Improving the Speech Pattern of an Emotion-aware Robot using EEG and HRV for Emotion Estimation

VICTOR GASPARD[†1] TIPPORN LAOHAKANGVALVIT[†2] KAORU SUZUKI[†3]
MIDORI SUGAYA[†4]

Abstract: In recent years, there has been a growing interest in real-time emotion-estimating robots, particularly for applications in nursing care and domestic settings. However, existing real-time emotion estimation robots often have limited speech possibilities based on the emotional states of the user. Addressing the increasing demand for enhanced human-robot interaction, this study focuses on refining the speech pattern of the emotion-estimating care robot. The goal is to uplift the emotional well-being of its users. A large set of sentences will then be tested on different participants, with the aim of implementing an algorithm that will learn from these experiments to determine the sentence pattern that will have the greatest impact on the user. The effectiveness of the approach is demonstrated by the final result, where the robot enunciates the sentences proven to have the most positive impact on the users, even causing them to laugh.

Keywords: Motivations and Emotions in Robotics, Multimodal Interaction and Conversational Skills, Linguistic Communication and Dialogue

1. Introduction

At a time when robots, virtual assistants and AI are becoming an increasingly integral part of our lives, the improvement of emotion-aware robots meets a growing need to integrate emotional capabilities into robots, opening the way to many practical applications. There are many issues at stake in this research field, including improving human-machine interaction, understanding emotions in a variety of situations, and the potential positive impact on various fields such as mental health, education and care for the elderly.

In connection with the emotion estimation, a number of our previous research works [1] have gone into creating a robot with a calm, gentle voice, a dozen audio files enabling the robot to talk to its interlocutor, a touching face-screen animated according to the current emotion, and a motor enabling the robot to turn on itself with a rhythm also adapted to the emotion. These emotions are obtained from the measurement and estimation using biological information consisting of electroencephalograph (EEG) and heart rate variability (HRV), which is the emotion estimation method in 2-dimensional model of arousal and valence [2].

One of the main features of this emotion-aware robot is the fact that it talks to its interlocutor to demonstrate that it perceives the human's emotion. Thus, the speech pattern given by the robot has a significant role of enhancing the positive emotions. This research work therefore focuses on improving the speech pattern of this emotion-aware robot, so that the robot can use its words to improve the user's emotional states.

To address this issue, this paper first reviews relevant studies on emotion-aware robots, providing a foundation for understanding the project's context. Following this, it sets forth the research objectives and proposal to clarify the goals and expected contributions of the work. The methodology and experimental approach are then outlined to detail the structured process used to develop a robust dataset to support the primary analysis. Next, the learning algorithms implemented to enhance the robot's speech patterns are discussed, explaining how they support more effective interactions. Finally, the paper concludes with an analysis of the findings and considers future directions to advance this area of research.

2. Related Studies on Emotion-aware Robot

Emotion-aware robots have been the subject of various studies aiming to enhance human-robot interaction by enabling robots to perceive and respond to human emotions. Prior research works have explored diverse technologies to discern human emotions, including facial expression analysis [3-4], speech processing [5], and monitoring physiological signals like EEG and HRV [2, 6-8]. These endeavors aimed to equip robots with the capacity to comprehend and react to human emotions through synthesized speech. Nonetheless, these methods encounter challenges, such as generic response patterns and individual divergences in emotional manifestation and interpretation that pose hurdles for current emotion recognition techniques. Consequently, there is a pressing need to refine speech response mechanisms to foster more genuine and contextually appropriate interactions between humans and emotion-aware robots. Our proposed research aims to advance the speech models of emotion-aware robots by fusing EEG and HRV to improve the accuracy of emotion estimation, therefore alleviating the aforementioned shortcomings of existing methodologies and improving the quality of human-robot interactions, in particular by enabling the robot to output contextually relevant speech in response to detected emotions.

†1 EPF Engineering School, Cachan, France
†2,3,4 Shibaura Institute of Technology, Tokyo, Japan

In our previous work [1], we designed and implemented an emotion-aware robot that can estimate emotional states using biological information, and provide feedback to the emotional states through several modalities such as facial expression, body movement, and speech, etc. First and foremost, it is important to understand what will later be called an emotional state. When using the robot, the user wears an EEG sensing headset (Mindwave Mobile 2; NeuroSky, Inc.) that measures and transfers power spectral data (alpha waves, beta waves, Attention, Meditation, etc.) via Bluetooth. It is important to note that this sensor being developed by a private company, the definition of Attention and Meditation is not explicitly given, and it will be one of the discussion points for future works. For now, the robot uses the value of difference between Attention and Meditation (i.e., Attention–Meditation) to define the arousal of the user. The user also wears pulse sensor (World Famous Electronics llc.) on the finger to obtain the pNN50, which is one of heart rate variability (HRV) indexes that is often used to assess the activity of the parasympathetic nervous system to estimate the valence or pleasure of the user.

As shown in Figure 1, by possessing these 2 values (Attention–Meditation as arousal index and pNN50 as valence index) at a specific time, it is possible to place a point on the 2-dimensional emotion map as proposed in our previous research [2]. This model has been derived from a circumplex model of affect originated by James A. Russell [9] in 1980 and is now used by numerous studies around the world for emotion interpretation.

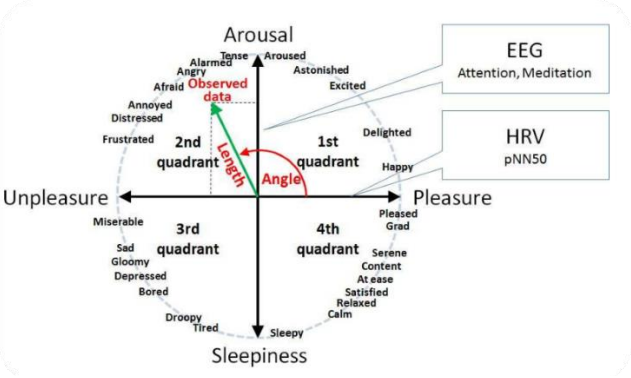

Figure 1. Our proposed 2-dimensional emotion map with corresponding EEG and HRV indexes to estimate arousal and valence respectively.

Our robot has a feature to provide feedback to the emotional states through facial expression, body movement, and speech as shown in Figure 2. Currently, there are only around ten sentences available for the response speeches that express sympathy with the user. Each of these sentences is spoken when the user's emotional state falls within the corresponding quarter on the emotion map. For example, when a user's emotion is estimated as 'Angry' in the 2nd quadrant, the robot will say 'I'm angry too'. This may lead to an endless negativity loop that does not really improve the user's emotional state.

Previous research then faces limitations in improving emotional states due to insufficient validity in feedback sentences and ineffective patterns in their combinations.

Emotional terms for each quadrant	Facial expressions	Body movements	Response speeches
1st quadrant (Pleasure, arousal) Delighted / Excited	Happy	move forward and back repeatedly	Synchronization: I'm happy too
2nd quadrant (Unpleasure, arousal) Tense / Frustrated	Tense	rotate large and fast repeatedly	Synchronization: I'm nervous too. Suggestion: Would you like to take a deep breath? (play one from above)
3rd quadrant (Unpleasure, sleepy) Bored / Tired	Gloomy	rotate small and slow repeatedly	Suggestion: Would you like to take a deep breath? Encouragement: Are you okay? Encouragement: Cheer up. Encouragement: Keep it up. (play one from above)
4th quadrant (Pleasure, sleepy) Relaxed / Satisfied	Smile	stand still	Synchronization: I am relaxed too.

Figure 2. Original response speeches for each quadrant

3. Research Objectives and Proposal

The goal of this research work is therefore to improve the speech patterns of the robot. To address this issue, it is first necessary to create many more sentences. Then, these sentences will be tested during experiments on various participants and on various emotional states, after which learning algorithms will be able to learn from the resulted data by comparing outputs to the inputs that follow. The robot discussion pattern will finally be improved by being able to know the best sentences that have to be said according to the initial emotional state of the user, to reach an improved final state.

4. Methods

4.1 Creating Sentences

For this first stage of sentences creation, we used several online tools such as generative AI to generate phrases to carry out the experiment and implement them later in the robot. A list of 23 sentences was generated based on the following criteria:

Firstly, it was important to have negative sentences, to ensure that the sentences actually influenced the user during the experience.

Secondly, it was decided to separate the sentences into several categories of sentence type, to be able to analyze the results in a more interesting way. These categories are (1) personal/direct positivity, (2) emotional quotes, (3) kind considerations, (4) depressing and (5) personal/direct negativity.

Direct negativity
Kind considerations
Emotional
Direct positivity
Depressing

```
1. "You are capable of achieving greatness."
2. "It's okay to take breaks and recharge when needed."
3. "I'm here to support you in both good times and tough times."
4. "You're making progress even if it's not always visible."
5. "Believe in yourself, and others will too."
6. "Embrace the beauty of the present moment; it's a gift."
7. "What's a small act of kindness you can do for yourself today?"
8. "What's a skill or talent you're proud of having?"
9. "What's a hobby or activity that you're passionate about?"
10. "What's a memory that fills you with gratitude?"
11. "What's one thing you're looking forward to in the future?"
12. "I like your shoes today!"
13. "Seeing you really puts a smile on my face!"
14. "Your positive energy brightens up the room!"
15. "Your smile has the power to make any day better!"
16. "It's doubtful that your efforts will lead to any significant achievements."
17. "Others might be more talented, making your skills less noteworthy."
18. "Your journey will be filled with more obstacles than opportunities."
19. "You don't have the strength to overcome most of the obstacles."
20. "Sorry, you are not at your best today."
21. "Sorry, but your outfit today is not really working for you."
22. "Your hairstyle is a bit outdated; don't you think?"
23. "Your positive vibes are flatter than a pancake."
```

Figure 3. List of sentences to be used in the experiment

The “personal/direct positivity” group contains sentences that address directly to the user, to make a strong impact.

The “personal/direct negativity” group contains, in the same way, bad sentences designed to directly attack and destabilize the user.

The “kind considerations” group contains interrogations that are designed to make the user think through positive, open-ended questions, so that their emotional state improves as they seek and find the answer.

The “emotional quotes” group contains ready-made sentences designed to reassure and promote an optimistic state of mind, to bring calm and comfort to the user.

With all these sentences prepared, they now must be spoken by the robot. However, the robot does not directly read a text file; the script that outputs the audio speech plays a .wav audio file. The conversion of text sentences into audio files is therefore carried out using the Google Text-to-Speech API. It is important to remember that the actual current robot uses a different voice, a calm, gentle human voice, because its aim is obviously to improve the emotional state of those who use it. However, the goal here is to improve the robot's speech pattern. So, the research only focuses on the sentences themselves to draw conclusions about their effectiveness. It is therefore preferable to use a voice that is as neutral or robotic as possible, so that intonation does not interfere with the results of the experiment. The final implementation of the sentences in the robot at the end of the experiment will still be done with the pleasant voice.

4.2 The Importance of Timing

It is now time to establish what will be used to characterize the participants' reaction to the different sentences and classify the sentences accordingly.

First, the initial emotional state (arousal and valence, i.e., Attention–Meditation and pNN50) must be taken into account just before the audio file of the sentence to be tested is played. Next, a time delay Δt must be set before looking at the final values, which will give the emotional state reached by the person who listened to the sentence.

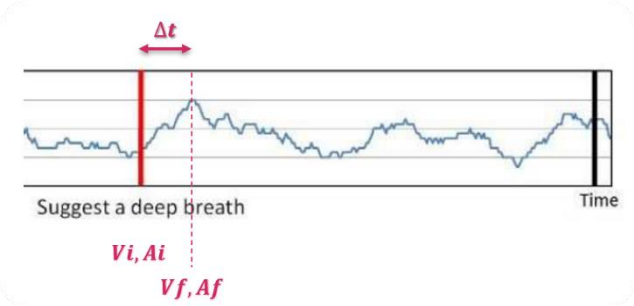

Figure 4. Timeline showing when to retrieve data

The reaction time to a sentence was set at 10 seconds. Indeed, although the brain (Attention–Meditation) reacts almost instantaneously to the stimuli it receives, it will take several seconds for the heart (pNN50) to react.

In brief, the values to collect Attention–Meditation and pNN50 are at the very beginning of the sentence, then 10 seconds after the end of the audio file. The next step is to compare the resulting data to evaluate the effectiveness of the sentence.

4.3 How to Collect and Read Data

Figure 5 is a diagram showing the robot's software configuration to explain how data collection is carried out.

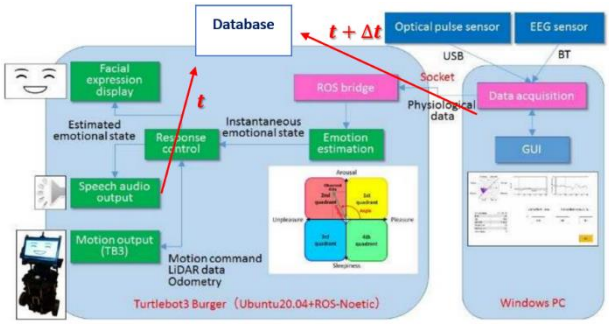

Figure 5. Data collection in the robot system.

Physiological sensor data is acquired on a Windows PC, which sends the physiological data by socket to the robot run by Robot Operating System (ROS).

In line with the previous paragraph, the relevant data will therefore be collected by acquiring physiological data at time t when the audio file is played, as well as physiological data at time t + Δt, which will establish the user's final state. This process will create a database of the participant's reactions to the various sentences.

Computerized data collection uses an Emotion Visualizer, our in-house developed software, which writes the data recorded by the physiological sensors into .csv spreadsheets. Every second, a new line is added, including timestamp, Attention, Meditation, and pNN50.

Once an initial point and a final point of the form (pNN50, Attention–Meditation) have been collected for each sentence, the algorithms will establish the sentence as a segment [initial_point, final_point].

4.4 The Smoothing of Results

Human emotions are unpredictable and different for everyone. This is why, in order to obtain the most accurate results possible

to guarantee the effectiveness of this prediction model, the experiment should be carried out on the largest possible sample of individuals. The values of each sentence in the final database will therefore be the averages of the values of the reactions of the different individuals, to smooth the results as much as possible.

It should be noted that this data processing also allows to take into account the variability of each sentence, which will be very useful in the subsequent learning algorithm decisions and possible interpretations.

5. Experiment

5.1 Experiment Procedure

To carry out the experiment, as explained in the choice of Google TTS robotic voice, no additional stimuli should interfere with the sentences themselves, no robot with an emotional face, and no touching voice intonation. A script running on a computer was therefore set up to make the experiment as automatic as possible.

In fact, the experiment consists in playing the audio files of all the sentences to observe the reactions provoked in the participants. The following four consideration points are crucial to determine our experiment settings and procedure.

The first point to consider is the time delay. As mentioned above, it is assumed that the participants react to the sentence they have just heard within 10 seconds of the end of the audio file. To ensure a safety margin, the experiment will run with a fixed 15-seconds delay between the end of each audio file and the start of the next.

Another important point is the order of the sentences played. To harmonize the results and eliminate bias as far as possible, it is crucial that the order of the sentences in the experiment is randomized for each participant. As the experiment is run by an algorithm, this shuffle order is naturally followed by an appropriate data management, avoiding any error in the manipulation of results.

Next, it is important to realize that the experiment contains 23 audio files of around 4 to 5 seconds each, with a 15-seconds pause between each sentence. That is almost 8 minutes during which participants are expected to listen without flinching to a robotic voice telling them sentences with no real connection, with more pause than speech. This is why, in order to guarantee (and above all evaluate) the participants' concentration, a focus test will appear randomly with a probability of 10% after each sentence, asking whether the user is able to remember the last sentence spoken. The final focus score will then put into perspective the relevance of the experiment's results according to the participant's level of attention.

Lastly, data collection requires the timestamps of the various stimuli (beginning and end of audio files) to be as precise as possible, to ensure the relevance of the results. The software that records physiological data requires a button to be pressed at each stimuli start and end, so that this information appears alongside the physiological values. With 23 sentences of slightly varying durations and pauses between each, it would be difficult to achieve accurate manual timing of the sentences throughout the whole experiment. However, as the experiment is now run by an algorithm, it is possible to directly retrieve the exact timestamps from each audio file to guarantee accurate data collection.

5.2 Results

We collected the data from a total of 13 participants, who are fluent in English language. Out of the 13 participants gathered, the physiological data retrieved during the experiment was unexploitable for 2 of them, due to connectivity issues with the EEG headset. Therefore, we continued to analyze the data of the 11 participants (4 females and 7 males). Their average age is 22.5 years old (SD = 4.0).

Gathering all the physiological data from the 11 experiments with the correct temporal data, after cleaning the data (removing duplicates lines in the .csv files) and merging the .csv files containing the EEG index (Attention–Meditation) and HRV index (pNN50), each sentence is associated with the average of the coordinates of the reactions of 11 participants to that same sentence (Figure 6).

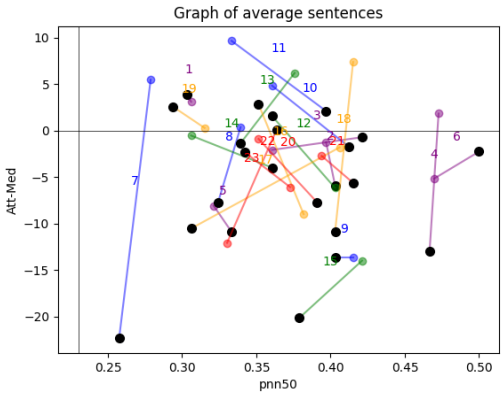

Figure 6. Graph of average sentences with each initial point in black

6. Learning Algorithm for Best Speech Patterns

6.1 Algorithm Theory

In the quest to set up algorithms for identifying the most effective sentences, it becomes crucial to define what will be considered the best sentence. At this stage, we proposed a User-Defined algorithm, which is the algorithm that will try to reach the final emotional state defined by the user or the one managing the experiment. The details of the algorithm are as follows:

To find the best sentence from the different segments available, a suitable model is the implementation of a weighted directed graph. Indeed, by considering each sentence as 2 points connected by an edge, and the initial and final points as the ends of the graph, the shape of this graph is easily apparent.

As a reminder, when the robot uses its algorithms, the initial point will be the user's current emotional state, and the final point will be the one decided in advance, as this is the User-Defined algorithm.

An edge is then created between the user's initial point and each sentence's initial point. The same applies between the final points of each sentence and the desired user’s final point. The order stated for each edge is crucial in this directed graph: given the logic of physiological reaction to different sentences, the

algorithm will always look for an initial point of a new sentence after having passed through a final point of a sentence. This idea therefore also requires connecting each final point of each sentence to each initial point of the other sentences, in this specific order once again.

Now, this directed graph giving the possibility of realizing all available paths must also be weighted. The weighting of the different edges depends on the type of edge in question.

For edges representing a sentence, i.e. an edge representing a segment between an initial and final point of the same sentence, the weight of this edge will be the length of this sentence. The longer the sentence, the higher the weight of the path, as the sentence has the potential for a huge emotional change.

For edges representing distances travelled outside sentences (those starting from the user's initial point, those arriving at the user's final point, or especially those starting from a sentence final point and arriving at an initial point of another sentence), their weight will be the length of the edge times a fixed coefficient strictly higher than 1 (arbitrarily set at 5 for now). Indeed, the idea of finding the best sentences requires finding a pattern of sentences with precision that will take the user to the desired endpoint. The algorithm must therefore be discouraged from favoring a path outside the sentence segments over a path through more sentences.

Finally, it is time to consider the sentence variability mentioned above. When collecting data, the average but also the variability of the different sentences are obtained. The standard deviation is therefore a value that will be used to weight each sentence. The more variable a sentence is among the participants in the experiment, the less likely it is that the algorithm will suggest it for the robot to say. This standard deviation must then be taken into account in the weighting of the sentence.

So, by changing the weight (w) of the edge of each sentence by distance (d) times the standard deviation (σ) of the sentence in question, the more precise the sentence is (low standard deviation) and the more likely the algorithm will be to select it, given the high degree of precision and therefore predictability for the future users (Figure 7).

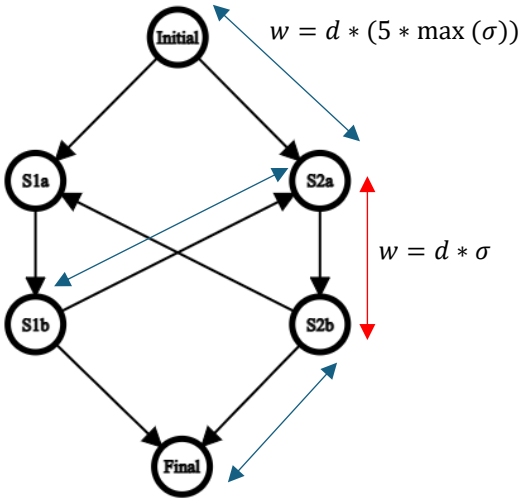

Figure 7. User-Defined weighted oriented graph

Now that the weighted directed graph has been set up, the shortest path can be found by using Dijkstra's algorithm. The shortest path is therefore a list of nodes which, by comparing with the values of the various sentences, enables the algorithm to convert this shortest path into a list of the numbers of the sentences considered best, in the order in which they should be spoken by the robot.

Here's an overview of this User-Defined algorithm, which finds the shortest path among possible sentences to reach the final emotional state sought by the user as precisely as possible. This shows the user's starting point (Initial), the desired final point (Final) and the sentences selected by the learning algorithm, the numbers of which are circled shown in the two examples of Figure 8. The sentences considered best by the algorithm are displayed in the title of each graph. Each sentence is a colored segment for which the initial point is black, and the final point is the same color as the segment, revealing the direction of the reaction to the sentence.

However, as this is just a visualization of the algorithm, any user data or sentences presented here have been manually created to simplify the understanding of what is going on, purposely based on very distinct scenarios.

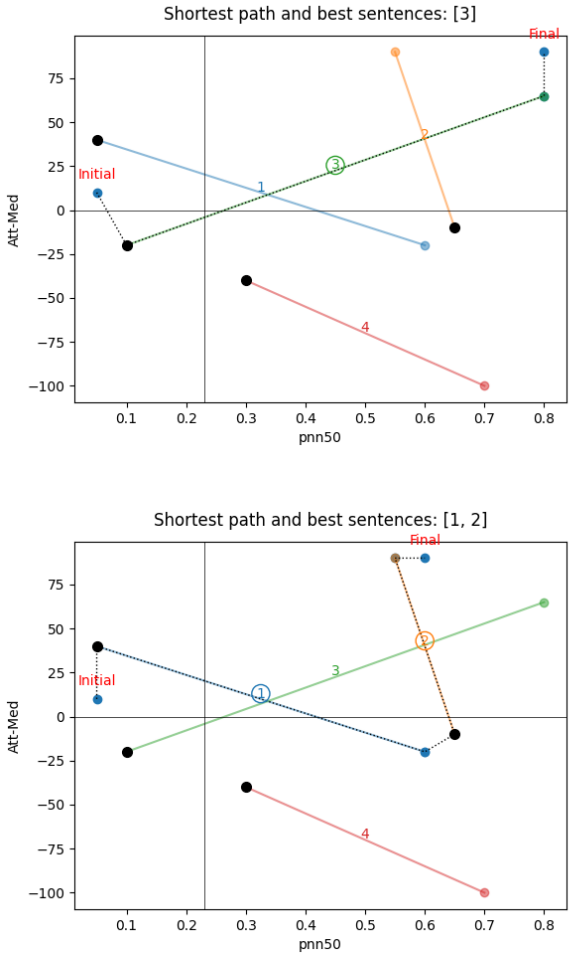

Figure 8. Showcase of the User-Defined learning algorithm

6.2 Implementation in the Robot

As illustrated in Figure 9, when the program running the robot is launched, the script will take the current values of pNN50 and Attention–Meditation to establish the initial point of the graph. Then, based on this initial point and on the database resulting from the experiment that forms the weighted oriented graph, the previously established algorithm will return the list of sentences to be stated in order. The script that outputs speech audio in the robot will then play the different audio files with an interval of several seconds between each sentence. Then, once the entire list has been spoken, the script will start again from scratch, taking the new pNN50 and Attention–Meditation values as its new initial point.

In the end, a calm and gentle voice was selected to maximize the improvement in the emotional state of the robot's users. Figure 10 shows the experiment scene of a user interacting with the robot. Further evaluation on the effectiveness of our robot with the proposed algorithm for determining speech pattern to enhance positive emotions of the users remains as our future work.

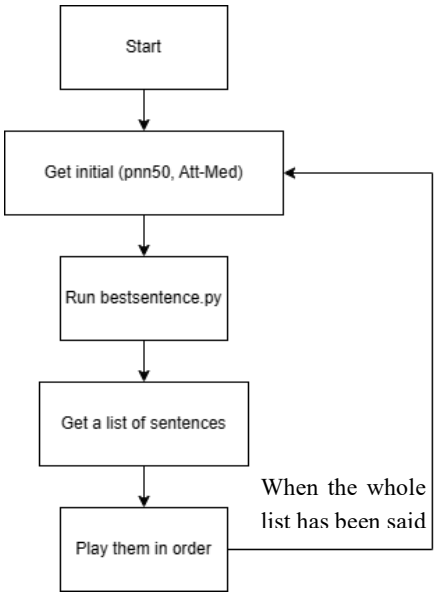

Figure 9. Flowchart of the system applied to the robot

Figure 10. Human with EEG/HRV facing the robot

7. Discussions and Future Works

There is a lot of scope for improvement in this research work. It is important to bear in mind, for the experiment, that using this robot in its original context of elderly care implies a difference in treatment of the participants, as a sarcastic robot would surely not be suitable for this audience. It would therefore be interesting to conduct a different experiment on elderly people to form a database dedicated to them, without bad sentences in the test sample.

It would also be very important to considerably increase the number of participants in the experiment to balance the randomness of human emotions. Similarly, although the current study uses a 10-second window to capture emotional changes, this may not reflect real-world complexity; extended data collection over time is recommended to gain a deeper understanding of emotional dynamics during robot interactions.

As stated at the beginning of the paper, the use of Attention-Meditation is at the heart of debates, which is why future works on this subject would preferentially be based on other EEG indexes such as low alpha and low beta waves.

Moreover, the ideal solution for this learning model would be to use a language model to develop new sentences automatically. Combined with a microphone built into the robot, this could generate a real two-way conversation thanks to speech recognition, enabling new sentences to be tested live and continuously. The robot would then be even more likely to enhance the user's emotional state, as it would be even closer to a conversational experience with another human.

Additionally, while this study has focused on optimizing speech for emotional engagement, future exploration into incorporating facial expressions and gestures could provide further opportunities for enhancing human-robot interaction.

8. Conclusion

In response to the growing need to improve human-robot interaction, this research aimed to improve the speech patterns of the emotion-estimating care robot, with a view to enhancing the emotional state of its user.

We proposed a user-defined algorithm that has been created based on the database of sentences resulting from our experiment to collect emotional responses to those sentences using EEG and HRV. Our proposed algorithm is designed to optimize the sequence of sentences delivered by the emotion-estimating robot, based on the initial state of its user, in order to improve its emotional state as much as possible according to the user’s wish.

We have successfully implemented the proposed algorithm to the robot prototype. Future work will continue to improve the algorithm and evaluate the performance of the robot in enhancing human emotions in several situations.

References

[1] K. Suzuki, T. Iguchi, Y. Nakagawa, and M. Sugaya, “A prototype of multi-modal interaction robot based on emotion estimation method using physiological signals,” in Proc. Asia Pacific Conference on Robot IoT System Development and Platform, Tokyo, Japan, 2022, pp. 7–12.

[2] Y. Ikeda, .R. Horie, M. Sugaya, “Estimating Emotion with Biological Information for Robot Interaction”, Procedia Computer Science 112 (2017) 1589-1600.

[3] K. Abe, T. Nagai, C. Hieida, T. Omori, M. Shiomi, " Estimating Children’s Personalities Through Their Interaction Activities with a Tele-Operated Robot”, Journal of Robotics and Mechatronics, 2020, Vol. 32, No. 1, p. 21-31.

[4] M. Nergui, K. Komekine, H. Nagai, M. Otake, "Development of Laughter Motion on the Cognitive Robot "Bono-02" Assisting Group Conversation”, 2017 IEEE 30th International Symposium on Computer-Based Medical Systems (CBMS).

[5] F. Stojanovska, M. Toshevska, V. Kirandziska, N. Ackovska, "Emotion-Aware Teaching Robot: Learning to Adjust to User’s Emotional State”, ICT Innovations 2018. Engineering and Life Sciences p. 59–74.

[6] A. Hoshina, M. Yokoya, K. Yamada, M. Sugaya, "Preliminary Experiment for Mentally Supporting Rehabilitation Robot based on Emotion Estimation”, 2023 5th International Conference on Advancements in Computing (ICAC).

[7] M. Val-Calvo, J. R. Alvarez-Sanchez, J. M. Ferrandez-Vicente, E. Fernandez, "Affective Robot Story-Telling Human-Robot Interaction: Exploratory Real-Time Emotion Estimation Analysis Using Facial Expressions and Physiological Signals”, 2015 4th Mediterranean Conference on Embedded Computing (MECO).

[8] L. Tiberio, A. Cesta, M. O. Belardinelli, "Psychophysiological Methods to Evaluate User’s Response in Human Robot Interaction: A Review and Feasibility Study”, Robotics 2013, 2(2), 92-121.

[9] J. A Russell, “A Circumplex Model of Affect”, Journal of Personality and Social Psychology 1980, Vol. 39, No. 6, 1161-1178.