Proceedings of the 52nd Hawaii International Conference on System Sciences | 2019

Emoty: An Emotionally Sensitive Conversational Agent for People with Neurodevelopmental Disorders

Fabio Catania
I3Lab - Department of Electronics, Information and Bio-engineering at Politecnico di Milano, Milano, Italy
fabio.catania@polimi.it

Nicola Di Nardo
Alten Research Center, Alten SA, Sophia Antipolis Cedex, France
nicola.dinardo@mail.polimi.it

Franca Garzotto
I3Lab- Department of Electronics, Information and Bio-engineering at Politecnico di Milano, Milano, Italy
franca.garzotto@polimi.it

Daniele Occhiuto
I3Lab-Department of Electronics, Information and Bio-engineering at Politecnico di Milano, Milano, Italy
daniele.occhiuto@polimi.it

Abstract

Our research aims at exploiting the advances in conversational technology to support people with Neurodevelopmental Disorder (NDD). NDD is a group of conditions that are characterized by severe deficits in the cognitive, emotional and motor areas and produce severe impairments in communication and social functioning. This paper presents the design, technology and exploratory evaluation of Emoty, a spoken Conversational Agent (CA) created specifically for individuals with NDD. The goal of Emoty is to help these persons enhancing communication abilities related to emotional recognition and expression, which are fundamental in any form of human relationship. The system exploits emotion detection capabilities based on the semantics of the speech by calling the IBM Watson Tone Analyzer API and from the harmonic features of the audio thanks to an "all-of-us" Deep Learning model. The design and evaluation of Emoty are based on the close collaboration among computer engineers and specialists in NDD (psychologists, neurological doctors, educators).

1. Background and Introduction

The general scope of our research is to exploit the advances in Conversational Technology to support persons with *Neurodevelopmental Disorder* (NDD). In particular, we investigate the use of *spoken Conversational Agents* to mitigate the impairments of these persons related to the difficulty of recognizing and expressing emotions – a problem clinically referred to as *Alexithymia*.

Conversational Technology is a general term for integrated technologies that encompass results in various fields: machine learning, natural language processing, speech recognition, dialog generation, and human–computer interaction, among others. A *Conversational Agent* is a system that exploits Conversational Technology to interpret and respond to statements made by the users in natural language.

Neurodevelopmental Disorder (NDD) denotes a group of conditions that are characterized by severe deficits in the cognitive, emotional and motor areas and produce severe impairments in social functioning. Its causes can be genetic or result from lesions or environmental factors. The range of developmental deficits varies from specific limitations of learning or control of executive functions to global impairments of social skills or intelligence. ID (Intellectual Disability), ADHD (Attention Deficit Hyperactivity Disorder) and ASD (Autistic Spectrum Disorder) are all classified as forms of NDD [1][2]. NDD affects at least 3% of the world population. The number of people with the most common form of NDD, autism, are currently 60 million all over the world. Just in the United States, the social cost in order to care of autistic people's needs amounted to $367 billion in 2015 and, if autism's prevalence continues the steep rise seen over the last decade, the projected costs will top $1 trillion by 2025 [3]. Most kinds of NDD are chronic, but early and focused interventions are thought to - at least - mitigate its effect [4][5].

Alexithymia is a personality construct characterized by the subclinical inability to identify and describe emotions in the self. The core characteristics of alexithymia – which often occur among persons with NDD - are marked dysfunction in emotional awareness, social attachment, and interpersonal relating. Furthermore, people with alexithymia have difficulty in distinguishing and appreciating the emotions of others, which is thought to lead to non-empathic and ineffective emotional responding. Alexithymia is traditionally treated with counselling or

URI: https://hdl.handle.net/10125/59641
ISBN: 978-0-9981331-2-6

talk therapies that involve various techniques: group conversations, individual reading of emotional stories, engaging in creative art, daily journaling, and relaxation techniques. While these methods are reported to bring some benefits in mitigating alexithymia effects, specialists are also looking for new approaches [13].

The role of interactive technology in NDD has been explored in several studies. On the one hand, the high children's exposure to chaotic sensory stimulations given, for example, by videogames and multimedia applications have been blamed as one of the causes of the increasing number of cases of cognitive disorders during the developmental age [6]. On the other hand, recent research about children's development has acknowledged interactive technology as a potentially useful tool to support existing therapies and new approaches to the improve the learning process [7][8][9][10][11][12]. Embracing this later vision, we have developed a novel conversational service called *Emoty* that plays the role of *emotional facilitator and trainer* for persons with NDD who manifest severe forms of Alexithymia.

Emoty is a voice-based Italian speaking Dialog System able to converse with the users in ordinary natural language and to entertain them with small talks and educational games. Emoty does not act as a virtual assistant for daily life support but aims at helping people with NDD to develop a better emotional control and self-awareness, which would lead them to enhance their communication capabilities and consequently to improve their quality of life. Emoty exploits conversational technologies, Machine Learning and Deep Learning techniques for emotion recognition from voice sentences based on the processing of user's audio pitch.

The project has been carried out in close collaboration with psychologists, neurological doctors and caregivers who actively participated in eliciting the key requirements, evaluating iterative prototypes, and performing an empirical evaluation.

To our knowledge, Emoty is the first conversational agent that addresses the needs of *emotional learning* among persons with NDD, with the hope of contributing to improve their communication skills. The originality of Emoty is also in the exploitation of audio pitch for emotion recognition from user's aural dialogues. From a more general perspective, our research might pave the ground towards a better understanding of the cognitive and emotional mechanisms associated to NDD and towards new forms of therapeutic interventions for these subjects.

The rest of the paper is organized as follows. In Section 2, we provide an overview of the state of the art about artificial conversational agents and automatic emotion recognition systems from the audio pitch, and then we give an overview of the existing technologies supporting people with NDD. In Section 3, we describe Emoty from a high-level point of view, considering functional and non-functional aspects. Section 4 illustrates the system architecture and its core modules. Particularly relevant are the conversational module and the Cognitive Computing unit, responsible for the emotion detection from the harmonic features of the audio. Section 5 and Section 6 describe the procedure that has been followed to collect and analyze data during the first exploratory study, and its earliest results. Section 7 provides the general conclusions and outline the following steps in our research.

2. Related Work

A Conversational Agent (CA), or dialog system, is a software program able to decode and understand the meaning of the natural language inputs by the user, to act consequently and in a consistent way and to finally render logical outputs that can be understood again by humans. Dialog systems have a long history: the first natural-language processing system, Eliza, was developed in the mid-sixties at MIT [17]. Since then conversational systems have become more mature by leveraging the advancements in Natural Language Processing (NLP) technology and Deep Learning techniques. More recently, many big companies have been investing strongly in the CA area, often with the aim of developing "utility-driven" chatbots to have natural digital conversations indistinguishable from human ones [18]. This strong interest is confirmed by the fact that voice-based assistants are now used by 46% of United States adults both to complete short tasks, as setting up an alarm clock or checking weather forecasts on their phones, PCs and IoT devices, as well as for multi-step tasks such as making a restaurant reservation or booking a taxi [19]. The dissemination of conversational agents in a broad range of applications in business enterprises, government, healthcare and entertaining has been helped by Google, Facebook, Amazon, IBM, Microsoft and many other companies providing to world-wide developers some interesting Conversational Agent development platforms such as Dialogflow [20], wit.ai [21], Alexa [22], Watson [23] and Azure bot service [24]. These tools facilitate the creation, development and deployment of dialog systems and are characterized by different natural language understanding capabilities, supported idioms, accepted programming languages, integration with other systems, and pricing.

Some cognitive-disability-specific Conversational Agents have been commercially developed in the last years. For example, three mobile and tablet applications have been recently launched with the aim to mitigate anxiety and support depression treatment by simulating the conversations with the therapist: Woebot [26], Tess [27] and Wysa [28]. Furthermore, a recent exploratory study used Amazon Echo in speech rehabilitation [25] for users with cognitive disorders.

Conversational Agents supporting skills related to communication, emotion expression and socialization, are still scarce [29], and we are still far away from having clear evidence of the therapeutic effectiveness of this technology in the NDD area. Rachel [30], for instance, is an embodied CA designed for autistic children's skills that aims to create semantically emotionful narratives. From some preliminary experiments, children who had severe impairments in interacting with other people were more likely to interact with Rachel and more motivated to improve their communication and emotion expression skills. The results suggest that conversational technology can create a comfortable, socially non-threatening environment in which autistic children behave without the anxiety caused by the conversation with humans, which requires interpreting the complexity of verbal and non-verbal human communication signals.

Emotions play a very important role in human-human communication, and in the last years researchers in conversational technology have started focusing on emotional human-machine interaction. There are two main branches of this research area: emotion recognition by machines, typically performed by means of Machine Learning algorithms, and emotion expression by machines. In our work we focus on the former one, specifically on the emotion recognition starting from voice sources.

According to [14], voice intonation is responsible for about 85% of the message perception in verbal information transmission while actual words account just for the 15%. These results emphasize the importance of the ability to express and to understand information not openly communicated as factual content. Emotions are among this kind of information and emotion recognition and expression are two of the most relevant skills of a good listener and speaker. It is important to underline that emotion recognition is not an exact science and therefore a perception error rate must be always taken for granted. In order to understand how difficult is for people to recognize emotions during a talk, we report some results in voice emotion recognition from previous experiments. For example, K. R. Scherer's review of approximately 30 studies conducted up to the early 1980s yielded an average accuracy percentage of about 60% for people attending emotion recognition problems from the speech among seven emotional classes [15]. A human classification accuracy of 67.3% and 84.3% is reported on the Danish Emotional Speech Corpus for the five-class problem and for the seven-class problem on the Berlin Emotional Speech Database [16].

At the moment of writing, the most advanced example of the state of the art in automatic emotion recognition is represented by Vokaturi [31]. Its algorithms can measure directly from your voice whether you are happy, sad, afraid, angry or have a neutral state of mind. Other relevant services do not detect actual emotions from the tone of voice, but they can recognize nuanced moods and cognitive states such as the relax level, the degree of laughing and the arousal status. This is the case for example of Affectiva [32], Good Vibrations [33] and Beyond Verbal [49].

Some technologies different from the conversational ones have been used to investigate the emotional sphere of individuals with NDD. For instance, in 2008 a research group of the MIT developed Emotional Bubble System [34], an emotion recognizer by facial expression that address people with autism. According to some preliminary studies, this tool facilitates the users' understanding of relations between their way of acting and the interpreted state of mind. Emotion Trainer [52] and Lets Face It [53] teach emotion and identity recognition from facial expressions. Junior Detective program [54] combines ICT with group training in order to teach social skills. ASC-Inclusion [50] [51] combines several state-of-the-art technologies in one comprehensive game environment, including analysis of users' gestures, facial and vocal expressions, training through games, text chatting, animation, video and audio clips. Early results from ASC-Inclusion's evaluation studies at clinical sites are promising, indicating that children with autism improved on emotion recognition after undergoing the intervention. Still, we could not find any published paper that explores dialog systems as a tool for people with NDD to improve emotion expression skills.

3. Design of Emoty

Emoty has been conceived as a complement to existing learning practices and interventions for individuals affected by NDD and it has been designed having in mind their special needs in the socialization and communication area. People with neurodevelopmental disorders are characterized by weak social interaction capabilities and impaired verbal and non-verbal communication. They have difficulties to identify, express and describe their own feelings. In addition, they need assistance in most of their activities and notably seek a comfortable and safe setting to stay.

To understand the specific needs of our target group, to identify possible solutions, and to elicit the functional and non-functional requirements of Emoty, we attended public events devoted to disability [35], and we worked in close collaboration with NDD experts for two months, through a series of weekly meetings at the care center "Fraternità e Amicizia" [36] in Milan, Italy.

Emoty is a goal oriented, domain restricted and proactive Dialog System. It plays the role of artificial emotional trainer and promotes a proper use of the tone of voice to express emotions and feelings. Its name is catchy and easy to remember and aims to create a direct link to the world of emotions. It talks to the user in ordinary natural language (so far only Italian) and entertains her/him with small talks and educational games in the area of emotions. It holds the initiative of the conversation, meaning that it completely controls the dialogue flow asking to the user a series of questions and steering her/him to familiar domains.

Emoty sessions and its associated dialogs are structured according to the Non-Formal Education model by the American theorist Kolb, who promotes education starting from your own experience [37].

In Kolb's theory, the learning activity evolves as a four stages cycle and it becomes effective when the learner progresses through the whole loop:

- experience, i.e., the concrete practical activity;
- reflection, when the learner is asked to reflect on the experience on personal basis;
- conceptualization, when the user is invited to generalize and make some conclusions starting from the experience and the formation of abstract generally valid ideas occurs;
- application, when the learner is requested to think about future situations to apply what she/he found out, which will lead to the next concrete experience.

After some short cheerful talks to make the user feel more comfortable and safer, Emoty proposes to her/him some emotion expression activities in the form of a game. The user experience unfolds according to Kolb's model, repeating a flow of tasks within the same session that can be performed at increasingly levels of difficulty. For example, in level 1:

- the user is asked to read an assigned sentence trying to express a given emotion with her/his tone of voice as if she/he were an actress/actor practicing for a show ("experience" stage);
- Emoty lets the user reflect about how she/he faced the assigned task ("reflection" stage)
- Emoty lets the user understand more in depth what was good and what was wrong in her/his performance thanks to some supportive acoustic and visual elements ("conceptualization" stage)
- finally, the CA invites her/him to think about common situations where she/he can feel and recognize that emotion ("application" stage).

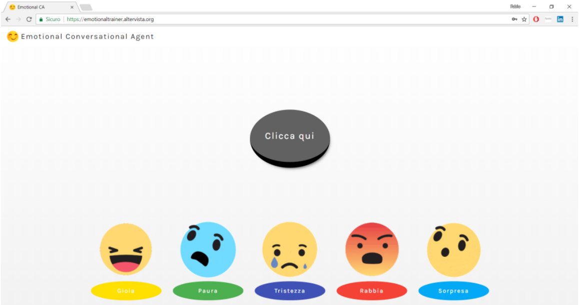
Emotional CA
Sicuro
https://emotionaltrainer.altervista.org
Emotional Conversational Agent
Clicca qui
Gioia
Paura
Tristezza
Rabbia
Sorpresa

Figure 1. Emoty's game view

Some studies have observed that the "invisible" nature of spoken conversational technology might make it difficult to learn and use [38]. To address this concern, in addition to the verbal channel Emoty exploits the visual one to improve usability. Visual stimuli are also useful to enforce the communication of emotion semantics, and to help users with NND to maintain their attention on the exercise.

The GUI shows a big and attractive button that must be clicked by the user before speaking. Emoty uses emojis to facilitate the conceptualization of the emotions and to help reinforce their meaning [39], and support in emotion recognition and expression. To our knowledge, there is no material about the effectiveness of emojis among people with NDD. After considering the results of existing studies on emojis among neurotypicals [39], and discussing them with the NDD experts, we chose five basic emotions - joy, sadness, anger, surprise and fear - and used the visual elements in Figure 2 to exemplify them.

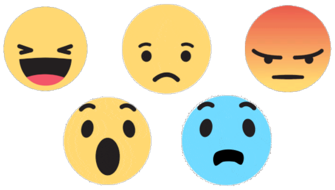

Figure 2. The emojis

To facilitate the emotion representation even further, we decided to use colors as reinforcement in the system's spoken feedbacks. The combination emotion-color is common in literature, and in our work we chose to follow Plutchik's emotions circumplex notation [40]. In his model, joy is represented as yellow, fear is dark green, surprise is cyan, sadness is blue, and anger is red. We associated neutrality to light grey.

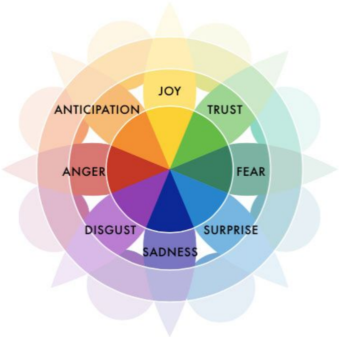

Figure 3. Plutchik's wheel of emotions

Emoty addresses the customization requirements of any technology for NDD that is induced by the wide spectrum of impairments in the NDD spectrum. The application offers to the caregivers the opportunity to personalize the conversational tasks proposed to the user, and to tailor the interaction according to the needs and profile of a specific person.

Finally, Emoty offers some functionality in the caregiver's interface that enables therapists to observe the data automatically collected about each session and get statistical insights from them.

4. Technology

The first prototype of the system has been realized as a web application because of web apps' pervasiveness, their ease of use and the absence of any installation and configuration. Web apps enable both vocal and visual interaction with the user, for example, through the screen, microphone and speakers of the device (both standalone and mobile).

The whole software architecture is organized in functional modules integrating a number of different Smart Services and follows the Model-View-Controller (MVC) pattern. This allows to easily integrate the CA in digital devices such as tablets or smart phones as-is or to embed it in everyday physical objects (e.g. toys, home equipment) just changing the client interface.

The controller of the entire system lies on the Cloud and is accessible through serverless functions to guarantee a fair pricing, a safe execution environment and high-level scalability and availability. On every conversation step, the engine receives an HTTPS request by the client with the user replicas recorded in audio format. The audio input is processed and generates a consistent output, as explained in the following paragraph. The model component manages the data storing them in a dedicated database. The database holds all the sessions' transcriptions, the received audio tracks (which can be reused in the successive steps to refine the emotion recognition service), a list of possible answers the system could gave w.r.t. the different dialog context, and finally the users' personal data to grant the user recognition and as a consequence the session customizations to set up the user-tailored interaction with our CA.

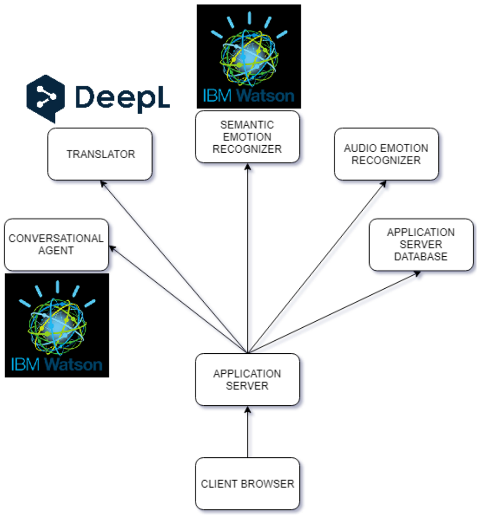

Figure 4. Functional view of the modules composing the system

4.1. Dialog System

The Dialog System has been implemented with the IBM Watson Assistant service, which combines Machine Learning, Natural Language Understanding and integrated dialog tools to create consistent conversation flows. Watson uses an intent-based approach, that means that, in every conversation step, it elaborates a reaction to the user's intention detected by the Natural Language Processing unit starting from the content of her/his message. Furthermore, the system proposes a default fallback intent to deal with requests that do not match any user intent. In our application, in addition to this first elaboration, the final answer given to the user depends also from the emotion that the user transmitted with the contents and with her/his voice to our original emotion recognizer.

According to David Berlo's SMCR Model of Communication [41], there exist four main factors in the communication process:

- the Source, who is the sender of the message;

- the Message, that includes the content of the communication, its structure, the context and the form the message is sent;
- the Channel, that is the medium used to send the message (looking, listening, tasting, touching and smelling);
- the Receiver, who is the person who gets the message and tries to understand what the sender wants to convey in order to respond accordingly.

After studying Berlo's theory, we decided to implement our application's flow as follows: every time the system transmits a voice message to the user, it attaches also an object representing the context it refers to. The context is a state information describing the conversation up to that time. When the user answers, she/he sends back the message in audio format and the same context. At this point, the IBM Watson Speech to Text service decodes what the user said, and the Watson Assistant service detects her/his intention according to her/his request (message plus context). Then, when emotional analysis from the semantics and from the harmonic features of the audio is completed, the system merges both outputs and picks up a consistent answer from the pool of possible responses to return to the user. Doing so, it acts in a consistent way according to both the voice intonation and the actual words by the user and the same replica in different situation or pronounced with different tones will provoke different reactions.

4.2. Semantic Emotion Recognizer

The semantic analysis of the input contents is delegated by our system to the IBM Watson Tone Analyzer [42]. This service uses linguistic analysis to detect emotional tones in written text at both document and sentence levels and returns a dictionary that maps anger, fear, joy and sadness to the probability that the author is expressing the respective emotion. Watson Tone Analyzer provides also the interlocutor's analytical attitude, the level of self-confidence and the degree of inhibition, but we are not interested in this information because, despite they describe the user's inclination, they cannot be considered actual emotions. At the time of developing, there is no emotion analysis API working in Italian language and this constrains us to translate our Italian inputs to English and then to let them be processed by the emotion recognizer. We opted for the DeepL Translator API [43] to mechanize the translation process. In the automatic translation there could be a loss of information to consider, but after a preliminary evaluation the results are satisfying and not too far apart from the ones obtained analyzing human translated texts.

4.3. Audio Emotion Recognizer

The emotion recognizer from the harmonic features of the audio has been developed all by us as a standalone API service in order to be independent from any specific application. The decision not to use one of the already existing emotion detection services (e.g., Vokaturi) was to be freely and totally in control of this software module and to truly experiment on it. Indeed, in this way we could openly choose the set of emotions to be detected by the system, we researched and investigated about different Artificial Intelligence classification algorithms and solutions, and finally we trained our Machine Learning model with the dataset of our choosing.

The general idea of this component is to be fed with an audio file as input and to give back a vector of probabilities associated to the considered emotions as output. The emotions recognized by our system are joy, fear, anger, sadness and surprise. Several techniques have been evaluated in order to achieve our goal during the early research stage, which resulted highlighting Neural Networks as the most promising model to exploit [44].

Our final Machine Learning model has been trained with an open source and free Italian dataset called Emovo [45]. It is a corpus of recordings by 6 actors playing 14 different sentences simulating the five emotion categories mentioned above. The evaluation of the model on the Emovo dataset shows an accuracy between 53% and 57%. There is certainly room for improvement, but this result can be considered adequate for an exploratory study.

Concerning the emotion recognizer, the pipeline of the recognition process consists of two macro-steps falling one after the other: harmonic feature extraction and classification.

4.3.1. Features extraction. As first step, the process requires to extract low-level features from the audio signal given as input. First of all, it extracts samples and frequency of the signal which will be further aggregated to compute both temporal and spectral characteristics. Temporal features, which fall in the time domain, are the easiest characteristics to extract and have an actual physical interpretation. These comprehend Zero Crossing Rate (ZCR), Energy and Entropy of Energy. Instead, spectral features fall in the frequency domain and are: Spectral Centroid, Spectral Spread, Spectral Entropy, Spectral Flux, Spectral Roll-off, 13 Mel-Frequency Cepstral Coefficients (or MFCCs) and Chroma Vector, a vector consisting of 12 coefficients which are closely related to the 12 traditional pitch classes.

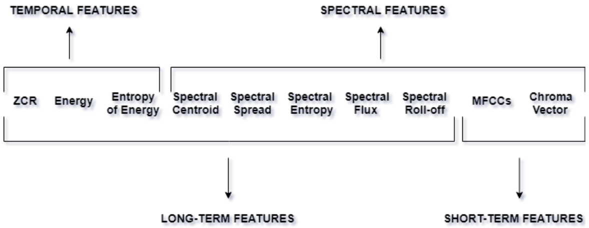

Figure 5. The extracted audio features

In our model design, we took into account an additional binary categorization of the features, namely long-term and short-term features. Such a partition considers the number of samples on top of which a specific property of the signal is computed. Being more detailed, long-term features regard the sampling of a given speech signal in its entire length; whereas, short-term ones are computed taking side-by-side subsets (windows) of those samples.

4.3.2. Classification. Once features have been extracted, the actual classification into a vector of probabilities can take place. Six classes of emotions are considered: anger, joy, surprise, sadness, fear and neutrality. While experimenting, it turned out that traditional supervised machine learning algorithms were not able to effectively classify the given input due to their weakness in the internal representation of complex data, such as pitch signals. As a consequence, we decided to exploit a more powerful algorithm in terms of internal representation of data. We opted for Convolutional Neural Networks, even though they are mostly used in different domains, such as image recognition. As commonly acknowledged, the strength of CNNs dwells into the principle of spatial locality, which led scientists to build high accuracy models for image classification tasks. Based on this background, we decided to shift the classification problem into the time domain, exploiting the time locality instead of the spatial one and significantly raising the accuracy of our experiments.

It should be noticed that we have developed a light preprocessing step in order to remove outliers before entering the network and therefore to make it less susceptible to overfitting during the training phase. This step simply represents a dimensionality reduction phase working at row-level based on the principle of locality. Being more specific, the system uses the Local Outlier Factor algorithm (or LOF), an unsupervised outlier detection technique which computes the local density deviation of a given sample w.r.t its neighbors and then, filters out those samples based on an empirically chosen threshold.

The final and resulting network is a hybrid network, also called wide and deep NN, which is half traditional and half Convolutional. Clearly, from a structural point view, it reflects the aforementioned partition into long-term and short-term features. Indeed, only short-term features, on which the principle of temporal locality can be applied, will be used to feed the convolutional part of the networks; whereas, long-term ones will skip the first convolutional layers of the network joining the computation only as input of the traditional NN.

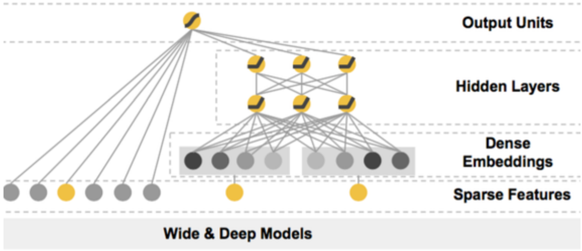

Figure 6. Wide and deep Neural Network [46]

5. Empirical Study

The exploratory study has been designed as a six-months experimentation organized in weekly scheduled sessions taking place during the existing activities of the care center “Fraternità e Amicizia” in Milan, Italy. The aim of the experimentation was to evaluate the usability of Emoty and to measure the improvements in performance (i.e., the completion time for the activities) of users across sessions after a six months period of regular use of the application.

The participants consist of ten people with NDD at mild or moderate severity level that are in a range between 16 and 45 years old. During each session, every user used the application for 10’-15’ under the supervision of a therapist.

The following categories of data were collected during the study:

- Information about users (personal details, type and level of disability) – These data are inserted into the system by the therapist at registration time;
- Information about the sessions (duration, completion percentage, rate of correct/wrong answers for each emotion) – These data are quantitative and are automatically retrieved by the system;
- Qualitative data about user experience in the use of the product (attention, attitude, communication skills, comprehension, etc.) – These data are collected by the therapists and the designers of the application supervising the sessions and by directly asking questions to the user.

To analyze the data, we adopted a quantitative approach even though the corpus of retrieved data was a mixture of qualitative and quantitative information. In particular, we followed the guidelines stated by Glaser and Strauss [47] and their grounded theory about analyzing qualitative data in a systematic way. The analysis started from a set of empirical observations and aimed to develop a theory grounded on top of them.

6. Preliminary Results

At the time of writing, we have performed only the first three sessions of the whole experimentation and the results are still preliminary; it is obviously too early to compute significant statistics on the collected quantitative data. The rest of the section describes the most relevant highlights emerged so far.

Before starting the activity, every participant wanted to play with the application and expressed curiosity and interest in the topic of emotions. During each session everybody was totally engaged in the game. In agreement with the caregivers, we enabled for this test just the first and simplest level of the game. According to the attending therapists, some users tended to be more eager to interact with our device rather than to speak to other people. It was nice to see that all the participants were silent and respectful while a mate was playing and that, in many cases, some of them helped and supported each other during the session.

Figure 7. A user playing with the application

Figure 8. Drawing by a user as a side activity

At the end of the game, everybody was excited and wanted to repeat the experience with the Conversational Agent and a week apart everyone remembered about the game and its rules. It was notable how everybody facing the computer accepted critics and feedbacks completely trusting the machine because, from their point of view, "it cannot commit mistakes". The fact that they could communicate to the system in natural language and that the agent answered back calling them by name, in a gentle way and with continuous repetitions and explanations of the concepts created since the begin a comfortable environment and was incredibly impressive and catchy for them. During the conversation, the agent was perceived as a person so much that, when it was not able to understand what they said, they asked to us "it is deaf, isn't it?"

With the help of the therapist's User Interface, it has been fast to aggregate data across users and sessions. The graphical representation of the performances by three users across the three sessions is reported in Figure 9.

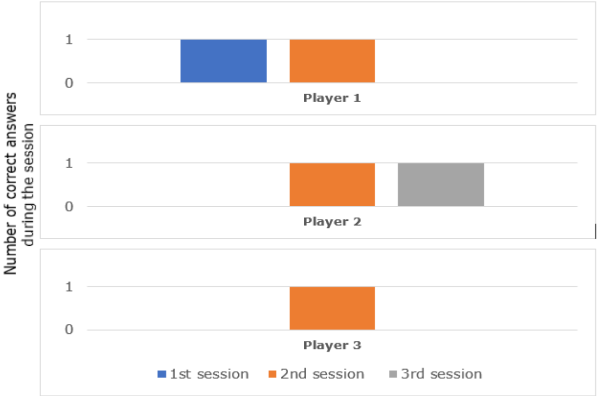

Figure 9. Users' progress across the sessions

Regarding the effectiveness of the application as a supporting tool for individuals with NDD, the opinion of the therapists being aware of the users' background is very important, in principle even more than the detached information extracted from the graphs.

An evidence of the therapeutic significance of the developed system is the application by the users in their daily life of what they learned during the sessions. For now, during the three weeks of practice with the emotional trainer, the caregivers observed a growing awareness by the users of their own feelings. This let us think that more frequent and continue sessions could lead to even more impressive results.

Regarding the usability of the application, during the first session, 6 people out of 11 had difficulty to press the main button on the center of the screen before to start to speak. This issue, as discussed with the therapists, is probably related to the fact that the action of pressing a button is completely uncorrelated from the action of speaking. Fortunately, the interaction with the application has become smoother session by session and, consequently, also the performance within the game improved. Finally, we have noted that Emoty was able to correctly perform the speech-to-text operation despite the elocution problems by some of the participants. Our users often talked to the speech interface with short and directive statements but were rarely ambiguous as observed instead by Derboven et al. in their exploratory study [48].

7. Conclusion and future works

Emoty is a spoken, emotion-sensitive Conversational Agent that has been specifically designed as emotional facilitator and trainer for individuals with NDD. It promotes skills of emotion expression and recognition and has been designed in cooperation with psychologists and therapists as a supporting tool for regular interventions. The system exploits capabilities of emotion detection from the semantics of the speech, by calling the IBM Watson Tone Analyzer API, and from the harmonic features of the audio thanks to an all-of-us Deep Learning model. The evaluation of the Artificial Intelligence component trained on the open source Italian Emovo dataset shows an accuracy between 53% and 57%. Although Emoty is still a prototype, the initial results of our empirical study indicate that Emoty can be easily used by therapists and persons with NDD and has some potential to mitigate Alexithymia and its effects. After only three weeks of practice with the emotional trainer, the caregivers observed a growing awareness by the users of their own feelings.

The first step in our future research agenda is to complete the empirical study and provide stronger evidence of the therapeutic effectiveness of Emoty for persons with NDD. In parallel, we will work to improve the accuracy of the emotion recognition Machine Learning ML model. To do so, we will create and tag a larger, proprietary emotional dataset collecting speech-based conversations and dialogues with the collaboration of local theatre companies and acting schools. Finally, we will improve and extend the functionality for NDD specialists and caregivers. We will develop a more user-friendly personalization interface for therapists to enable them to autonomously extend the body of conversations. We will add new features for data gathering and analysis. We will include functions to support the query and exploration of the most relevant quantitative data on the user's performance in emotion expression and recognition, to enable therapists monitor the improvements of their patients. We also plan to integrate Emoty with wearable biosensors. They will provide an additional source of emotional measures extracted from physiological data. The integration of these devices will enable us to cross-compare data coming from distinct source channels and validate our emotion recognition model.

8. Acknowledgements

The authors are grateful to the persons from Center “Fraternita' ed Amicizia” (Milan) who contributed to the design and the empirical study. This research is partially supported by IBM Italia and EIT Digital, under Project “Cognitive Computing for Cognitive Disability”.

9. References

[1] American Psychiatric Association, 2013, “Diagnostic and statistical manual of mental disorders - V”.

[2] H. Sontheimer, 2015, “Neurodevelopmental Disorders”, in Diseases of the Nervous System: 319-347.

[3] www.autismspeaks.org/news/autism%E2%80%99s-costs-us-economy-estimated-top-265-billion-2015.

[4] G. Cioni et al., 2016, “Early intervention in neurodevelopmental disorders: underlying neural mechanisms”, Dev Med Child Neurol, 58: 61-66.

[5] M. J. Guralnick, 2011, “Why early intervention works: a systems perspective”, Infants and young children, 24(1): 6.

[6] H. K. Kabali et al., 2015, “Exposure and Use of Mobile Media Devices by Young Children”, Pediatrics.

[7] S. Ghavifekr et al., 2015, “Teaching and learning with technology: Effectiveness of ICT integration in schools”, IJRES: 175-191.

[8] K. Tanner et al., 2010, “The Digital Technology in the Learning of Students with Autism Spectrum Disorder in Applied Classroom Settings”, in Proc. of the World Conf. on Educational Media and Technology: 2586-2591.

[9] E. Bouck et al., 2014, “High-tech or low-tech? Comparing self-monitoring systems to increase task independence for students with autism”, Focus on Autism and Other Developmental Disabilities, 29(3): 156-167.

[10] K. Rising, 2017, "Use of Classroom Technology to Promote Learning Among Students with Autism".

[11] N.O. Obiyo et al., 2013, “The Use of Ict as an Integral Teaching and Learning Tool for Children with Autism: A Challenge for Nigeria Education System”, Journal of Education and Practice, 4(23).

[12] K.I. Boser et al., 2014, “Technology Tools for Students with Autism. Innovations that Enhance Independence and Learning”, Canadian Journal of Communication, 41(3).

[13] M. A. Lumley et al., 2007, “The Assessment of Alexithymia in Medical Settings: Implications for Understanding and Treating Health Problems.”, Journal of personality assessment, 89(3): 230-246.

[14] C. C. Chibelushi et al., 2003, “Facial expression recognition: A brief tutorial overview”, CVonline: On-Line Compendium of Computer Vision, 9.

[15] K. R. Scherer, 2003, “Vocal communication of emotion: A review of research paradigms”, Speech communication, 40(1-2): 227-256.

[16] S. Steidl, 2009, “Automatic classification of emotion related user states in spontaneous children's speech”: 1-250.

[17] J. Weizenbaum, 1966, "ELIZA- a computer program for the study of Natural Language Communication between man and machine", Communications of the ACM, 9(1): 36-45.

[18] M. Jain et al., 2018, "Convey: Exploring the Use of a Context View for Chatbots", in Proc. of the 2018 CHI Conference on Human Factors in Computing Systems: 468.

[19] K. Olmstead, 2018, "Voice assistants used by 46% of Americans, mostly on smartphones", [Online]. Available: https://pewrsr.ch/2l4wQnr.

[20] Dialogflow, "Dilogflow", [Online]. Available: https://dialogflow.com/.

[21] Facebook, "Wit", [Online]. Available: https://wit.ai/.

[22] Amazon, "Alexa", [Online]. Available: www.developer.amazon.com/alexa.

[23] IBM, "Watson", [Online]. Available: www.ibm.com/watson.

[24] Microsoft, "Azure", [Online]. Available: www.azure.microsoft.com/en-us/services/botservice.

[25] A. Pradhan et al., 2018, "Accessibility Came by Accident: Use of Voice-Controlled Intelligent Personal Assistants by People with Disabilities", in Proc. of the 2018 CHI Conf. on Human Factors in Computing Systems: 459.

[26] Woebot, US, 2017, [Online].Available: www.woebot.io.

[27] Tess, Canada, 2018, [Online].Available: www.x2ai.com.

[28] Wysa, India, 2018, [Online]. Available: www.wysa.io.

[29] S. Robertson, 2018, "Designing Pronunciation Learning Tools: the Case for Interactivity against Over-Engineering", in Proc. of the 2018 CHI Conference on Human Factors in Computing Systems: 356.

[30] E. Mower et al., 2011, "Rachel: Design of an emotionally targeted interactive agent for children with autism", in Multimedia and Expo (ICME), 2011 IEEE International Conference: 1-6.

[31] Vokaturi, [Online]. Available: www.vokaturi.com.

[32] Affectiva, [Online]. Available: www.affectiva.com.

[33] Good Vibrations, [Online]. Available: www.good-vibrations.nl/.

[34] M. Madsen et al., 2008, "Technology for Just-In-Time In-Situ Learning of Facial Affect for Persons Diagnosed with an Autism Spectrum Disorder", in Proc. of the 10th international ACM SIGACCESS conference on Computers and accessibility: 19-26.

[35] Festival Informatici Senza Frontiere, Rovereto, 13-15 October 2017, [Online]. Events scheduling available: www.isf-festival.it/events/.

[36] Fraternità e Amicizia, Milan non-profit organization, [Online]. Available: www.fraternitaeamicizia.it/.

[37] D. Kolb, 1984, "Experiential learning: Experience as the source of learning and development".

[38] C. Myers et al., 2018, "Patterns for How Users Overcome Obstacles in Voice User Interfaces", in Proc. of the 2018 CHI Conference on Human Factors in Computing Systems (CHI '18). ACM, New York, NY, USA, Paper 6.

[39] D. Derks et al., 2007, "Emoticons and Online Message Interpretation", in SSCR, 26(3): 379-388.

[40] R. Plutchik, 2001, "The nature of emotions", in American Scientist, 89(4): 344-350.

[41] R. S. Croft, 2004, "Communication Theory", [Online]. Available: https://bit.ly/2PJlnDS.

[42] [Online]. Available: www.ibm.com/watson/services/tone-analyzer/.

[43] DeepL translator, [Online]. Available: www.deepl.com/translator.

[44] Y. LeCun et al., 2015, "Deep learning", in Nature 521: 436-444.

[45] G. Costantini et al., 2014, "Emovo Corpus: an Italian Emotional Speech DB", in Proc. of the 9th International Conf. on Language Resources and Evaluation Iceland, 3501-3504.

[46] Tensorflow, [Online]. Available: www.tensorflow.org/.

[47] J. Lazar et al., 1988, "Research Methods in Human-Computer Interaction", in Handbook of Human-Computer Interaction, Chapter 42: 905-928.

[48] J. Derboven et al., 2014, "Designing voice interaction for people with physical and speech impairments.", in Proc. of the 8th Nordic Conference on Human-Computer Interaction: Fun, Fast, Foundational (NordiCHI '14). ACM, New York, NY, USA, 217-226.

[49] Beyond Verbal, [Online]. Available: www.beyondverbal.com/

[50] B. Schuller et al., 2013, "ASC-Inclusion: Interactive emotion games for social inclusion of children with Autism Spectrum Conditions.", in Proc. of IDGEI 2013, Chania, Greece, no pagination.

[51] B. Schuller et al., 2015, "Recent developments and results of ASC-Inclusion: An integrated internet-based environment for social inclusion of children with autism spectrum conditions", in Proc. of IDGEI 2015, Atlanta, GA, no pagination.

[52] M. Silver et al., 2001, "Evaluation of a new computer intervention to teach people with autism or asperger syndrome to recognize and predict emotions in others.", in Autism 5, Issue 3: 299 - 316.

[53] J. Tanaka et al., 2010, "Using computerized games to teach face recognition skills to children with autism spectrum disorder: the Let's face it! program.", in Journal of Child Psychology and Psychiatry, 51: 944-952.

[54] R. Beaumont, 2008, "A multi-component social skills intervention for children with Asperger syndrome: The junior detective training program", in Journal of Child Psychology and Psychiatry, 49: 743-753.