DIGITAL SOUL: Multi-Modal Personality Prediction from Speech

DSI442 / Thammasat University

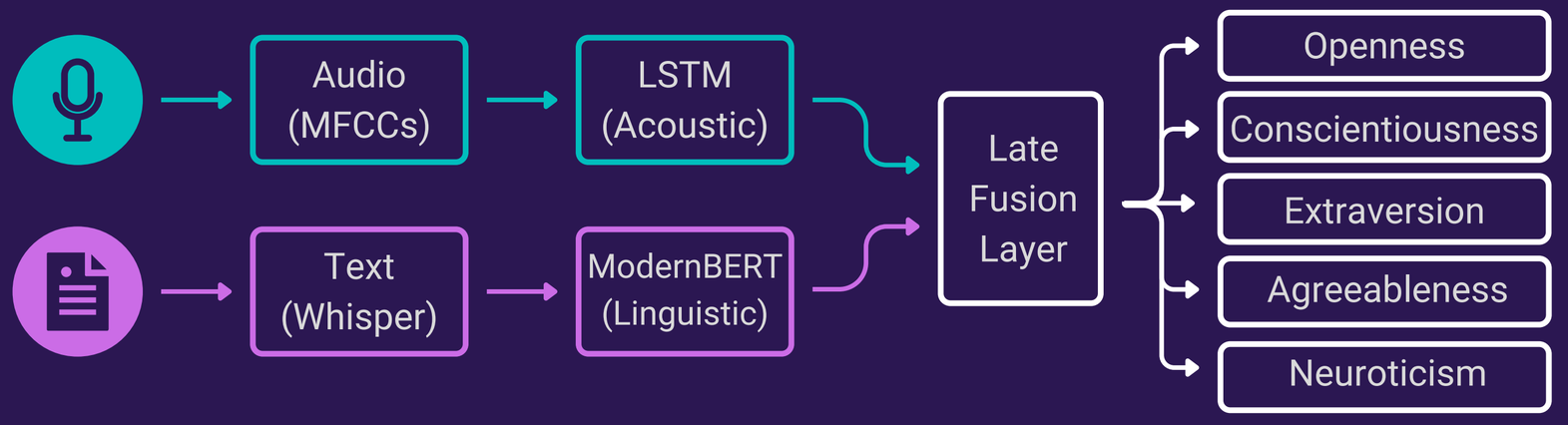
Motivation & Objective

1. **Background**
- Personality Computing: Understanding human personality is key to improving Human-Computer Interaction (HCI).
- The Missing Link: Traditional AI relies heavily on text analysis (what is said). However, psychology tells us that paralinguistics—how something is said (tone, energy, hesitation)—reveal emotional states that text misses.
- The Gap: Current models often treat these modalities separately or ignore audio entirely, leading to an incomplete "Digital Fingerprint."

2. **Problem Statement**
- How can we effectively fuse disparate data streams—discrete text tokens and continuous audio signals—to predict complex psychological traits?
- Can a machine learn to "hear" personality from just a 15-second audio clip?

3. **Objectives**
- Develop a Multi-Modal Deep Learning system that predicts the Big Five Personality Traits (Openness, Conscientiousness, Extraversion, Agreeableness, Neuroticism).
- Implement a Late Fusion architecture combining state-of-the-art Linguistic (ModernBERT) and Acoustic (LSTM) models.
- Evaluate the system against a random baseline to prove that fusing audio and text yields higher accuracy than either modality alone.

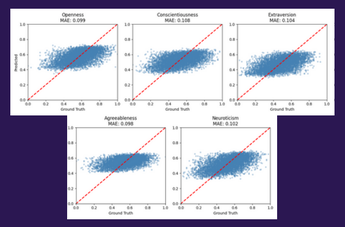
Methodology: Multi-Modal Fusion



Experiments & Results

1. **Quantitative Performance**
- Metric: Mean Absolute Error (MAE) on Test Set.
- **Score: 0.1076**

2. **Visual Evaluation**

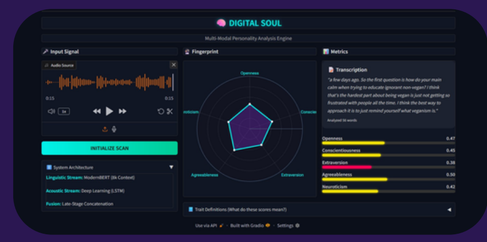


Predicted vs. Actual scores across all 5 traits.

3. **Key Findings**
- Fusion Works: Combining modalities reduces error by ~12% compared to using text alone.

Multi-Modal Fusion yields the lowest error compared to Text-only or Audio-only models.

Analysis & Conclusion



Limitations & Future Work
- Language Barrier: Incorporate XLM-RoBERTa to support Thai and other languages.

Conclusion
- This project successfully validates that a Late Fusion Deep Learning Architecture can predict personality traits from short audio clips with a 0.1076 MAE, significantly outperforming random baselines and demonstrating the viability of automated personality profiling.