Emotion Recognition Based on Audio Data Using Sensor-Driven AI

Dr. Honghui Xu

Danielle Del Barrio

Sadie Shudde

Garvin Francois

Kennesaw State University

Abstract

This project explores the development of an emotion recognition system that utilizes sensor-driven artificial intelligence (AI) to analyze audio data. Further addressing challenges such as environmental noise, speaker variability, and contextual cues for emotion expression. This project will utilize low-cost microphone sensors and deep learning models, as well as the capturing of temporal speech patterns and contextual dependencies. The methodology includes training on controlled datasets and evaluation on speech, with a focus on enabling generalization across diverse speakers and environments. In regards to AI use, this work will contribute to the development of real-time emotion recognition systems grounded in sensor-based AI. Overall, this project will help in advancing applications used in mental health monitoring, human–device interactions, and personalized assistance use.

As robotics and artificial intelligence (AI) systems become increasingly rampant in our daily lives, understanding and responding to human emotion is critical for effective consumer use. Emotion-aware AI can significantly enhance user experience, trust, and safety, especially in diverse settings where systems must adapt to real-time human behavior. In such contexts, emotion recognition serves as a useful tool for devices, enabling AI systems to interpret subtle emotional cues and respond accordingly.

Recent advancements in sensor-driven AI have opened new possibilities for emotion analyzation through non-intrusive methods, like audio-based analysis'. Unlike facial expression or biological monitoring, audio signals offer a unique set of emotional data through the diverse variations of tone, pitch, and intensity. Furthermore, analyzing audio datasets with sensor-driven AI can be cost effective, as this project can simply rely on microphone sensors. On the other hand, some challenges this project can face, include different environmental noises, speaker variability, and the cultural contexts of human expression. To conclude, this literature review will discuss past methodologies, models, and datasets used in audio-based emotion recognition, with a focus on sensor-driven, deep learning approaches aimed at improving robustness and generalization in real-world settings.

Review of Literature

In modern times, the growing complexity of daily life and its emotional demands have intensified interest in affective computing. Affective computing is an interdisciplinary field that focuses on recognizing and responding to human emotions using technology, using external stressors that can distinguish one's physical and mental well-being. Researchers and engineers in particular, have turned to medical and sensor-based technologies to detect and interpret emotional states. "With uncontrollable external factors escalating the emotional states in our daily life, there is no denying that our health and fitness, both physically and mentally, would be influenced. Thus, the rise of affective computing, aiming at the application of the sensing technology to assess human emotions, has intrigued the interest of researchers and engineers to utilize various medical technology for the detection and recognition of different affective manifestations" (Basha, Shaik).

Emotion recognition systems, particularly those based on audio data, rely heavily on established psychological frameworks. One common model identifies six basic human emotions, which are anger, disgust, fear, happiness, sadness and surprise. These emotions can then be classified using arousal and valence. This is described in the literature, which states that these emotions can then be stimulated with varying arousal and valence levels. "The quadrants of arousal and valence axes are universally divided into joy (positive valence and positive arousal), pleasure (positive valence and negative arousal), anger (negative valence and positive arousal), and sadness (negative valence and negative arousal)" (López-Pérez, Belén). Understanding these quadrants is essential for designing a machine learning model that can classify emotions using audio signals like tone, pitch, and rhythm. Additionally, studies have shown that anger is more reliably detected than emotions like fear or sadness, possibly due to clearer linguistic patterns. As stated in the text where, “Some of the results pointed to higher sensitivity and specificity for anger scenarios compared to the other emotions, especially fear. The classification challenges observed in the fear and sadness scenarios can be due to participants frequently reporting strategies that elaborate on the target’s emotional experience” (López-Pérez, Belén).

Given our team’s goal of building a sensor-driven, audio-based emotion recognition system, we plan to implement Long Short-Term Memory (LSTM) networks. LSTM is a specialized type of Recurrent Neural Network (RNN), that we believe is best suited for analyzing quick inputted data such as speech. One paper that supports this idea, states that “Since the RNN is a simpler system, the intuition gained by analyzing the RNN applies to the LSTM network as well. Importantly, the canonical RNN equations, which we derive from differential equations, serve as the starting model that stipulates a perspicuous logical path toward ultimately arriving at the LSTM system architecture” (Scott, Jano). This shows how the LSTM would better support design decisions in preparing our dataset and model pipeline. Furthermore, “The advantage of this path is that it affords an opportunity to build a certain degree of intuition that can prove beneficial during all phases of the process of incorporating an open source model” (Scott, Jano). This would be important as the development of our project must be adaptable as we input important information such as speech, and must take out information quickly in real-time environments.

As we move forward with our project, recognizing what can evoke certain emotions is also important. Music in particular, is often used in experiments to provoke distinct emotional states because of its strong psychological impact. This is supported by the idea that, “Music is one of the most popular choices of stimuli to induce the affective states… Another form of art including paintings and images have been shown to induce emotions depending on the observers" (López-Pérez, Belén). These methods of stimulation would help in creating robust training data for models that aim to detect nuanced emotional cues.

Finally, sensor-driven AI systems often incorporate cognitive modeling to simulate human emotion processing. As "Models like this can be used to simulate or predict human behavior and performance on activities similar to the ones being modeled, leading to better human-computer interaction" (Leelaarporn, Pitshaporn). By combining audio signal processing with cognitive models, researchers can create more adaptive and responsive AI systems capable of real-time emotional analysis.

To conclude, the integration of psychological theory, sensor-based data, and advanced AI models has driven meaningful progress in the field of audio-based emotion recognition. However, continued research is needed to refine these systems for broader emotional accuracy, personalized responsiveness, and real-world usability.

References

Basha, Shaik Abdul Khalandar, and P. M. Durai Raj Vincent. “DHERF: A Deep Learning Ensemble Feature Extraction Framework for Emotion Recognition Using Enhanced-CNN.” *Journal of Advances in Information Technology*, vol. 15, no. 7, 2024, pp. 853–861. doi:10.12720/jait.15.7.853-861.

Leelaarporn, Pitshaporn, et al. “Sensor-Driven Achieving of Smart Living: A Review.” *IEEE Sensors Journal*, vol. 21, no. 9, 1 May 2021, pp. 10369–10376. doi:10.1109/JSEN.2021.3059304.

López-Pérez, Belén, et al. “Exploring the Potential of Large Language Models to Understand Interpersonal Emotion Regulation Strategies From Narratives.” *Emotion*, American Psychological Association, Apr. 2025, https://doi.org/10.1037/emo0001528.

Scott, Jano, and Sotirios Diamantas. “Vocal Verification.” *Mayfield College of Engineering, Tarleton State University*, 2025. Poster.

Zhang, Xingxia, et al. “Psychological Mechanism of Language Cognition to ‘Awaken’ Artificial Intelligence.” *Psychological Trauma: Theory, Research, Practice, and Policy*, vol. 16, no. 5, 2024, pp. 825–836. doi:10.1037/tra0001305.

Brunnlieb, Cornelia, et al. “Understanding Investment Behavior: Evidence from Neuroeconomics.” *Journal of Economic Behavior & Organization*, vol. 170, 2020, pp. 353–372. Elsevier, https://doi.org/10.1016/j.jebo.2019.12.020.