Nazarbayev University

CSCI 409 Senior Project II – Final Project Report
Spring 2025
Project Title: AI-Based Multimodal Emotion Recognition System
Group Members: Ilyas Sultangazy, Nurasyl Sapa, Anuar Ospanov, Temirlan Zhexembeyev
Advisor: Professor Khalil Khan

1. Executive Summary

The project built an AI system for emotion recognition which integrated video data-processing with audio analysis as well as textual information assessment. Real-time facial emotion detection through Vision Transformers (ViT) and speech emotion recognition through Wav2Vec2 made up the core targets of the project with the aim of their integration. The project overcame dataset problems along with scope adjustments by concentrating on processing video and audio content instead of text analysis. The ultimate version of the prototype shows **90% accuracy** in detecting emotions across high-definition video material thus creating a new framework which benefits applications in the areas of service interaction and psychiatric assessment.

2. Introduction

Emotion recognition systems are important for human computer interaction and are known to be generally inaccurate under different situations. The issues present here are attacked by this project in a multimodal fashion using the latest and greatest in deep learning models. It processes video frames for facial expressions, audio for speech emotion analysis, and integrates the results for knowing the whole norms and behaving accordingly.

3. Background and Related Work

Key Prior Work:

- **Facial Emotion Recognition:** The FER-2013 dataset and ViT architecture (trpakov/vit-face-expression) improved accuracy over traditional CNNs [1].
- **Speech Emotion Recognition:** Wav2Vec2 (superb/wav2vec2-base-superb-er) outperforms RNN-based models in noisy environments [2].
- **Multimodal Fusion:** Prior systems like OpenFace [3] lack audio-text integration, which our pipeline addresses.

Our system merges different approaches to analyze multiple modalities without necessitating retraining of the architectural framework.

Users must input information through the system in the English language at this time. The testing of non-English languages has not been conducted yet and model retraining with language-specific datasets is needed to achieve satisfactory performance. Unfortunately, no qualitative datasets in Kazakh languages exist, so there was not enough data to train the model.

4. Project Approach

Technical Methodology:

- **Facial Emotion Recognition:**
  - **Model:** `trpakov/vit-face-expression` (Vision Transformer fine-tuned on facial expressions).
  - **Workflow:** Video frames extracted → face detection → emotion classification.
- **Speech Emotion Recognition:**
  - **Model:** `superb/wav2vec2-base-superb-er` (Wav2Vec2 fine-tuned for emotion detection).
  - **Workflow:** Audio extracted from video → segmented into clips → emotion classification.
- **Integration:**
  - Parallel processing of video and audio streams.
  - Results combined into a unified report with confidence scores.

Tools & Pipeline:

- **Frontend+Backend:** No heavy frameworks were used because we decided it was excessive for the ML project, so we stuck with HTML+CSS+JS.

5. Project Execution

Challenges & Adjustments:

- **Dataset Limitations:** Public datasets lacked diversity; frames were manually filtered for quality.
- **Scope Reduction:** Text sentiment analysis was deprioritized to focus on video/audio.
- **Leadership Shift:** Temirlan Zhexembeyev assumed ML development leadership due to expertise in optimizing ViT/Wav2Vec2.

Team Roles:

- **Temirlan Zhexembeyev:** ML pipeline optimization.
- **Ilyas Sultangazy:** Frontend and user authentication, ViT model integration.
- **Nurasyl Sapa:** Deployment and integration of the model into the backend.
- **Anuar Ospanov:** Audio processing and API development.

Language Support Limitations

1. The team attempted to introduce Kazakh language capabilities but faced three essential problems during their analysis.
2. The available Kazakh emotion datasets exhibited insufficient availability because they had small sample quantities and audio files of inconsistent quality.
3. The team encountered training limitations because applying their available resources to tune their speech/text models for Kazakh proved insufficient.
4. Preliminary testing with Kazakh Speech yielded no better results than the standard English model because of unclear audio data.
5. The project team decided to narrow down the scope by focusing on making English sentences correct.
6.

6. Evaluation

Metrics:

- **Accuracy:**
    - **Facial Recognition:** 87% (average), 90% (high-quality videos).
    - **Speech Recognition:** 82% (neutral tones), 78% (noisy environments).
- **User Feedback:** 85% of testers rated the interface intuitive via post-analysis questionnaires.

Limitations:

- Performance drops with low-resolution videos.
- No temporal analysis between frames.

7. Conclusion & Future Work

- The support of Kazakh language acquisition in their system encountered numerous hurdles due to the scarcity of high-quality datasets and minimal computing power along with poor audio quality throughout Kazakh speech corpora. Local institutions need to work together with the project to create domain-focused data collections and secure cloud-based training capabilities.