Speech Emotion Recognition System (SERS)

Integrating Signal Processing and Deep Learning for Emotional Insight

Presented by: Umesh Sheela (23123048)

Project Overview: The End-to-End SERS Pipeline

The Speech Emotion Recognition System (SERS) is a sophisticated, integrated pipeline engineered to accurately classify human emotional states from speech signals. This interdisciplinary project synthesizes foundational principles of Signals and Systems with state-of-the-art Machine Learning techniques, creating a robust framework for emotional insight extraction.

01

Audio Preprocessing & Feature Extraction

Raw audio undergoes preprocessing and transformation into numerical feature vectors using Short-Time Fourier Transform (STFT) and Mel-Frequency Cepstral Coefficients (MFCCs), capturing the acoustic essence of emotional expression.

02

Deep LearningClassification

A Convolutional Neural Network (CNN) architecture is employed for robust pattern recognition, trained on diverse datasets to establish sophisticated mappings from acoustic features to emotional categories.

03

Real-Time Emotion Detection

The system delivers real-time detection across eight primary emotions—Angry, Calm, Disgust, Fearful, Happy, Neutral, Sad, and Surprised—achieving approximately 85% accuracy on benchmark datasets.

Dataset Foundation: RAVDESS

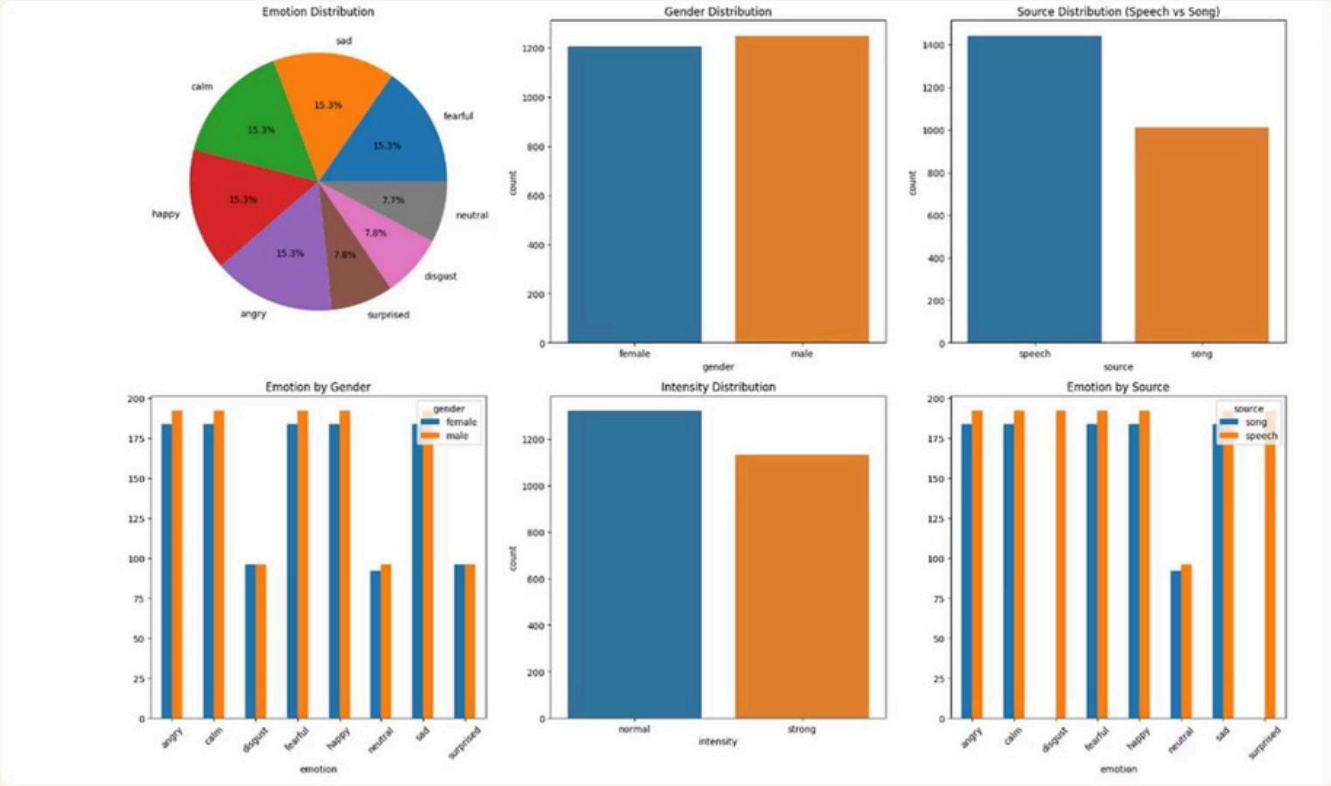
We leverage the Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS), a meticulously curated, balanced dataset that serves as the cornerstone for robust model training and validation.

Key Dataset Characteristics

- Total Files:7,356 high-fidelity professional recordings
- Actors:24 professional performers (12 male, 12 female) ensuring comprehensive speaker diversity
- Content:Two standard English statements performed across various emotions and intensity levels
- Modalities:Speech and song available in audio-only, video-only, and audio-visual formats
- Emotion Categories:Neutral, Calm, Happy, Sad, Angry, Fearful, Disgust, Surprised

The structured naming convention (e.g., `02-01-06-01-02-01-12.mp4`for a fearful speech sample) enables precise, automated label and metadata extraction—essential for supervised learning workflows.

EDA: The Feature Engineering Process



This architectural diagram illustrates the core components of the Speech Emotion Recognition System, tracing the journey from raw speech signal acquisition through sophisticated feature extraction pipelines to final emotion classification outputs.

The exploratory data analysis reveals critical patterns in acoustic features across emotional categories, guiding our feature selection and model architecture decisions for optimal classification performance.

Signals and Systems: The Foundation

The interpretation of speech as a signal is fundamentally rooted in classical Signals and Systems theory. Understanding both frequency and time-domain characteristics of the human voice provides the mathematical foundation for extracting meaningful emotional featuresfrom acoustic data.

Acoustic Signal

Speech is modeled as a non-stationary, quasi-periodic signal where rapid temporal changes in frequency and amplitude convey both linguistic content and paralinguistic information such as emotional state.

Transform Domain

The Fourier Transformconverts time-domain signals into frequency domain representations, revealing the spectral energy distribution across different pitches (formants)—key indicators of emotional expression.

Windowing & Filtering

Short-Time Fourier Transform (STFT) applies windowing functions to analyze small, quasi-stationary signal segments, providing localized spectral content views essential for emotion detection.

Key Feature Extraction Techniques

To enable effective machine learning, we transform raw audio signals into compact, discriminative feature vectors. Our approach focuses on two powerful, industry-standard techniques that capture complementary aspects of the acoustic signal.

Mel-Frequency Cepstral Coefficients (MFCCs)

MFCCs form a coefficient vector representing the vocal tract's spectral envelope. They are widely adopted because they mimic the nonlinear human perception of sound (Mel scale), focusing on formants critical for both emotion and speech recognition tasks.

- Robust against environmental noise and speaker variability
- Captures perceptually relevant spectral information
- Represents short-term power spectrum characteristics

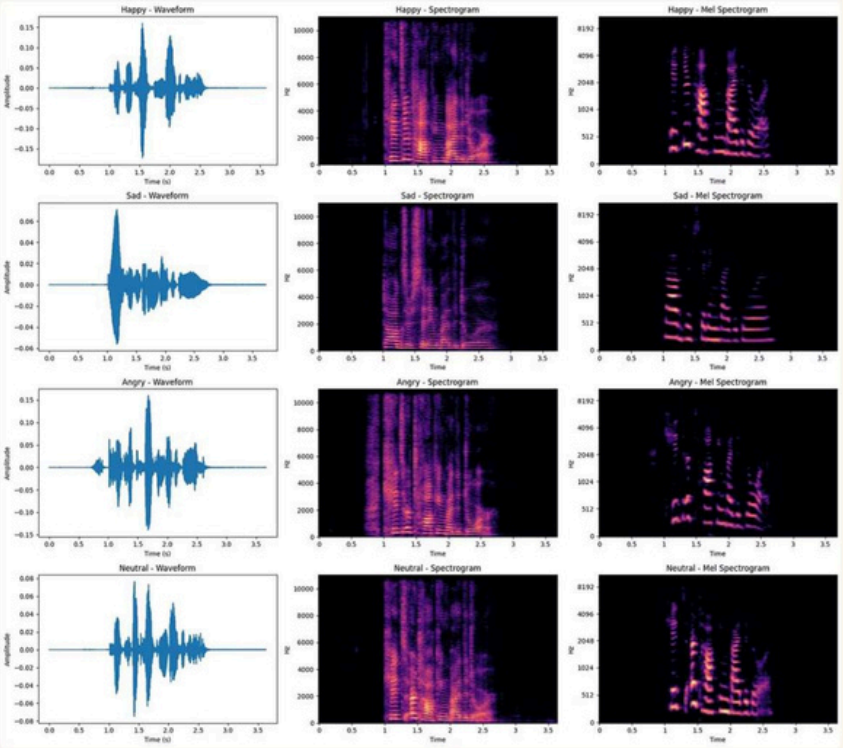
Short-Time Fourier Transform (STFT)

STFT segments the signal into smaller temporal frames and computes the Fourier Transform for each segment. This process generates spectrograms—time-frequency representations that visually capture dynamic spectral content evolution.

- Essential for analyzing non-stationary speech signals
- Provides time-frequency resolution trade-offs
- Foundation for generating mel-spectrograms

Visualizing Acoustic Features

Spectrograms and Mel-Spectrograms transform one-dimensional audio signals into two-dimensional image representations, enabling us to leverage powerful convolutional neural networks (CNNs) originally designed for image recognition tasks.

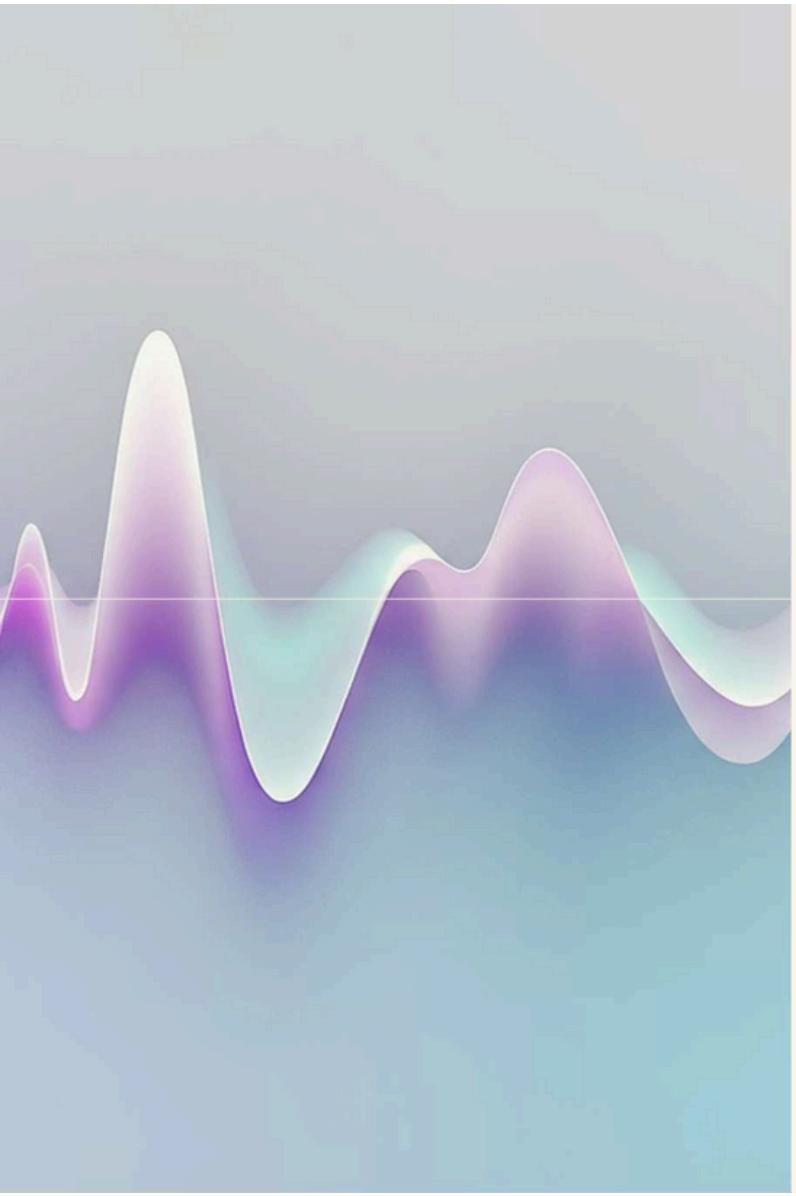


Spectrogram

A visual representation of the frequency spectrum as it varies with time. Frequency is plotted on the y-axis, time on the x-axis, and amplitude is encoded through color intensity, creating a heat map of acoustic energy distribution.

Mel-Spectrogram

Similar to a standard spectrogram, but the frequency axis is scaled according to the Mel scale—a perceptual scale that aligns with human auditory perception. This transformation significantly improves emotion detection accuracy.



Zero Crossing Rate (ZCR)

Function Implementation

librosa. feature.zero_c rossing_rate(y)

Definition

The number of times the discrete audio signal crosses the zero amplitudeaxis per analysis frame, indicating instantaneous frequency characteristics.

Mathematical Formula

$$ZCR = \frac{1}{N-1} \sum_{n=1}^{N-1} 1_{\{x[n]x[n-1]<0\}}$$

Physical Meaning

- Quantifies signal noisiness and frequency content
- High ZCR→High frequency or noisy characteristics (e.g., angry/shouted speech)
- Low ZCR→Smoother, lower frequency signal (e.g., sad/soft speech)

Signal Processing Concept

Relates to instantaneous frequency changesin the time domain. A high ZCR approximates high average frequency content, providing a computationally efficient measure of spectral characteristics without explicit frequency domain transformation.

Spectral Centroid

The spectral centroid represents the "center of mass" of the frequency spectrum, providing insight into the perceived brightnessor sharpness of a sound signal.

Function Implementation

```
librosa.feature.spectral_centroid(y=y, sr=sr)
```

Mathematical Formula

$$C = \frac{\sum f_k \mid X(f_k) \mid}{\sum \mid X(f_k) \mid}$$

Interpretation and Meaning

- Represents the energy-weighted mean frequency from the Fourier spectrum
- Higher centroid valuesindicate bright or sharp sounds with emphasis on higher frequencies
- Lower centroid valuessuggest dull or muffled sounds dominated by lower frequencies
- Angry/happy emotional tonestypically exhibit higher centroid values
- Sad emotional expressionsgenerally show lower centroid values

Signal Processing Context

The spectral centroid is calculated from the Fourier spectrum, measuring the frequency point around which spectral energy is balanced. This feature provides a single-value summary of the spectral shape, making it highly useful for distinguishing between different emotional states in speech.

Spectral Bandwidth

Spectral bandwidth quantifies the width of the frequency distribution around the spectral centroid, describing the range and concentration of spectral energy in the signal.

f(x)

Function Implementation

```
librosa.feature.spectral_bandwidth(y=y, sr=sr)
```

Mathematical Formula

$$B = \sqrt{\frac{\sum (f_k - C)^2 \mid X(f_k) \mid}{\sum \mid X(f_k) \mid}}$$

Where Cis the spectral centroid

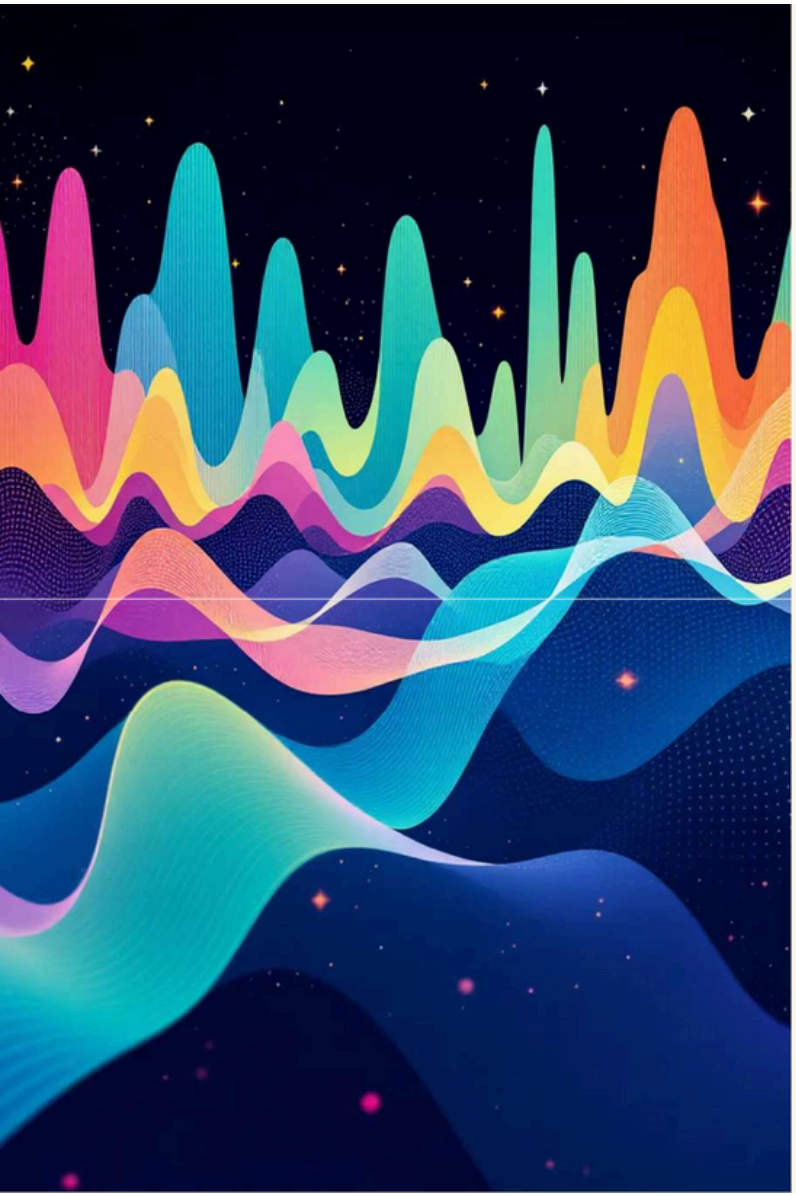
Physical Interpretation

- Describes the spread of frequency energy around the spectral centroid
- Wide bandwidth→Complex or high-pitched sounds with energy distributed across many frequencies
- Narrow bandwidth→Pure tones or flat spectral characteristics with concentrated energy
- Acts as the standard deviation of the spectral distribution

Systems Theory Connection

Spectral bandwidth relates directly to the concept of system bandwidthin Signals and Systems theory. It provides a measure of spectral complexity and can differentiate between tonal and noisy speech components—a crucial distinction for emotional classification tasks.

Key Insight:Together with spectral centroid, bandwidth forms a powerful two-dimensional descriptor of spectral shape, enabling robust emotion discrimination across diverse speakers and recording conditions.



Spectral Rolloff: Quantifying Sound Brightness

Spectral Rolloff is a fundamental audio feature extraction technique that pinpoints the frequency below which 85% of the total spectral energyis concentrated. This metric serves as a crucial indicator of a sound's spectral shape, specifically measuring how rapidly high-frequency components decay.

High Rolloff Frequency

Slower decay of high-frequency energy resulting in a **brighter, more energetic sound** with prominent treble characteristics.

Low Rolloff Frequency

Energy concentrated in lower frequencies, producing a **duller or darker sound**with muted high-end response.

Function: `librosa.feature.spectral_rolloff(y=y, sr=sr)`

This metric effectively represents the cumulative energy distributionwithin the frequency domain, offering insights into the timbral characteristics of audio signals.

MFCCs: Mel Frequency Cepstral Coefficients

The Processing Pipeline

01

Fourier Transform

Convert time-domain signal to frequency domain

02

Mel Filter Banks

Apply perceptually-motivated frequency scaling based on human auditory response

03

Logarithmic Compression

Map energy to logarithmic scale mimicking human loudness perception

04

Discrete Cosine Transform

Decorrelate coefficients for efficient representation

Key Characteristics

MFCCs represent the spectral envelope—the overall shape of the frequency spectrum—making them exceptionally powerful for capturing timbre and vocal texture.

`Function:librosa.feature.mfcc(y=y, sr=sr, n_mfcc=20)`

Critical for differentiating emotions in speech and music analysis tasks.

Connection to Signal Processing Theory

- Filter bank analysis:Mel filters act as band-pass filters, a core LTI system concept
- Perceptual modeling:Logarithmic energy mapping aligns with psychoacoustic principles
- Decorrelation:DCT transforms responses into efficient, uncorrelated representations

Delta MFCCs: Capturing Temporal Dynamics

First-Order Derivatives of MFCCs

Delta MFCCs represent the rate of change or velocity of MFCC features, revealing how sound evolves over time. This temporal derivative captures the dynamic characteristics and transient behavior of audio signals.

Function: `librosa.feature.delta(mfcc)`

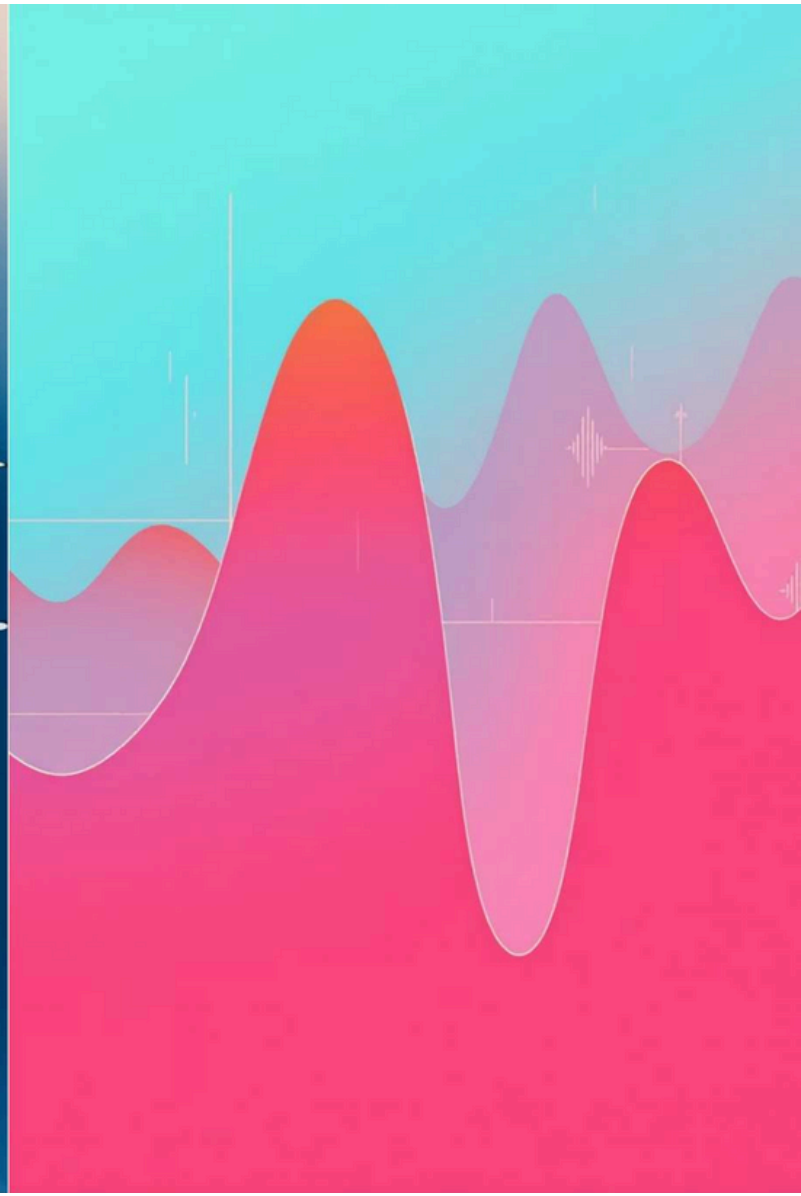
Mathematical Foundation

The discrete derivative is calculated as a first-order difference:

$$\Delta c[n] = c[n] - c[n-1]$$

Signal Processing Interpretation

This captures the dynamics and transient behavior of the signal, analogous to observing how a system's response changes over time—a fundamental concept in temporal signal analysis.



Traditional Audio Features: A Comprehensive Overview

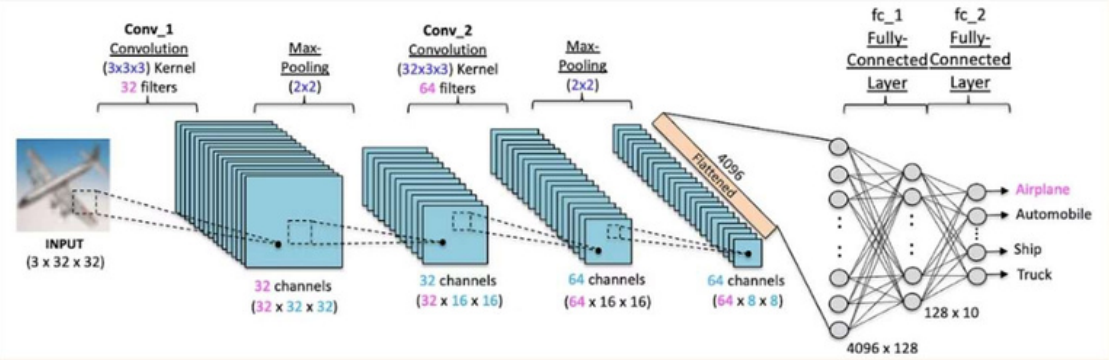
While deep learning architectures can learn features directly from raw audio, many high-level tasks still benefit from pre-calculated features derived from established DSP techniques. These features bridge the gap between signal processing theory and machine learning practice.

#	Feature	Domain	Description
1	ZCR	Time	Zero-crossings per frame, associated with signal frequency, roughness, or noisiness characteristics
2	RMS Energy	Time	Root Mean Square energy—a measure of signal power or loudness over time
3	Spectral Centroid	Frequency	The "center of mass" of the spectrum, indicating the average frequency present and overall brightness
6-8	MFCCs & Deltas	Cepstral	Mel-Frequency Cepstral Coefficients and their derivatives, widely used for speech recognition due to robustness
9	Chroma	Harmonic	Pitch class profile representation, excellent for music analysis and harmonic structure identification
14	Tempo	Time	Measures perceived speed (BPM) or beat interval periodicity in musical signals
16	Poly Features	Frequency	Spectrum shape and slope modeling features, useful for texture description and timbre analysis

Convolutional Neural Networks for Audio

CNNs: Spatial Feature Extraction from Spectrograms

Convolutional Neural Networks excel at extracting hierarchical spatial features from data, making them exceptionally well-suited for image-like audio representations such as spectrograms and mel-spectrograms. By treating time-frequency representations as 2D images, CNNs can learn complex patterns across both temporal and spectral dimensions.



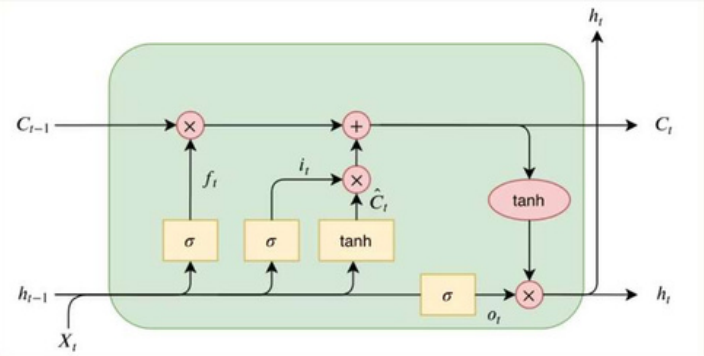
Convolutional Layers

Identify local patterns in frequency and time domains, detecting specific tones, harmonic structures, and transient events

Pooling Operations

Reduce dimensionality and provide translation invariance, making models robust to slight shifts in time or frequency

LSTM Networks: Modeling Temporal Dependencies



Contextual Temporal Analysis

For sequential data like audio signals, Long Short-Term Memory (LSTM) networks are crucial for modeling long-range dependencies—capturing how events early in the signal influence interpretation much later in the sequence.

1

Gate Mechanisms

Internal gates (input, forget, output) control information flow, allowing the network to selectively "remember" or "forget" previous context based on relevance

2

Sequence Modeling

Ideal for recognizing patterns that unfold over time, such as spoken phrases, musical rhythm, or prosodic features in speech

3

State Management

Maintains both short-term hidden state and long-term cell state, enabling robust memory across extended sequences without vanishing gradient issues

CNN Architecture for Spectrogram Classification

A typical CNN structure for audio classification employs stacked convolutional and pooling layers to progressively refine feature representations from spectrogram input, culminating in dense layers for final classification.

CNN Model Architecture:

Model: "sequential"

Layer (type)	Output Shape	Param #
dense (Dense)	(None, 1024)	263,168
batch_normalization (BatchNormalization)	(None, 1024)	4,096
dropout (Dropout)	(None, 1024)	0
dense_1 (Dense)	(None, 512)	524,800
batch_normalization_1 (BatchNormalization)	(None, 512)	2,048
dropout_1 (Dropout)	(None, 512)	0
dense_2 (Dense)	(None, 256)	131,328
batch_normalization_2 (BatchNormalization)	(None, 256)	1,024
dropout_2 (Dropout)	(None, 256)	0
dense_3 (Dense)	(None, 128)	32,896
batch_normalization_3 (BatchNormalization)	(None, 128)	512
dropout_3 (Dropout)	(None, 128)	0
dense_4 (Dense)	(None, 64)	8,256
batch_normalization_4 (BatchNormalization)	(None, 64)	256
dropout_4 (Dropout)	(None, 64)	0
dense_5 (Dense)	(None, 8)	520

Total params: 968,904 (3.70 MB)

Trainable params: 964,936 (3.68 MB)

Non-trainable params: 3,968 (15.50 KB)

Input Layer

Spectrogram representation (e.g., Mel-scale, log-scale, or STFT magnitude)

Stacked Conv/Pool Layers

Multiple convolutional and pooling blocks learn increasingly complex, high-level spatio-temporal features

Dense Layers

Fully connected layers combine extracted features and prepare for classification decision

Output Layer

Softmax activation for multi-class or Sigmoid for binary classification tasks

Bidirectional LSTM Architecture

For optimal performance in audio tasks requiring full sequence context, Bidirectional LSTMs (Bi-LSTMs) process sequences in both forward and backward directions, capturing temporal dependencies from past andfuture context.

Forward Pass

Analyzes input sequence from start to end, building context progressively from past information and temporal precedents

Backward Pass

Analyzes input sequence from end to start, integrating future context and information that follows each time step

Output Integration

Outputs from both directions are concatenated at each time step, forming a richer, bidirectionally context-aware representation

This bidirectional approach is particularly powerful for tasks where future context matters—such as speech recognition, where later phonemes can disambiguate earlier ones, or music analysis, where harmonic resolution depends on what follows.

LSTM Model Architecture:

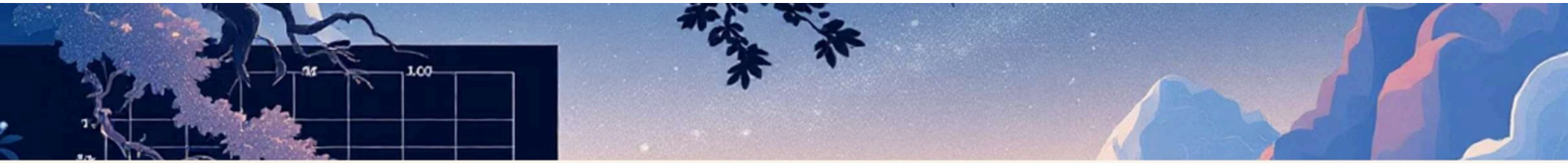
Model: "functional_1"

Layer (type)	Output Shape	Param #
input_layer_1 (InputLayer)	(None, 256, 1)	0
lstm (LSTM)	(None, 256, 128)	66,560
lstm_1 (LSTM)	(None, 64)	49,408
dense_6 (Dense)	(None, 64)	4,160
dropout_5 (Dropout)	(None, 64)	0
dense_7 (Dense)	(None, 32)	2,080
dropout_6 (Dropout)	(None, 32)	0
dense_8 (Dense)	(None, 8)	264

Total params: 122,472 (478.41 KB)

Trainable params: 122,472 (478.41 KB)

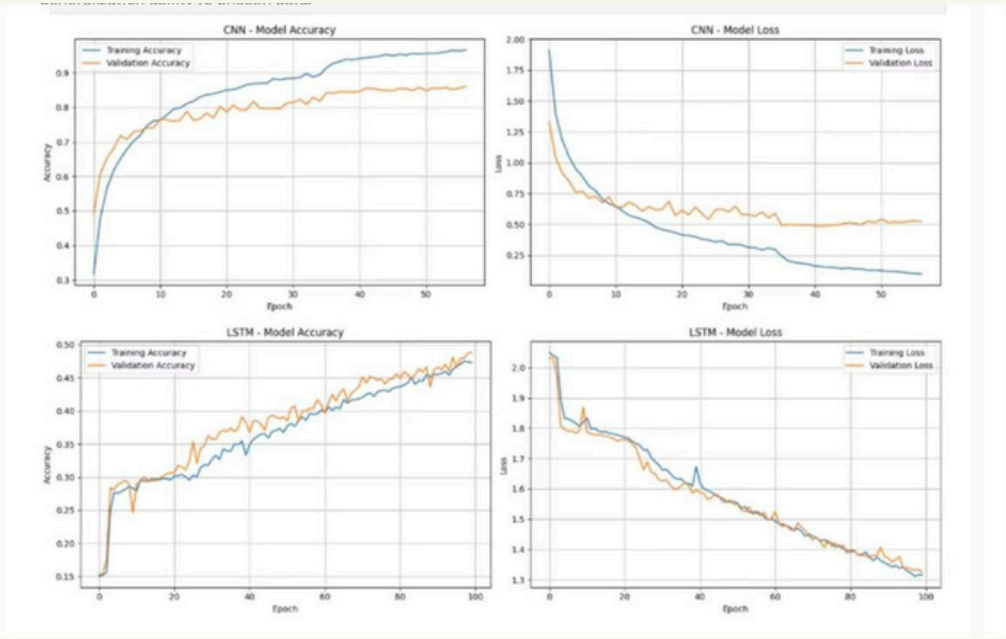
Non-trainable params: 0 (0.00 B)



Model Training Dynamics: Monitoring Convergence

Interpreting Training Curves

Monitoring training and validation loss curves is critical for diagnosing model performance, detecting overfitting, and assessing generalization ability to unseen data.



1

Convergence

Both training and validation curves should decrease steadily, indicating successful learning and parameter optimization

2

Generalization

Validation loss tracking closely with training loss demonstrates the model adapts well to unseen data without memorization

3

Overfitting Risk

A widening gap where training loss drops but validation loss plateaus or rises signals overfitting—employ regularization techniques

Comparative Performance Analysis

Selecting the optimal architecture depends heavily on the specific audio task, computational constraints, and temporal dependency requirements. This analysis highlights typical trade-offs among key deep learning architectures used in audio processing.

```
CNN Results:
Accuracy: 0.8552
F1 Score: 0.8550

CNN Classification Report:
              precision    recall  f1-score   support

       angry       0.92      0.91      0.91       301
        calm       0.89      0.90      0.90       301
     disgust       0.80      0.83      0.81       153
     fearful       0.84      0.80      0.82       301
       happy       0.88      0.85      0.87       301
     neutral       0.81      0.91      0.86       150
         sad       0.80      0.78      0.79       301
   surprised       0.85      0.89      0.87       154

    accuracy                           0.86      1962
   macro avg       0.85      0.86      0.85      1962
weighted avg       0.86      0.86      0.86      1962


LSTM Results:
Accuracy: 0.4883
F1 Score: 0.4704

LSTM Classification Report:
              precision    recall  f1-score   support

       angry       0.60      0.65      0.63       301
        calm       0.49      0.75      0.59       301
     disgust       0.37      0.39      0.38       153
     fearful       0.43      0.43      0.43       301
       happy       0.46      0.42      0.44       301
     neutral       0.38      0.09      0.14       150
         sad       0.52      0.36      0.42       301
   surprised       0.51      0.66      0.57       154

    accuracy                           0.49      1962
   macro avg       0.47      0.47      0.45      1962
weighted avg       0.48      0.49      0.47      1962
```

Future Directions and Applications

The established SERS pipeline has broad implications across various fields demanding automated emotional intelligence.

1. Customer Service Analytics:Real-time monitoring of customer frustration or satisfaction to prioritize urgent calls or improve agent training protocols.
2. Mental Health & Well-being:Screening tools to detect early indicators of emotional distress, anxiety, or depression through voice analysis.
3. Driver and Operator Safety:Detecting fatigue, stress, or anger in drivers or heavy machinery operators to prevent accidents

Thank You!