proceedings

MDPI

Proceedings

MAmIoTie: An Affective and Sensorized Toy to Support Emotion Perception †

Raúl García-Hidalgo [1], Esperanza Johnson [1], Ramón Hervás [1,*], Iván González [1], Tania Mondéjar [1,2] and José Bravo [1]

[1] MAmI Research Laboratory, University of Castilla-La Mancha, 13071 Ciudad Real, Spain; raul.garciahidalgo@alu.uclm.es (R.G.-H.); mesperanza.johnson@uclm.es (E.J.); Ivan.GDiaz@uclm.es (I.G.); Tania.mondejar@esmile.es (T.M.); jose.bravo@uclm.es (J.B.)

[2] eSmile, Psychology for Children & Adolescents, 13003 Ciudad Real, Spain

* Correspondence: ramon.hlucas@uclm.es; Tel.: +34-926295300 (ext. 6356)

† Presented at the 12th International Conference on Ubiquitous Computing and Ambient Intelligence (UCAmI 2018), Punta Cana, Dominican Republic, 4–7 December 2018.

Published: 23 October 2018

Abstract: Affective Computing aims at developing systems to recognize, process and interpret emotions. This paper presents a sensorized toy with affective functionalities through cognitive services based on IBM Watson technology. The purpose of this research is to improve the quality of life through the assistance oftherapies with children and preadolescents to support emotion perception. This is focused from three points of view: (a) self-perception, (b) empathy and, (c) social-emotional skills. MAmIoTie was evaluated with 10 healthy preadolescent subjects to assess how effectively it analyzes users' emotional perception. The results were generally positive in terms of analysis, though there were aspects that behaved in a way we did not expect.

Keywords: affective computing; cognitive computing; assistive technologies; IOT; sensorized toys; emotion perception

1. Introduction

In the last few years, Affective Computing and Cognitive Computing, working together, have been proven to be effective for achieving a natural interaction with users. Affective computing applies technologies that show, recognize, process, interpret or influence emotion to coach human behavior regarding communication through emotions. On the other hand, Cognitive Computing, in a general view, aims to mimic the functioning of the human brain. This work applies advances on cognitive and affective computing through the implementation of features in terms of speech recognition, natural language processing and sentiment analysis. With these technologies in mind, the goal of this paper is to develop an assistive system to assist therapies with children and preadolescents to support emotion perception. The system is implemented into a sensorized toy (MAmIoTie, the stuffed mammoth) with affective functionalities through cognitive services based on IBM Watson technology [1]. The emotion perception is focused from a triple perspective (perception triad): (a) self-perception, one's emotions self-awareness, (b) empathy, understanding of another person feelings and, (c) social-emotional skills.

The proposed system models the styles that underlie emotional behavior and its influence on mental processing and external behavior. The neural basis of human emotions is difficult to study, because emotions are primarily subjective and nondeterministic. To find basic principles of emotions and their underlying mechanisms, neuroscientists typically study specific emotions, using specific situations [2]. For this work we have selected the emotions happiness, fear, sadness and anger, being the most significant in the developmental age, that are contextualize in prototypic situations. The

conversational agent implemented into MAmIoTie guides the conversation to collect information about the emotional perception of users considering the triad explained above.

The paper is organized as follows: Section 2 explains the background of this paper where we describe neuropsychological fundamentals of this work and explore related work. Section 3 describes the developed system distinguishing the hardware infrastructure and software components. The study results are described in Section 4. Finally, Section 5 concludes the paper with the summarized findings, strengths, and limitation of the proposed system. Moreover, the dialog graph of the conversational agent is described in Appendix A.

2. Background

2.1. Neuropsychological Fundamentals

Emotions are very important in our life because of their adaptive function, i.e., they are the base of interactions with the context and the relation with other people. Emotions coordinate sets of responses to internal or external events that have a particular significance for the organism [3]. In order to successfully manage emotions, named emotional regulation, we have extrinsic and intrinsic processes responsible for monitoring, evaluating, and modifying reactions, especially to accomplish one's goals [4]. Individuals differ in their ability to regulate emotions, some choosing more successful strategies than others [5]. Emotion regulation is crucial for developing emotional intelligence because it occurs every time one (consciously or unconsciously) activates the goal to influence the emotion-generative process [6]. Each individual can thus be characterized by a certain emotion regulation style, which contributes to make him/her predictable in the eyes of others and also carries certain consequences for long-term adaptation [7,8]. The same emotion-regulation strategy can thus be adaptive or maladaptive, depending on the specific individual, the emotion, its intensity, and the context [9–11].

Moreover, emotion sharing refers to expressing one's emotions in a socially shared language [12]. The speech signal is the fastest and the most natural method of communication between humans [13]. This ability to perceive and understand emotions influences social interaction more indirectly, by helping people interpret internal and social cues and thereby guiding emotional self-regulation and social behavior [14]. Thus, emotion sharing is beneficial to mental health due to several indirect effects such as the construction or reinforcement of social bonds and the transference of affection and warmth [12].

Emotion regulation and emotion sharing are closely associated with the quality of social functioning among children and adolescents [15]. In general, all of us learn to regulate ourselves by experience, but there are some people that present disabilities in this area. Based on that idea, we are focusing on the classification of the Diagnostic and Statistical Manual for Mental Disorders [16]. Specifically, we are focused on people with neurodevelopmental disorders that affect intelligence and communication (e.g., autism spectrum disorder, Down Syndrome, etc.).

2.2. Related Work

Human-Computer Interaction has been changing as technology evolves, and its integration in our environment has become greater. These advances, along with the implementation of Ambient Intelligence, allows the users to interact with digital systems through natural interfaces.

There are many examples of natural interfaces, such as touchscreens, though there are also alternatives, like the human voice. The way it is used as a medium is having tools that allow for its recognition, as well as its processing so that the voice can be used as input. This can be used from simple instructions to be performed, to more advanced options, such as the simulation of a human conversation [17].

Being able to simulate a human conversation would allow us to face problems that would otherwise be more complicated to approach. An example would be to evaluate and diagnose communication disorders or alterations. These can manifest in problems when expressing and

manifesting ideas, understanding messages, etc. They can be divided into motor dysfunction (aphasia, deafness, etc.) [18], or those related to cognitive functions (autism, intellectual deficiency, etc.) [19]. This work focuses on the latter group, as its diagnosis in early stages is more complex.

Developmental disorders, denominated and categorized by the DSM-5 in Autism Spectrum Disorders (ASD), are the most persistent deficits in communication and social interaction throughout multiple situations, including the ability to maintain a conversational flow, difficulty to fit within the social context, or aversion towards sharing feelings and emotions, among other symptoms [16]. These symptoms mean that the diagnosis in these cases is complicated, due to failure in communication and empathy between patient and therapist. Because of this, new research lines in ASD therapies are complemented with robotsthat interact with the patients. Because robots present behavioral patterns, which are easy to interpret, with clear dialog, they make interaction with people with ASD easier, and allow them to reach higher level of empathy than when interacting with other people.

For more than 30 years there has been research pertaining the introduction of robots to education, started by Seymour Papert in 1980 [20]. This was later applied to therapy for ASD, given its potential to educate people. There are many examples nowadays, like Nao [21], a general-purpose humanoid robot, or Milo [22], by the Robots4Autism, which is programmed to educate children with ASD to show emotions and facial expressions that transmit feelings. For this reason, we decided to approach this with a sensorized animal-toy, which we have called MamIoTie, to facilitate the interaction with the users.

3. System Description

This section describes the developed system to create MAmIoTie that consists of a hardware infrastructure (Section 3.1) based on single-board minicomputer with several sensors and actuators, and several software components to perform the conversation and analyze emotions (Section 3.2).

3.1. Hardware Infraestructure

Simple and affordable hardware is required to transform a traditional toy into a sensorized one. Figure 1 illustrates the set of prototyping modules and the necessary wired up scheme to achieve this goal. As introduced in Section 1, we intend to add sensorization capabilities to a stuffed animal, particularly a mammoth. The combination of hardware and software modules that makes up the described system allows us to compose a conversational agent capable of acquiring human voice, processing it and providing natural responses in contextualized dialogues.

From the hardware perspective, a Raspberry Pi Zero WH single-board computer (Raspberry Pi Foundation, Cambridge, UK) is used for interface management with a couple of dedicated modules that compose the interaction layer (microphone and amplifier).

Pi Zero WH single-board was not selected by chance. This device has an embedded BCM2835 System on Chip (SoC) developed by Broadcom (Broadcom Inc., San José, CA, USA). This SoC integrates a 1 GHz Arm11 [23] single-core processor (ARM1176JZ-F), which is combined with 512 MBytes RAM. This configuration is enough to run a GNU/Linux distribution (Raspbian) compiled for ARMv6 architectures. Raspbian provides support for Python 3 programming language, which is used to implement the software modules (Section 3.2) and to enable communication with the IBM Watson REST services that afford the affective perception functionalities.

The BCM2835 SoC also incorporates 802.11 n Wi-Fi and Bluetooth Low Energy wireless transceivers. Full TCP/IP stack support is provided by the Pi Zero WH running Raspbian and Wi-Fi communications capabilities through the connection to a close WLAN access point are crucial aspects that allow software modules to send service requests to IBM Watson and receive relative responses through Internet.

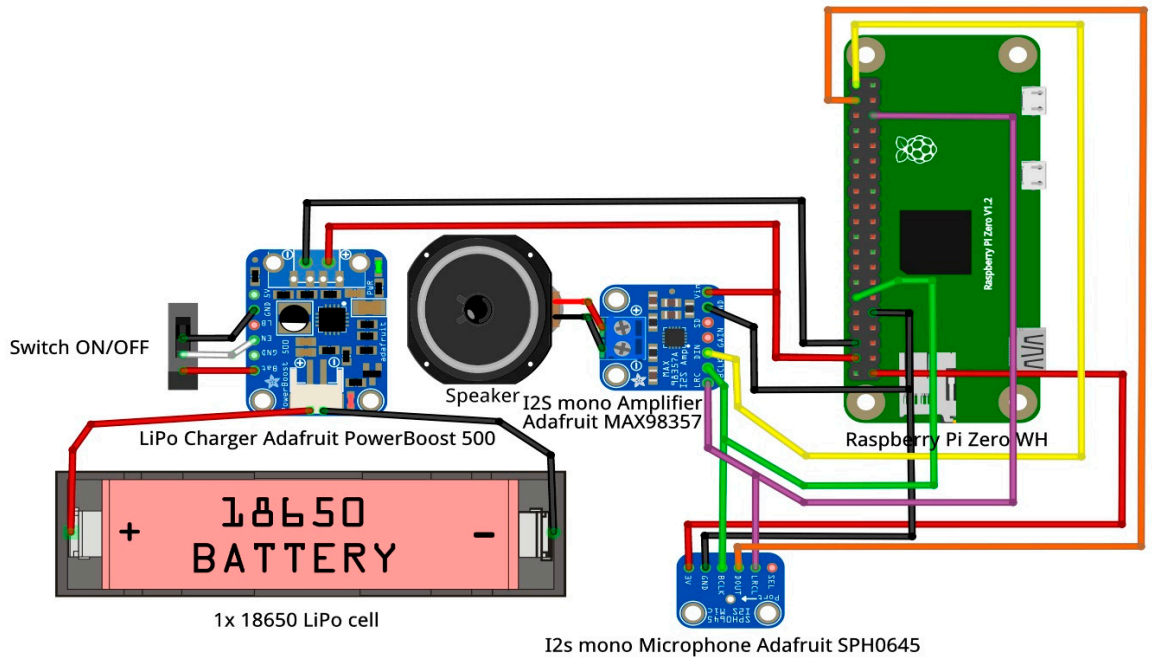


Figure 1. Hardware setup embedded in MAmIoTie sensorized toy.

Another key feature of the Raspberry Pi Zero WH is the Inter-IC Sound (I2S) [24] bus support. The I2S interface provides a bidirectional, synchronous and serial bus to off-chip audio devices. In our case, it makes it possible to directly transmit/receive PCM (Pulse-Code Modulation is a quantization method used to digitally represent sampled analog signals) (digital) audio in both directions, from the I2S mono microphone (audio capture) to the BCM2835 SoC; and from the latter to the I2S mono amplifier (audio playback). Both external modules get direct I2S communication through the GPIO (General-Purpose Input/Output are generic pins on a computer board whose behavior—including whether they work as inputs or outputs—is controllable at run time by the control software) header on the Pi Zero WH board. The I2S bus separates clock and serial data signals, resulting in a lower jitter than is typical of other digital audio communications interfaces that recover the clock from the data stream (e.g., S/PDIF (Sony/Philips Digital Interface)). According to the I2S specification [24], three pins/wires are required to communicate every two ICs (Integrated Circuits):

- LRC (Left/Right Clock)—This pin indicates the slave IC when the data is for the left/right channel, as I2S is designed to work with stereo PCM audio signals. It is typically controlled by the I2S master (the BCM2835 in this case).
- BCLK (Bit Clock)—This is the pin that indicates the I2S slave when to read/write data on the data pin. Also, it is controlled by the I2S master (BCM2835 in this case).
- DIN/DOUT (Data In/Out)—This is the pin where actual data is coming in/out. Both, left and right data are sent on this pin, LRC indicates when left or right channel is being transmitted. If the I2S slave is a microphone or a transductor, it will have a Data Out behavior. Conversely, if the I2S slave is an amplifier wired to a speaker, the pin will be of Data Input type.

Pi Zero WH enables two I2S stereo channels, one as input and one as output. Thus, a configuration with a I2S stereo amplifier and a couple of I2S mono microphones could be made. However, in order to get MAmIoTie as affordable as possible, we set up a prototype with a single I2S mono MEMS (MicroElectroMechanical Systems) microphone (SPH0645) and a I2S 3.2 Watts mono amplifier (MAX98357), both from Adafruit (Adafruit Industries LLC., New York, NY, USA). In fact, the MAX98357 amplifier is a DAC (Digital-to-Analog Converter) module whose analog output is directly wired to a 4-ohm impedance speaker to emit sound. The SPH0645 microphone, for its part, acts like an ADC (Analog-to-Digital Converter) transforming analog audio signal into a digital one.

Thanks to these integrated DAC and ADC converters, digital signals are transmitted through the I2S interface instead of using old schemes transmitting pure analog signals.

Figure 1 shows at a glance how I2S modules have been connected to the Pi Zero WH board. The hardware setup is completed with a 3000 mAh 3.7V 18650 Li—ion battery and a Adafruit PowerBoost 500 module that charges the battery and protects it from overloads and over-discharges.

Lastly, Figure 2 shows a picture of the MAmIoTie working prototype.

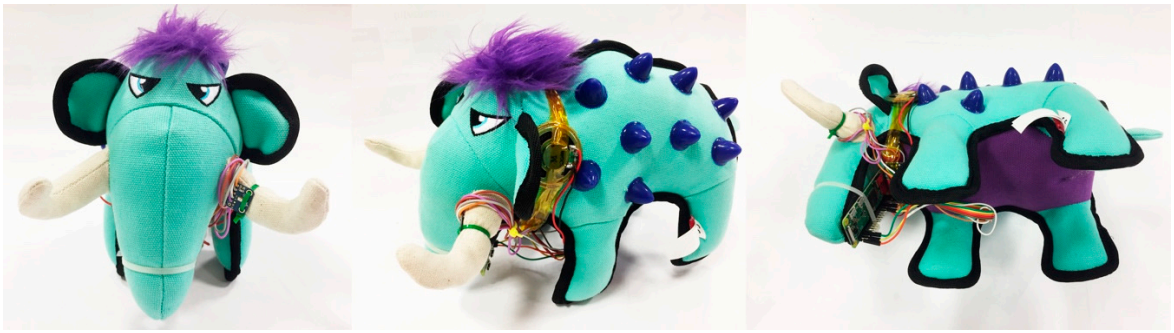


Figure 2. MAmIoTie prototype and internal hardware elements used during the development and evaluation of this work (hardware will be embedded into the toy in the production phase).

3.2. Software Infraestructure

The proposed system uses IBM Watson Cognitive services to provide several functionalities to the application, such as speech recognition, emotional analysis and conversational agent framework. The main programming language chosen for the development of the program is Python, since it will facilitate the implementation of the components in the Raspberry Pi thanks to its compatibility. Also, IBM Watson provides a useful SDK for this language to make calls to its services.

The user interacts with the program through their voice, answering the questions defined by the iterations of the dialogue graph (see Appendix A). The output of the program is a CSV file representing the phases of the therapy session with information about the user's transcriptions, the emotions detected in them and latency of the services, used for performance measurements.

3.2.1. Software Elements

The main software components that are part of the system are shown in Figure 3 and are explained as follows:

- **Noise detector**: The method is run at the beginning of the session. It captures a number of audio samples through the microphone, analyzing its volume. It defines a noise threshold with the average of the volume values of the samples, using 30 percent of the highest values (threshold experimentally obtained).
- **Real-time voice recorder**: The program, through the Python audio libraries, records the audio detected by the microphone and then analyzes its volume to discard silence and noise. When the threshold defined by the noise detector is exceeded, it starts to save the recorded audio, until it detects a range of silence samples (also calculated experimentally), which will determine the end of the user's sentence. It generates an audio file containing the recording.
- **Recorded audio editor**: Once the audio is generated, silence intervals are added to the beginning and end of the file and the volume of the whole sample is normalized. This is done to improve the efficiency of the speech recognition
- **Speech-To-Text**: The final audio file is sent to the Watson Speech to Text service. It returns a text transcription of the voice detected in it. If the service doesn't recognize any coherent word, the program will return to the audio capture state.
- **Tone analyzer**: The text obtained from the transcription is introduced as input to the Watson Tone Analyzer service in order to get the emotions (joy, anger, sadness or fear) corresponding

to the textual input. At the moment this service is not available in Spanish, so it is necessary to translate the text into a supported language such as English. This is done through the Watson Language Translator service, and only affects this module, as the Chatbot and other services work in Spanish.

- **Chatbot:** It defines the conversation graph that models the flow of therapy (we will describe it in detail later). It is implemented by the Watson Assistant service, using its on-line tool to define and model intentions, entities and conversation flows. During the program's iterations, the service will be called and will receive the textual transcription as user response and the main emotion detected in it as conversation context. The service will return a textual answer based on the user's input, the detected emotion and the current state of the conversation graph.
- **Voice synthesizer**: Through the Watson Text To Speech service voice is generated from the textual output of the conversational agent, returning an audio file. Then sound effects (reverb and lower pitch change) are applied to it in order to dehumanize the voice and simulate a more characterized voice of the toy animal. The obtained file is played by the Python audio libraries through the speakers.

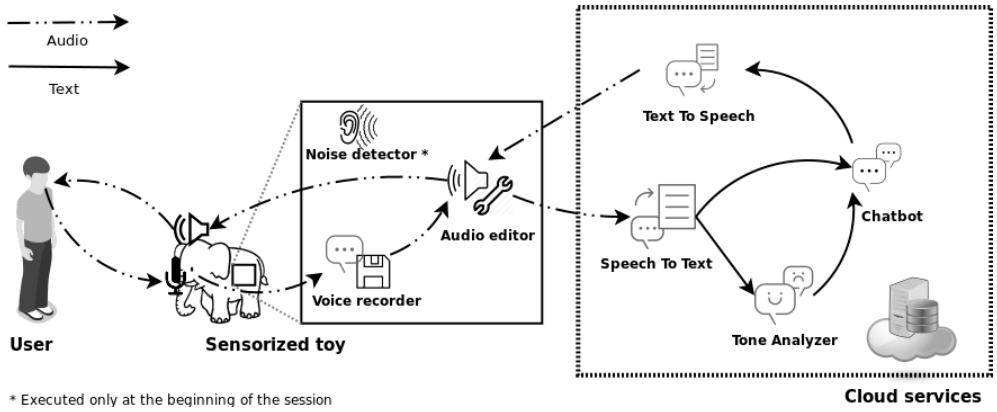


Figure 3. General view of MAmIoTie system including implemented services (Noise detector, Voice recorder and Audio Editor) and cognitive cloud services from IBM Watson.

3.2.2. Dialog Design in the Evaluation Sessions

The evaluation was performed in a laboratory, by 10 children aged 10–15, who performed the experiment as follows: an evaluation session is divided into four phases, one for each main emotion (joy, anger, sadness and fear). There are 24 conversation iterations in total (dyads of answers-response). Each phase consists of 6 questions about emotional states in different situations, divided into groups of two kinds of approaches, one direct and one indirect. In the direct approach MAmIoTie asks directly what emotion would the user feel in a specific situation. On the other hand, the indirect questions consist in asking the last situation in which the user felt a given emotion. The groups of questions will refer to three types of subjects, in this order: the receiver (the user himself), the interlocutor (sensorized toy) and a third neutral person. These three subjects correspond to the perception triad: one's emotions self-awareness, empathy and social-emotional skills, as previously explained.

A first emotion will be used for the first question about the user, depending on the answer given to MamIoTie's first question, and randomly changing the emotions for the questions referring to the second (the toy) and third person, to add variety and avoid monotony. The emotions are changed until all possible person-emotion combination have been explored.

At the beginning of the session, a noise test is performed to measure the silence threshold that will be used to recognize the start and end of the user's speech. After that, the user is asked for their name, which will be used to identify the results generated after the session.

The conversation graph represents all the possible response paths that can occur during the session. As it can be seen in the diagram in Figure 4, the conversation is forced to continue the flow

of questions defined by emotional therapy, without multiple choices or parallel paths for the sake of simplification. The conversation flow depends on the emotion detected in the user responses. The conversation states, conditions and examples of MAmIoTie questions and responses are deeply explained in Appendix A.

At the beginning, MAmIoTie invokes the Watson Assistant, which sends a welcome message and asks the user's name, starting the session. The name is saved as a context variable for later use in stored results. Then MAmIoTie asks about the current emotional state of the user, obtaining the main emotion that will be used for the questions of the first stage. In this way, the conversation always starts exploring the current user emotion. This emotion determines the order in which MAmIoTie will complete the study of the four target emotions. The secondary emotion associated to each user response is also saved to guide this conversation and to add variety and avoid monotony during the session. The final objective is to address the analysis of the four emotions, directly and indirectly, and attending to the perception triad, all in a natural and non-forced way. Moreover, all the graph nodes contain a set of multiple answers based on the main emotion of each phase, increasing the conversation variety and avoiding having to define different nodes for each emotion in the conversation graph. For that, the main and secondary emotions detected in the answers to the questions of each phase are stored in context variables to be able to use them as a parameter of the multiple answers and to store their values. At the end of each phase the context is modified to indicate to the program the saving of the actual state of the emotional variables within the context.

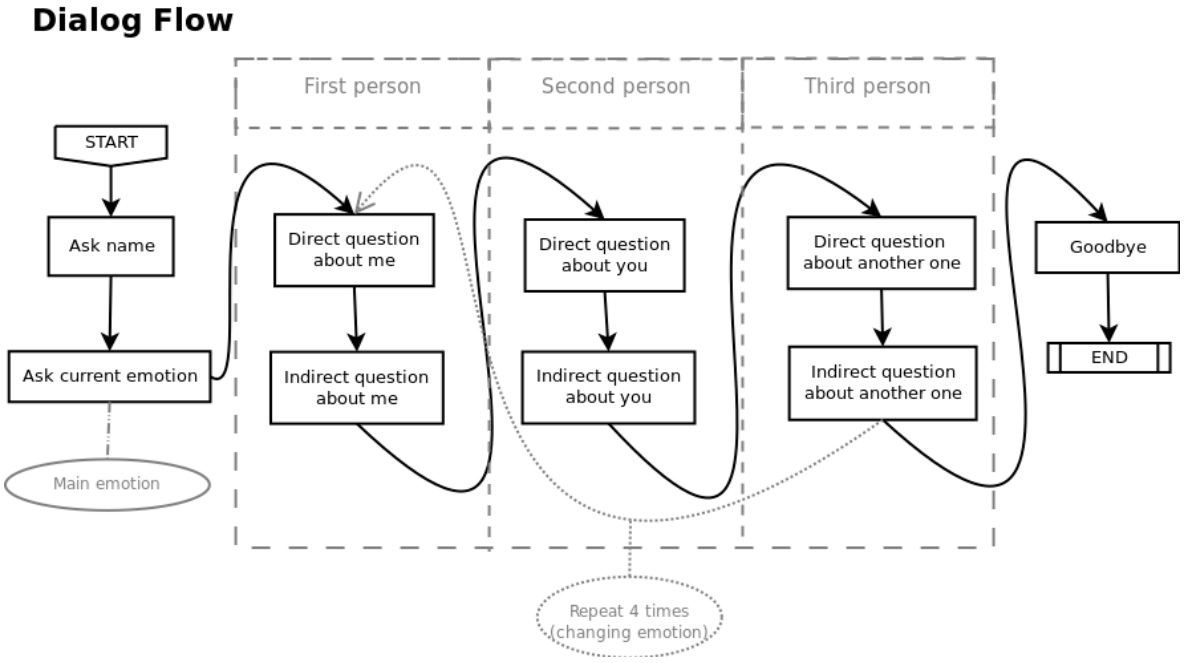


Figure 4. Conversation flow where each main phase consists of 6 questions about emotional states in different situations, including direct and indirect questions, apart from the starting and ending phase.

When all the phases are done, MAmIoTie will say goodbye to the user and notify the program to store the final results. During the session, the results are saved in a CSV file for later revision. This CSV is the data that is analyzed in Section 4. The generated dataset is described in more detail in the section Supplementary Materials.

4. Results

Overall, there are a total of 120 interactions (question-response pairs). During the interactions, several aspects of it are saved into a CSV file (as detailed in the Supplementary Materials). Among those, the expected and detected emotions (joy, sadness, fear and anger) were saved, as well as the transcription of the answers given, result of Watson's Speech-to-Text service. Therefore, we will begin

by analyzing the overall performance (Section 4.1), followed by focusing on the emotional perception (Section 4.2) and the transcription analysis and latency (Section 4.3). We will end by listing the lessons learned from this process (Section 4.4).

4.1. System General Performance

From the analysis of the expected and detected emotions, we applied the formula seen in (1).

$$X= (N_{total} - D_{differ}/N_{total})*100, \quad (1)$$

where N_{total} is total number of interactions, and D_{differ} the times where the expected and detected emotions were different. It is important to remark the fact that Watson's Tone Analyzer service sometimes returns "none" as a detected emotion, if the sentence seems neutral. One example was that users sometimes answered questions with time-related answers, such as "yesterday", "a few weeks ago", etc. These answers have no discernable emotion attached to them, hence the "none" emotion returned by the service. Focusing on those responses that actually have an emotional sense, we saw that in 58.4% of the studied interactions, the expected and detected emotions coincided.

4.2. Emotional Perception Analysis

We expected specific emotions as an answer to the questions posed, which were selected with the help of a psychology expert, and taking into account everyday situations that could be easily understood by the user. In the confusion matrix from Table 1, we can observe how often the expected emotion was detected as such, and how often it was confused with a different emotion

Table 1. Confusion matrix, where the first row contains the expected emotion, and the first column contains the detected emotion.

	Joy	**Sadness**	**Anger**	**Fear**	**None**
joy	15	3	0	0	15
sadness	3	11	2	2	8
anger	1	9	10	1	12
fear	2	9	1	9	4

The emotional perception analysis was performed on the transcribed text, so sometimes, due to several reasons, the audio transcription did not coincide with the user's answer. This meant that on those occasions, a given answer aligned with the emotion we expected, but the transcription was different, and its analysis differed. This accounts for some of them, but it is interesting to observe the results in more depth and explain the other reasons for which the expected and detected emotions differ.

As it can be seen, it is often when the expected emotion is then detected as "none", and in some cases, such as in "anger", the amount of times it is confused with no emotion surpasses the amount of times it predicted correctly. This can be due to two different reasons. The first one, mentioned in the previous paragraph, is that the answers to some of the questions were time-based. This meant that we expected the user to elaborate on what happened the last time they were angry, but they answered when they last felt angry, skewing the results we expected. Another reason as to why angry has worse results than the others, and is confused more often, is also tied to its confusion with sadness. Many of the times, the posed question expected anger from the user, but they answered with sadness instead. This is likely due to the fact that in one given situation, different people can feel different things, and we observed that many felt sadness instead of anger. As an example, to the question "Today I was punished without peanuts, how do you think I feel?" posed by MamIoTie, many answered "sad" or something similar, when the system expected "anger". These two factors contributed to the low levels of prediction of anger in comparison with the prediction of the other emotions. Something similar happened with fear, though it more often got confused with sadness, and it rarely received an answer which had no emotional connotation.

Both mentioned problems can easily be fixed, firstly ensuring more specific questions to obtain a more detailed situation instead of time-based responses, and secondly, considering several negative emotions as valid associations for a given situation. Though a manual analysis of responses, we have identified which are false-negatives. Omitting them, the success rate identifying emotions during the evaluation was 84.4%.

4.3. Transcription and Latency Analysis

As previously stated, one of the reasons for the confusion was the fact that the Speech-To-Text service from Watson did not always accurately transcribe the words.

In order to get a better understanding, we performed a WER (Word Error Rate) analysis, to measure the performance. The formula we applied can be seen in (2).

$$\mathrm{WER} = \mathrm{S} + \mathrm{I} + \mathrm{E/N}, \tag{2}$$

where *S* are substitutions, *I* are insertions and *E* are eliminations performed by the system, and *N* the number of words resulting from the transcription of Watson's Speech-To-Text service.

From this formula, we observed a WER value of 0.33519, with N = 481. Upon further observation, it was seen that the main contributor to the value obtained for WER was the amount of substitutions performed by the service, which also affected the results of the emotion detection by the system. It was rare to have the service perform an insertion or a deletion, although there was a very small amount of deletions observed in the results. These deletions could be attributed to the latency of the system overall, since it sometimes took longer than the users expected, and they started speaking too soon. The average latency of the system was 5489.2755 milliseconds, which translates to approximately one bit over 5 s.

4.4. Additional Learned Lessons

Some of the lessons we have learned relate to the performance of the experiment. Firstly, we posed the questions as "When was the last time you felt ...?", which received time-based answers from the users. This meant that we could not gather proper information on the emotion attached to whatever answer we expected the users to give. As such, the questions will be posed with "What happened the last time you felt ...?", which should fix this particular issue.

We also observed that, as time went on, the users elaborated more on the answers, but after a certain time, users got tired. This is probably due to the latency and the time between each question; and the overall length of the experiment: over 9 min in total.

Some of the users commented, upon completion of the experiment, on MamIoTie's voice, saying they found it "too deep", or "a bit scary", although they still found it amusing. However, if we were to work with younger children, this should be taken into account.

5. Conclusions and Future Work

The main goal of this research is to support therapies in terms of emotion perception and management through a combination of cognitive and affective technologies using a friendly sensorized toy named MAmIoTie. This kind of assistive technology can support people with particular disabilities or pathologies related to social communication, to lead more independent lives, by helping them with difficulties by improving their social skills and understanding of their emotional state. In the current state, this work does not aim to validate the therapeutic use of MAmIoTie. This paper is focused on the technological validation in terms of addressing a natural conversation that permits the analysis of emotion perception from three points of view: one's emotional self-awareness, empathy, social-emotional skills.

In general, the system performance was promising in terms of transcriptions success and emotion detection. However, the results presented in this paper show the importance of a specific formulation of questions, to get the desired type of answers. If the users feel that they should answer with a time-based answer, we will not get any results in terms of what things make them feel certain

emotions. At the same time, it is important to take into account the length of time of the experiment, since some users got tired. Improving the latency of the system is also a key element, as it helps with the sensation of a more natural flow to the conversation, and would help the users communicate. Another consideration is the high latency observed from IBM Watson services, which make the conversation arduous.

Before using MamIoTie within therapy sessions, we identified several technological improvements to address. One of the critical functions of the system is the efficient detection of emotions. At this moment, they are detected by means of the user's transcriptions and Watson Tone Analyzer, in a textual way. An improvement could be to analyze voice recordings and recognize emotions directly in the audio recordings, which would allow consideration of more variables such as intonation or laughter. Another problem to solve is the multiple interpretation of emotions for particular situations, typically in negative emotions that may be correct in certain cases. For example, as it could be seen in the confusion matrix, several times the users answered with sadness instead of anger to certain situations. Both are perfectly valid responses to some situations, as they can be expected to be felt in the example that we used in the experiment (getting grounded), but the current system did not account for that, and so it indicated a disparity between expected and detected emotions. Finally, considering latency aspects, the delay between interactions, caused by the latencies of calls to services, can break the flow of the conversation. To make these calls more efficient, concurrent programming with threads could be applied to the program, making calls to the services concurrently and controlling the coherence of the program flow.

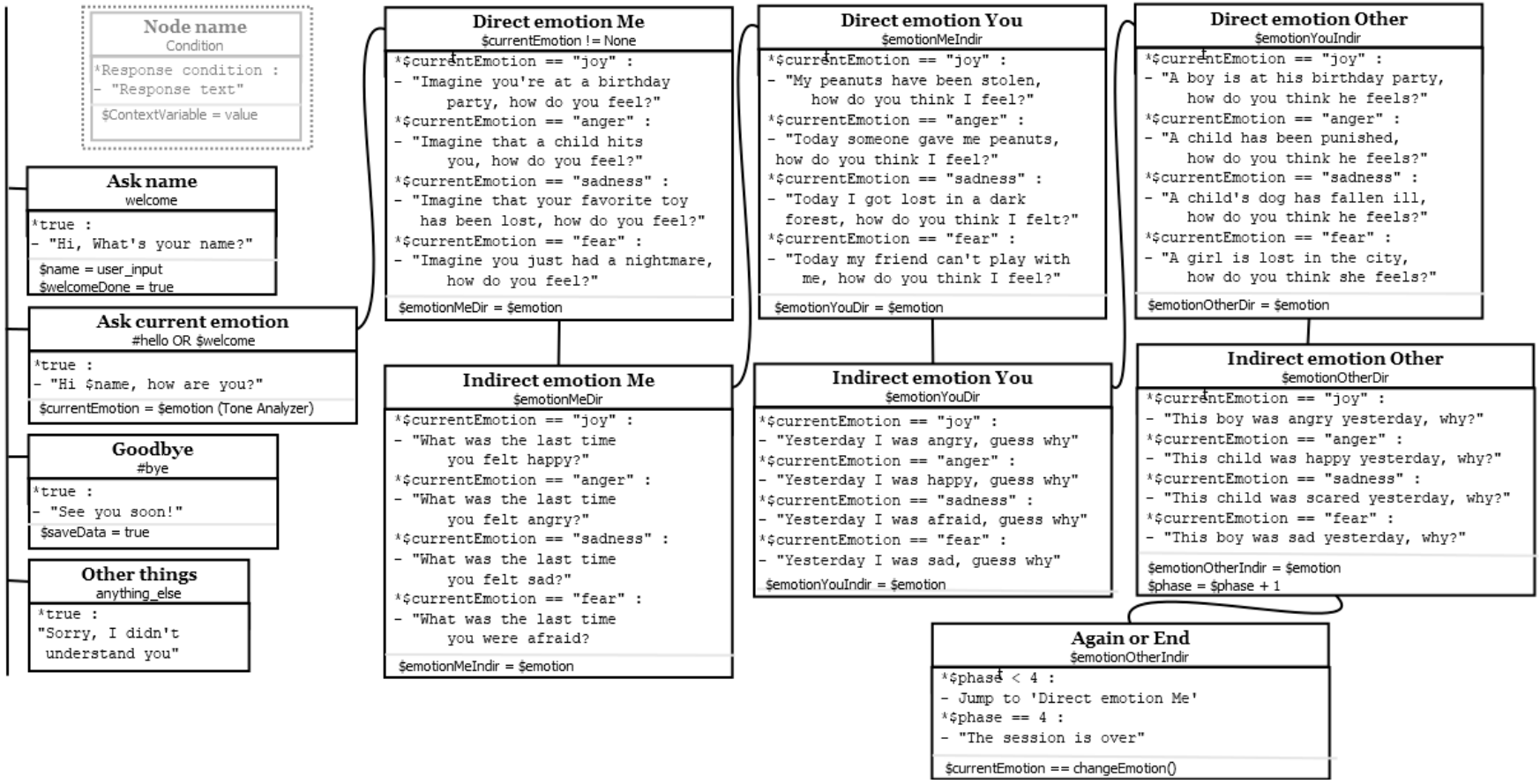
Supplementary Materials: The dataset generated and analyzed for this study can be found in [http://www.esi.uclm.es/www/mami/web/index.php/datasets]. The dataset is published in one CSV per participant with 24 rows (one per dyad conversation, plus the header) and 10 columns for the different variables. The description of the 10 columns is as follows. The first column contains the iteration number that corresponds to the rest of the data. The second has the subject of the question, as per the emotion perception tryad (one's emotional self-awareness, empathy and social-emotional skills), and the third shows if it was a direct or indirect question. The fourth column shows the expected emotion that is returned by the user's answer, while the fifth shows the actual emotion detected. The sixth column contains the transcription by Watson's Speech-To-Text of what the user said, the seventh column contains the translation to English of that transcription, and the eighth the question that was asked by the system. Finally, the ninth column is blank, to be filled with expert observations, and the tenth has the latency for each iteration. The dataset does not contain any identifying aspects of the users

Author Contributions: Conceptualization, R.H. and I.G.; Methodology, R.H. and T.M.; Software, R.G.-H.; Hardware, I.G.; Validation, R.G.-H., E.J. and T.M.; Formal Analysis, R.G.-H. and E.J.; Investigation, R.G.-H., R.H., T.M. and I.G.; Resources, R.G.-H.; Data Curation, E.J.; Writing-Original Draft Preparation, R.G.-H and R.H.; Writing-Review & Editing, E.J. and J.B.; Supervision, J.B. and R.H.; Project Administration, J.B.; Funding Acquisition, J.B.

Acknowledgments: We want to especially thank the participants who helped us by testing and appreciating our cherished MAmIoTie. This work is supported by the "*Plan Propio de Investigación*" from Castilla-La Mancha University.

Conflicts of Interest: The authors declare no conflict of interest.

Appendix A. MAmIoTie Conversation Graph



References

1. IBM Watson APIs. Available online: https://watson-api-explorer.ng.bluemix.net/ (accessed on 25 May 2018).
2. Fellous, J.M. Emotion: Computational modeling. In *Encyclopedia of Neuroscience*; Elsevier Ltd.: Amsterdam, The Netherlands, 2010; pp. 909–913.
3. Lazarus, R.S. From psychological stress to emotions. *Ann. Rev. Psychol.* **1993**, *44*, 1–21.
4. Thompson, R.A. Emotion regulation: A theme in search of a definition. *Monogr. Soc. Res. Child Dev.* **1994**, *59*, 25–52.
5. Mayer, J.D.; Salovey, P. What is emotional intelligence? In *Emotional Development and Emotional Intelligence*; Salovey, P., Sluyter, D.J., Eds.; Basic Books: New York, NY, USA, 1997; pp. 3–31.
6. Gross, J.J.; Sheppes, G.; Urry, H.L. Emotion generation and emotion regulation: A distinction we should make (carefully). *Cogn. Emot*. **2011**, *25*, 765–781. doi:10.1080/02699931.2011.555753.
7. Bar-On, R. *Bar-On Emotional Quotient Inventory: Technical Manual*; MHS (Multi-Health Systems): Toronto, ON, Canada, 1997.
8. Gross, J.J.; John, O.P. Individual differences in two emotion regulation processes: Implications for affect, relationships, and well-being. *J. Pers. Soc. Psychol.* **2003**, *85*, 348–362. doi:10.1037/0022-3514.85.2.348.
9. Bonanno, G.A.; Papa, A.; Lalande, K.; Westphal, M.; Coifman, K. The importance of being flexible: The ability to both enhance and suppress emotional expression predicts long-term adjustment. *Psychol. Sci.* **2004**, *15*, 482–487. doi:10.1111/j.0956-7976.2004.00705.x.
10. Sheppes, G.; Scheibe, S.; Suri, G.; Gross, J.J. Emotion-regulation choice. *Psychol. Sci.* **2011**, *22*, 1391–1396. doi:10.1177/0956797611418350.
11. Aldao, A.; Nolen-Hoeksema, S. When are adaptive strategies most predictive of psychopathology? *J. Abnorm. Psychol.* **2012**, *121*, 276–281. doi:10.1037/a0023598.
12. Rimé, B. The social sharing of emotion as an interface between individual and collective processes in the construction of emotional climate. *J. Soc. Issues* **2007**, *63*, 307–322.
13. Ayadi, M.M.; Kamel, M.S.; Karray, F. Survey on speech emotion recognition: Features, classification schemes, and databases. *Pattern Recogn.* **2011**, *44*, 572–587.
14. Lopes, L., Salovey, P., Cote, S. y Beers, M., Emotion regulation abilities and the quality of social interaction. *Emotion* **2005**, *5*, 113–118.
15. Eisenberg, N.; Fabes, R.A.; Guthrie, I.K.; Reiser, M. Dispositional emotionality and regulation: Their role in predicting quality of social functioning. *J. Personal. Soc. Psychol.* **2000**, *78*, 136–157.
16. American Psychiatric Association. *Diagnostic and Statistical Manual of Mental Disorders*, 5th ed.; American Psychiatric Association: Washington, DC, USA, 2013.
17. Buenaño Proaño, A.C.; Veloz, H.; del Carmen, M. Reconocimiento de voz. Bachelor's Thesis, Pontificia Universidad Católica del Ecuador, Quito, Ecuador, 2003.
18. Hattori, M., Sumita, Y.I., Kimura S., & Taniguchi. Application of an automatic conversation intelligibility test system using computerized speech recognition technique. *J. Prosthodont. Res*. **2010**, *54*, 7–13. doi:10.1016/j.jpor.2009.07.004.
19. Minshew, N.J.; Goldstein, G.; Siegel, D.J. Speech and language in high-functioning autistic individuals. *Neuropsychology* **1995**, *9*, 255.
20. Papert, S. *Mindstorms: Children, Computers, and Powerful Ideas*; Basic Books, Inc.: New York, NY, USA, 1980.
21. Tapus, A.; Peca, A.; Aly, A.; Pop, C.; Jisa, L.; Pintea, S.; Rusu, A.S.; David, D.O. Children with autism social engagement in interaction with Nao, an imitative robot: A series of single case experiments. *Interact. Stud*. **2012**, *13*, 315–347. doi:10.1075/is.13.3.01tap.
22. Robins, B.; Dautenhahn, K.; Wood, L.; Zaraki, A. Developing Interaction Scenarios with a Humanoid Robot to Encourage Visual Perspective Taking Skills in Children with Autism–Preliminary Proof of Concept Tests. In Proceedings of the International Conference on Social Robotics, Tsukuba, Japan, 22–24 November 2017; pp. 147–155.

23. Arm11 Family Processors. Available online: https://developer.arm.com/products/processors/classic-processors (accessed on 25 May 2018).
24. I2S Bus Specification. Available online: https://www.sparkfun.com/datasheets/BreakoutBoards/I2SBUS.pdf (accessed on 25 May 2018).