Toward Empathetic Human–Robot Interaction: A Multimodal Framework Integrating Psychological Profiling and Emotion Recognition*

Alberto Borboni[1,*,†], Luca Ragno[1,*,†] and Fabio Zanoletti[1,†]

[1] *Università degli Studi di Brescia, Department of Mecahnical and Industrial Engineering, Via Branze 38 25123 Brescia Italy*

Abstract

This paper introduces a multimodal framework designed to enhance empathetic communication in human–robot interactions. Our approach integrates psychological profiling—leveraging publicly available social network data and the DISC personality framework—with real-time emotion recognition across facial, audio, and textual modalities. Facial expressions are analyzed using convolutional neural networks and MTCNN-based face detection, while audio signals are processed through Mel-Frequency Cepstral Coefficients and support vector machines. Textual inputs are evaluated via sentiment analysis using advanced language models. These individual emotional assessments are fused through a fuzzy aggregation method, emphasizing non-verbal cues to derive a comprehensive and adaptive emotional profile. The resulting digital agent tailors its verbal responses and interaction style to align with the user's personality traits and current emotional state, offering a more natural and supportive communication experience.

Keywords

Empathetic communication, Human–robot interaction, Psychological profiling, DISC personality framework, Multimodal emotion recognition, Facial expression analysis, Audio emotion recognition, Text sentiment analysis, Fuzzy aggregation, Adaptive dialogue systems

1. Introduction

Empathetic communication is a fundamental aspect of human interaction, fostering social bonding, trust, and cooperation. In recent years, advancements in artificial intelligence and robotics have enabled the development of socially interactive robots that can engage in empathetic communication with humans. Such robots have the potential to revolutionize various domains—including healthcare, education, and social companionship—by providing emotional support and enhancing human well-being [1,2].

In human–robot interaction (HRI), empathy is typically defined as the robot's ability to recognize, interpret, and appropriately respond to human emotions through both verbal and non-verbal cues [3,4]. Prior studies have shown that robots employing empathy-related gestures (such as nodding, gazing, and leaning) can positively influence human–human interactions by improving interpersonal evaluations and increasing emotional support [5,6]. These findings suggest that, rather than replacing human interaction, robots may act as effective mediators in emotionally charged conversations [7,8].

Despite these promising outcomes, significant challenges remain in designing robots that can truly engage in empathetic communication. One key issue is determining the appropriate degree of anthropomorphism required for effective empathy expression. While some research suggests that human-like facial expressions are essential for conveying empathy [9], other studies indicate that empathetic responses can be perceived through contextual verbal communication alone [10]. Moreover, different age groups appear to have varying expectations regarding robotic

[*] martIndustry-2025: 2nd International Conference on Smart Automation & Robotics for Future Industry, April 03-05, 2025, Lviv, Ukraine

1 Corresponding author.

[†] These authors contributed equally.

alberto.borboni@unibs.it (A. Borboni); luca.ragno@unibs.it (L. Ragno); fabio.zanoletti@unibs.it (F. Zanoletti)

0000-0001-7069-1095 (A. Borboni); 0009-0004-4524-3599 (L. Ragno); 0009-0006-6192-5746 (F. Zanoletti)

empathy—older adults tend to prioritize emotional adaptation, whereas younger individuals emphasize functional interaction aspects such as gaze and response timing [11]. Beyond cognitive and expressive factors, a robot's movement and actuation are also crucial in creating a comfortable and engaging interaction for different mechatronic systems [12]. The ability to generate subtle facial expressions and dynamically adjust postural changes, i.e. with micro or smart actuators [13], enhances the perception of empathy. Studies have demonstrated that optimized motion profiles [14, 15] and reduced residual vibrations [16] contribute to a more natural and fluid interaction, ensuring both user comfort and system efficiency. In medical applications, for instance, engagement through empathetic communication has been shown to improve therapeutic outcomes by fostering adherence to rehabilitation programs and increasing patient motivation [17]. Furthermore, recent developments in artificial intelligence and speech synthesis have enhanced robots' ability to generate natural empathetic responses [18-20]. Advances in deep learning models —such as bidirectional LSTMs and HMM-based speech synthesis—have significantly improved the expressiveness of robotic speech, making it more aligned with human expectations [21,22]. Additionally, integrating multimodal emotional prediction systems now allows robots to anticipate user emotions and adjust their communication styles accordingly [23,24]. These improvements contribute to more effective human–robot interactions by increasing the robot's capacity to recognize and respond appropriately to human emotional states [25]. The application of machine learning in empathetic dialogue generation has also seen significant progress, with methods that leverage intention recognition from social network messages to enhance contextual awareness [26]. Similarly, studies of discourse relations and speech synthesis have enabled better comprehension of adversative and emphatic speech patterns, further improving human–robot dialogue interactions [27,28]. Moreover, recent research on emotional enhancement and cognitive adaptation has been instrumental in developing robots that can personalize their interactions to better accommodate individual user needs [29,30]. This paper aims to propose a preliminary engineering approach to facilitate empathetic communication in human–robot emotional interaction. It can enhance human well-being through empathetic verbal responses, and adaptive behaviors fostering meaningful connections.

2. Materials and Methods

This study proposes a method for identifying the human agent interacting with the digital agent using visual data and verbal inquiries. Thereafter, by examining publicly available information online, the digital agent constructs a psychological profile of the human agent. The psychological profile is essential for a more advanced contextualization of communication. The digital agent evaluates visual, auditory, and textual data to comprehensively ascertain the emotional condition of the human agent. The emotional state facilitates the refinement of communication context to enhance empathetic alignment with the human agent. The ultimate stage of contextualization entails employing psychological methods to actualize this empathy in the human agent, thereby reinforcing the empathetic connection. The accurate definition of the context enables the digital agent to completely adjust to the human agent by selecting an appropriate communicative register and enhancing vocal expression to convey the correct emotions, preserving every nuance of meaning and formality within the empathetic communication established by the human agent. The following subsections address the tracking of the time-invariant psychological profile and its applications, the analysis of emotions through facial, vocal, and textual expressions, and the identification of the instantaneous time-variant emotional profile.

2.1. Psychological profiling of the human agent

Traditional methods of personality assessment often require self-reported questionnaires, which can be time-consuming and subject to biases. Advances in natural language processing (NLP) and personality analytics have led to automated profiling systems that infer personality traits from online data. One such tool is Crystal Knows@, a personality prediction engine that utilizes the DISC framework to categorize individuals based on their publicly available professional information.

This study presents a Python-based approach to interfacing with Crystal's API for automated psychological profiling. The script operates by collecting a LinkedIn profile URL extracting the human agent and using it to query the Crystal API, which then returns a structured personality profile based on the DISC classification. To achieve this, the methodology follows a series of steps. First, the human agent provides information to obtain a LinkedIn profile URL along with a valid Crystal API key for authentication. The script then processes an HTTP POST request containing the URL and sends it to the API using the `requests` library. Upon successful retrieval, the API returns a JSON response containing the subject's personality classification, which is then parsed and displayed in a structured format for easy interpretation. To ensure robustness, the script includes error handling mechanisms that display appropriate messages in case of failures, such as invalid API credentials, insufficient data, or server errors.

Crystal's DISC framework categorizes personality into 16 distinct types, each defined by varying degrees of Dominance (D), Influence (I), Steadiness (S), and Conscientiousness (C). These types are represented by specific archetypes that offer insights into individual behaviors and communication styles. The Dominance (D) category includes the Captain (assertive and ambitious), Driver (decisive and persuasive), Initiator (charismatic and resourceful), and Influencer (energetic and adventurous). The Influence (I) category features the Motivator (enthusiastic and outgoing), Encourager (warm and light-hearted), Harmonizer (patient and accommodating), and Counselor (empathetic and supportive). The Steadiness (S) category comprises the Supporter (calm and respectful), Planner (predictable and detail-oriented), Stabilizer (reserved and cautious), and Editor (meticulous and independent). Lastly, the Conscientiousness (C) category includes the Analyst (methodical and private), Skeptic (logical and efficient), Questioner (competitive and strategic), and Architect (strong-willed and purposeful). By understanding these personality types, the digital agent can enhance self-awareness, improve team dynamics, and refine interpersonal relationships.

To effectively adapt to DISC profiles, communication should be tailored to each personality type. When interacting with individuals high in Dominance (D), it is best to be direct, concise, and focused on results, avoiding unnecessary details that may be seen as time-wasting. For those with Influence (I) traits, engaging with enthusiasm, recognizing their contributions, and allowing space for social interaction fosters better communication. Individuals with Steadiness (S) prefer a patient approach, reassurance, and stability, so sudden changes should be minimized to avoid discomfort. Finally, those high in Conscientiousness (C) value detailed and accurate information, requiring time to process and respond thoughtfully.

The psychological profile of the human agent is associated to a reference composed by name, surname and face to store it for future use.

2.2. Human agent identification

The proposed face identification system utilizes the `face_recognition` library to detect and recognize human faces in real-time video streams. Initially, a dataset of known faces is loaded from a predefined directory, where each image is processed to extract facial encodings using deep learning-based feature extraction. During execution, frames from a video source are captured and preprocessed by resizing and converting them to RGB format. The system detects facial landmarks, extracts encodings, and compares them against the stored database using a distance-based similarity metric. If a match is found, the system labels the detected face accordingly; otherwise, it assigns an "Unknown" label and prompts the human agent to provide their name and surname. This information, along with the extracted facial encoding, is then stored in the database for future recognition. The results, including bounding boxes and names, are overlaid onto the video stream and displayed in real time. The system operates continuously, allowing for dynamic face recognition, and can be terminated by the user via a key press.

Leveraging emotional insights in conjunction with a subject's psychological profile, such as one derived from the DISC model, can significantly enhance the effectiveness of your interactions.

2.3. Face emotion recognition

For more detailed instructions, the proposed system is built upon the Facial Expression Recognition (FER) framework, leveraging convolutional neural networks (CNNs) for the classification of emotions from facial images. The implementation follows a structured pipeline comprising image acquisition, face detection, emotion classification, and real-time visualization. A modular architecture is adopted, wherein face detection is performed using the Multi-Task Cascaded Convolutional Networks (MTCNN) detector to ensure robust localization, while the FER model classifies emotions into seven predefined categories: Angry, Disgust, Fear, Happy, Sad, Surprise, and Neutral. Real-time processing is facilitated through the OpenCV library, enabling frame capture via a webcam and subsequent display of annotated results. The system is implemented in Python, employing OpenCV for video acquisition and the FER library for emotion classification. The workflow consists of initializing the FER detector with MTCNN, capturing frames, detecting faces and extracting bounding boxes, applying the FER classifier for emotion prediction, overlaying the detected emotion onto the video frame, and displaying the annotated frame in real time, with termination triggered by user input.

2.4. Audio emotion recognition

The proposed approach consists of three main steps: feature extraction, classification, and prediction, following a supervised learning paradigm in which a machine learning model is trained to recognize emotions based on extracted speech features. The initial phase involves feature extraction from the raw speech signal, utilizing Mel-Frequency Cepstral Coefficients (MFCCs), which are widely adopted in speech processing due to their effectiveness in capturing the perceptual characteristics of human hearing. The extraction process comprises audio preprocessing, where the input audio file is loaded using the Librosa library while maintaining its original sampling rate, followed by the computation of 40 MFCC coefficients from the signal, and statistical aggregation through the calculation of mean MFCC values across time frames to generate a fixed-length feature vector. The extracted features are then used to train a supervised classifier, specifically a Support Vector Machine (SVM) with a radial basis function (RBF) kernel, chosen for its robustness in speech emotion recognition tasks. The approach is evaluated using the Toronto Emotional Speech Set (TESS) dataset, which contains 2800 speech samples spoken by two actresses and classified using the same emotional scale applied in facial emotion analysis, ensuring consistency in multimodal emotion recognition.

2.5. Text emotion recognition

In the proposed system, emotion analysis is conducted by leveraging OpenAI’s ChatGPT-4o API, which processes textual input to generate a statistical evaluation of emotional content. The methodology involves sending a structured request containing the target text to the API, which then analyzes linguistic features, semantic context, and sentiment polarity to classify the underlying emotions. The model provides a probabilistic distribution of detected emotions, allowing for a quantitative assessment rather than a binary classification.

2.6. Global emotional state

The emotional state of the human agent is evaluated in a fuzzy manner as an overlap of the assessments obtained from facial, audio, and text emotional analyses with appropriate weights. To appropriately define these weights, it is assumed that the human agent may attempt to mask their emotional state; therefore, greater weight is given to the communication that is more difficult to mask, namely non-verbal communication, then to the tone of voice, and finally to the text chosen for communication, according to the formula indicated in (1), where **e**(t), $\mathbf{e}_f$(t), $\mathbf{e}_a$(t), and $\mathbf{e}_t$(t) are, respectively, the vectors of the overall emotional state, the one obtained from facial analysis, the one obtained from audio analysis, and the one obtained from the neutral language analysis of the text at time t for the human agent.

$$\boldsymbol{e}(t) = 0.5 \cdot \boldsymbol{e}_f(t) + 0.34 \cdot \boldsymbol{e}_a(t) + 0.16 \cdot \boldsymbol{e}_t(t) \tag{1}$$

The vector of emotional states is composed of seven probability values ranging from 0 to 1 associated with the seven emotions: Fear, Angry, Disgust, Sad, Neutral, Happy, (positive) Surprise, as shown in (2), where the time dependence is omitted for the sake of simplicity. It would be possible to defuzzify the result by identifying a single most probable emotion, but it was preferred to maintain the emotion defined in a fuzzy manner. In general, the first four emotions are considered indicative of negative feedback, while the last three emotions are considered indicative of positive feedback.

$$\begin{aligned}
\boldsymbol{e} &= \langle e_{fear} \quad e_{angry} \quad e_{disgust} \quad e_{sad} \quad e_{neutral} \quad e_{happy} \quad e_{surprise} \rangle \\
\boldsymbol{e}_f &= \langle e_{f,fear} \quad e_{f,angry} \quad e_{f,disgust} \quad e_{f,sad} \quad e_{f,neutral} \quad e_{f,happy} \quad e_{f,surprise} \rangle \\
\boldsymbol{e}_a &= \langle e_{a,fear} \quad e_{a,angry} \quad e_{a,disgust} \quad e_{a,sad} \quad e_{a,neutral} \quad e_{a,happy} \quad e_{a,surprise} \rangle \\
\boldsymbol{e}_t &= \langle e_{t,fear} \quad e_{t,angry} \quad e_{t,disgust} \quad e_{t,sad} \quad e_{t,neutral} \quad e_{t,happy} \quad e_{t,surprise} \rangle
\end{aligned} \tag{2}$$

2.7. Emphatic prompt modifications due to emotional state

The communicative form of the digital agent is guided by the psychological profile of the human agent, but some behavioral changes can be expected to vary with the emotional state of the human agent.

A first category of strategies consists of acknowledging emotions. In particular, two approaches are used: verbal validation and reflective listening [31, 32]. With verbal validation, when the digital agent detects signs of stress or frustration in the emotional state of the human agent (negative emotional state) explicitly acknowledges these feelings. For instance, it might say, “I can see that this situation is really overwhelming for you,” or “It sounds like you’re feeling quite frustrated right now.” This not only validates the human agent’s emotional experience but also creates an opening for them to elaborate if they wish. With reflective language, the digital agent mirrors back what it observes. For example, it can say “It seems like the recent changes have been really challenging,” or “I understand that this topic is difficult for you.” This technique shows that the digital agent is actively listening and empathizing with human agent state of mind.

A second category of strategies consist in modifying tone and pace according to the emotional state of the human agent [33, 34]. If the digital agent notice that human agent is under stress (negative emotional state), it can slow down the speech to induce a calming effect. A slower communication helps ensure that the message is clear and gives the human agent time to process the response. Similarly, if the human agent's tone is tense or rapid, the digital agent might consciously lower its voice, and speak in a measured tone, using softer intonations. If the human agent is expressing distress, a deliberate pause after acknowledging their feelings gives them space to process their thoughts and may encourage further sharing. This also demonstrates that digital agent is not rushing the conversation and is tuned into human agent's emotional needs.

3. Results

3.1. General decision algorithm

The flowchart in Figure 1 illustrates the proposed automated system for personalized human-robot interaction based on facial recognition, emotional analysis, and psychological profiling. The process begins with face detection and recognition; if the person is unidentified, their name and surname are collected, stored in a database, and their LinkedIn profile is retrieved. If a LinkedIn profile exists, the system uses the Crystal API to generate a psychological profile, which is also stored. Subsequently, face and audio inputs are analyzed for emotional content using multiple modalities: acoustic features, textual context, and facial expressions. A fuzzy aggregation method integrates these emotional cues with the psychological profile to determine the user's state. This state is then used to generate a context-aware response by querying ChatGPT, which is converted into a vocal message using a text-to-speech module and delivered to the user. The flowchart outlines a structured pipeline for adaptive and emotionally intelligent human-computer dialogue.

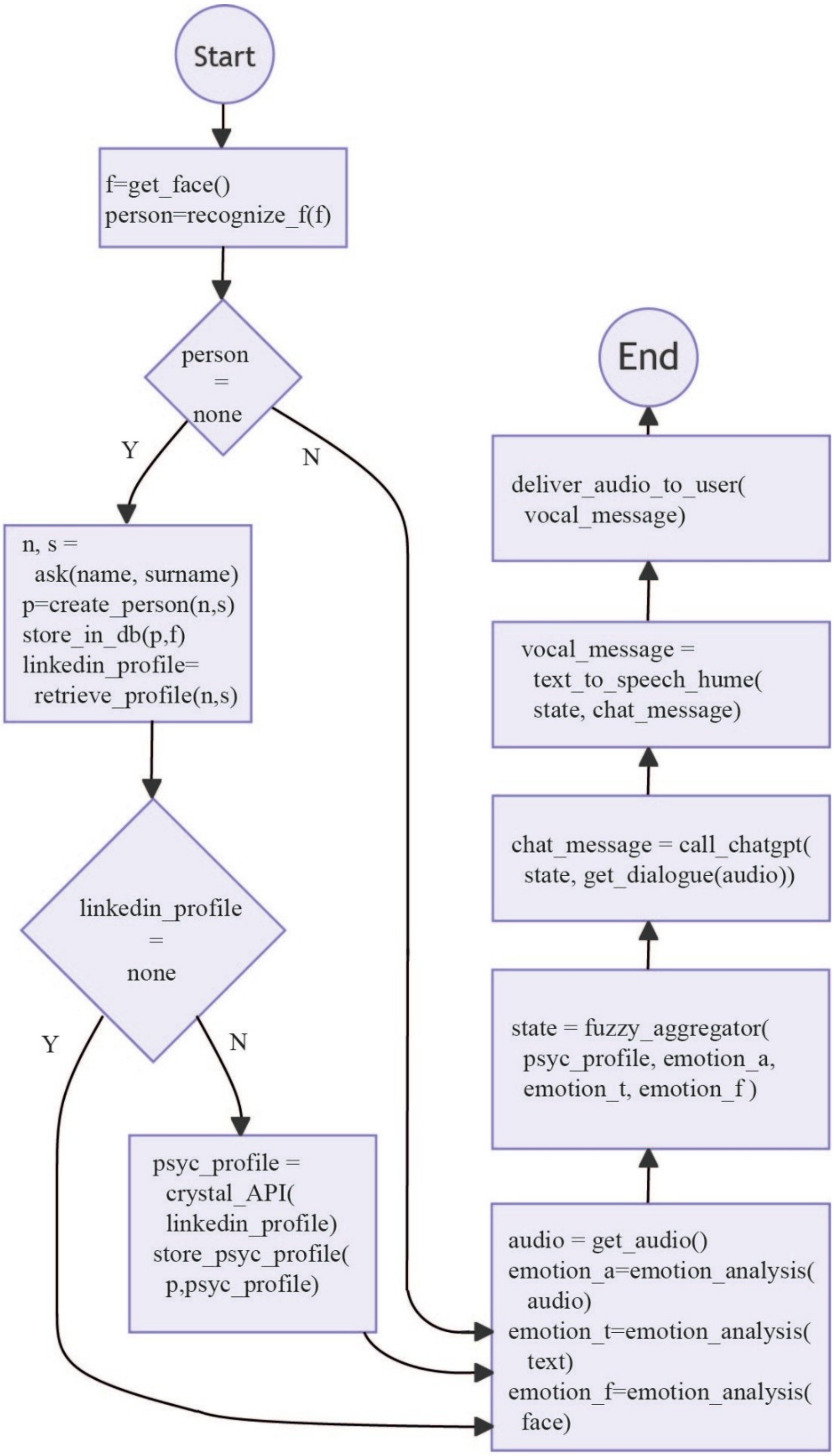
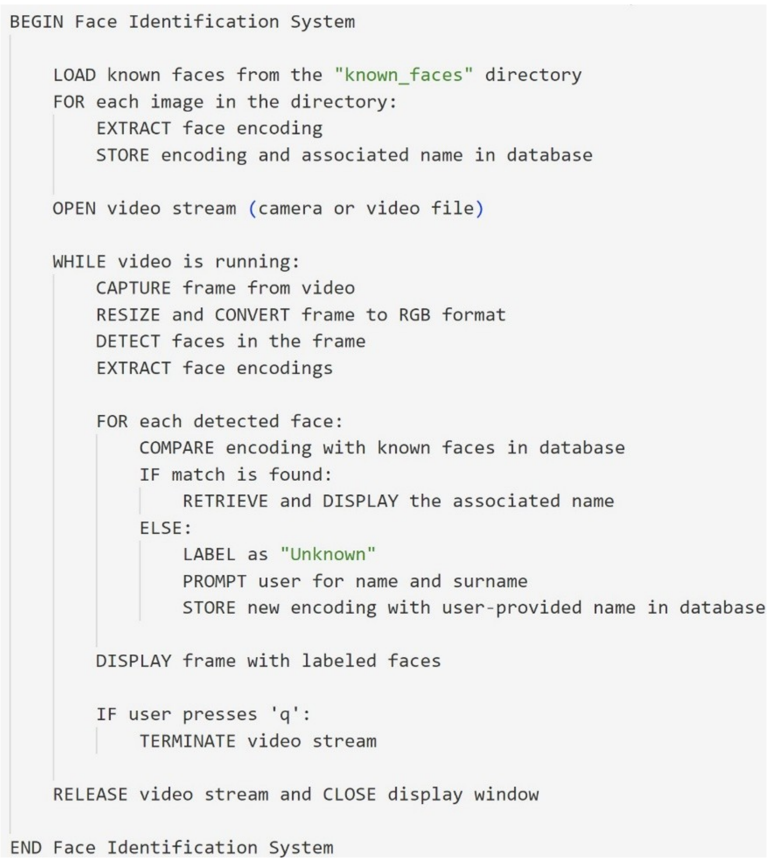

Start
f=get_face()
person=recognize_f(f)
person = none
Y
N
n, s = ask(name, surname)
p=create_person(n,s)
store_in_db(p,f)
linkedin_profile= retrieve_profile(n,s)
linkedin_profile = none
Y
N
psyc_profile = crystal_API(linkedin_profile)
store_psyc_profile(p,psyc_profile)
audio = get_audio()
emotion_a=emotion_analysis(audio)
emotion_t=emotion_analysis(text)
emotion_f=emotion_analysis(face)
state = fuzzy_aggregator(psyc_profile, emotion_a, emotion_t, emotion_f)
chat_message = call_chatgpt(state, get_dialogue(audio))
vocal_message = text_to_speech_hume(state, chat_message)
deliver_audio_to_user(vocal_message)
End

Figure 1: Flowchart of the proposed automated system for emphatic human-robot interaction.

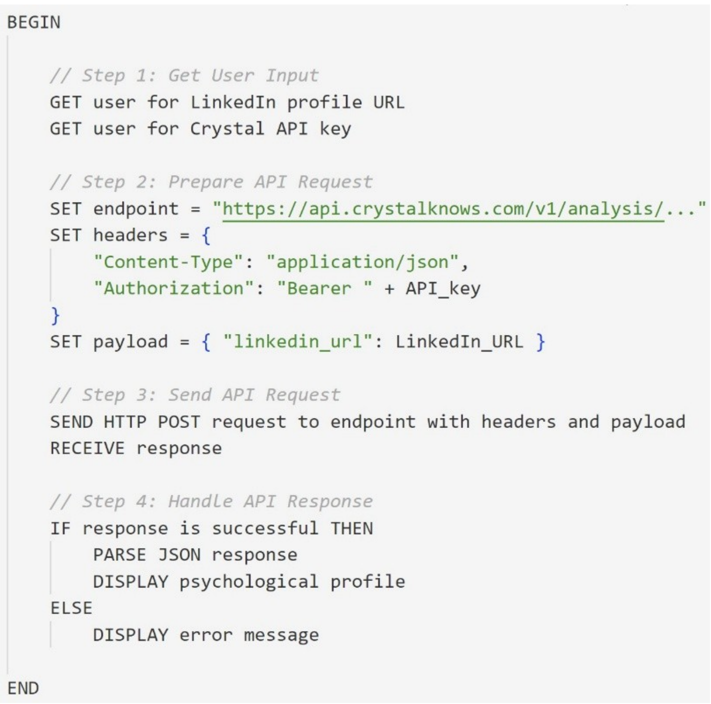
3.2. Pseudocodes of principal modules

Figure 2 presents the pseudocode for face recognition. Figure 3 presents the pseudocode for psychological profile extraction. Figure 4 presents the pseudocode for face emotion recognition. Figure 5 presents the pseudocode for audio emotion recognition. Figure 6 presents the pseudocode for text emotion recognition.



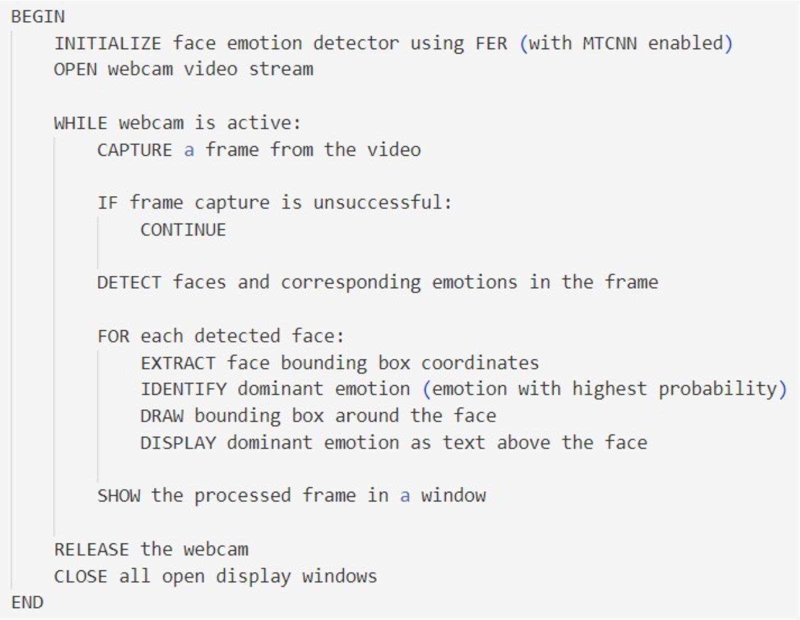
```
BEGIN Face Identification System

    LOAD known faces from the "known_faces" directory
    FOR each image in the directory:
        EXTRACT face encoding
        STORE encoding and associated name in database

    OPEN video stream (camera or video file)

    WHILE video is running:
        CAPTURE frame from video
        RESIZE and CONVERT frame to RGB format
        DETECT faces in the frame
        EXTRACT face encodings

        FOR each detected face:
            COMPARE encoding with known faces in database
            IF match is found:
                RETRIEVE and DISPLAY the associated name
            ELSE:
                LABEL as "Unknown"
                PROMPT user for name and surname
                STORE new encoding with user-provided name in database

        DISPLAY frame with labeled faces

        IF user presses 'q':
            TERMINATE video stream

    RELEASE video stream and CLOSE display window

END Face Identification System
```

Figure 2: Pseudocode for face recognition.



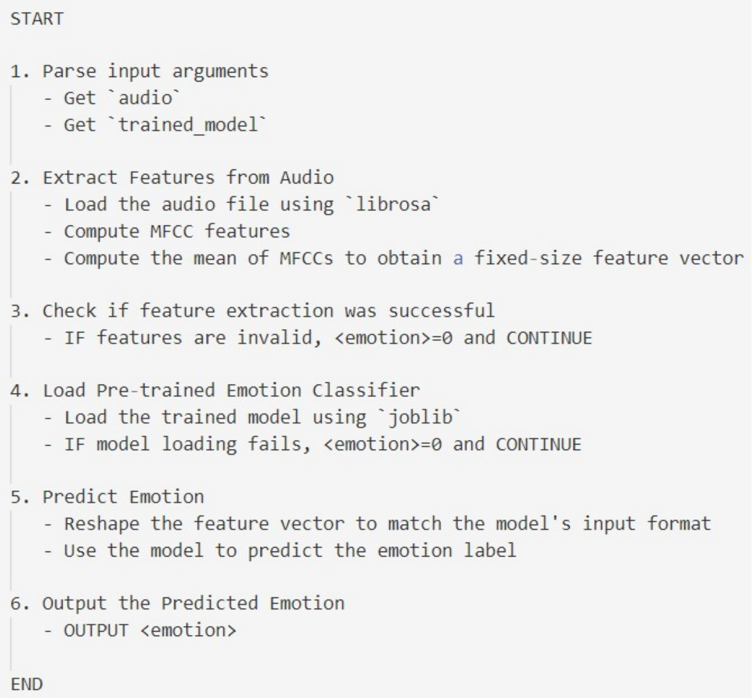
```
BEGIN

    // Step 1: Get User Input
    GET user for LinkedIn profile URL
    GET user for Crystal API key

    // Step 2: Prepare API Request
    SET endpoint = "https://api.crystalknows.com/v1/analysis/..."
    SET headers = {
        "Content-Type": "application/json",
        "Authorization": "Bearer " + API_key
    }
    SET payload = { "linkedin_url": LinkedIn_URL }

    // Step 3: Send API Request
    SEND HTTP POST request to endpoint with headers and payload
    RECEIVE response

    // Step 4: Handle API Response
    IF response is successful THEN
        PARSE JSON response
        DISPLAY psychological profile
    ELSE
        DISPLAY error message

END
```

Figure 3: Pseudocode for psychological profile extraction.



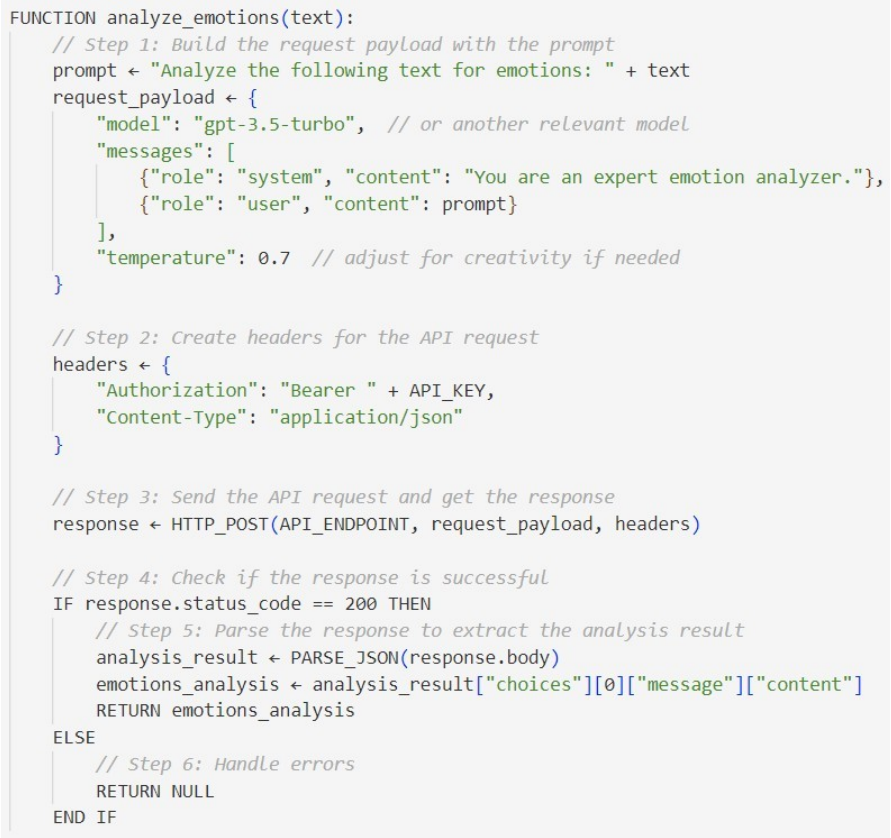
```
BEGIN
    INITIALIZE face emotion detector using FER (with MTCNN enabled)
    OPEN webcam video stream

    WHILE webcam is active:
        CAPTURE a frame from the video

        IF frame capture is unsuccessful:
            CONTINUE

        DETECT faces and corresponding emotions in the frame

        FOR each detected face:
            EXTRACT face bounding box coordinates
            IDENTIFY dominant emotion (emotion with highest probability)
            DRAW bounding box around the face
            DISPLAY dominant emotion as text above the face

        SHOW the processed frame in a window

    RELEASE the webcam
    CLOSE all open display windows
END
```

Figure 4: Pseudocode for face emotion recognition



```
START

1. Parse input arguments
   - Get `audio`
   - Get `trained_model`

2. Extract Features from Audio
   - Load the audio file using `librosa`
   - Compute MFCC features
   - Compute the mean of MFCCs to obtain a fixed-size feature vector

3. Check if feature extraction was successful
   - IF features are invalid, <emotion>=0 and CONTINUE

4. Load Pre-trained Emotion Classifier
   - Load the trained model using `joblib`
   - IF model loading fails, <emotion>=0 and CONTINUE

5. Predict Emotion
   - Reshape the feature vector to match the model's input format
   - Use the model to predict the emotion label

6. Output the Predicted Emotion
   - OUTPUT <emotion>

END
```

Figure 5: Pseudocode for audio emotion recognition



```
FUNCTION analyze_emotions(text):
    // Step 1: Build the request payload with the prompt
    prompt ← "Analyze the following text for emotions: " + text
    request_payload ← {
        "model": "gpt-3.5-turbo",  // or another relevant model
        "messages": [
            {"role": "system", "content": "You are an expert emotion analyzer."},
            {"role": "user", "content": prompt}
        ],
        "temperature": 0.7  // adjust for creativity if needed
    }

    // Step 2: Create headers for the API request
    headers ← {
        "Authorization": "Bearer " + API_KEY,
        "Content-Type": "application/json"
    }

    // Step 3: Send the API request and get the response
    response ← HTTP_POST(API_ENDPOINT, request_payload, headers)

    // Step 4: Check if the response is successful
    IF response.status_code == 200 THEN
        // Step 5: Parse the response to extract the analysis result
        analysis_result ← PARSE_JSON(response.body)
        emotions_analysis ← analysis_result["choices"][0]["message"]["content"]
        RETURN emotions_analysis
    ELSE
        // Step 6: Handle errors
        RETURN NULL
    END IF
```

Figure 6: Pseudocode for text emotion recognition

4. Discussion

Based on the proposed methodology, a digital agent capable of empathetic interaction in the verbal domain with a human agent has been developed. To achieve this result, the human agent is analyzed and decomposed into a time-invariant component during the single interaction, the psychological profile, and a time-variant component during the single interaction, the emotional state. There are works in the literature on profiling through social networks [35]; the present work is based on Crystal Knows@ profiling techniques which rely on the LinkedIn database.

A primary issue may be related to the fact that an individual may not have a LinkedIn profile or there may be multiple profiles associated with the same name and surname. In the first case, the digital agent behaves neutrally, whereas in the second case, it poses additional questions to differentiate between the various profiles.

A second critical issue is related to the fact that individuals present an altered image of themselves on social networks. It would be appropriate to implement a profile-building method based solely on direct interaction; however, this approach might also prove ineffective if the human subject exhibits different psychological aspects when interacting with a digital agent compared to another human agent. On the other hand, there are various other psychological profiling software through social networks, including, for example, IBM Watson Personality Insights, Apply Magic Sauce API, Portrait, or Humantic AI and Kosinski et al [36] demonstrated that accessible digital records of behavior like social networks can be used to accurately predict personal traits in the same way of standard psychological tests with a statistical evidence.

There are several works in the literature on multimodal emotion analysis, both hierarchical [37] and non-hierarchical [38]; the present work also uses various libraries and commercial systems, integrating them hierarchically. An example of an advanced application of this type is Google@'s PaliGemma Mix; which presents an excellent level of maturity compared to the present work.

Instead, no work was found in the literature that simultaneously uses both psychological profiling and emotion analysis to promote empathetic interaction.

As indicated in Figure 1, an empathetic Text-to-Speech (TTS) was used to communicate emotions through an appropriate tone and timbre of voice, speed, cadence, and rhythm. In particular, the Hume.AI service was used, as it is one of the most advanced empathetic TTS systems currently available. This technology has reached a high level of maturity.

This study has several limitations that should be acknowledged. First, the use of LinkedIn and the Crystal API for creating psychological profiles may introduce inaccuracies, as the information provided in these profiles can be selective or distorted. Users often curate their professional profiles to present a particular image, which may not fully reflect their actual personality traits or behavioral tendencies. Future work should consider integrating additional data sources or validation methods to improve the reliability of personality assessments.

Another limitation is the emphasis placed on non-verbal signals in emotional expression. While non-verbal cues play a significant role in human communication, their interpretation varies across cultures and individuals. This variability may lead to inconsistencies in how emotions are perceived and responded to by the system. Further studies should investigate the influence of cultural and personal differences on the effectiveness of non-verbal cues in human-robot interaction. Then the response is slow due to the hardware used and the type of cloud-based implementation: it might be interesting to develop an edge computing application to reduce latency.

Finally, although the system is designed to adapt the tone and pace of speech according to the user's emotional state, the impact of these modifications on interaction quality remains unclear. The study does not provide an evaluation of whether these adjustments enhance engagement or lead to unnatural behavior in the robot. Future research should assess user perceptions and the overall effectiveness of vocal adaptation in improving interaction dynamics.

The effectiveness of the approach in real cases was not explored in this preliminary work, which should be addressed in future developments.

It is also important to note that psychological profiling, although widely used and applicable in various professional and non-professional contexts, presents delicate aspects related to privacy and the decisions of both the subject interacting with the digital agent and those experienced by the subject themselves. To avoid improper behavior by the digital agent, it is possible to refer to various ethical guidelines or advanced regulations valid in certain geographical areas, such as the European General Data Protection Regulation (GDPR) generally for all subjects, but especially for cognitively weaker subjects.

5. Conclusions

In this work, we presented a novel multimodal framework for enhancing empathetic communication in human–robot interaction. By integrating psychological profiling derived from social network data with real-time emotion recognition from facial, audio, and textual inputs, our system offers a comprehensive understanding of a user's emotional state. This integration allows the digital agent to adapt its communication style dynamically, delivering responses that are not only context-aware but also emotionally resonant. Our architecture leverages advanced face recognition techniques, robust audio and text analysis, and a fuzzy aggregation method to synthesize these diverse modalities. The resulting framework has shown promising potential in improving interaction quality across various applications—from healthcare and education to social companionship. By prioritizing non-verbal cues, which are often more difficult to mask, the system effectively tailors its behavior to meet the emotional needs of users. Despite these advances, several challenges remain. Issues such as processing latency, the reliability of social network-based profiling, and the optimization of multimodal fusion require further exploration. Future work will focus on refining these components, incorporating more sophisticated machine learning techniques, and conducting comprehensive user studies to validate the system's effectiveness in real-world scenarios. In conclusion, our framework lays the groundwork for next-generation human–robot interactions that are not only functionally efficient but also emotionally intelligent, paving the way for more natural, supportive, and adaptive digital communication interfaces.

Declaration on Generative AI

The author(s) have not employed any Generative AI tools.

References

[1] M. De Gennaro, E. Krumhuber, G. Lucas, Effectiveness of an Empathic Chatbot in Combating Adverse Effects of Social Exclusion on Mood. Frontiers in Psychology 10, 2020. doi:10.3389/fpsyg.2019.03061.

[2] Y. Gao, Y. Fu, M. Sun, F. Gao, Multi-Modal Hierarchical Empathetic Framework for Social Robots With Affective Body Control. IEEE Transactions on Affective Computing 15, 2024, pp. 1621-1633. doi:10.1109/TAFFC.2024.3356511.

[3] S. Park, M. Whang, Empathy in Human–Robot Interaction: Designing for Social Robots. International Journal of Environmental Research and Public Health 19, 2022. doi:10.3390/ijerph19031889.

[4] R. De Kervenoael, R. Hasan, A. Schwob, E. Goh, Leveraging human-robot interaction in hospitality services: Incorporating the role of perceived value, empathy, and information sharing into visitors' intentions to use social robots. Tourism Management 78, 2020. doi:10.1016/j.tourman.2019.104042.

[5] N. Tuyen, A. Elibol, N. Chong, Learning Bodily Expression of Emotion for Social Robots Through Human Interaction. IEEE Transactions on Cognitive and Developmental Systems, 13, 2020, pp. 16-30. doi:10.1109/TCDS.2020.3005907.

[6] X. Li, Human-robot interaction based on gesture and movement recognition. Signal Process. Image Commun. 81, 2020. doi:10.1016/j.image.2019.115686.

[7] Y. Noguchi, H. Kamide, F. Tanaka, Weight Shift Movements of a Social Mediator Robot Make It Being Recognized as Serious and Suppress Anger, Revenge and Avoidance Motivation of the User. Frontiers in Robotics and AI 9, 2022. doi:10.3389/frobt.2022.790209.

[8] C. Fu, Q. Deng, J. Shen, H. Mahzoon, H. Ishiguro, A Preliminary Study on Realizing Human–Robot Mental Comforting Dialogue via Sharing Experience Emotionally. Sensors (Basel, Switzerland) 22, 2022. doi: 10.3390/s22030991.

[9] J. Chen, C. Guo, R. Xu, K. Zhang, Z. Yang, H. Liu, Toward Children's Empathy Ability Analysis: Joint Facial Expression Recognition and Intensity Estimation Using Label Distribution Learning. IEEE Transactions on Industrial Informatics 18, 2021, pp. 16-25. doi: 10.1109/TII.2021.3075989.

[10] C. Regenbogen, D. Schneider, R. Gur, F. Schneider, U. Habel, T. Kellermann, Multimodal human communication — Targeting facial expressions, speech content and prosody. NeuroImage 60, 2012, pp. 2346-2356. doi:10.1016/j.neuroimage.2012.02.043.

[11] X. Hu, S. Tong, Effects of Robot Animacy and Emotional Expressions on Perspective-Taking Abilities: A Comparative Study across Age Groups. Behavioral Sciences 13, 2023. doi:10.3390/bs13090728.

[12] F. Aggogeri, A. Borboni, R. Faglia, Reliability Roadmap for Mechatronic Systems. Applied Mechanics and Materials, volume 373–375, 2022, pp. 130–133. doi:10.4028/www.scientific.net/amm.373-375.130.

[13] A. Borboni, F. Aggogeri, N. Pellegrini, R. Faglia, Innovative Modular SMA Actuator. In Advanced Materials Research, volume 590, Trans Tech Publications Ltd, 2012, pp. 405–410. doi: 10.4028/www.scientific.net/amr.590.405.

[14] A. Borboni, R. Faglia, Stochastic evaluation and analysis of free vibrations in simply supported piezoelectric bimorphs. Journal of Applied Mechanics, Transactions ASME, volume 80, issue 2, art. no. 021003, 2013. doi:10.1115/1.4007721.

[15] F. Aggogeri, A. Borboni, R. Faglia, A. Merlo, S. de Cristofaro, Precision Positioning Systems: An Overview of the State of Art. Applied Mechanics and Materials, volume 336–338, 2013 pp. 1170–1173. doi:10.4028/www.scientific.net/amm.336-338.1170.

[16] A. Borboni, M. Lancini, Commanded motion optimization to reduce residual vibration. Journal of Vibration and Acoustics, volume 137, issue 3, article no. A1, 2015. doi:10.1115/1.4029575.

[17] C. Sconza, F. Negrini, B. Di Matteo, A. Borboni, G. Boccia, I. Petrikonis, E. Stankevičius, R. Casale, Robot-Assisted Gait Training in Patients with Multiple Sclerosis: A Randomized Controlled Crossover Trial. Medicina, volume 57, issue 7, article no. 713, 2021. doi:10.3390/medicina57070713.

[18] J. James, B. Balamurali, C. Watson, B., MacDonald, Empathetic Speech Synthesis and Testing for Healthcare Robots. International Journal of Social Robotics 13, 2020 pp. 2119-2137. doi:10.1007/s12369-020-00691-4.

[19] F. Efthymiou, C. Hildebrand, Empathy by Design: The Influence of Trembling AI Voices on Prosocial Behavior. IEEE Transactions on Affective Computing, volume 15, 2024, pp. 1253-1263. doi:10.1109/TAFFC.2023.3332742.

[20] Q. Ren, Y. Hou, D. Botteldooren, T. Belpaeme, No More Mumbles: Enhancing Robot Intelligibility Through Speech Adaptation. IEEE Robotics and Automation Letters, 9, 2024, pp. 6162-6169. doi:10.1109/LRA.2024.3401117.

[21] S. Rathor, S. Agrawal, A robust model for domain recognition of acoustic communication using Bidirectional LSTM and deep neural network. Neural Computing and Applications, 33, 2021, pp. 11223 - 11232. doi: 10.1007/s00521-020-05569-0.

[22] Z. Ling, S. Kang, H. A. Zen, M. Schuster, X. Qian, H. Meng, L. Deng, Deep Learning for Acoustic Modeling in Parametric Speech Generation: A systematic review of existing techniques and future trends. IEEE Signal Processing Magazine, 32, 2012, pp 35-52. doi:10.1109/MSP.2014.2359987.

[23] A. Hong, N. Lunscher, T. Hu, Y. Tsuboi, X. Zhang, S. Alves, G. Nejat, B. Benhabib, A Multimodal Emotional Human–Robot Interaction Architecture for Social Robots Engaged in Bidirectional Communication. IEEE Transactions on Cybernetics, 51, 2020, pp. 5954-5968. doi:10.1109/TCYB.2020.2974688.

[24] Z. Liu, F. Pan, M. Wu, W. Cao, L. Chen, J. Xu, R. Zhang, M. Zhou, A multimodal emotional communication based humans-robots interaction system. 2016 35th Chinese Control Conference (CCC), 2016, pp. 6363-6368. doi:10.1109/CHICC.2016.7554357.

[25] T. Applewhite, V. Zhong, R. Dornberger, Novel Bidirectional Multimodal System for Affective Human-Robot Engagement. 2021 IEEE Symposium Series on Computational Intelligence (SSCI), 2021, pp. 1-7. doi:10.1109/SSCI50451.2021.9659935.

[26] G. Singh, M. Firdaus, A. Ekbal, P. Bhattacharyya, EmoInt-Trans: A Multimodal Transformer for Identifying Emotions and Intents in Social Conversations. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 31, 2023, pp. 290-300. doi:10.1109/TASLP.2022.3224287.

[27] J. Crumpton, C. Bethel, C. A Survey of Using Vocal Prosody to Convey Emotion in Robot Speech. International Journal of Social Robotics, 8, 2015, pp. 271 - 285. doi:10.1007/s12369-015-0329-4.

[28] C. Li, X. Zhang, D. Chrysostomou, H. Yang, ToD4IR: A Humanised Task-Oriented Dialogue System for Industrial Robots. IEEE Access, 10, 2022, pp. 91631-91649. doi:10.1109/ACCESS.2022.3202554.

[29] S. Shenoy, Y. Jiang, T. Lynch, L. Manuel, A. Doryab, A Self Learning System for Emotion Awareness and Adaptation in Humanoid Robots. 2022 31st IEEE International Conference on Robot and Human Interactive Communication (RO-MAN), 2022, pp. 912-919. doi:10.1109/RO-MAN53752.2022.9900581.

[30] O. Nocentini, L. Fiorini, G. Acerbi, A. Sorrentino, G. Mancioppi, F. Cavallo, A Survey of Behavioral Models for Social Robots. Robotics, volume 8, article no. 54, 2019. doi:10.3390/ROBOTICS8030054.

[31] E. Kim, C. Kim, Comparative effects of empathic verbal responses: reflection versus validation. Journal of counseling psychology, volume 60, issue 3, 2013, pp. 439-44. doi:10.1037/a0032786.

[32] E. Lavee, G. Itzchakov, Good listening: A key element in establishing quality in qualitative research. Qualitative Research, volume 23, issue 3, 2023, pp. 614-631. doi:10.1177/14687941211039402.

[33] E. Rodero, L. Cores-Sarría, Best Prosody for News: A Psychophysiological Study Comparing a Broadcast to a Narrative Speaking Style. Communication Research, 50(3), 2023, pp. 361-384. doi:10.1177/00936502211059360.

[34] J. Rodd, H. Bosker, M. Ernestus, P. Alday, A. Meyer, T. Bosch, Control of speaking rate is achieved by switching between qualitatively distinct cognitive "gaits": Evidence from simulation. Psychological review, 2019. doi:10.1037/rev0000172.

[35] E. Ishukova; V. Salmanov; A. Kalyabin; A. Antonenko, Approaches to Construct a Psychological Portrait of Users Based on Analysis of Data in Open Profiles of Social Networks: Proceedings - 2019 1st International Conference on Control Systems, Mathematical Modelling, Automation and Energy Efficiency, SUMMA 2019, 2019, pp. 537–539

[36] M. Kosinski, D. Stillwell, T. Graepel, Private traits and attributes are predictable from digital records of human behavior. Proceedings of the National Academy of Sciences of the United States of America, volume 110, issue 15, 2013, pp. 5802-5805. doi:10.1073/pnas.1218772110

[37] C. Li, L. Xie, X. Wang, H. Pan, Z. Wang, A twin disentanglement Transformer Network with Hierarchical-Level Feature Reconstruction for robust multimodal emotion recognition. Expert Systems with Applications, volume 264, article no. 125822, 2025. doi:10.1016/j.eswa.2024.125822.

[38] E. Boitel, A. Mohasseb, E. Haig, MIST: Multimodal emotion recognition using DeBERTa for text, Semi-CNN for speech, ResNet-50 for facial, and 3D-CNN for motion analysis. Expert Systems with Applications, 270, art. no. 126236, 2025. doi:10.1016/j.eswa.2024.126236.