Speech Emotion Recognition

Presented by Group-5

Team Members

Dhruv Varshney

Gautam Arora

Sumit Kumar

Dwij Om Oshoin

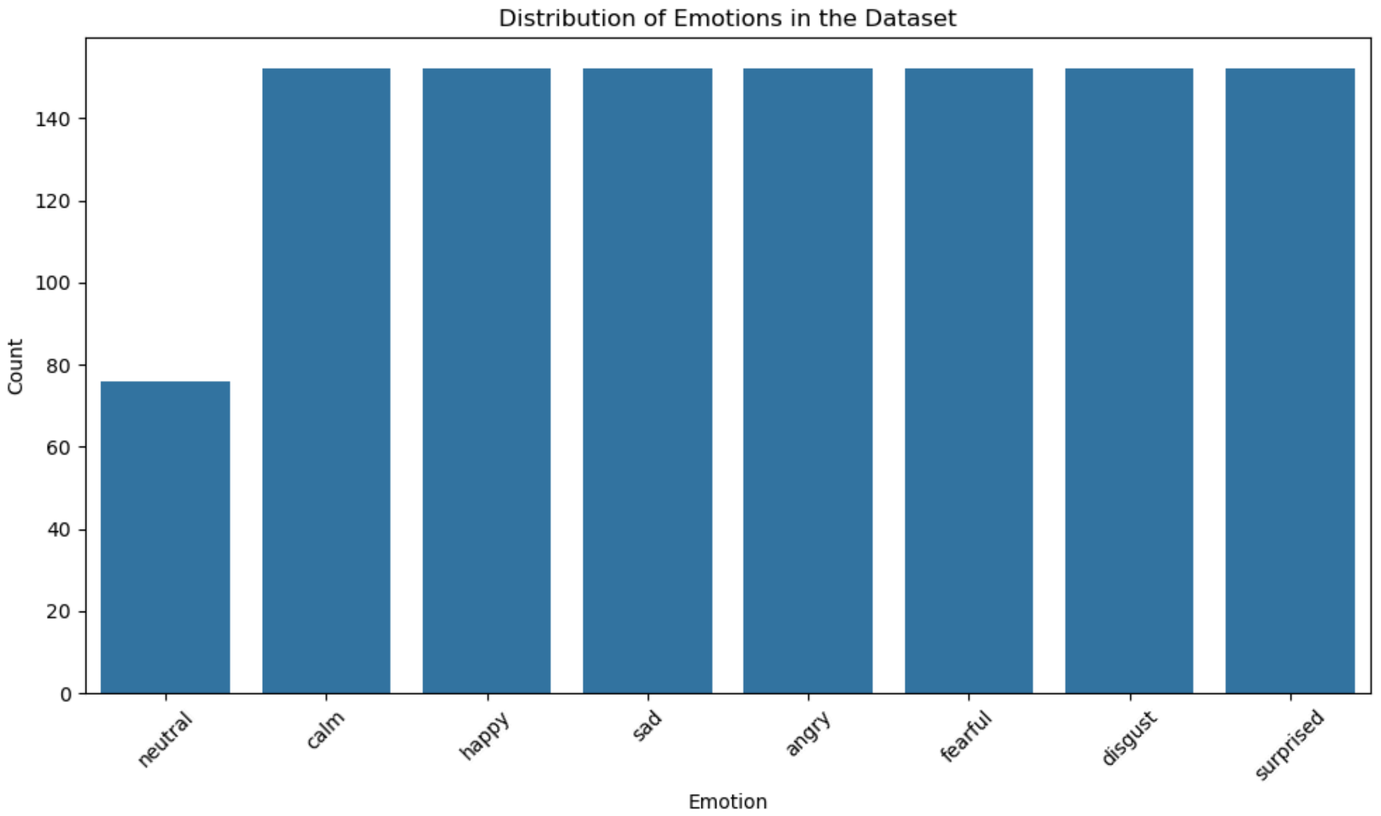
Manas Ranajn

Objective and Dataset

- **Objective:** Recognize emotions from speech audio files by analyzing vocal characteristics and patterns.
- **Dataset:** Dataset (Actor folders, ~19 speakers, 8 emotions)

Attributes	Details
Samples	1140 - audio files
Emotions	8 classes (neutral, happy, sad, angry, fearful, disgust, surprised, calm)
Key Features	• 20 MFCCs + Δ/ΔΔ • Spectral descriptors (chroma, ZCR, contrast) • Time-domain features
Class Balance	Upsampled minority classes ("neutral")

Emotion Distribution Bar Chart



Feature Extraction FlowChart

[Audio File] → [Trim Silence] → [MFCCs + Spectral Features] → [Normalize] → [Model Input]

Features & Processing

Final Features

- MFCCs
- MFCC-Delta
- MFCC-Delta-Delta
- Spectral Contrast
- Spectral Centroid
- Spectral Bandwidth
- Spectral Rolloff
- ZCR
- Chroma

Features Preprocessing

- Librosa-based pipeline (trim silence, pre-emphasis, 22.05kHz sample rate)
- **Normalization**: StandardScaler per feature dimension

```
# Code snippet highlight
mfcc = librosa.feature.mfcc(y=y, sr=sr, n_mfcc=20)
features = np.concatenate([mfcc, chroma, spectral_contrast], axis=0)
```

Data Balancing

```
Checking training data balance...
Class 0: 110.0 samples
Class 1: 109.0 samples
Class 2: 109.0 samples
Class 3: 110.0 samples
Class 4: 110.0 samples
Class 5: 55.0 samples
Class 6: 110.0 samples
Class 7: 110.0 samples
WARNING: Dataset is imbalanced. Consider resampling or stronger class weights.
Upsampling all classes to 110 samples
Balanced dataset shape: (880, 83, 64, 1), (880, 8)
After balancing:
Class 0: 110.0 samples
Class 1: 110.0 samples
Class 2: 110.0 samples
Class 3: 110.0 samples
Class 4: 110.0 samples
Class 5: 110.0 samples
Class 6: 110.0 samples
Class 7: 110.0 samples
Model: "functional"
```

CNN Model Architecture

1.Input Layer

- Shape: 82 (MFCCs + spectral features) × 64 (time steps) × 1 (channel)
- Purpose: Accepts processed audio features as 2D spectrogram-like input.

2.Convolution(Conv2D) Layers:-

- First Conv Layers (32 filters):
  - Scans the "sound picture" with 3×3 magnifying glasses to find small patterns (e.g., sudden pitch changes).
  - Each filter learns to detect different features (like angry shouts vs. calm hums).
- Deeper Conv Layers (64/128 filters):
  - Combine small patterns into bigger ones (e.g., a rising-then-falling tone = surprise).

3. **Batch Normalization (BN)**

- Keeps volume consistent across all audio clips (like auto-adjusting microphone levels).
- Prevents one loud emotion (e.g., angry) from drowning out others.

4.**ReLU Activation**

- Turns negative values to zero ("mutes" irrelevant noise).
- Keeps only the noticeable features (e.g., keeps loud spikes, removes background hiss).

5.**Max Pooling**

- Shrinks the "sound picture" by keeping only the loudest/most noticeable parts every 2×2 area.
- Makes the model focus on important features (like a highlight reel).

6.Dropout

- Randomly ignores 20-40% of detected features during training.
- Prevents over-reliance on any single clue (e.g., not just using "loudness" to detect anger).

7.**Residual(Skip) Connection**

- Provides a shortcut for the original sound features to bypass some layers.
- Ensures the model doesn’t "forget" basic info (like whether the speaker is male/female).

8.**Global Average Pooling**

- Averages all features into one summary number per pattern.
- Simpler and less prone to errors than flattening.

9.**Dense(Fully Connectred)Layers**

- First Dense (256 neurons):
  - Combines all detected patterns ("high pitch + fast tempo + shaky voice = fear").
- Second Dense (128 neurons):
  - Refines the decision with fewer but more precise connections.

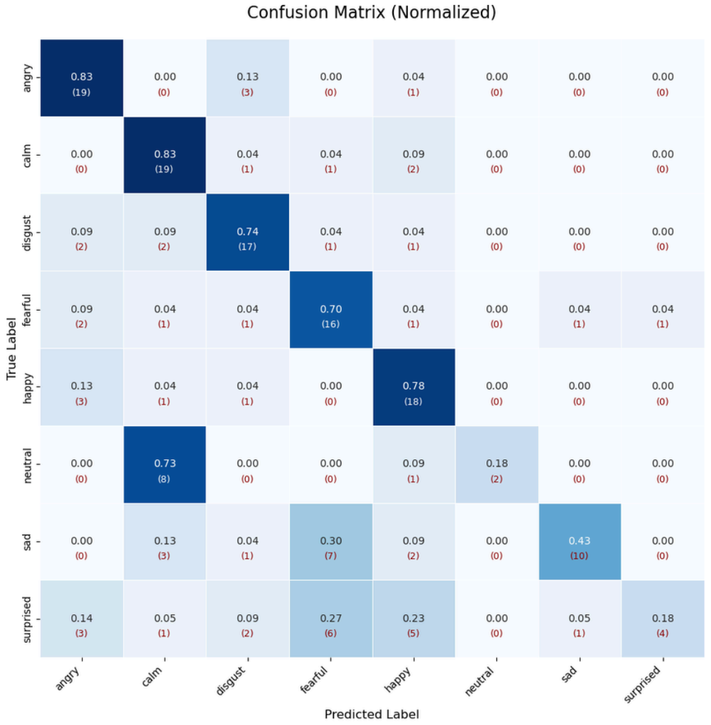
10.**Softmax Output**

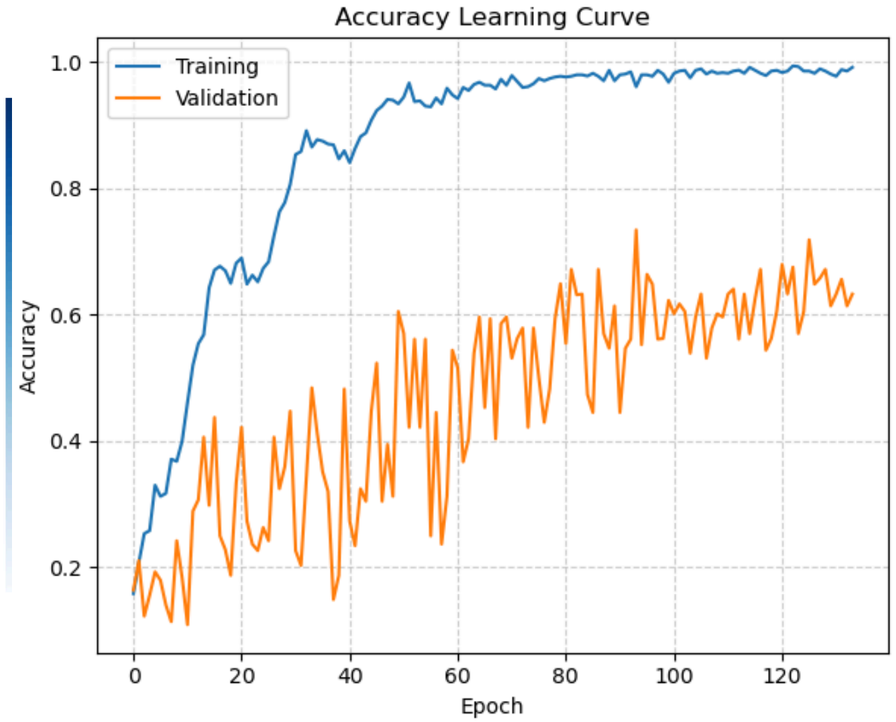
- Converts scores into probabilities (e.g., 85% angry, 10% happy, 5% neutral).
- Ensures all probabilities add up to 100%.

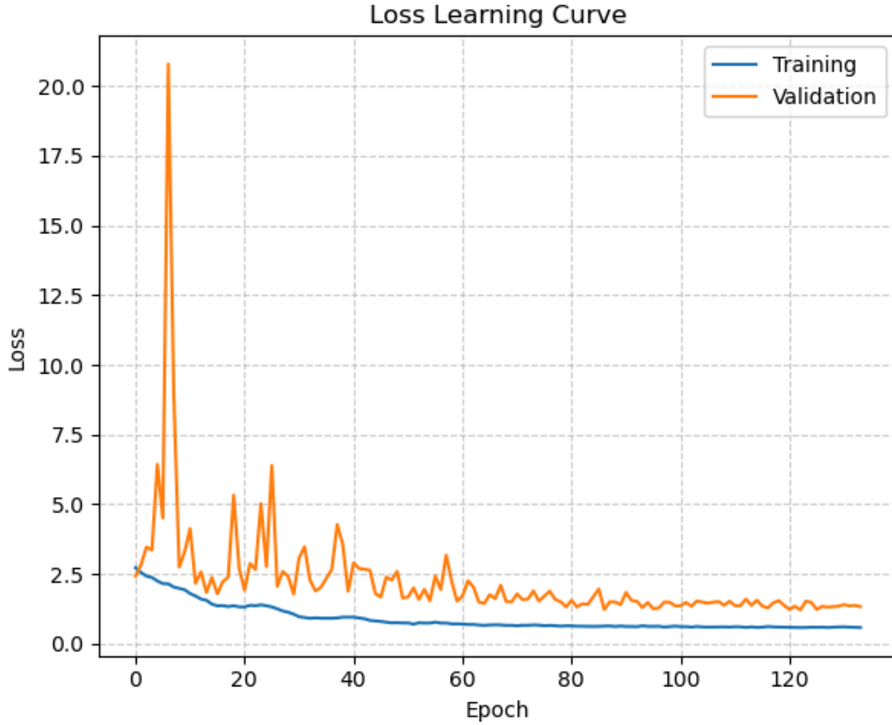
Why This Works for Emotions?

1. **Early Layers**: Detect sound ingredients (pitch, tone).
2. **Middle Layers**: Recognize combinations (shouting = anger, trembling = fear).
3. **Final Layers**: Weigh all clues to pick the dominant emotion.

Performance







The validation accuracy of our model is 71.88%

Thank you very much!