A Speech and Facial Information based Emotion Recognition System of Collaborative Robot for Empathic Human-Robot Collaboration

Jianna Loor, Jordan Murphy, and Rui Li*, *Member, IEEE*

***Abstract*— A robot's ability to effectively recognize human emotions is critical in human-robot collaboration. However, most of the current collaborative robots were designed to improve productivity. Few of these robots consider human emotions. This situation would cause humans to be unwilling to work with robots for a long time. Motivated by this gap, this research developed a human emotion recognition system for enhancing the interaction abilities of collaborative robots. In this project, both speech and facial information were analyzed for robust human emotion recognition in complex working environments like manufacturing assembly environments. In the experiment, the developed system has been tested through three different human-robot co-assembly scenarios: (1) the robot effectively assists humans in finishing the task, (2) the robot is slow in response, leading to the task failing, (3) the robot frequently picks up wrong tools leading to task failure. Experimental results have demonstrated the effectiveness of the developed system in recognizing human co-worker's emotions when the human and the robot were working in the above scenarios. It further shows the developed system has the potential to contribute to the development of an empathic collaborative robot companion in the manufacturing area.**

I. Introduction

Collaborative robots have wide applications in smart manufacturing areas. Though collaborative robots are proposed to work closely with human workers, these robots do not consider human factors well. Human factors are important for human-robot collaboration as well as can be complex and involve multiple aspects. Among these aspects, human emotions are one of the very important factors that are usually overlooked and could potentially impact the collaboration process. However, integrating emotion recognition in practical human-robot co-assembly remains unresolved due to the current collaborative robots in manufacturing were mainly designed to be productivity-centered only. Moreover, the complexities of the manufacturing environment also led to difficulties in implementing emotion recognition in collaborative robots. These questions will further limit the application of collaborative robots in working closely with humans.

To solve the questions, this paper develops a multimodal emotion recognition method that can be implemented on collaborative robots to better understand human status during co-assembly tasks. Such recognition results have the potential to be further used for guiding the robot behaviors for better human-robot collaboration in finishing assembly tasks. The developed system was tested and evaluated in an experimental setup, designed to mimic realistic industrial scenarios, offers a safe and controlled platform to explore the abilities of collaborative robot's in understanding human emotions. In the experiment, the collaborative robot was tasked with improving task efficiency and identifying human emotion changes during the progress of the tasks.

The contribution of this paper can be described as: (1) developing a multimodal emotion recognition system that can be integrated into manufacturing collaborative robots. Using multimodal information (speech and facial) can help avoid the instability caused by complex environments and relying on a single modality. (2) designing three different real-world collaborative assembly scenarios for naturally eliciting human different emotions. (3) validating and analyzing the effectiveness and performance of the developed system through the designed real-world co-assembly tasks. The findings obtained from the validation and analysis can offer insight into the study of the dynamic interplay between human emotions and robot-assisted tasks.

II. Related Works

Emotion recognition has made significant strides in recent years, harnessing the power of multimodal information processing. For human emotion recognition, facial expression recognition (FER) and speech-emotion recognition (SER) have been widely used.

Chowdary et al. [1] demonstrated a 96% accuracy in facial expression recognition (FER) using transfer learning with pre-trained networks like ResNet50 and Inception V3 on the CK+ database. Ko et al. [2] reviewed the shift from conventional FER methodologies to deep-learning-based approaches, highlighting the effectiveness of convolutional neural networks (CNNs) and long short-term memory (LSTM) networks. Jain et al. [3] presented an advanced deep neural network (DNN) model that surpassed state-of-the-art FER approaches with deep residual blocks and convolution layers. Li et al. [4] surveyed deep learning techniques in FER, addressing challenges like overfitting and expression-unrelated variations. Pramerdorfer et al. [5] and Mehendale [6] explored the use of CNNs for recognizing facial expressions, with Mehendale introducing the FERC technique for improved accuracy. Canal et al. [7] comprehensively analyzed FER strategies, distinguishing between classical and neural network-based approaches. Swain et al. [8] emphasized the

Jianna Loor, Jordan Murphy, and Rui Li are with the School of Computing, Montclair State University, Montclair, NJ 07043 USA.
*Corresponding author: Rui Li (email: liru@montclair.edu)

expressive capacity of speech and the development of various databases and classifiers. Issa et al. [9] introduced an architecture that significantly improves SER accuracy using mel-frequency cepstral coefficients and other acoustic features. Khalil et al. [10] highlighted the use of attention based DNNs for mining information from speech signals. Aouani et al. [11] proposed a two-stage emotion recognition approach using autoencoders for feature extraction and SVM for classification in SER. Jahangir et al. [12] contributed to advancing SER through various deep learning approaches, emphasizing automatic feature extraction, affective gap bridging, and the consideration of emotional and neutral speech parts, respectively. Bertero et al. [13] introduced a real-time CNN model for emotion and sentiment recognition from raw speech, bypassing traditional feature engineering. Patel et al. [14] showed that dimensionality reduction via autoencoders can enhance emotion detection accuracy. Khan et al. [15] described a feature fusion approach using a deep stride CNN combined with bi-directional LSTM, significantly improving accuracy on the RAVDESS dataset. In addition, the synergy between facial expression and speech analysis has led to innovative multimodal approaches that enhance interaction experiences between humans and machines. Liu et al. [16] demonstrated the efficacy of transfer learning from FaceNet for speech emotion recognition. Cai et al. [17] proposed a method merging speech and facial expression features that significantly improved the IEMOCAP dataset. Guanghui et al. [18] emphasized the strength of combining speech and facial data, improving accuracy and robustness in emotion recognition. Lastly, Siddiqui et al. [19] presented a framework that fuses speech with visible and infrared images, achieving high accuracy across diverse environments. In the realm of collaborative robots, Heredia et al. [20] proposed an adaptive and flexible emotion recognition architecture utilizing EmbraceNet+ for social robots. Antonelli et al. [21] developed an emotion recognition system for collaborative robots, focusing on enhancing the safety and efficiency of human-robot interactions by implementing emotional intelligence.

Existing works for emotion recognition often do not fully address the complexities of real-world applications, especially in dynamic and noisy environments such as manufacturing. Our work aims to solve this with a multimodal emotion recognition system developed and tailored for collaborative robots in manufacturing.

III. Mathematical Model

A. Speech Recognition

In the domain of emotion recognition from speech data, the approach begins with a process of augmenting the audio data to improve the model's robustness against variations in speech speed and pitch. Specifically, time-stretching is used to adjust the pace of the audio signal. It can be represented using:

$$y_{fast} = time\_stretch(y, A), \quad (1)$$

$$y_{slow} = time\_stretch(y, B), \quad (2)$$

where A=1.25 for speeding up the audio and B=0.75 to slow down the audio. Additionally, pitch shifting modifies the pitch of the audio signal upwards and downwards by two semitones:

$$y_{pitch\_up} = pitch\_shift(y,\ sr,\ 2), \quad (3)$$

$$y_{pitch\_down} = pitch\_shift(y,\ sr, -2), \quad (4)$$

The feature extraction process is critical for translating the raw audio data into a more analytically useful form. Mel-frequency cepstral coefficients (MFCCs) are central to this process, capturing the timbral aspects of the speech. The calculation of MFCCs involves mapping the power spectrum of the audio signal onto the mel scale, closely approximating the human auditory system's response. The MFCCs:

$$M = DCT(\log(\varphi(y))), \quad (5)$$

where M signifies the MFCCs, DCT represents the discrete cosine transform, and $\varphi(y)$ represents the Mel Spectrogram function applied to the audio signal y. The Mel Spectrogram, another pivotal feature, is computed to provide a logarithmic amplitude representation of the sound, given as

$$S_{\text{mel}} = \log(1 + \phi \cdot \varphi(y)), \quad (6)$$

where S_{mel} represents the logarithmic amplitude representation of the sound after mapping onto the mel scale, and ϕ is a constant equal to 10000, improving the precision of the equation. Furthermore, spectral contrast and Tonnetz features offer additional insights into the emotional content of speech. Spectral contrast delineates the difference in amplitude between the spectrum's peaks and valleys, while Tonnetz features, capturing changes in harmonic content, can indicate different emotions conveyed through speech.

The cornerstone of the emotion recognition model is the use of Long Short-Term Memory (LSTM) networks, which are adept at processing data sequences and capturing temporal dependencies within them. The model employs a Bidirectional LSTM layer to analyze the sequence of forward and backward speech features, enhancing its capacity to understand the context and nuances of emotional expressions. The hidden state updates in the LSTM layer are governed by:

$$H_t = \sigma(W_{ih}X_t + b_{ih} + W_{hh}H_{t-1} + b_{hh}), \quad (7)$$

where H_t represents the hidden state at time t, σ represents the sigmoid function to ensure non-linearity, W_{ih} and W_{hh} represent the weight matrices that transform inputs and relay information across time steps, respectively. X_t is the input at a time, and b_{ih} and b_{hh} are the bias terms.

The model includes Dense layers to classify emotions from the processed features, where the final classification is performed through a combination of linear and non-linear transformations. Each Dense layer's output is calculated by:

$$Y = \gamma(WX + b), \quad (8)$$

where Y represents the output, W and b denote the weight matrix and bias term, respectively, and γ is the activation function, such as ReLU for intermediate layers and softmax for the output layer, ensuring probabilities for each emotion class are obtained.

The training of this model is suitable for multi-class classification scenarios and utilizes the categorical cross entropy loss function, defined as

$$L = -\sum_{e=1}^{T} y_{o,c} \log p_{o,c} \quad (9)$$

where T represents the total number of classes, $y_{o,c}$ a binary indicator of whether the class c is the correct classification for

o, $p_{o,c}$ represents the predicted probability that observation o belongs to class c, as output by the model. This function quantifies the difference between the predicted probabilities and the actual distribution of the labels, guiding the model towards minimizing this discrepancy over the training process.

The LSTM model was trained on the RAVDESS dataset [22], which includes a diverse range of emotion expressions in speech. The training resulted in an accuracy of 89.6%, demonstrating the model's effectiveness in recognizing a variety of emotions. The spectral contrast graph in Fig. 1(a) highlights the distinction between high-frequency peaks and low-frequency valleys, suggesting harmonic structures or noise indicative of robotic activity or interaction. The Tonnetz graph in Fig. 1(b) shows fluctuations in tonal centroids that may reflect changes in the participant's vocal pitch and harmony in response to the robot's behavior. In contrast, the MFCC graph in Fig. 1(c) indicates a stable vocal tract configuration, implying controlled emotional expression during the interaction. The Mel spectrogram in Fig. 1(d) outlines energy distribution across frequencies, with bright bands indicating speech or high-frequency noise and darker areas suggesting less activity.

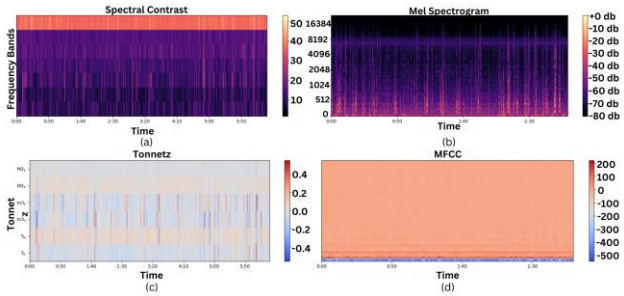

Figure 1: Graphs of Voice Features: (a) Spectral Contrast, (b) Tonnetz, (c) Mel-frequency cepstral coefficients, (d) Mel Spectrogram.

B. Facial Expression Recognition

Considering the complexities of the working environment may impact the stability of speech signals, facial information is also considered for robust emotion recognition results. The DeepFace library based on facial expression data is integrated into the system. The choice of DeepFace is attributed to its high accuracy and efficiency in processing diverse facial expressions, making it ideal for real-time emotion recognition tasks. Based on a robust convolutional neural network (CNN) model, DeepFace excels in processing and interpreting diverse facial expressions with high accuracy and efficiency, making it well-suited for real-time emotion recognition tasks. The library's core functionality involves extracting and analyzing frames from continuous video capture, utilizing its analyze function to detect facial expressions within each frame, as shown in Fig. 2. This process leverages a pre-trained CNN to identify the dominant emotion presented based on a interpretation of facial features and expressions.

The technical prowess of DeepFace lies in its ability to provide detailed insights into the emotional dynamics of individuals, offering a window into the range of emotional expressions exhibited over time. By accumulating and assessing the dominant emotions detected across fixed

Figure 2: Example of emotion recognition by using DeepFace library.

intervals, DeepFace employs a systematic aggregation method. This involves calculating the frequency of each detected emotion within the interval and identifying the most prevalent emotion as the period's dominant emotional state. Such an approach allows for understanding emotional trends and patterns. Under the framework of Deepface, the VGG-Face model was further utilized due to its high performance in facial emotion recognition tasks, achieving an accuracy of 95% upon testing, making it an ideal choice. It also works well for face recognition with complex backgrounds.

C. Speech and Facial Information for Emotion Recognition

This section will outline the method for combining both speech and facial information for emotion recognition. Given a video segment, the system generates a set of emotions, E_{video}, from facial analysis, and an E_{audio}, from audio analysis. The facial emotion recognition, powered by DeepFace, yields a set of emotions for each frame, which are then aggregated to identify the most frequent (dominant) facial emotion, E_{vd}, within the segment, with vd representing the video dominant emotion. This process can be mathematically represented as:

$$E_{vd} = mode\{E_{video}\} \quad (10)$$

where the mode function identifies the most frequently occurring emotion in the set of emotions detected in the video frames. Simultaneously, the audio emotion recognition system processes the corresponding audio segment to produce an emotion prediction, E_{audio}. This prediction is based on the features extracted from the audio signal and analyzed through an LSTM model. The emotion prediction from audio can be defined as:

$$E_{audio} = \arg\max(P_{audio}) \quad (11)$$

where P_{audio} represents the set of probability distributions over possible emotions obtained from the LSTM model, and arg max selects the emotion with the highest probability.

Integrating of audio and video emotional cues into a unified dominant emotion for each segment involves a decision rule prioritizing the non-neutral emotion from the video analysis, resorting to the audio-derived emotion if the video emotion is deemed neutral. Mathematically, this decision rule can be expressed as:

$$E_{dom} = \begin{cases} E_{video_{dominant}} \; if \; E_{video_{dominant}} \neq "neutral" \\ E_{audio} \qquad\qquad otherwise \end{cases} \quad (12)$$

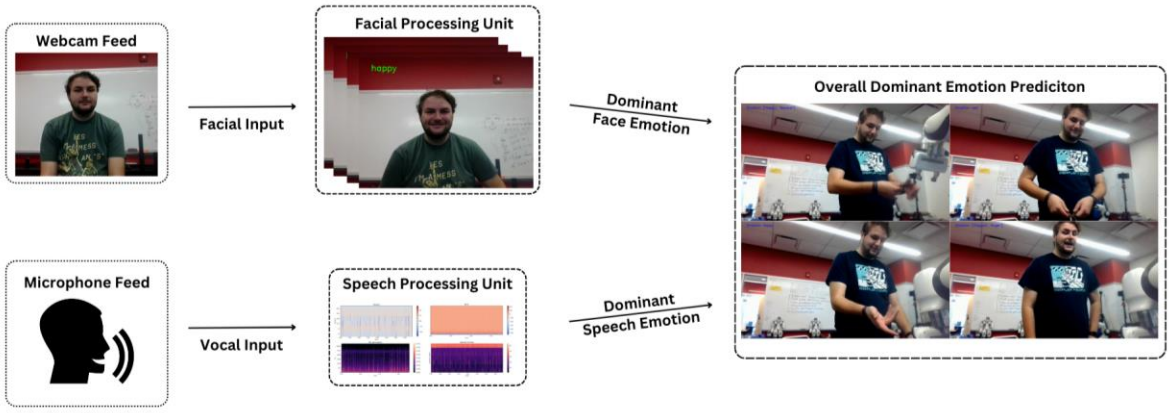

Figure 3: System Overview, illustrating the multimodal emotion recognition system's workflow. This includes (a) facial input that is analyzed using the DeepFace framework and (b) vocal input that is analyzed via an LSTM network, focusing on acoustic features. The overall dominant emotion prediction (c) combines the dominant face and speech emotion to determine the dominant emotion.

Implementing this combined approach requires careful synchronization between the video and audio processing pipelines to ensure that the emotions are aligned temporally. Each video segment's dominant emotion is then annotated onto the video to provide a comprehensive emotional narrative of the observed scene. This integrated model leverages the strengths of both audio and visual cues to overcome the instability caused by complex environments and relying on a single modality.

IV. System Overview

The system overview in Fig. 3 describes the developed emotion recognition system for interpreting human emotional states by analyzing multimodal data (speech signals and facial information). The workflow initiates with simultaneous data collection from facial and voice inputs. Facial expressions are captured and recognized in real-time. Concurrently, voice input is meticulously preprocessed to highlight emotional indicators, then parsed into acoustic features like MFCCs and Mel spectrogram. These features are analyzed using an LSTM network, specifically tuned to decode the complex emotional cues embedded within speech dynamics. In addition, a rule-based model is employed to determine the dominant emotion, favoring the output from facial analysis unless a 'neutral' emotion is detected, at this point, the speech analysis takes precedence. This stratified approach ensures the system's acute sensitivity to overt emotional expressions and adeptness at discerning subtler nuances from speech when facial cues are not definitive. Fig. 3 also demonstrates the output phase of the system, where the integrated emotion data is ready to be annotated onto the video feed or utilized in interactive applications.

V. Experiments and Analysis

A. Experimental Setup

The core of the experimental design revolves around co-assembly tasks, where a human and a robotic assistant co-assemble an object that necessitates the utilization of a screwdriver, an Allen wrench, and a piler fundamental to the assembly process. Fig. 4 shows an example of our experimental setup. The Franka Emika panda is implemented to collaborate with the human worker. A laptop's built-in webcam is utilized for video capture of humans during assembly tasks. Audio is captured using an Audio-Technica AT2020USB+ microphone connected to the laptop, ensuring high-quality audio recording for further analysis.

Three task scenarios are carefully designed for observing the impact of various robotic actions—namely, the robot's response speed, its precision in selecting the correct tools, and the velocity at which it delivers these tools—on the emotions of the human. Specifically, in scenario (1), the robot promptly provides the appropriate tools. This scenario (1) is hypothesized to elicit the satisfaction or happiness emotion of the human. Conversely, two other scenarios characterized by the robot's delayed actions (2) or incorrect tool selection (3) are hypothesized to elicit negative emotions, including frustration or impatience. To ensure a comprehensive assessment, the experiment is structured around several vital phases, beginning with the initial setup, where the human participant and the robotic assistant are positioned at the assembly station, equipped with all necessary tools and the components to be assembled as shown in Fig. 4. Following the setup, the experiment proceeds with executing the predefined task scenarios. These scenarios encompass a spectrum of interaction dynamics, ranging from effective collaboration, characterized by the robot's accurate and timely assistance, to challenging situations where the robot's performance is intentionally compromised, such as by introducing delays in tool delivery or by selecting incorrect tools, to simulate

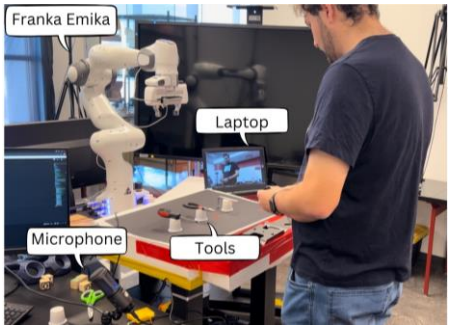

Figure 4: Experimental setup.

potential real-world complications and observe their effect on the human worker's emotional state and task performance.

B. Experimental Results and Analysis

In the first scenario, where the robot provided tools in an accurate and timely manner, the observed emotional outcomes predominantly matched the expected happy or neutral states. Participants exhibited clear signs of satisfaction, such as brief smiles and positive verbal feedback, highlighting the efficiency and smoothness of the task completion process as shown in Fig. 5(a). The system's accuracy in recognizing these emotions was 88%. Notably, despite its predictive interval of every 5 seconds, the emotion recognition system captured these positive responses consistently. For example, in Fig. 5(a-1), the participant is seen requesting a tool with a satisfied demeanor. The prompt delivery of the tool is captured in Fig. 5(a-2), where the participant's satisfaction continues. Fig. 5(a-3) and 5(a-4) show the participant engaging in the assembly task with evident contentment, indicative of the smooth interaction with the robot. In Fig. 5(a-5), the participant is captured returning the tool to the robot, with his expression and body language consistently conveying a happy or neutral state. Fig. 5(a-6) depicts the participant receiving pliers from the robot. The emotion recognition system fluctuates between happy and neutral, reflecting the subject's contentment with the interaction. The set concludes with Fig. 5(a-7), where the participant is seen stowing away the tool, the emotion recognition system affirming a sustained happy or neutral demeanor throughout the task. During the second scenario, characterized by deliberate delays in robotic tool assistance, participants' responses were anticipated to be sad or neutral. The observed outcomes generally aligned with these expectations. The system's accuracy in recognizing these emotions was 81%. The delays in assistance led to moments of visible impatience, such as shifts in posture or sighs, reflecting a slight dip in participant satisfaction as shown in Fig. 5(b). For example, Fig. 5(b-1) captures the participant's frustration due to the robot's slow pace. The anticipation of assistance is visually apparent in Fig. 5(b-2), with the participant showing impatience while waiting for the robot. Fig. 5(b-3) depicts the participant's dissatisfaction upon finally receiving the tool, and Fig. 5(b-4) shows the participant continuing the assembly with a sense of annoyance, underscored by the delay in assistance. Fig. 5(b-5) illustrates the participant using the screwdriver with visible signs of annoyance, indicating the emotional shift due to the robot's delayed assistance. In an instance of the system's limitations, Fig. 5(b-6) shows the participant agitated while requesting a tool, yet the predicted emotion is incongruently registered as happy/neutral. Fig. 5 (b-7) demonstrates the complexity of human emotion. The participant is laughing, a response to frustration, which the system misinterprets as a happy state due to the smile. However, these moments did not escalate to overtly negative emotional displays, indicating a resilience or understanding toward the experimental constraints. The emotion recognition system's intervallic predictions captured these subtler emotional shifts but were also susceptible to noise. At times, the system interpreted non-emotional cues (e.g., looking away in thought or adjusting seating position) as emotional responses, adding a layer of complexity to interpreting these neutral to mildly negative states as shown in Fig. 5(c-3). The third scenario introduces delays and inaccuracies in tool assistance, with an expected outcome of anger or fear. The system's accuracy in recognizing these emotions was 90%. The observed emotional responses included visible frustration, such as frowning and negative verbalizations, which aligned with the anticipated angry or fearful states as shown in Fig. 5(c). For example, Fig. 5(c-1) shows the participant with a clear expression of anger while requesting a tool. In Fig. 5(c-2), the reception of an incorrect tool further exacerbates the participant's frustration. Fig. 5(c-3) is notable as it depicts a seemingly happy emotion; however, this is identified as 'noise' within the data since the subject is notably frustrated. Lastly, Fig. 5(c-4) shows continued anger as the participant receives another incorrect tool, emphasizing the negative impact of compounded delays and inaccuracies in robotic assistance on the participant's emotional state. Figure 5(c-5) shows the participant as he angrily replaces the tool, a gesture that suggests dissatisfaction with the robot's performance. A moment of heightened tension is visible in Figure 5(c-6), where the participant vehemently expresses frustration at the robot for selecting the incorrect tool. Finally, Figure 5(c-7) portrays the participant in a state of exasperation, waiting for the robot to hand over the tool, underscoring the emotional toll of compounded delays and inaccuracies in robotic assistance. And the robot recognized the negative emotion effectively.

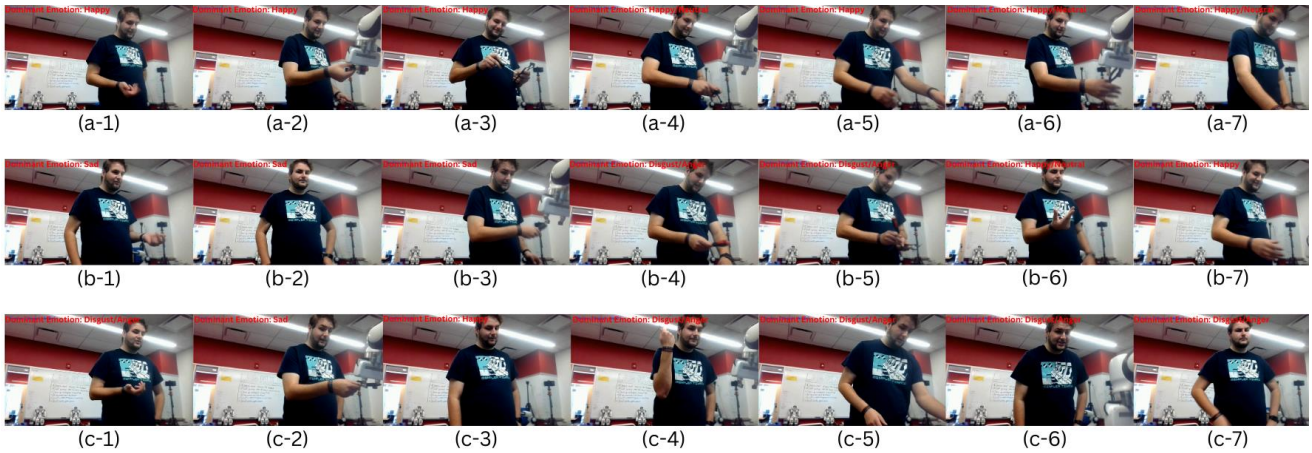

Figure 5: Experimental Video Results. (a) depicts positive reactions to timely and accurate robotic assistance; (b) shows mild frustration from delays; and (c) captures pronounced dissatisfaction from incorrect tool delivery.

C. Comparison Analysis

Our work addresses the nuanced challenges highlighted by Kim et al. [24] and Cai et al. [26] in multimodal emotion recognition. By integrating facial and speech cues, we aim to understand emotional states during real-time human-robot interactions comprehensively. As we build upon the technical prowess of Fayek et al. [13], our study contributes to this growing field by showcasing the operational impact of emotion recognition on collaborative task performance and efficiency. By integrating these insights with our observations on robotic assistance's effect on worker satisfaction, our research underscores the need for robotic systems that are both technically efficient and emotionally intelligent. The challenges we've encountered, particularly the noise in emotional data recognition, shed light on the intricate balance required in system design—balancing accuracy in emotion detection with the unpredictability of human behavior. Addressing this balance is paramount for applying emotion recognition in robotics, a concern that parallels the technical considerations of SER methodologies reviewed by Ko et al. [2]. In conclusion, while applying emotion recognition systems in human-robot collaboration offers promising benefits in improving understanding, engagement, and task efficiency, it also presents significant challenges. Overcoming these hurdles involves enhancing the technical capabilities of such systems and ensuring they are developed and used ethically and responsibly, with a keen awareness of their impact on human emotional well-being.

VI. Conclusion

In conclusion, this paper has developed a multimodal emotion recognition system for manufacturing collaborative robots, leveraging both speech and facial information to enhance stability in complex environments. Three real-world collaborative assembly scenarios were designed to naturally elicit various human emotions and validate the system's effectiveness and performance through these tasks. The experimental results have demonstrated the effectiveness of the developed system. Future work will focus on improving the existing system for emotion recognition in more complex environments.

Acknowledgment

This work is supported by the National Science Foundation under Grant CMMI-2301678 and Grant CNS-2117308.

References

[1] M. K. Chowdary, T. N. Nguyen, and D. J. Hemanth, "Deep learning-based facial emotion recognition for human–computer interaction applications," *Neural Comput Appl*, vol. 35, no. 32, 2023, doi: 10.1007/s00521-021-06012-8.

[2] B. C. Ko, "A brief review of facial emotion recognition based on visual information," *Sensors*, vol. 18, no. 2, p. 401, 2018.

[3] D. K. Jain, P. Shamsolmoali, and P. Sehdev, "Extended deep neural network for facial emotion recognition," *Pattern Recognit Lett*, vol. 120, 2019, doi: 10.1016/j.patrec.2019.01.008.

[4] S. Li, and W. Deng, "Deep facial expression recognition: a survey," *IEEE Trans Affect Comput*, vol. 13, no. 3, 2022, doi: 10.1109/TAFFC.2020.2981446.

[5] C. Pramerdorfer, and M. Kampel, "Facial expression recognition using convolutional neural networks: state of the art," *arXiv preprint arXiv:1612.02903*, 2016.

[6] N. Mehendale, "Facial emotion recognition using convolutional neural networks (FERC)," *SN Appl Sci*, vol. 2, no. 3, 2020, doi: 10.1007/s42452-020-2234-1.

[7] F. Z. Canal, T. R. Müller, J. C. Matias, G. G. Scotton, A. R. de Sa Junior, E. Pozzebon, and A. C. Sobieranski, "A survey on facial emotion recognition techniques: A state-of-the-art literature review," *Inf Sci (N Y)*, vol. 582, 2022, doi: 10.1016/j.ins.2021.10.005.

[8] M. Swain, A. Routray, and P. Kabisatpathy, "Databases, features and classifiers for speech emotion recognition: a review," *Int J Speech Technol*, vol. 21, no. 1, 2018, doi: 10.1007/s10772-018-9491-z.

[9] D. Issa, M. Fatih Demirci, and A. Yazici, "Speech emotion recognition with deep convolutional neural networks," *Biomed Signal Process Control*, vol. 59, 2020, doi: 10.1016/j.bspc.2020.101894.

[10] R. A. Khalil, E. Jones, M. I. Babar, T. Jan, M. H. Zafar, and T. Alhussain, "Speech emotion recognition using deep learning techniques: a review," *IEEE Access*, vol. 7, 2019, doi: 10.1109/ACCESS.2019.2936124.

[11] H. Aouani, and Y. Ben Ayed, "Speech emotion recognition with deep learning," *Procedia Comput Sci*, vol. 176, pp. 251–260, 2020.

[12] R. Jahangir, Y. W. Teh, F. Hanif, and G. Mujtaba, "Deep learning approaches for speech emotion recognition: state of the art and research challenges," *Multimed Tools Appl*, vol. 80, no. 16, 2021, doi: 10.1007/s11042-020-09874-7.

[13] D. Bertero, F. Bin Siddique, C. S. Wu, Y. Wan, R. H. Y. Chan, and P. Fung, "Real-time speech emotion and sentiment recognition for interactive dialogue systems," in *EMNLP 2016 - Conference on Empirical Methods in Natural Language Processing, Proceedings*, 2016. doi: 10.18653/v1/d16-1110.

[14] N. Patel, S. Patel, and S. H. Mankad, "Impact of autoencoder based compact representation on emotion detection from audio," *J Ambient Intell Humaniz Comput*, vol. 13, no. 2, 2022, doi: 10.1007/s12652-021-02979-3.

[15] W. A. Khan, H. ul Qudous, and A. A. Farhan, "Speech emotion recognition using feature fusion: a hybrid approach to deep learning," *Multimed Tools Appl*, 2024, doi: 10.1007/s11042-024-18316-7.

[16] S. Liu, M. Zhang, M. Fang, J. Zhao, K. Hou, and C.-C. Hung, "Speech emotion recognition based on transfer learning from the FaceNet framework," *J Acoust Soc Am*, vol. 149, no. 2, 2021, doi: 10.1121/10.0003530.

[17] L. Cai, J. Dong, and M. Wei, "Multi-modal emotion recognition from speech and facial expression based on deep learning," in *Proceedings - 2020 Chinese Automation Congress, CAC 2020*, 2020. doi: 10.1109/CAC51589.2020.9327178.

[18] C. Guanghui, and Z. Xiaoping, "Multi-modal emotion recognition by fusing correlation features of speech-visual," *IEEE Signal Process Lett*, vol. 28, 2021, doi: 10.1109/LSP.2021.3055755.

[19] M. F. H. Siddiqui, and A. Y. Javaid, "A multimodal facial emotion recognition framework through the fusion of speech with visible and infrared images," *Multimodal Technologies and Interaction*, vol. 4, no. 3, 2020, doi: 10.3390/mti4030046.

[20] J. Heredia, E. Lopes-Silva, Y. Cardinale, J. Diaz-Amado, I. Dongo, W. Graterol, and A. Aguilera, "Adaptive multimodal emotion detection architecture for social robots," *IEEE Access*, vol. 10, 2022, doi: 10.1109/ACCESS.2022.3149214.

[21] M. G. Antonelli, P. Beomonte Zobel, C. Manes, E. Mattei, and N. Stampone, "Emotional intelligence for the decision-making process of trajectories in collaborative robotics," *Machines*, vol. 12, no. 2, 2024, doi: 10.3390/machines12020113.

[22] S. R. Livingstone, and F. A. Russo, "The ryerson audio-visual database of emotional speech and song (ravdess): A dynamic, multimodal set of facial and vocal expressions in north American english," *PLoS One*, vol. 13, no. 5, 2018, doi: 10.1371/journal.pone.0196391.