad_collate function ensures spectrograms can be batched together by padding to uniform width.
- Data Splitting: The dataset is split 80:20 for training and validation using stratified sampling to maintain class distribution across sets.

For initial models, we implemented an alternative feature extraction approach that extracted 64 MFCCs using librosa (with hop length of 1% and FFT window of 2% of sample rate), computed additional spectral features (centroids and bandwidth), and aggregated these into 62-dimensional feature vectors. This approach included validation steps for data integrity checking.

III. Previous Models

A. Decision Tree

Hyperparameter tuning was performed on the decision tree classifier using a grid search with splitting criterion fixed to 'gini'. We iterated over candidate values for parameters such as max depth, min samples required to split a node, and min samples at a leaf and the model's performance was evaluated using cross-validation accuracy came out to be 42%. It struggled with subtle differences in data, and relationship between features and classes being non-linear.

B. Random Forest

This model is set up with Gini score for splitting evaluation, square-root feature choice at each node, and a set of 5000 decision trees in ensemble. Bootstrapped sampling is used in training, to promote diversity and minimize overfitting. The model is then tested on the validation set which resulted in 64% Validation accuracy, which would have improved if neutral emotion also had equal number of audio files but was still pretty good as compared to the Decision Tree model.

C. MLP

The MLP model consists of two hidden layers (128 and 256 neurons, ReLU activation) and a softmax output layer. It was trained using Adam optimization with sparse categorical cross-entropy loss over 100 epochs with accuracy being 53%. Deeper architectures or CNNs would better capture the speech patterns.

IV. Final Model Architecture

A. Input Specifications

The model processes mel spectrograms with dimensions: $64 \times 400 \times 3$ derived from audio signals sampled at 48 kHz.

Key parameters:

- 1024-point FFT with 512-hop length
- 64 mel bands converted to decibel scale
- Batch size: 32 samples

B. Feature Extraction Backbone

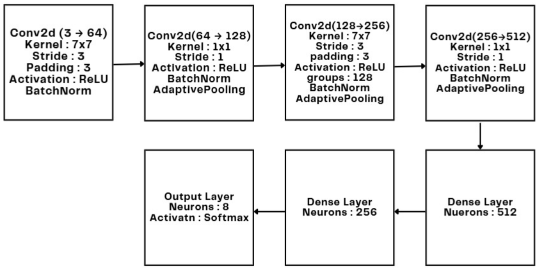


Fig. 1. CNN Architecture

Architecture components shown in Fig. 1.

1) Initial Convolutional Block:

- Layer 1: 2D convolution with 64 filters
  - Kernel size: 7×7, stride = 2, padding = 3
  - Output: $64 \times 200 \times 200$
- Batch normalization and ReLU activation
- Max pooling:
  - Kernel size: 3×3, stride = 2
  - Output: $64 \times 100 \times 100$

2) Depthwise Separable Blocks:

Three blocks progressively increase channel depth while reducing spatial dimensions:

- **Block 1:**
  - Depthwise convolution:$64 \times 3 \times 3$ kernel, groups= 64
  - Pointwise convolution:1×1 kernel, 128 out channels
  - Max pooling: 2×2
  - Output: $128 \times 50 \times 50$
- **Block 2:**
  - Depthwise convolution:$128 \times 3 \times 3$kernel, groups=128
  - Pointwise convolution:1×1 kernel, 256 out channels
  - Max pooling: 2×2
  - Output: $256 \times 25 \times 25$
- **Block 3:**
  - Depthwise convolution:$256 \times 3 \times 3$kernel,groups= 256
  - Pointwise convolution:1×1 kernel, 512 out channels
  - Global average pooling: 1×1
  - Output: $512 \times 1 \times 1$

Each block includes batch normalization and ReLU activation. Total parameters in the feature extractor: **1.2M**.

C. Classification Head

- Flatten operation → Dropout(0.5)
- FC(512,256) + ReLU → Dropout(0.5)
- Output layer: FC(256,8) + Softmax

Total parameters 1.3M(Feature extractor 1.2M,Classifier 132k)

D. Key Architectural Features

1) Depthwise Separability:

$$\text{Params}_{dw} = C_{in}(K^2 + C_{out}) \quad (1)$$

Reduces parameters by 58% compared to standard convolutions.

2) Progressive Downsampling:

- 16× spatial reduction ($200 \times 200 \rightarrow 12 \times 12$)
- Channel depth increases 8× ($64 \rightarrow 512$)

3) Regularization Strategy:

- Dual dropout (p=0.5)
- L2 weight decay ($\lambda = 0.0001$)
- Batch normalization after each convolution

E. Training Configuration

TABLE I
Training Parameters

Parameter	Value
Optimizer	Adam ($\beta_1 = 0.9$, $\beta_2 = 0.999$)
Initial LR	0.001
LR Schedule	ReduceOnPlateau (factor=0.1, patience=5)
Loss	Categorical Cross-Entropy
Augmentation	SpecAugment (time warp ±30%, freq masking)
Epochs	70

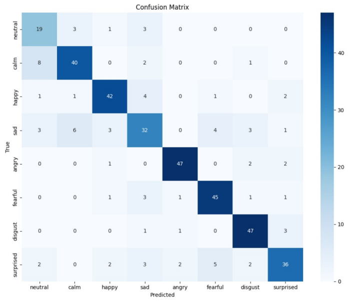
TABLE II
Performance Comparison

Model	Params	Acc (%)
Decision Tree	-	42
Random Forest	-	64
MLP	128k	50
Proposed CNN	**1.3M**	**79.4**

V. Model Performance

Final CNN model achieved following performance metrics:

Validation Accuracy: 79.4% Precision (weighted): 0.79
Recall (weighted): 0.79 F1 Score (weighted): 0.79



The confusion matrix Fig. 1 showed that the model performed particularly well in distinguishing between distinct emotions like ”happy” and ”sad,” but had some confusion between similar emotions like ”neutral” and ”calm.” Per-class performance analysis revealed that ”angry” and ”surprised” emotions were recognized with highest accuracy 85%, while ”neutral” and ”calm” had relatively lower recognition rates (approximately 70%, likely due to their subtle acoustic differences.)

References

[1] Tanvi Puri, Mukesh Soni, Gaurav Dhiman, Ehtiram Raza Khan, Osama Ibrahim Khalaf, Malik Bader Alazzam , ”Detection of Emotion of Speech for RAVDESS Audio Using Hybrid Convolution Neural Network” , Journal of Healthcare Engineering , February , 2022, [Online] , Available : here

[2] K.Simonyan and A. Zisserman, ”Very Deep Convolutional Networks for Large-ScaleImage Recognition” , Available : here

[3] M. Takalkar, M. Xu,Q. Wu, and Z. Chaczko, ”A survey: facial emotion recognition using deeplearning,” Smart Computing and Communication, ,2018, Available here

[4] Jianzhu Guo, Dingyun Zhang, Xiaoqiang Liu, Zhizhou Zhong, Yuan Zhang, Pengfei Wan, Di Zhang. LivePortrait: Efficient Portrait Animation with Stitching and Retargeting Control. July, 2024, Available here