Project Report: A Scalable Architecture for Multimodal Analysis of Meeting Dynamics

October 6, 2025

Abstract

To gain a competitive edge in understanding and optimizing professional communication, I am proposing the development of a sophisticated, AI-driven platform for analyzing human behavior in meetings. The primary objective is to process user video and audio streams to extract granular data on emotional expression and nonverbal cues ("mimics"). This data will then be synthesized into high-level, quantifiable metrics that can be benchmarked against established values for a successful meeting.

The core tenets of this project are **accuracy, scalability, and expandability**. To achieve this, I have designed a modular, multi-pipeline architecture that leverages state-of-the-art deep learning techniques. The system will operate by deconstructing the input data into three parallel streams—**Visual (facial analysis), Vocal (acoustic analysis), and Linguistic (textual analysis)**—before fusing them to create a holistic and accurate interpretation of the user's state.

This document outlines the complete system architecture, details the end-to-end data processing pipeline, specifies the structure of the final exported data, and provides a curated list of the foundational research and tools that inform this design.

Contents

1 Executive Summary

To gain a competitive edge in understanding and optimizing professional communication, I am proposing the development of a sophisticated, AI-driven platform for analyzing human behavior in meetings. The primary objective is to process user video and audio streams to extract granular data on emotional expression and nonverbal cues ("mimics"). This data will then be synthesized into high-level, quantifiable metrics that can be benchmarked against established values for a successful meeting.

The core tenets of this project are **accuracy, scalability, and expandability**. To achieve this, I have designed a modular, multi-pipeline architecture that leverages state-of-the-art deep learning techniques. The system will operate by deconstructing the input data into three parallel streams—**Visual (facial analysis), Vocal (acoustic analysis), and Linguistic (textual analysis)**—before fusing them to create a holistic and accurate interpretation of the user's state.

This document outlines the complete system architecture, details the end-to-end data processing pipeline, specifies the structure of the final exported data, and provides a curated list of the foundational research and tools that inform this design.

2 Proposed System Architecture

The architecture is designed as a series of interconnected, scalable microservices, allowing each component to be developed, updated, and scaled independently. This modularity ensures that the system can easily incorporate new analytical models or additional data modalities in the future.

The system is composed of four primary stages: **Data Ingestion, Unimodal Processing, Multimodal Fusion, and Metrics Generation.**

2.1 Stage 1: Data Ingestion

This initial module is responsible for capturing and pre-processing the raw audio and video streams from the user.

- **Input:** Live or pre-recorded video and audio feed.
- **Process:** The module will handle stream synchronization, segmenting the data into manageable chunks (e.g., 1-5 second intervals) for parallel processing, and ensuring consistent data formats for the downstream analytical engines.

2.2 Stage 2: Unimodal Processing Pipelines

The ingested data is simultaneously fed into three specialized pipelines, each focused on a single data modality.

2.2.1 Visual Pipeline (Facial Expression & Mimic Recognition - FER)

- **Objective:** To analyze facial cues with the highest possible accuracy. This pipeline will use computer vision models to detect faces, track landmarks, and classify expressions [1, 2].
- **Technology:** I will implement a deep Convolutional Neural Network (CNN) trained on extensive facial expression datasets [3]. The process involves:
  1. **Face Detection:** Isolate the face in each frame using robust algorithms like MTCNN [2].
  2. **Landmark Detection:** Identify 68 key facial landmarks (corners of eyes, mouth, etc.) using a library like **Dlib** [4, 5].
  3. **Facial Action Unit (AU) Recognition:** Instead of basic emotion labels, the core of this pipeline will be the detection of AUs—the fundamental muscle movements of the face [6]. This provides an objective, granular analysis of "mimics" and is a key differentiator for accuracy. The **OpenFace 2.0** toolkit provides a strong foundation for this.

2.2.2 Vocal Pipeline (Speech Emotion Recognition - SER)

- **Objective:** To analyze the paralinguistic, non-verbal aspects of speech (tone, pitch, energy).
- **Technology:** This pipeline will employ a hybrid deep learning model, likely a combination of a CNN and a Long Short-Term Memory (LSTM) network, which has proven highly effective for SER.
  1. **Feature Extraction:** Convert raw audio into a spectrogram and extract key acoustic features, with a primary focus on **Mel-Frequency Cepstral Coefficients (MFCCs)**, which are superior for capturing emotional nuances. The **librosa** Python library is ideal for this task [7].
  2. **Emotion Classification:** The CNN-LSTM model will classify the feature set into emotional categories (e.g., happy, sad, angry, neutral) with associated confidence scores.

2.2.3 Linguistic Pipeline (Speech-to-Text & Sentiment Analysis)

- **Objective:** To analyze the semantic content of what is being said.
- **Technology:** This pipeline will leverage a state-of-the-art Automatic Speech Recognition (ASR) model to transcribe the audio into text.
  1. **Transcription:** Utilize a powerful ASR engine like **OpenAI's Whisper** model for high-fidelity transcription, even in noisy environments. The **SpeechRecognition** Python library provides a convenient wrapper for this [8].
  2. **Text Emotion/Sentiment Analysis:** The resulting transcript will be processed by a Transformer-based Natural Language Processing (NLP) model (e.g., fine-tuning a model from **Hugging Face**) to determine the sentiment and emotional content of the words themselves [9, 10, 11, 12].

2.3 Stage 3: Multimodal Fusion Engine

This is the core of the system, where the outputs from the three unimodal pipelines are intelligently integrated to produce a single, coherent emotional assessment. Relying on a single modality is insufficient; fusion is critical for accuracy and robustness.

- **Technology:** I will implement an advanced fusion strategy. While early- and late-stage fusion are options, a **hybrid fusion model using an attention mechanism** is the state-of-the-art choice [13, 14, 15]. A Transformer-based architecture can learn to dynamically weigh the importance of each modality (visual, vocal, linguistic) in real-time, which is crucial for interpreting complex social cues like sarcasm, where vocal tone contradicts linguistic content [16, 17, 18, 19].

2.4 Stage 4: Metrics Generation & Data Export

The final stage translates the fused, low-level data into the high-level, actionable metrics required for comparison against success benchmarks.

3 Data Export Pipeline and Format

The system is designed to ingest raw data and export structured, analyzable metrics.

3.1 Pipeline Flow

1. **Input:** A user initiates a session, providing a live video/audio stream.
2. **Processing:** The stream is fed into the architecture described above. Each unimodal pipeline processes the data in real-time, generating time-stamped feature vectors.
3. **Fusion:** The Multimodal Fusion Engine combines these vectors at each time step (or over short windows) to produce a unified emotional state prediction.

4. **Aggregation:** The Metrics Generation module aggregates the time-series data to calculate overall scores for Engagement, Confidence, and Active Listening for the entire session.

5. **Export:** Upon session completion, the system generates a comprehensive JSON file containing all raw and aggregated metrics.

3.2 Exported Data Format (JSON Structure)

The output will be a structured JSON object designed for easy parsing and analysis. It will contain a session summary and a detailed time-series log.

```
{
  "session_id": "unique_session_id_123",
  "session_duration_seconds": 1800,
  "session_summary": {
    "overall_engagement_score": 0.85,
    "overall_confidence_score": 0.92,
    "overall_active_listening_score": 0.78,
    "dominant_emotion": "Neutral",
    "emotion_distribution": {
      "Neutral": 0.60,
      "Happy": 0.25,
      "Surprise": 0.10,
      "Sad": 0.05
    }
  },
  "time_series_data": [
    {
      "timestamp": "00:00:01.000",
      "fused_emotion": {
        "label": "Neutral",
        "confidence": 0.98
      },
      "visual_cues": {
        "face_detected": true,
        "head_pose": {"pitch": 0.05, "yaw": 0.01, "roll": -0.02},
        "action_units": [
          {"au": "AU01_r", "intensity": 0.0},
          {"au": "AU12_c", "intensity": 0.1}
        ]
      },
      "vocal_cues": {
        "is_speaking": false,
        "emotion": {"label": "Neutral", "confidence": 0.99},
        "pitch_hz": 150.5,
        "energy": 0.2
      },
      "linguistic_cues": {
        "transcript_segment": "",
        "sentiment": "Neutral"
      }
    },
    {
      "timestamp": "00:00:02.000",
      "fused_emotion": {
        "label": "Happy",
        "confidence": 0.88
      },
      "visual_cues": {
        "face_detected": true,
        "head_pose": {"pitch": 0.12, "yaw": -0.04, "roll": 0.03},
        "action_units": [
          {"au": "AU06_c", "intensity": 0.8},
          {"au": "AU12_c", "intensity": 0.9}
        ]
      },
      "vocal_cues": {
        "is_speaking": true,
        "emotion": {"label": "Happy", "confidence": 0.92},
        "pitch_hz": 180.2,
        "energy": 0.6
      },
      "linguistic_cues": {
        "transcript_segment": "That’s a great idea",
        "sentiment": "Positive"
      }
    }
  ]
}
```

4 Key Findings and Supporting Resources

The architectural decisions outlined in this report are grounded in established and emerging academic work. A curated list of key findings and relevant open-source tools is provided for reference. A full bibliography is available at the end of this document.

Table 1: Key Findings and Supporting Resources

Category	Finding / Tool	Description & Relevance
System Architecture	Multimodal Emotion Recognition (MER)	Integrating facial expressions, voice, and text provides a more accurate and robust analysis of emotional states than any single modality alone.
Visual Analysis	OpenFace 2.0	An open-source toolkit for real-time facial landmark detection, head pose estimation, and Facial Action Unit (AU) recognition. Essential for granular "mimic" analysis.
Visual Analysis	Dlib	A powerful C++ and Python library known for its highly accurate 68-point facial landmark detector, which is foundational for analyzing facial geometry [4, 5].
Vocal Analysis	librosa	A Python library for audio analysis that is the industry standard for extracting features like MFCCs and spectrograms needed for SER models [7].
Linguistic Analysis	SpeechRecognition	A Python library that provides a simple interface for various ASR engines, including OpenAI's Whisper, for accurate transcription [8].
Model Repositories	Hugging Face & TensorFlow Hub	These platforms host thousands of pre-trained models for vision, audio, and text, which can be fine-tuned to accelerate development and achieve state-of-the-art performance [20, 21, 11, 12].
Fusion Techniques	Attention-Based Fusion	Recent research (2024-2025) confirms that fusion methods using attention mechanisms and Transformer architectures are most effective for MER, as they dynamically weigh modalities [13, 14, 15, 16, 17, 18, 19].
Social Signal Processing	SSP Frameworks	The field of Social Signal Processing provides the theoretical basis for translating low-level cues (e.g., head nods, smiles) into high-level social metrics like engagement and confidence.

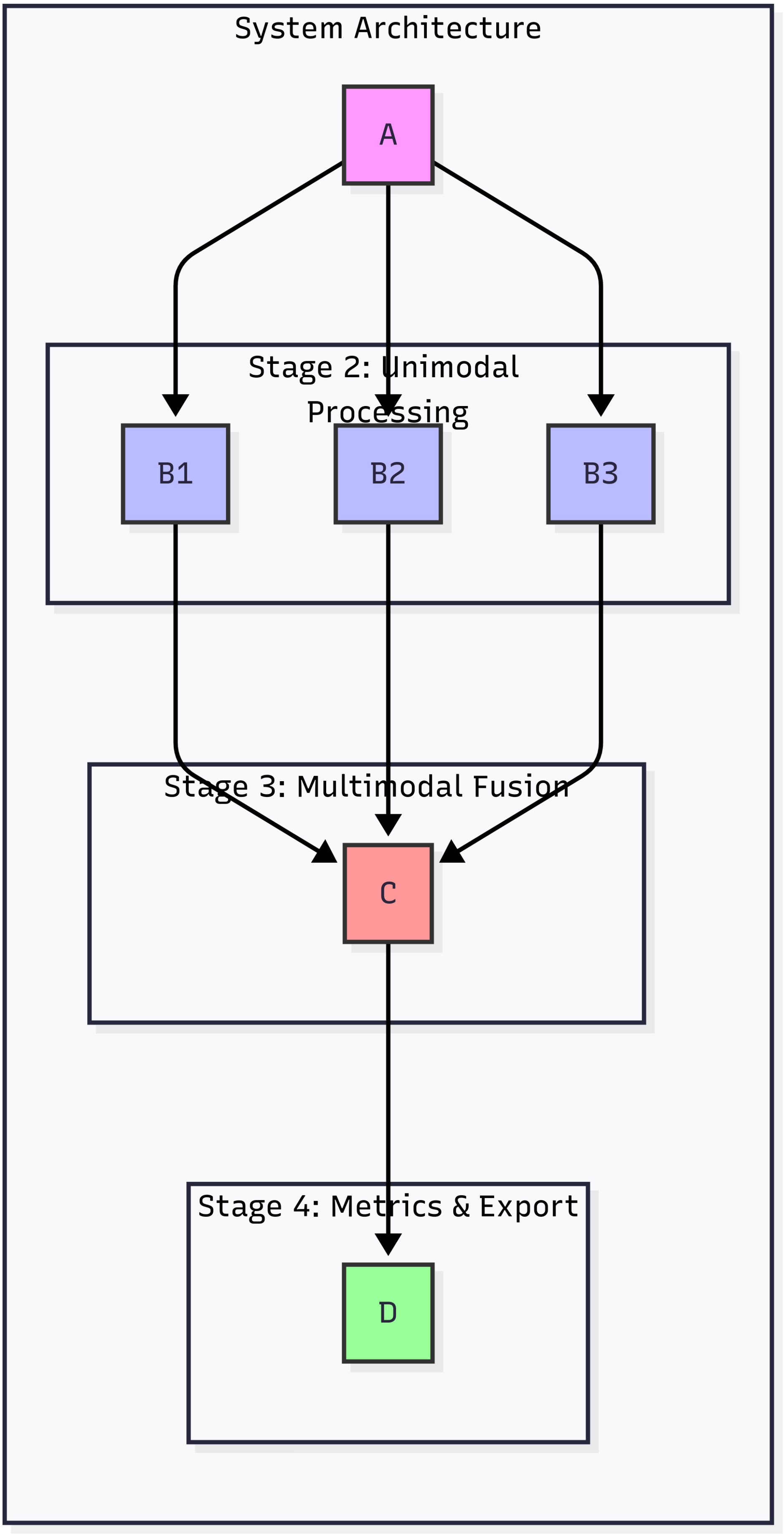
System Architecture
A
Stage 2: Unimodal Processing
B1
B2
B3
Stage 3: Multimodal Fusion
C
Stage 4: Metrics & Export
D

Figure 1: High-Level System Architecture for Multimodal Analysis.

References

[1] A Survey on Facial Expression Recognition (FER) within computer vision and pattern recognition.

[2] A study on emotion detection in video content using AI algorithms, including Multitask Cascade Convolutional Networks (MTCNN).

[3] A comprehensive guide to building a facial emotion detection model using a Convolutional Neural Network (CNN).

[4] An overview of facial landmark detection with dlib, OpenCV, and Python.

[5] An overview of the Dlib C++ toolkit and its Python bindings for machine learning and computer vision tasks.

[6] A review of popular feature extraction techniques in Facial Expression Recognition (FER) systems, including Action Units (AUs).

[7] An introduction to librosa, a python package for music and audio analysis.

[8] Documentation for the SpeechRecognition Python library, which supports several engines and APIs, including OpenAI Whisper.

[9] A fine-tuned Wav2Vec 2.0 model for Speech Emotion Recognition (SER) available on Hugging Face.

[10] An overview of SpeechBrain, an open-source conversational AI toolkit based on PyTorch, with models available on Hugging Face.

[11] A tutorial on using Hugging Face Transformers for emotion detection in text.

[12] A guide to pre-trained models for automatic speech recognition from the Hugging Face Audio Course.

[13] A study on cross-modal fusion techniques for emotion detection from spoken audio and corresponding transcripts.

[14] A Master's Thesis on Hierarchical Fusion Approaches for Enhancing Multimodal Emotion Recognition in Dialogue-Based Systems.

[15] A framework to evaluate fusion methods for multimodal emotion recognition, highlighting self-attention and weighted methods.

[16] Rizaldy, A., et al. (2025). HyperPointFormer: Multimodal Fusion in 3D Space with Dual-Branch Cross-Attention Transformers. *arXiv:2505.23206.*

[17] Xie, G., et al. (2025). HTMNet: A Hybrid Network with Transformer-Mamba Bottleneck Multimodal Fusion for Transparent and Reflective Objects Depth Completion. *arXiv:2505.20904.*

[18] Shihata, Y. (2025). Gated Recursive Fusion: A Stateful Approach to Scalable Multimodal Transformers. *arXiv:2507.02985.*

[19] Haque, M. R., et al. (2025). MMFformer: Multimodal Fusion Transformer Network for Depression Detection. *arXiv:2508.06701.*

[20] An introduction to TensorFlow Hub, a repository of trained machine learning models.

[21] A catalog of pre-trained models available on Kaggle, with a redirect from TensorFlow Hub.