Multimodal Video Sentiment Analysis System

End-to-end system analyzing speech, video, and audio to detect emotions and sentiment with deep learning fusion.

Core Modalities and Encoders

Text Encoder

BERT-based, frozen model with projection head

- Pre-trained semantic understanding
- Processes speech transcripts

Video Encoder

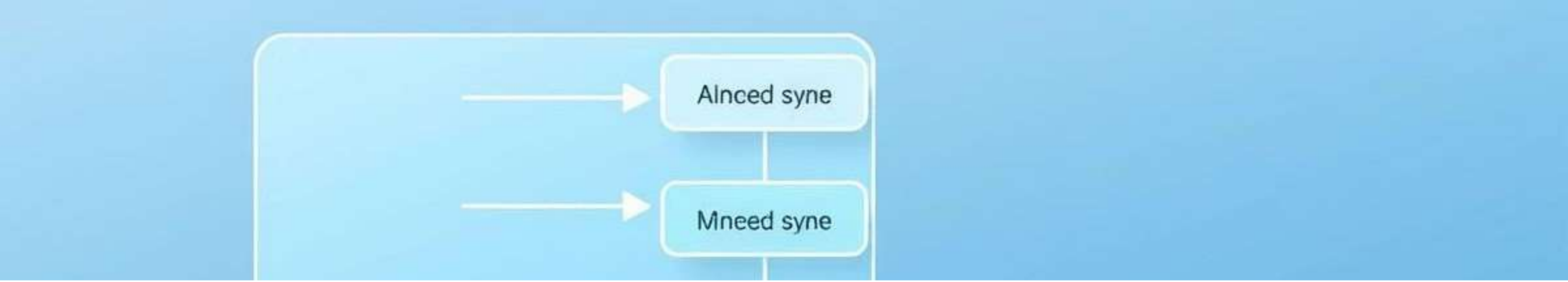
Modified 3D ResNet (R3D-18)

- Captures spatiotemporal visual features
- Processes normalized RGB frames

Audio Encoder

Custom 1D CNN extracting paralinguistic cues

- Handles variable-length audio
- Robust acoustic feature extraction



Multimodal Fusion Architecture

Late Fusion Layer

Combines features from all three encoders

Equal Feature Dimensions

128-dim vectors ensure balanced modality influence

Dual Classification Heads

- Emotion Recognition (7 classes)
- Sentiment Analysis (3 classes)

Emotion & Sentiment Classes

Emotion Recognition

- Anger
- Disgust
- Fear
- Joy
- Neutral
- Sadness
- Surprise

Sentiment Analysis

- Positive
- Negative
- Neutral

Technical Highlights

1

Temporal Alignment

Synchronizes video segments with speech transcripts

2

Efficient Inference

Frame sampling and ONNX runtime compatibility

3

Balanced Features

Ensures no modality dominates

4

Batch Processing

Parallel GPU acceleration across modalities

Training and Optimization

Loss Function

Dual cross-entropy for emotion and sentiment

Regularization Techniques

- Dropout 20-30%
- Batch Normalization
- Frozen encoder backbones

Optimizer

AdamW with weight decay

Interactive Dashboard Features

Video Upload & Preview

Supports MP4 processing and playback preview

Visualization

Emotion confidence timeline and sentiment radar

Advanced Metrics

Segment-level confidence and modality contributions

Pipeline Architecture Overview

Input Processing

Video segmentation via FFmpeg

Speech Recognition

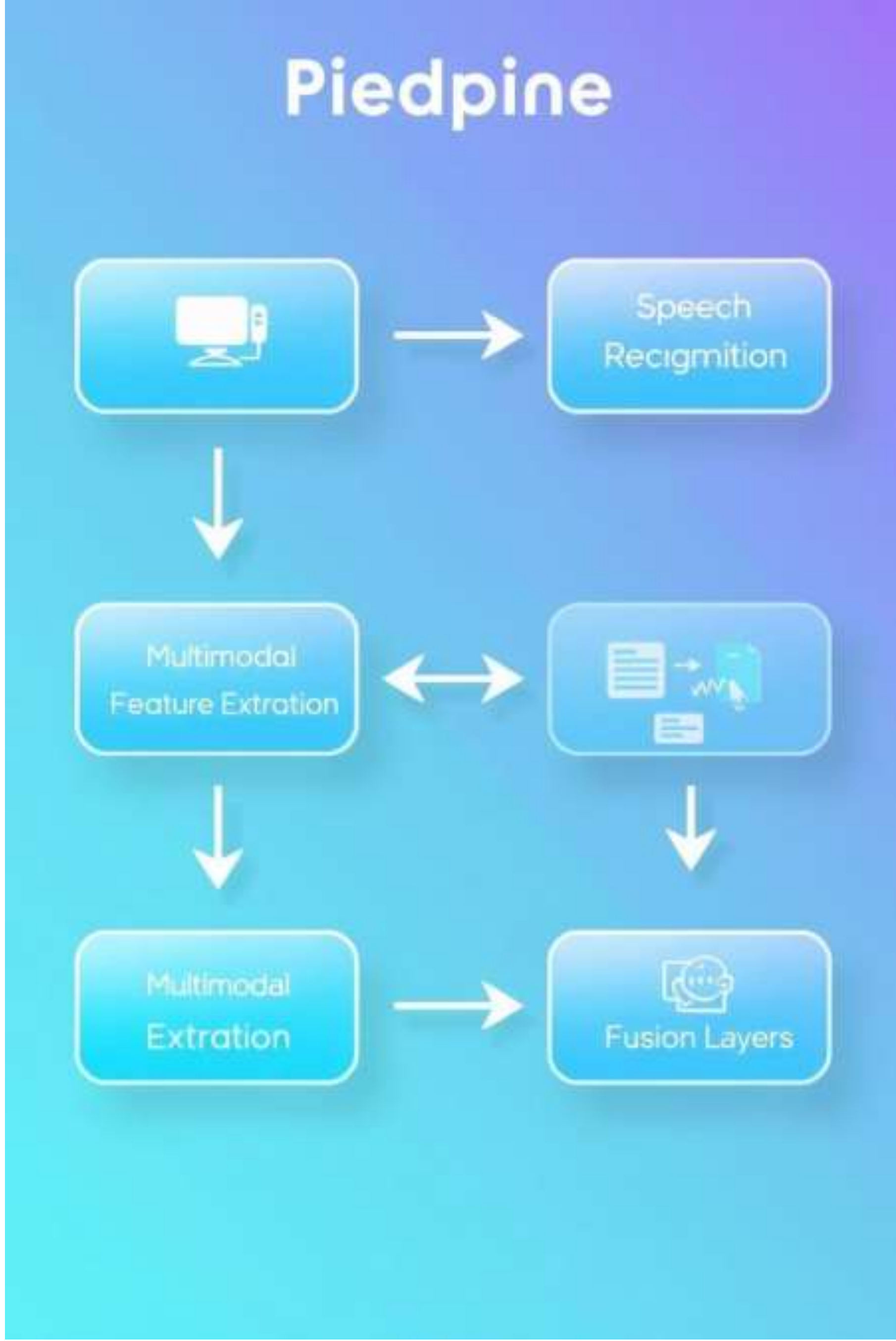
Whisper ASR for text extraction

Feature Extraction

Parallel GPU-accelerated modality processing

Fusion and Postprocessing

Feature concatenation, softmax, top-k aggregation



Summary and Impact

- **Advanced multimodal fusion for sentiment analysis**
- **Balances features to improve accuracy**
- **Efficient inference enables real-time analysis**
- **Interactive dashboard enhances user insights**

Empowers emotion detection in various applications: media, marketing, human-computer interaction.

Thank You

We appreciate your time and interest in our multimodal sentiment analysis system.