Human-Robot Interaction in Educational and Healthcare Service Robots

Georgi Angelov
Institute of Robotics
Bulgarian Academy of Sciences
Sofia, Bulgaria
george@robotic.bg

Maya Dimitrova
Institute of Robotics
Bulgarian Academy of Sciences
Sofia, Bulgaria
m.dimitrova@ir.bas.bg

***Abstract* — The paper presents an approach to the design of collaborative service robots in education and healthcare, based on the ability of the service robot to predict the positive or negative effects on the people they interact with at the systems design stage, rather than after their ad hoc implementation. The design focuses on 2 main functional characteristics, which represent communication functionalities - human-robot verbal dialogue and detection by the robot of negative/ambiguous emotional reactions of the human. Some aspects of the future implementation are outlined in the paper.**

Service robots, Collaborative robots, Cyber-physical nurse, Functional characteristics, Compassionate presence of the robot

I. Introduction

The idea of the feasibility of the so-called "cyber-physical nurse", which is a technical system (service robot) used not only in a medical, but also in a social context to help "alleviate suffering through compassionate presence" [1] (p. 6) is further developed in the present paper towards its main communication functionalities like human-robot verbal dialogue, on the one hand, and detection by the robot of negative/ambiguous emotional reactions of the human, on the other.

The adopted approach emphasizes the ability of the collaborative service robot to predict the positive or negative effects on the people they interact with at the systems design stage [2, 3], rather than after their ad hoc implementation. In the present paper the commonalities of some recent implementations of dialogue and mood detection in educational and healthcare service robots are outlined and discussed.

Two distinct, yet overlapping, cases of situations, demanding "compassionate presence" on behalf of the robot are outlined. The first one is the situation of special education when children are involved in learning activities with multiple repetitions and attempts to keep focused attention. It is very difficult to maintain positive emotionality during these rehearsals by the child. Social robots have revealed their potential to entertain the child and help implicitly learn new knowledge by preserving the cognitive resource of the child [4, 5, 6]. One cyber-physical system of the kind is presented in [4, 5], where learning is mediated by the mini robot BeBot, which supports the teacher and maintains the attractiveness of the learning situation in the class. The potential of BeBot for entertaining patients in hospitals is discussed in [4].

Section II outlines the main features of an educational scenario for helping the child learn mathematics. The implementation of an inclusive class with robots, performing diverse tasks like detection of attention shift of the child, or expressed negative emotion, is depicted in figure 1.

Fig. 1. *Cyber-Physical Classroom Concept (generated with the help of SORA AI [7]).*

The second situation is a rehabilitation setting where a person has to communicate information about their condition, feelings of discomfort or current need. The main tool for this is the ability of the service robot to understand the utterances of the patient in a safe manner in order to provide an efficient support. Examples are reporting thirst, discomfort, attempts to attract attention, reporting a negative emotion, etc.

Section III presents a voice interaction architecture for a collaborative robot in relation to the verbal communication and activation of some immediate actions towards the patient.

Section IV covers a concept of the described voice human interaction architecture in a healthcare scenario. The Conclusion section outlines possible future research.

II. Learning Mediated by the Mini Robot BeBot

The concept of "compassionate presence" in learning was piloted in an educational experiment carried in the Day center for children and youth with disabilities "Sveta Nedelya" (Sandanski, Bulgaria) [4, 5]. Figure 1 represents the concept of a classroom where teachers and robots pay attention to each individual child to support their learning need.

Fig. 2. *Robot-Assisted Education - Learning Basic Math.*

The robot used in this scenario - BeBot - has the following capabilities:

✓ to do precise movements and rotations of the mobile base;

✓ to speak using a TTS engine that supports 127 languages;

✓ to project images and videos on the wall (the robot head is a high resolution DLP projector);

✓ to express different emotions using specially designed emoticon images and the head-projector.

The robot programming and choreography are done using the software framework, called ERICS part of the OPERA platform.

The educational experiment in general is a robot assisted math lesson, where the children can be taught to sum the numbers from 1 to 10. The robot projects a movie with the lesson, speaks and makes funny moves. The teacher guides the entire process and gives confidence to the student so that the acquired knowledge is retained. The BeBot robot with its projection capabilities captures the interest of the child and makes the teaching process much easier. The teacher has the main and leading role in the teaching process and the robot is assisting the process, greatly improving the attention focus of the child on the educational topic. According to the teachers, involved in the pilot of this scenario, the proposed robot-assistive education enhances the learning of concepts of the child, their attention span and positive mood during the lesson. The received feedback from the children, participating in the study, was strongly positive.

III. Rehabilitation Setting for Communication Information about Health Condition

The core system architecture for voice human-robot interaction (VHRI) in service robots comprises a voice interface pipeline coupled with an emotion detection module. Figure 3 presents an overview of the voice recognition and emotion detection pipeline. The audio input from the user is processed in order to understand what is being said (speech content and intent) and how it is being said (emotional tone), enabling the robot to decide on an appropriate response or action. The software framework is modular, allowing components to be improved or replaced as needed (for instance, swapping in a new speech recognizer or emotion classifier without overhauling the entire system).

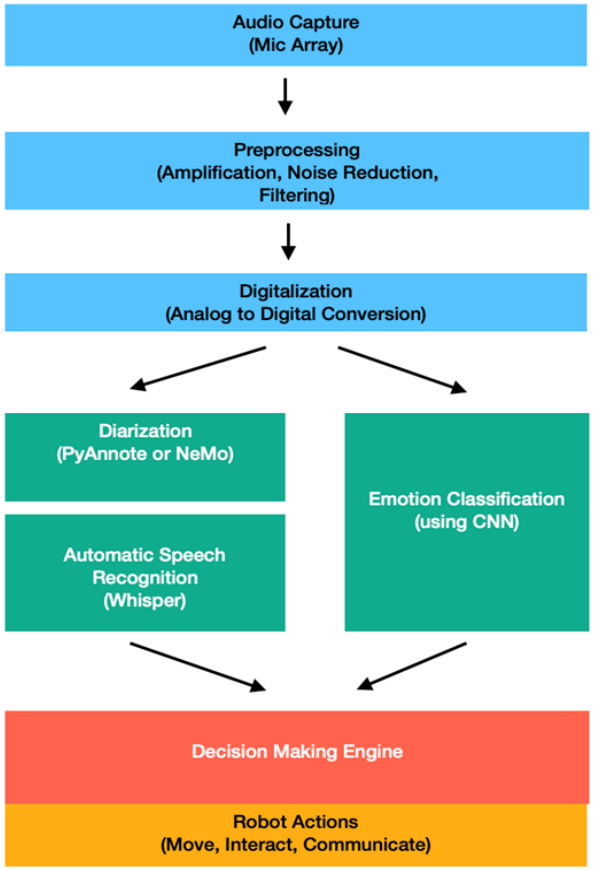

Fig. 3. *Human-Robot Voice Interaction Architecture.*

The key components of the architecture include:

Audio Capture and Preprocessing: Sound from the human is captured via the robot's microphone array. The raw audio signal first undergoes preprocessing steps such as amplification, voice activity detection (to detect when a person is speaking), normalization of volume levels, and noise reduction to filter out background noise. These steps ensure that the subsequent analysis receives a clear and consistent audio input. Techniques like spectral subtraction or adaptive filtering may also be beneficial to suppress ambient noise, which is especially important in busy environments like classrooms or hospitals.

Automatic Speech Recognition (ASR) and Diarization: The cleaned audio is fed into a speech recognition (ASR)

engine to transcribe spoken words into text. In our current setup, this is handled by OpenAI's Whisper model, running locally on Mac Silicon hardware [8]. Whisper is a state-of-the-art transformer-based ASR model with relatively good accuracy even in noisy conditions. Trained on 680,000 hours of multilingual data, Whisper can transcribe speech even in challenging acoustic environments and across multiple languages [8, 9]. This multilingual ability is beneficial in both education and healthcare contexts (e.g., a robot that can understand English, Bulgarian, Spanish, etc., out of the box). Alongside transcription, the system can perform speaker diarization – determining "who spoke when" in the audio (currently for English Language only). Diarization (i.e. creating a diary) is important if multiple people (e.g., a patient and a nurse) are speaking near the robot, so that the robot can attribute statements or requests to the correct individual.

We have experimented with an open-source diarization tool - PyAnnote [10]. PyAnnote provides neural models for speaker segmentation and clustering, effectively separating different speakers in an audio stream. In practice, the audio pipeline uses PyAnnote to label segments by speaker ID, and Whisper then transcribes each segment. The combination yields transcripts with speaker tags, so the robot knows, for example, that Person A said "Where is my medication?" and Person B said "It's on the table." This information can be crucial for context in decision-making. Alternatively NVIDIA's NeMo Framework can be used for diarization. NeMo is a scalable and cloud-native generative AI framework built for researchers [11].

Natural Language Understanding and Decision Module: The transcribed text (along with speaker information and optionally punctuation or sentiment cues from the ASR) is passed to the robot's dialogue and action management module. This module interprets the user's intent and decides on the robot's response. Two approaches can be used here: a rule-based dialogue manager or a large language model (LLM). In simple command-and-control scenarios (especially in healthcare routines), a rule-based system might map specific phrases or keywords to actions – for example, if the text contains "Bring water!" the robot will execute a fetch-water routine. However, for more flexible and natural interaction, the system can leverage an LLM-based AI agent. Modern LLMs (like GPT-style models) can take the user's transcribed input and generate contextually appropriate responses or action plans. For instance, the robot could use an LLM to parse a complex request ("I'm feeling cold, could you close the window and maybe tell me a joke?") into actionable sub-tasks (adjust environment; engage in small talk). The combination of STT→LLM (and subsequently TTS for replying) is a growing trend in voice AI, enabling systems to listen, reason, and respond conversationally. The decision module may also incorporate additional logic to ensure safety and reliability – e.g., critical commands might require confirmation, and any ambiguous input might trigger a clarification question from the robot. The output of this module is a high-level decision: it could be a verbal response, a physical action, or a combination of both.

Emotion Detection Module: In parallel with speech-to-text processing, the architecture includes an emotion recognition pipeline operating on the user's voice. The same audio input (after preprocessing) is analyzed to infer the speaker's emotional state. This module produces an emotion label (such as "happy," "calm," "anxious," "confused," etc.) based on context and vocal intonation and other acoustic features. The emotion detection system can use machine learning models that have been trained on labeled emotional speech data. Generally this involves extracting features from the audio using Mel-Frequency Cepstral Coefficients (MFCCs) or learned representations from a neural network to classify the emotion. The resulting emotion information is fed into the decision module, providing important context for the robot's response. For example, if a patient's words say they are feeling "okay", but the emotion recognizer detects sadness or strain in the tone of voice, the robot can interpret that the patient might actually need help or comfort, prompting a more considerate response. It is a good approach if the emotion detection runs concurrently with speech recognition to minimize latency, and the decision module waits a short moment to receive both the transcribed content and the emotional cues before formulating a response.

Robot Response Generation: Finally, based on the decision module's output, the robot carries out the response. This can involve physical actions (moving, fetching an item, gesturing, etc.) and/or social actions like speaking to the user and expressing an emotion. For spoken responses, the system uses a TTS (text-to-speech) engine to synthesize the reply in a natural-sounding voice.

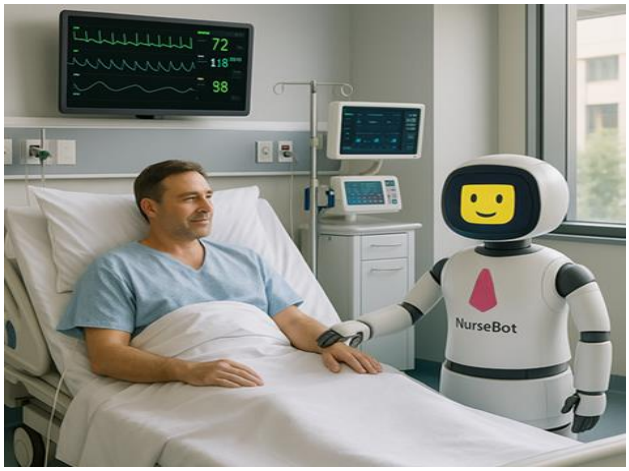

Fig. 4. *Cyber-Physical Nurse Concept (generated with the help of SORA AI [7]).*

In the future, if an improved speech engine is used, the robot might be able to modulate its tone or choose more gentle wording if it detects the user is upset, aligning with a more compassionate bedside manner. The closed-loop system then awaits further input, continually listening for voice commands or dialogue from humans.

In summary, the system architecture enables the robot to listen to what a user says, understand both the content and the emotional subtext, and act appropriately. This creates a more natural interaction flow, much like human conversation, where tone and words together inform how we respond. All the components such as Whisper ASR, PyAnnote diarization and emotion classifier can communicate asynchronously. The modular design allows swapping components - for example, updating the emotion model, without altering the rest of the system, and facilitates debugging.

Example of Emotion Extraction From Audio Data: CNN Classifier Trained on RAVDESS - The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS)

is a popular dataset, containing recordings of actors speaking sentences with various emotions (neutral, calm, happy, sad, angry, fearful, surprise, disgust) [12]. Using this dataset, a convolutional neural network classifies audio into one of these emotion categories. The approach involves converting the audio input into a suitable representation (Mel-spectrogram or a MFCC[1] matrix) and then feeding it into a CNN that learns to discriminate the emotion-related patterns in the voice. CNNs have been widely used in SER (Speech Emotion Recognition) because they can capture spectral-temporal features of audio effectively. In our case, the CNN model has a few convolutional layers for feature extraction followed by dense layers for classification. It achieves high accuracy on the RAVDESS test set - on the order of 70-80% classification accuracy, depending on the emotion, which is comparable to other research results [13]. Such performance is quite good considering human listeners also have trouble distinguishing certain emotions. The advantages of this CNN approach are its simplicity and speed: the model is relatively lightweight and can run in real time on the robot's onboard computer without requiring a GPU. It processes short audio chunks (e.g., 2-3 seconds of speech) and outputs an emotion label almost instantly, which is ideal for responsive interaction. Moreover, training on RAVDESS means the model is tuned for English speech with acted expressions – it recognizes exaggerated cues well (like a very angry tone or a very sad tone). However, there are limitations: since RAVDESS is acted, the model might not generalize perfectly to natural, subtle emotions in real conversations. It might misclassify emotions that are less pronounced or mixed. Also, because it's trained on English phrases, we must be cautious when dealing with other languages or cultural differences in emotional expression (though acoustic features of emotions tend to transfer across languages to some extent). If we needed the robot to detect emotions in another language, we'd ideally retrain or fine-tune the model on data in that language, or at least ensure non-language-specific features (tone, pitch) dominate the classification. Despite these caveats, this CNN (convolutional neural network) approach provides a solid baseline: it's easy to integrate and fast enough for continuous monitoring even on a Raspberry Pi 5 single board computer [13].

IV. A Robot-Assisted Healthcare Scenario

Hospital with Cyber-Physical Nurse Robots: NurseBot (the cyber-physical nurse robot) operates in a hospital ward, assisting human nurses in routine tasks and providing companionship to patients (figure 5).

Let's think about a scene during the morning round as depicted in figure 4: NurseBot enters a patient's room to check on them and greets him, "Good morning, how are you feeling today?" using a pleasant, gentle voice (the robot's TTS is configured to a calm tone appropriate for a healthcare setting). The patient responds, "I'm okay, I guess… just a bit sore." The voice recognition module transcribes this. At the same time, the emotion detector notes a strain in the patient's voice that correlates with discomfort or sadness.

NurseBot's decision module interprets that the patient might be downplaying his pain or mood. Instead of just logging vitals, the robot decides to show concern. It replies, "I'm sorry to hear that you're feeling sore. Is there anything I can do to help ease your discomfort?" This empathetic response is made possible by the emotion-aware system recognizing that "I'm okay" didn't sound genuinely okay. The patient might then admit, "Well, my back hurts quite a bit this morning." NurseBot can follow up with an offer: "Would you like me to alert the nurse to bring your pain medication, or perhaps adjust your pillows?" In this way, the robot facilitates communication between the patient and the healthcare staff, ensuring needs are met promptly

Fig. 5. *A Nurse, Assisted by NurseBot, BeBot and NAO (generated with the help of SORA AI [7]).*

Another scenario is during the robot's autonomous rounds: NurseBot moves through the hallway delivering items or doing room checks. It might encounter a nurse who calls out, "NurseBot, bring this file to Room 12 when you get a chance." Using voice recognition, it transcribes and confirms the task. If the nurse's tone has sounded urgent or stressed (perhaps it has detected urgency in the voice), NurseBot can prioritize that task sooner.

If the robot meets a patient who is anxious (imagine a patient in isolation who hasn't seen many people), and the patient engages it in conversation, the emotion detection subsystem might sense *loneliness* or *anxiety*. NurseBot could then spend a few extra minutes talking with that patient (within its permitted scope), offering comforting words or even playing soft music – acting as a *social companion*. In doing so, it provides not just medical assistance but also compassion and emotional support. This is crucial in healthcare, where emotional well-being significantly impacts recovery.

In a multi-person interaction, say if a doctor, a nurse, and a patient are discussing in the room while NurseBot is present, the robot uses diarization to follow the conversation. If the patient becomes confused about the medical information (detected via a hesitant or worried voice asking a question), NurseBot can later provide a simplified explanation or notify the nurse to clarify, ensuring the patient's understanding. All these interactions are made smoother by the robot's ability to understand natural language voice commands (no one has to fiddle with a touchscreen or a sophisticated user interface) and to detect unspoken cues of emotion. Nurses often say a kind word or show concern naturally; NurseBot, through its programming,

[1] Mel-frequency cepstral coefficients

may try to emulate some of that by detecting cues and responding with pre-programmed empathetic phrases or actions. It is by no means a replacement for human empathy, but it augments the care team by extending the reach of compassionate monitoring – for example, continuously listening for signs of patient distress when nurses are not immediately present. If it detects a patient crying or calling out with fear in their voice at night, it can automatically alert staff and go to that room to say "Help is on the way" thereby comforting the patient until a nurse arrives. These scenarios show that voice and emotion recognition significantly enhance the robot's *compassionate presence*.

In education, the robot becomes an engaging, responsive aide that can motivate and comfort students. In healthcare, the robot transforms from a simple delivery or measuring device into a pseudo-companion that can attend to patients' emotional needs in between the busy schedules of human caregivers. Importantly, the robots perform these roles without needing complex interfaces – speech is the primary mode of interaction, which is the most natural human modality of communication [14].

The frequency of human-robot interaction can be modulated according to the preferences of the patient. In a study of the effect of a humanoid robot teacher on the degree of student confidence in class, two types of students were identified – socially-oriented and socially-indifferent [15]. The socially-oriented students preferred the presence of a human during the lesson performed by a humanoid robot NAO (as in figure 5), whereas the socially-indifferent students did not mind the presence of the robot only during the lesson (as in figure 4). Therefore, by accounting for the preferences of the patients, the frequency and distribution of human nurse and robot tasks can be optimized.

By adjusting its responses to the personality as well as to the emotional state of the human, the cyber-physical nurse NurseBot interacts in a way that feels more human-like, promoting trust and likability. This ensures long-term acceptance of such robots in everyday environments.

V. Conclusion

Human-robot interaction in the realms of education and healthcare will greatly benefit from advances in voice recognition and emotion detection. In this paper, we discuss how educational service robots (exemplified by the BeBot and OPERA platform) and the cyber-physical nurse robot concept (NurseBot) can engage in natural, intuitive interactions by understanding human speech and emotions. We presented a system architecture that integrates state-of-the-art speech-to-text transcription with speaker diarization and a decision-making engine. Parallelly, an emotion recognition module analyzes vocal cues to provide the robot with awareness of the user's emotional state. This combination enables the robot to not only hear the words being spoken, but also to feel the tone behind them, leading to richer and more context-aware responses. We also explored the importance of emotion recognition, highlighting that modern transformer models offer strong performance at the cost of higher complexity. By implementing these components, the robots are able to maintain an interactive, compassionate presence – BeBot supports teachers by engaging students with responsive dialogue and emotional feedback, while NurseBot assists healthcare staff by communicating with patients in an understanding manner.

ACKNOWLEDGMENT

The Authors Acknowledge The Financial Support Of The Project With Administrative Contract № KP-06-H57/8 From 16.11.2021. "Methodology For Determining The Functional Parameters Of A Mobile Collaborative Service Robot Assistant In Healthcare", Funded By The "Competition For Funding Basic Research - 2021." From The Research Sciences Fund, Bulgaria.

References

[1] https://www.isu.edu/media/libraries/school-of-nursing/program-pdfs/bsn-completion-program/UG-Student-Handbook-2024-25.pdf

[2] Shaikh, T. A., Rasool, T., & Verma, P. (2023). Machine intelligence and medical cyber-physical system architectures for smart healthcare: Taxonomy, challenges, opportunities, and possible solutions. Artificial Intelligence in Medicine, 146, 102692.

[3] Dimitrova, M., Valchkova, N.. Feasibility of the Cyber-Physical Nurse (2025). Proceedings of Tenth International Congress on Information and Communication Technology ICICT 2025, Lecture Notes in Networks and Systems, London, Volume 1, In: Xin-She Yang, R., Sherratt, S., Dey, N., Joshi, A. (Eds.) Springer Nature Singapore (in print).

[4] Angelov, G. (2025). Modern Applied Service Robotics. ISBN 978-619-93266-0-2, ROBOTIC.

[5] Dimitrova, M., Bogdanova, G., Noev, N., Sabev, N., Angelov, G., Paunski, Y., ... & Krastev, A. (2023). Digital accessibility for people with special needs: Conceptual models and innovative ecosystems. In 2023 8th International Conference on Smart and Sustainable Technologies (SpliTech) (pp. 1-5). IEEE.

[6] Dimitrova, M., Wagatsuma, H., Tripathi, G. N., & Ai, G. (2015). Adaptive and intuitive interactions with socially-competent pedagogical assistant robots. In 2015 International Conference on Information Technology Based Higher Education and Training (ITHET) (pp. 1-6). IEEE.

[7] Sora | OpenAI. openai.com. December 9, 2024.

[8] Introducing Whisper. https://openai.com/index/whisper /, September 21, 2022

[9] Gerganov G. (2024). A CPP Whisper Implementation. https://github.com/ggerganov/whisper.cpp

[10]. https://github.com/pyannote/pyannote-audio/blob/develop/FAQ.md

[11]. https://docs.nvidia.com/nemo-framework/user-guide/latest/nemotoolkit/starthere/intro.html

[12]. Livingstone, S., Russo, F. (2018). The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS), https://doi.org/10.5281/zenodo.1188976

[13]. Mountzouris, K., Perikos, I., Hatzilygeroudis, I., (2023). Speech Emotion Recognition Using Convolutional Neural Networks with Attention Mechanism. Electronics, https://doi.org/10.3390/electronics12204376

[14] Angelov, G., Paunski, Y. (2024). Voice Controlled Interface for ROS Service Robot in Healthcare. Complex Control Systems, Vol. 8, ISSN 2603-4697 (Online), http://ir.bas.bg/ccs/2024/08/3.pdf

Dimitrova, M., Wagatsuma, H., Tripathi, G. N., & Ai, G. (2019). Learner attitudes towards humanoid robot tutoring systems: Measuring of cognitive and social motivation influences. In: Dimitrova & H. Wagatsuma (Eds.) Cyber-physical systems for social applications (pp. 1-24). IGI Global.