Multimedia Tools and Applications (2023) 82:12797–12828
https://doi.org/10.1007/s11042-022-14135-w

1224: NEW FRONTIERS IN MULTIMEDIA-BASED AND MULTIMODAL HCI

A conversational agent for emotion expression stimulation in persons with neurodevelopmental disorders

Fabio Catania[1] · Franca Garzotto[1]

Received: 10 January 2022 / Revised: 13 October 2022 / Accepted: 25 October 2022 /
Published online: 9 November 2022

Abstract

Difficulty in emotion expression and recognition is typical of the personality trait known as alexithymia, which is often observed in persons with neurodevelopmental disorders (NDD). Past research has investigated various forms of conversational technology for people with NDD, but only a few studies have explored the use of conversational agents to reduce alexithymia. This paper presents Emoty, a speech-based conversational agent designed for people with NDD to train emotional communication skills. An original characteristic of this agent is that it exploits the emotional expression power of the *voice*. Emoty engages users in small conversations during which they are asked to repeat sentences and express specific emotions using the appropriate vocal tone. We ran an empirical study to evaluate the usability and effectiveness of our conversational agent. The study involved 19 Italian individuals with NDD and alexithymia aged from 29 to 45 (10 women and 9 men). They used Emoty in five individual sessions over two and a half months. The results showed that two subjects encountered problems using the system because they had difficulty verbalizing the sentences and were not understood by Emoty. The others performed the assigned tasks with the agent. Their capability to express emotions with the voice consistently improved, and other benefits were observed in other social and communication skills.

Keywords Human-computer interaction · Conversational technology · Speech emotion recognition · Computer-assisted therapy · Neurodevelopmental disorder · Alexithymia

✉ Fabio Catania
fabio.catania@polimi.it

Franca Garzotto
franca.garzotto@polimi.it

[1] Department of Electronics, Information and Bioengineering, Politecnico di Milano, Via Golgi, 39, Milano, 20133, MI, Italy

1 Introduction

Neurodevelopmental disorder (NDD) is a group of conditions with onset in the developmental period and is characterized by severe deficits in cognitive, social, and emotional areas [2]. *Emotions* are episodes of coordinated changes in neurophysiological activation, motor expression, and subjective feeling in response to external or internal events [79]. People with NDD often show impaired awareness of their and others' emotions and struggle to manifest and describe their feelings, which are typical conditions described under the name of *alexithymia* [43, 50, 73]. Traditional interventions for alexithymia involve various techniques such as group conversations, individual reading of emotional stories, engaging in creative art, daily journaling, and relaxation techniques [25]. While these methods are reported to bring some benefits, researchers and clinicians also look for new approaches to better understand and treat alexithymia. The use of interactive technology is thought to facilitate people in learning to feel, identify, and express emotions [48, 58]. Indeed, the technological interaction space may be perceived as more predictable, "safer", and simpler compared with the complexity of interacting with other humans [57, 59, 85, 86]. *Conversational agents* are software that interacts with humans in natural language [35, 44]. Some scholars have recently investigated their potential to complement traditional interventions for people with NDD [19], and only a few studies (focusing on persons with ASD) explored the use of conversational agents to mitigate the traits associated with alexithymia [34, 64, 66]. Still, our knowledge in this field is minimal: performing user studies with people with NDD and achieving empirical evidence is complex and challenging, given the broad spectrum of special needs that characterize this population and the multidisciplinary aspects to consider when designing and evaluating interactive technologies for them. Current results on the *usability* of conversational agents for people with NDD and their *effectiveness* to support emotional training are very preliminary and deserve further studies.

This paper describes Emoty, an emotion-aware conversational agent that acts as an *emotional facilitator* and *trainer* for persons with NDD, and reports an empirical study devoted to investigate the potential of this tool to promote the ability of people with NDD to express emotions. Emoty is based on the *Big Six* model of emotions, which is the categorical theory proposed by Paul Ekman suggesting that anger, fear, disgust, joy, sadness, and surprise are universally experienced in all human cultures [31]. While most tools for emotional training in people with NDD focus on the use of verbal, facial, and body expressions, Emoty aims at stimulating the capability of expressing emotions using the *voice*. This approach is grounded in the 7-38-55 rule of human-human communication [63], which indicates that words deliver 7% of the total information content, while non-verbal channels convey most meaning, i.e., the vocal tone (55%) and the body language (38%). The training tasks in Emoty were designed in collaboration with psychologists and therapists who are experts in NDD. The agent entertains users with small talks, asks them to verbalize sentences expressing specific emotions with the appropriate voice, and provides feedback about their "acting performance", i.e., the degree to which their verbalizations express those emotions. Acting performance is evaluated in two ways: (i) automatically by the agent, using a novel emotion recognizer that extracts emotion-related features from the analysis of audio recordings by the user; (ii) manually by the caregivers, who use a tool on a separate device to assign emotional tags (e.g., joy, sadness) to the user's speech.

Our empirical study focused on two main research questions:

RQ1. Is Emoty *usable* by persons with NDD?
RQ2. Is Emoty *effective* for people with NDD to improve emotion expression?

The study involved 19 participants with NDD who performed five individual sessions using Emoty over two and a half months. Our findings indicate that Emoty was generally usable by participants; in addition, the users' ability to express emotions with the voice during the interaction with the agent and the amount and correctness of their verbal communication increased.

Our empirical results are still limited and are not generalizable to all conversational agents and all individuals with NDD because of the different functional, psychological, and emotional characteristics of each person with NDD and the diversity of today's conversational agents in terms of design and natural language communication power. Nevertheless, our research contributes to the current state of the art not only in emotion-aware technology for persons with NDD but also from a broader perspective, confirming past research that indicates (i) that interaction with interactive technology in natural language is not an apriori barrier for persons with NDD, and (ii) that conversational agents have a potential utility for people with NDD, particularly concerning the improvement of language skills [69, 76]. Finally, our study highlights some crucial design issues - from both a UX perspective and a technological one - that might emerge when creating conversational agents for persons with NDD and provides insights that can benefit designers and developers in this area.

The remainder of this paper is organized as follows. Section 2 presents a literature review of previous studies involving conversational technology for training the emotional skills of people with NDD. Section 3 explains the design process, the user experience, and the enabling technology of Emoty. Section 4 provides an overview of the methodology of our empirical study, describing the participants, setting, procedure, and data gathering and analysis approach. Section 5 reports the qualitative and quantitative results of our analysis, which are discussed in Section 6. Finally, Section 7 draws the conclusions and outlines the next steps of our research.

2 State of the art

In the last two decades, an emerging theme of study has been the integration of interactive technologies into traditional cognitive and behavioral therapies for people with NDD [33, 86], and ad-hoc conversational technology for this population has been increasingly investigated [19, 69, 76]. We completed a systematic survey of the studies concerning conversational agents for people with NDD, and we highlighted the crucial design features of these systems, the therapeutic goals they address, and the empirical methods adopted for their evaluation [19].

There are several challenging aspects to take into account during the design process of a conversational agent for NDD, mainly related to the wide range of symptoms, severity levels, attitudes, and capability of the target users [65]. At the moment of writing, there is no standard method universally accepted in the HCI community for the design. Many studies pinpoint the difficulty of the direct involvement of persons with NDD in the process. For this reason, most research includes only a few subjects [15, 83, 87] and never describes as straightforward the data gathering process for technology evaluation. For example, questionnaire-based approaches [42, 46] involve critical and self-critical skills that are often lacking in people with NDD [2]. At the same time, task-based methods (e.g., cognitive walk-through [54]) assume a level of cognitive functioning that these persons might not have [2], e.g., to understand the tasks to perform on the system and the associated sequences of actions to be executed. To address these issues, previous research recommends including different stakeholders in the design process, such as the parents, psychologists, special educators, designers, and HCI engineers [12, 53, 80]. The input from these people enables the so-called *"proxy participation"*, where designers rely on the decision-making by the stakeholders as representing the user's choices.

Many researchers have explored the effectiveness of conversational agents in improving the communication skills of the participants with NDD such as joint attention, imitation, turn-taking, and communicative behaviors [10, 20, 28, 41, 49, 67, 72, 77]. To the best of our knowledge, only seven publications presented speech-based interactive agents to improve the emotional skills of people with NDD and empirically evaluated their usability or effectiveness. Before providing details on each of these papers, we offer a general analysis of their main features, considering different dimensions: the profile and number of participants in the empirical study, and the embodiment, conversational capability, and training goals of the agent. In all studies, the subjects were children diagnosed with ASD with an average number of participants equal to 11.14 (*SD* = 10.04). Five studies employed a social robot, while the others used a virtually embodied conversational agent (integrated into a web app or a virtual reality app). In five papers, the conversational systems were able to understand and respond in natural language, while two systems could produce speech-based output but could interpret touch-based input only [55, 78]. In five studies, the primary therapeutic goal of the conversational agent was to stimulate emotion *recognition* skills from *facial expressions*. In one study, the training was on emotion *expression* skills and again used facial expressions. In four studies, the assigned tasks addressed the ability to associate different contexts with different emotional states. Other types of tasks have been proposed in the remaining studies.

In [66], the authors described a conversational social robot for improving facial emotion recognition combined with eye contact, joint attention, and imitation capability. They used it in an ABA-like treatment with three high-functioning children with autism aged 8 to 13 (no info was provided about the gender). When the participants and the robot looked at each other, the robot stimulated them with a facial expression (anger, fear, joy, and sadness) and said: "I am (angry, happy, sad, scared), you do it!". If the children correctly imitated the facial expressions, the robot produced positive feedback. Statistical differences in eye contact and facial expression imitation behaviors after using the system were reported as preliminary results.

In [34], the researchers compared the interaction between four male children with autism aged 6 to 12 and an agent who could be a human being or a conversational socially assistive robot. During traditional and technology-based sessions, agents trained users' understanding and comprehension of emotions, their recognition and expression through face, their association with social situations, and empathy in a broad sense. The results indicated a higher incidence of eye contact, proximity, and verbal interaction during sessions with the robot than during sessions with the teacher. Additional behaviors (reduced fidgeting and increased attention and ability to follow instructions) improved during interaction with conversational technology.

In [55], the researchers combined a teacher-teleoperated robot and a tablet system to help teach facial emotions to users. An empirical study with 31 participants with ASD aged 4 to

14 (8F, 23M) showed that they could use the tablet interface to respond to the conversational robot and perform the proposed activities.

In [64], the researchers employed the digitally embodied conversational agent Rachel as an emotional coach to create semantically emotional narratives through four specific activities: basic emotional face recognition, emotional storytelling with an in-order and out-of-order stimuli presentation, missing emotional face identification, and mismatched face identification. Two male children with ASD aged 6 to 12 interacted with Rachel using speech, gestures, and touch through a screen interface. Early results suggested that the agent could create a comfortable, socially non-threatening environment in which children behaved without the anxiety caused by the conversation with humans, which requires interpreting the complexity of verbal and non-verbal human communication signals. They also suggested that digitally embodied conversational agents have the potential to provide an effective platform for eliciting and analyzing children's communicative abilities and encouraging affective and social behaviors.

In [7], the authors presented a virtual reality-based conversational agent capable of collecting eye-tracking, psychophysiological data, and EEG data while the users were involved in an emotion recognition task from facial expressions. Users saw an oval occlusion on the face of an avatar in the virtual world, and the occlusion was progressively revealed as they scanned the face by looking at relevant areas for emotion recognition, such as the eyes and mouth. A usability study investigated the performance of six male teens with ASD aged 12 to 18 with a gaze-sensitive version of the system providing an adaptive and individualized gaze feedback mechanism to users and a control group of six other male teens with ASD aged 12 to 18 who participated without the online gaze adaptation and occlusion paradigm. The results suggested that the gaze-sensitive system enabled the former group to improve better over time compared with the control group.

In [60], the researchers used a conversational social robot to support the therapy of fourteen children with autism aged 3 to 12 (2F, 12M) addressing the teaching of context-emotion association. Pre- and post-intervention assessments were conducted, and substantial improvements in contextualized emotion recognition, comprehension, and emotional perspective-taking through the use of human-assisted social robots were attained.

Analogously, another social robot was proposed in [78] to address the emotional understanding and perspective-taking skills of twelve children with autism aged 6 to 12 (5F, 7M). The robot provided a social situation, displayed as cartoon-like images on the touch screen, and asked children to choose what they thought the story character was feeling at different points in the story by selecting one of the multiple options displayed on the screen. This activity allowed participants to develop social skills and generalize them in human-human contexts.

To the best of our knowledge, none of the aforementioned studies have exploited speech-based conversational technology to train people with NDD with the ability *to express emotions with the voice*. The focus on this specific aspect is the most original feature of our conversational agent Emoty. In a previous publication [16], we presented the enabling technology of the conversational agent and a preliminary study exploring its usability with people with low severity of NDD. Another research [15] was published and is about the evaluation of Emoty's anthropomorphic perception by users with NDD (i.e., the degree to which they perceived the agent as a human being or an artificial entity). This paper extends past publications by discussing the UX design of Emoty, and by reporting a new larger empirical study focusing on its usability and therapeutic effectiveness.

3 Emoty

Emoty is a speech-based Italian conversational agent designed to improve the ability of individuals with NDD to express emotions with the voice. Emoty is integrated into a web application that also provides a control interface for caregivers. In this section, we describe the process that brought to the development of the system and discuss its UX design and technological choices.

3.1 Design process

In the design of Emoty, we proceeded along an iterative three-stage process including *exploration*, *prototyping*, and *validation*.

Exploration focuses on eliciting the emotional communication needs of the target population. We followed the examples of previous research on interactive technology for persons with NDD [12, 80]. We adopted an approach based on *proxy participation*: we did not include any representative of this population in the data gathering process, but relied on other stakeholders' points of view (the ones of professional caregivers) as representing the needs of the primary end-users. Focus groups were organized every two weeks for two months, involving a psychologist, a neurologist, a linguistic expert, and two therapists from a local daycare center. Each session lasted for one hour, used a semi-structured protocol, and focused on one of the following questions:

1. the characteristics of people with NDD in terms of their cognition and attention levels;
2. their communication and socialization challenges, particularly in the emotional domain;
3. the activities that are normally proposed to them during therapeutic interventions;
4. the attitude of the target population toward the technology.

We transcribed the focus group recordings, cleaned them up by stripping off nonessential words, annotated them based on the asked questions (set a priori) and the emerged topics (not set a priori). Finally, we organized the obtained information to inspire our technology design choices.

Once we implemented the prototype, we invited the same stakeholders to use it and give us opinions and suggestions for improvement. This knowledge informed the redesign of the advanced prototype described in Sections 3.2 and 3.3. A first exploratory evaluation of the prototype produced positive results in terms of usability [16]. In Section 4 we describe a further empirical evaluation we conducted involving some persons with NDD and their caregivers.

3.2 User experience overview

3.2.1 UX for people with NDD

Conversational experience with the agent The conversational agent Emoty was designed as *goal-oriented* [45] and *domain-restricted* [29]. It plays the role of an artificial trainer that promotes the proper use of the voice to express emotions and feelings. It entertains users with small talks and emotion expression tasks, but cannot sustain any other form of conversation. Each session with Emoty is structured as a single conversation with a starting and an ending point. The agent holds the initiative of the conversation [75], meaning that it completely controls the dialog flow asking the user a series of questions to lead them into a comfort zone. Emoty's dialogs were scripted by a linguistic expert paying particular attention to the conversational style. The agent calls the user by name, speaks gently, and continues repetitions and explanations of the concepts to create a comfortable and safe environment. Users can ask the agent to repeat the last produced utterance or to quit the session at any time. When this happens, the conversation ends with some greetings ("It was nice to see you. Goodbye!"). All these features are to facilitate the development of a long-term social relationship with users [14], and in this respect, Emoty falls into the category of *relational* conversational agents.

As depicted in Fig. 1, each session with Emoty is split into two phases: *"onboarding"* and *"acting activity"*.

The goal of onboarding is to get users familiar with the interaction paradigm, introduce them to the conversational experience and the coming acting activity. During this phase, Emoty welcomes the user and asks three questions to be potentially answered with short, simple responses such as "yes" or "no", or "good" or "bad" to prevent misunderstanding because of any possible speech impairment of the user and reduce their potential frustration.

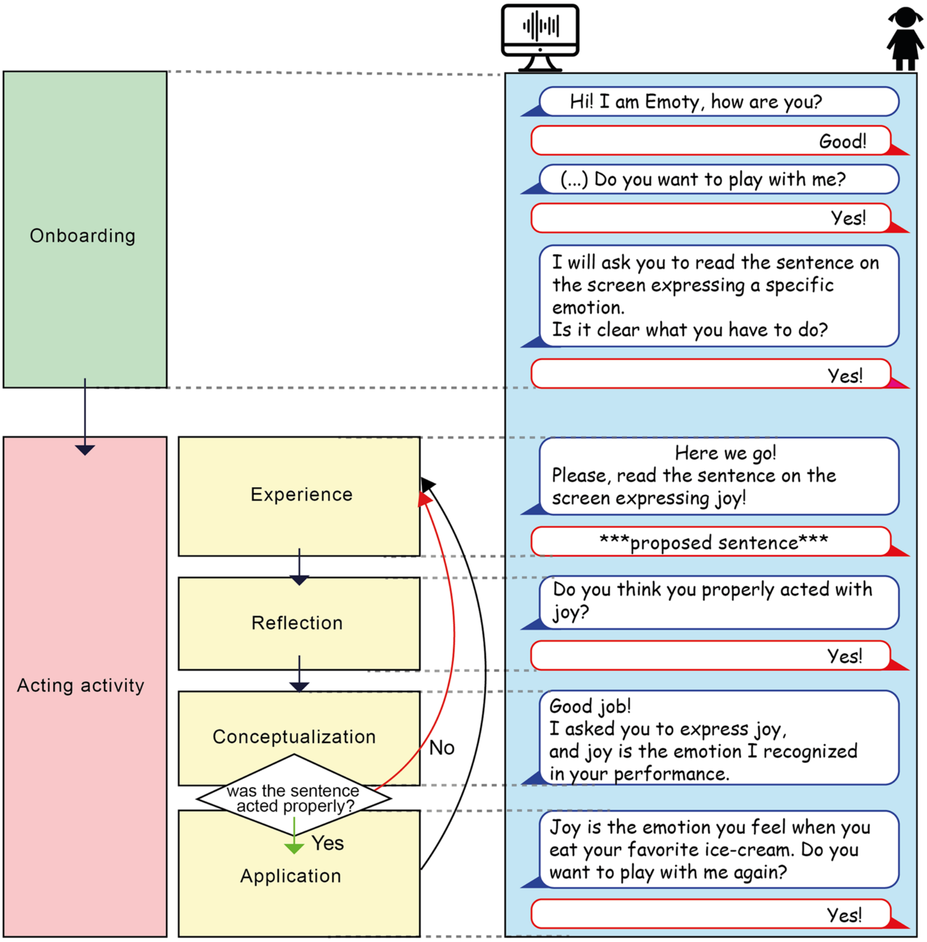
Onboarding
Acting activity
Experience
Reflection
Conceptualization
was the sentence acted properly?
No
Yes
Application
Hi! I am Emoty, how are you?
Good!
(...) Do you want to play with me?
Yes!
I will ask you to read the sentence on the screen expressing a specific emotion.
Is it clear what you have to do?
Yes!
Here we go!
Please, read the sentence on the screen expressing joy!
proposed sentence
Do you think you properly acted with joy?
Yes!
Good job!
I asked you to express joy, and joy is the emotion I recognized in your performance.
Joy is the emotion you feel when you eat your favorite ice-cream. Do you want to play with me again?
Yes!

Fig. 1 An exemplification of the dialog flow with Emoty

The acting activity was designed following the principles of *gamification*, which is meant as the use of game design elements in non-game contexts [27]. From the “ten ingredients of great games” identified in [71], we implemented a *narrative context* to link different tasks during the session, some explicit *rules* that govern the conduct within the activity, and a set of different types of *feedback* to the user to return corrective information about their actions. We opted for taking advantage of gamification since it is known to bring potential benefits relevant to users’ cognitive, motivational, social, and emotional development [39]. The dialog flow is structured according to the four stages of the non-formal education model by the American theorist Kolb [51].

1. *Experience*: the user is asked to read an assigned sentence trying to express a given emotion with the voice as if they were actors practicing for a show. At this point, the user sees on the screen both the sentence to verbalize and the name and emoji of the emotion to express (see Fig. 2A). Sentences are picked up in a randomized order from a pool of very short and easy-to-pronounce utterances chosen by psychologists and therapists. Emotions are randomly selected among joy, sadness, fear, anger, and surprise. We opted for this set of emotions because they are part of the Big Six emotions (joy, sadness, anger, surprise, fear, and disgust) recognized by the psychologist Ekman as universal to every human being, regardless of culture or education [31] and because they are the de-facto classification system used by emotion detectors in the affective computing field [88]. Following the advice of the NDD experts, we decided not to include disgust in the prototype of Emoty as previous studies have shown that people with NDD particularly struggle to recognize and express this emotion [3, 84]. Also, the NDD experts suggested not to include the possibility of accessing any emotion expression example at the beginning of each acting task to reduce the risk of “mechanical” repetition and imitation.

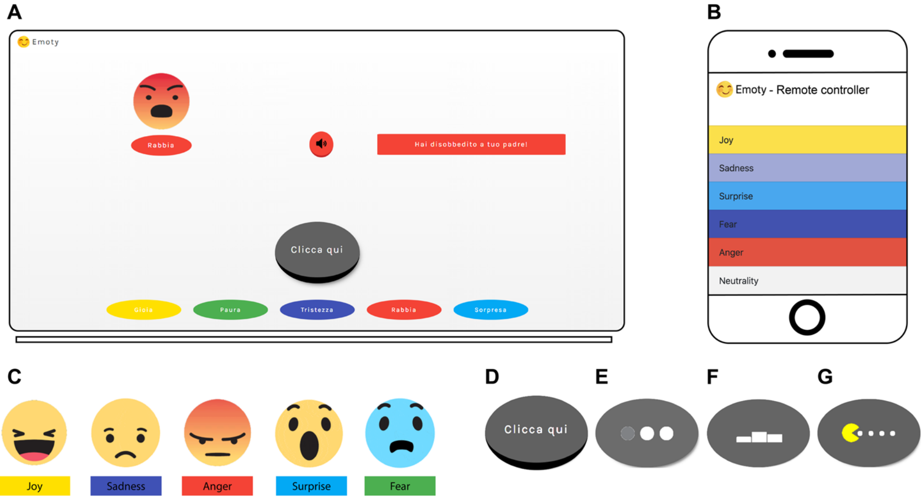

Fig. 2 The graphical user interfaces of Emoty. (A) The web page for the users with NDD showing an emotion and a sentence to be verbalized. (B) The web app for the therapists to force the emotion recognized by Emoty. (C) Emojis and colors used as a visual support. (D) Idle, (E) listening, (F) speaking, and (G) pending buttons describing the status of the agent

2. *Reflection*: Emoty stimulates the user to reflect on how they faced the assigned task by asking "Do you think you properly expressed the emotion I asked you?".
3. *Conceptualization*: The agent evaluates the user's performance based on the emotion detected in their voice. Acting performance may be evaluated in two ways: (i) automatically by the agent, using an original speech emotion recognition system that extracts emotion-related features from the analysis of audio recordings by the user; (ii) manually by the caregivers, who use a tool on a separate device to assign emotion tags to the user's sentences (see Section 3.2.2). Once the emotion is obtained, Emoty provides feedback on their performance to the user. If they properly expressed the requested emotion, the agent gives them visual and acoustic rewards ("raining" emojis, changed background color, and congratulations by voice), and then proceeds to Step 4. Otherwise, Emoty plays the audio recording of an actor correctly reading the sentence. At this point, the user can try again to act out the sentence (as in Step 1) up to two more times. When the user cannot complete the task for 3 times, Emoty jumps back to Step 1, generating a different sentence and requiring the expression of a different emotion.
4. *Application*: The agent invites the user to think about everyday situations where they can feel and recognize the emotion they have just expressed; it helps them in this task by listing some of the possible scenarios identified by psychologists and therapists. Finally, Step 1 is repeated with a new sentence and a new emotion.

General interaction features The conversational agent is integrated into a web application that enables the user's speech- and touch-based interactions through the microphone, touchscreen, or mouse. The phone, tablet, or PC speakers allow the conversational agent to produce audio prompts for the user while the screen acts as visual support for verbal communication. Emoty shows a big button on the screen and requires users to click on it each time they want to speak. In other words, clicking on the button is the *wake action* of Emoty: the agent wakes up every time the button is clicked and starts listening to the user; next, it stops listening when it recognizes a pause that marks the end of the user's speech. In addition, the agent provides the users with visual feedback about its status (idle, listening, speaking, or pending - see Fig. 2E-G) to help them to handle the interaction and to better understand the system. To prevent users with NDD from being distracted by Emoty's appearance, we designed the agent as *disembodied* [29], which means that it has no virtual or physical representation. Visual stimuli, such as emojis, are used to help users maintain their attention on the agent and to facilitate their conceptualization of emotions [26]. To our knowledge, the literature lacks information about the communicative effectiveness of emojis among people with NDD. After considering the results of existing studies on emojis among neurotypicals [26] and after discussing these results with the NDD experts, we chose to use the five visual elements in Fig. 2C to exemplify joy, sadness, anger, surprise, and fear. To further reinforce the visual communication features of Emoty, we matched colors with emotions, adopting Plutchik's emotions circumplex model [68]: yellow is associated with joy, dark green with fear, cyan with surprise, blue with sadness, red with anger (see Fig. 2C), and light gray with neutral feeling.

3.2.2 UX for caregivers

A simple graphical interface integrated with Emoty enables caregivers to *classify* the emotions expressed in the utterances produced by the user. Specialists on NDD requested this functionality after evaluating the system's initial prototype [16]. They considered the accuracy of our speech emotion recognition model to be too low and were concerned that the model had never been tested in the wild. In our application context, low accuracy in recognizing emotions would risk providing incorrect feedback to the users too frequently, affecting the effectiveness of the emotional training with Emoty. As shown in Fig. 2B, the control interface provides six buttons corresponding to the six emotional states considered by Emoty. When a button is selected, the corresponding emotion is automatically assigned to the user's sentence, overwriting the one recognized by Emoty.

3.3 Technology overview

As depicted in Fig. 3, Emoty has a client-server architecture and leverages the Model-View-Controller architectural pattern to support flexibility, robustness, and scalability. A state machine handles the logic of the conversational agent. The state represents a piece of contextual information describing the previous conversational experience up to that moment, as in David Berlo's SMCR model of communication [9]. For each step of the conversation, the controller tier

1. receives the user's audio input from the view tier;
2. selects a behavior tree that defines the pipeline of specific analysis operations (speech to text, natural language understanding, speech emotion recognition) to be performed on the input depending on the current state;
3. performs these operations;
4. produces a synthesized speech output that is consistent with the analysis results. The same sentence produced by the user in different situations or pronounced with different tones may generate different outputs from the system.

The audio output is played back from the view tier and saved with the input and the analysis results in the model tier.

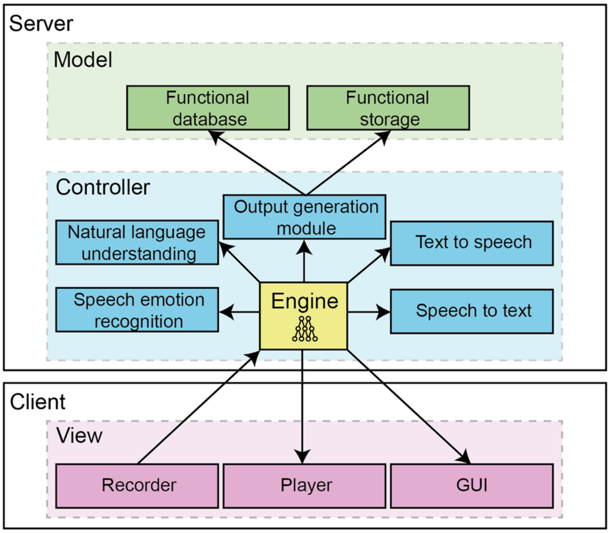

Fig. 3 Functional view of the software architecture of Emoty. Arrows show the data flow when the system is running (from caller to callee)

The controller is composed of independent modules to increase the system's reusability, scalability, and maintainability. The main modules involved in the analysis are as follows:

- the *speech to text module* receives as input a prerecorded audio containing some spoken words and returns them as a string. Emoty exploits Google Speech-to-Text [37] for this task because it represents one of the cutting-edge technologies owing to the large amount of data used to train their advanced deep learning and machine learning models;
- the *natural language understanding module* receives as input a string of text representing the transcription of the user's speech and the context it refers to. This module analyzes, understands, and returns the user's intent by exploiting domain knowledge and natural language comprehension capabilities. Intents are links between what the user says and its interpretation by the system. Contexts help differentiate requests that might have different meanings depending on previous requests. Emoty outsources the execution of natural language understanding and exploits Google Dialogflow API [36]. As for speech to text, the large amount of data Google has at their disposal to train the models makes their NLU unit one of the state-of-the-art technologies for this task;
- the *speech emotion recognition module* receives an audio recording containing human speech as input and returns the emotion perceived by analyzing the harmonic features of the audio (mfccs, spectrogram, chromagram). For this task, Emoty exploits an original deep learning model based on the fusion of a vanilla neural network and a convolutional one to detect joy, sadness, anger, surprise, fear, and neutrality [16]. The model was trained on the open-source Italian dataset called Emovo [23]. The evaluation of the model on a portion of Emovo that was not used for training showed an overall accuracy between 53% and 57% [16]. In addition, the speech emotion recognition module offers an endpoint that allows manual forcing of the recognized emotion.

The main modules involved in the generation of the output for the user include:

- the *output generation module*, which produces the response depending on the state of the conversation, the detected intention of the user, and the emotion recognized from the pitch of their voice. Emoty is a *retrieval-based* [45] conversational agent, which means that the output generation module selects the response from a predefined repository and does not generate conversational responses during the conversation. Responses may contain placeholders that are replaced with custom information in real-time to make dialogs user-centered (for example, by calling the user by name);
- the *text-to-speech module*, which is responsible for human voice synthesis, that is the artificial production of human speech. Compared to recorded human speech, the advantage of synthesized voice is that its content can change and be customized at runtime. Emoty exploits Google Text to Speech [38] with a female Italian voice.

4 The present study

4.1 Research questions and hypotheses

In this study, we investigated the following research questions.

RQ1. Is Emoty usable by participants with NDD?

RQ2. Is Emoty effective in helping people with NDD improve their expression of emotions with the voice?

To address our questions, we conducted an empirical study to accept or reject these hypotheses:

H1a. Participants can easily interact with Emoty since the first usage.
H1b. Interaction with Emoty becomes easier over time.
H1c. Users with NDD are not always understood by Emoty because of their possible speech and cognitive difficulties.
H1d. Misunderstandings decrease over time.
H2. Users improve over time their ability to express emotions with the voice.

In Section 4.6, we report the research variables we defined and analyzed to evaluate all hypotheses.

4.2 Participants

We managed to recruit 19 Italian-speaking people from two daycare centers we collaborate with on several projects. The centers provide various care services to persons with NDD to promote their occupational skills, social inclusion, autonomy, and well-being. People with NDD may receive therapeutic interventions, job training classrooms, operative laboratories (such as arts and crafts), individual sessions, and psychological support.

The profiles of the participants included in this study are reported in Table 1. The population was almost equally split between women and men (10 women and 9 men). The subjects ranged in age from 29 to 45 years ($M = 37$, $SD = 5.3$). All participants were diagnosed with NDD and coexisting mild (3), moderate (8), and severe (8) levels of cognitive impairment. The severity level of a cognitive impairment is defined based on the adaptive functioning of individuals because it determines the level of the support they require [2]. With mild cognitive impairment, people may still be able to perform everyday activities. Severe levels of cognitive impairment can lead to a loss of the ability to understand the meaning of something and the ability to talk, resulting in an inability to live independently [2]. P4, P14, and P17 were also diagnosed with Down syndrome. Additionally, all participants had traits associated with alexithymia, but none had an official diagnosis. It is common that this personality trait is not adequately diagnosed because it is usually difficult to properly assess its diagnostic criteria or because alexithymia features are overshadowed by other comorbid conditions in the primary diagnosis [11, 58]. Of the 19 participants, only P3 offered information about the daily use of a laptop. However, all of them already knew what a computer was, and most of them had a smartphone. Nobody said they had interacted with any conversational agent (e.g., Google Assistant, Siri, Alexa, and Cortana) previously.

In addition, we recruited a neurotypical person who performed a session with Emoty to provide us with a minimal baseline (that we consider our best-case scenario in terms of the number of sentences over time, quality of interaction, and acting performance) against which to compare the results of the participants with NDD. This person (female, aged 27, no previous experience with Emoty) interacted with our agent for one session, according to the same protocol defined for the entire study. In the rest of the paper, we called her *"control user"*.

The study protocol was approved by the Ethics Committee of our university. Participants' personal data were processed in compliance with the provisions of the current legislation on the protection of personal data. Parents or legal guardians of participants with NDD signed an informed consent dealing with the acceptance of the experimental procedure and the use of data.

Table 1 The profile of the participants with NDD and traits associated with alexithymia

ID	Age	Gender	Cognitive impairment level
P1	34	F	Mild
P2	44	F	Severe
P3	35	M	Mild
P4	42	M	Mild
P5	30	F	Mild
P6	38	M	Severe
P7	37	F	Severe
P8	29	F	Middle
P9	34	M	Middle
P10	37	F	Middle
P11	35	M	Middle
P12	35	F	Middle
P13	45	F	Middle
P14	44	F	Severe
P15	42	M	Severe
P16	35	M	Severe
P17	31	M	Severe
P18	44	M	Middle
P19	35	F	Severe

4.3 Setting

The study was conducted in a small quiet room with a stable Internet connection at each daycare center to help participants feel as comfortable as possible in a familiar space. In each room, there was a laptop on a desk and a chair in front of it. On one laptop's side, there was a desktop microphone connected to the laptop via a cable. On the other side, a wireless mouse allowed users to move freely during the experiment. In addition, it was an ambidextrous version to make the experience comfortable for both left- and right-handed participants. To create a more welcoming setting for the user and avoid making them feel under examination, we decided with the therapists not to use any camera to record the experiment.

In addition to participants, people in the experimentation room were:

- the *facilitator*, who was a psychologist or a therapist known by the participants managing the experiment at the forefront. Later, we refer to F1 when we talk about the facilitator in *L'Impronta* and F2 when we talk about the facilitator in *Collage*;
- the *test observer*, who was a member of our team silently observing the experiment from the background and taking notes about interaction challenges, requests for help, and commentaries by participants and observations aloud by the facilitator.

  To mitigate any possible impact of the presence of the test observer on the participants' experience, he sat on the sideline in the room and did not interact with the subjects. In addition, prior to starting the study, the observer introduced himself to the subjects and took part in some activities in the centers without using technology (drawing and painting).

4.4 Procedure

We designed a longitudinal study with sessions scheduled every two weeks for two months and a half. Each participant was involved five times. The duration of each session varied based on the assessment by the facilitator, who was aware of the capabilities, needs, and weaknesses of each participant. We set a maximum of twelve minutes for the conversational experience, or four different sentences that could be acted out during each session (*max tasks number* = 4). NDD experts placed these constraints based on the abilities of the participants.

During the sessions, the participants showed up one at a time to interact with Emoty. The facilitator welcomed the participant in the room, introduced them to the experience, and explained how to interact with the system (i.e., clicking on the button before speaking). Next, the facilitator performed the login on behalf of the participant to enable automatic data gathering and a customized experience (such as Emoty calling users by name). At this point, participants interacted with Emoty, following the instructions received directly from the conversational agent. If they could not interact with the system on their own, they were helped by the facilitator (for example, she clicked for them with the mouse when they were not able to do it). Moreover, when the instructions of the acting task were not clear, the facilitator stepped in to explain again and ensure a correct understanding. Every time the user performed an acting attempt, the facilitator was also in charge of using the side-app for the remote control of the emotional feedback by Emoty to select the emotion they recognized in the user's voice (see Fig. 2B). Indeed, automatic speech emotion recognition was disabled for this study as requested by the stakeholders during the design process (the accuracy of the model was considered too low) and worked just manually. Still, the system was semi-automated since speech recognition worked fully automatically. Emoty proposed the emotions to express in a random order to reduce the biases attributed to the experimental procedures. Also, the sentences to verbalize came up in random order and without repetition within the same session.

4.5 Data gathering

During each session, Emoty automatically collected some quantitative measures. As explained in Section 4.6, we used these quantitative measures to compute some variables and address the different research questions and hypotheses. Measures include:

- *onboarding utterances*, that is the number of utterances pronounced by the user during onboarding;
- *activity utterances*, that is the number of utterances by the user during the acting activity;
- *onboarding time*, that is the amount of time spent by the user to complete the onboarding;
- *activity time*, that is the amount of time spent by the user to perform the acting activity;
- *tasks*, that is the number of sentence-emotion pairs acted out by the user;
- *attempts(task)*, that is the number of attempts made by the user for the specific acting task;
- *success(task)*, that is 1 if the specified task has been successfully completed (i.e., Emoty has recognized the requested emotion) and 0 otherwise;
- *onboarding low quality utterances*, that is the number of "low quality" sentences pronounced by the user during onboarding, i.e., the sentences that were misunderstood or not recognized as speech by the automatic speech recognition system and triggered the

default fallback intent by Emoty ("I am sorry, I didn't understand that, can you repeat, please?");

- *activity low quality utterances*, that is the number of "low quality" sentences pronounced by the user during the acting activity, i.e., the sentences that have been misunderstood or not recognized as speech by the automatic speech recognition system.

During the entire session, the test observer took notes about interaction challenges, requests for help, commentaries, observations aloud by participants, and reports by the facilitator. At the end of each session, the test observer asked the facilitator to make a brief verbal report on her feelings about the participants' behaviors while interacting with Emoty. At the end of the experiment, the facilitator was also asked to comment on the perceived potential of adopting Emoty in regular interventions and list any possible user experience improvements from an expert perspective. We did not interview participants directly because question-based approaches involve critical and self-critical skills [47] which are often lacking in people with NDD [2].

4.6 Data analysis

To verify our hypotheses and gain a complete view of how participants with NDD interacted with Emoty during the sessions, we conducted qualitative and quantitative data analyses. We analyzed the qualitative and quantitative data separately, but we discussed them together.

For qualitative data, we merged a top-down approach and a bottom-up approach to group the notes and observations collected by the test observer [13, 22]. In the first stage, we divided the data into two categories based on the topic. Topics were set a priori and were related to our research questions: (RQ1) usability and (RQ2) effectiveness supporting people with NDD. In the second stage, we organized the data of each category into subcategories according to more specific sub-topics. In this case, the sub-topics were not set a priori but were obtained starting with simple themes and gradually imposing meanings and connections in an inductive manner. For example, observations about the impact of colors and emojis on the user experience could be included in the category "usability" and in the sub-category "visual support".

For quantitative data, we used the measures automatically collected by Emoty. We tailored some research variables on the agent and the characteristics of the users to numerically describe the interaction with the technology and obtain insights about its (RQ1) usability and (RQ2) effectiveness (see Table 2).

We compared all variables in sessions 1 and 5 to describe how participants interacted with Emoty and their acting performance at the beginning and at the end of the study. Moreover, when applicable, we compared variables in the onboarding phase and in the acting activity to study the system's usability at different stages. Finally, we compared all variables of the participants with NDD during sessions 1 and 5 against the variables of the control user in a single session. Statistical tests were performed using the IBM SPSS [52]. The results of the analyses are presented in the next section.

5 Main results

Participants P12 and P16 attended all five sessions but performed the onboarding phase only, without proceeding to the acting activity. Indeed, during the three onboarding questions, the severe language impairments of these two participants prevented them from being

Table 2 The variables used to address the different hypotheses

Research question	Hypotheses	Variable	Formula
RQ1	H1a, H1b	total utterances over total time	$\frac{onboarding\ utterances + activity\ utterances}{onboarding\ time + activity\ time}$
		onboarding utterances over onboarding time	$\frac{onboarding\ utterances}{onboarding\ time}$
		activity utterances over activity time	$\frac{activity\ utterances}{activity\ time}$
		onboarding utterances over total utterances	$\frac{onboarding\ utterances}{onboarding\ utterances + activity\ utterances}$
		onboarding time over total time	$\frac{onboarding\ time}{onboarding\ time + activity\ time}$
		tasks over total time	$\frac{tasks}{onboarding\ time + activity\ time}$
		attempts over total time	$\frac{\sum attempts(task)}{onboarding\ time + activity\ time}$
	H1c, H1d	*total low quality utterances over total utterances*	$\frac{onboarding\ low\ quality\ utterances + activity\ lowquality\ utterances}{onboarding\ utterances + activity\ utterances}$
		onboarding low quality utterances over onboarding utterances	$\frac{onboarding\ low\ quality\ utterances}{onboarding\ utterances}$
RQ2	H2	*acting performance*	$\frac{\sum \frac{success(task)}{attempts(task)}}{max\ tasks\ number}$

understood by the automatic speech recognition system. For this reason, the data about P12 and P16 were not considered in the data analysis; the results reported in the rest of this section concern only the other 17 participants.

5.1 Qualitative results

Overall, the facilitators declared that they would be happy to introduce Emoty during regular therapy despite some difficulties that persons with NDD encountered in the use of Emoty. The collected notes highlighted several insights related to the usability and effectiveness of Emoty and provided design suggestions to be addressed. We classified the qualitative findings by sub-themes as follows.

5.1.1 Usability

Misunderstanding issue Many participants struggled to make themselves understood by Emoty. In some cases, this happened because participants did not address the question they were asked or spoke about subjects outside the understandable domain of the agent. This occurred, for example, when Emoty asked the user how they were doing. P19 continued to repeat "ciao" (in English, "hello"), ignoring the agent's question until she got tired of not being understood and asked the facilitator to leave the session. Other participants (P3, P2, and P8) told Emoty about personal events and thoughts (for example, the fear of the dentist the next day, how pleasant the Easter holiday was, the nephew's birth, the defeat of the favorite team). In other cases, participants were not understood by the agent because they tried to interact with it by exploiting nonverbal communication channels. For example, P14 moved her finger and head instead of saying "no" with her voice. In many cases, the users responded verbally and consistently with the conversational context, but Emoty did not understand them because of their difficulty articulating words and structuring sentences. Still, all participants (including P12 and P16, who did not proceed beyond the onboarding stage) enjoyed interacting with the agent despite not always being understood. According to the facilitators, the limitations of Emoty in automatic speech recognition could be considered an opportunity to stimulate persons with NND to focus on voice communication and improve their pronunciation instead of compensating for speech impairments by using other communication channels. For example, P14 and P15 repeated the same sentence several times in a row, trying to make themselves understood by Emoty by putting effort into improving their pronunciation. P7 was self-critical when the agent did not understand her and, during a session, said aloud: "I should try to speak slowly and use a simpler language to make me understood by Emoty". P10 had some difficulty pronouncing the S, and, consequently, was not understood by the system every time she said: "sì" ("yes" in Italian). After several attempts, she finally understood that she could express the same meaning in alternative terminology (for example, by saying "certo", which in Italian means "of course"). Afterward, she said happily to the facilitator: "It took me a while to understand how to speak to Emoty, but now it is fine!". The same attitude of looking for alternative expressions to communicate with Emoty was also observed in P18. Sometimes, the conversation was stuck because of continuous misunderstandings, and facilitators wished they could move the conversation forward to prevent any frustration in the user.

The facilitators suggested implementing a way for them to control the flow of the conversation in real-time and force the responses to be understood by the agent, similar to how they did with the emotions recognized on behalf of the agent during the acting activity.

Wake action Participants perceived as complex the practice of clicking on the digital button before speaking. During the first session, only P4 could interact with the agent autonomously. Other participants were helped by facilitators, who mediated the interaction by triggering the agent on their behalf. Some people, for example P3, struggled to understand that they had to speak only after clicking on the button and started speaking before clicking, ending up not being understood by Emoty. P13 and P18 could not control the mouse well because of their reduced motor capabilities. For the others, the facilitators supposed that this usability issue was probably related to the fact that clicking is completely uncorrelated from speaking. For some participants, the interaction with Emoty became smoother session by session, and seven out of 17 participants interacted autonomously during the last session.

The facilitators suggested a different wake-up action to be implemented in the future to overcome the barriers observed during this study and meet users' special needs. Instead of clicking on the button at every conversational step, both F1 and F2 would prefer to trigger the agent just once at the beginning of the interaction and let it active for a while, enabling free-form speech.

Visual support The facilitators valued the presence of visual support for the participants. They found it effective to communicate the system's status and provide participants with reinforcements, feedback, and hints about the ongoing activity via emojis and colors. Notably, emojis and colors did not even have to be explained to participants because their meaning seemed clear right away. F1 reported that "for some participants, voice-only communication may not be enough to get their attention in the mid- to long-term." Once happened that P3 got stuck in staring into the distance during the session and returned to focus on Emoty after seeing something changing on the screen.

F2 suggested that "especially for users with a mild cognitive impairment, Emoty should have less basic and more animated visual communication to keep the user focused on the agent".

5.1.2 Effectiveness

Emotional training The facilitators positively evaluated Emoty as a tool for emotional training for participants with NDD, especially because participants accepted Emoty's feedback. In general, all participants showed weaknesses in expressing emotions with their voices. Some subjects initially could not modulate their voice to express emotions and read the sentence on the screen without any pathos. Other people, like P7 and P2, included the name of the emotion into the sentence they were asked to repeat instead of verbalizing the sentence with drama. P6 did not modulate his voice but appended words to the proposed sentence that were semantically consistent with the given emotion (for example, "help" when the emotion was fear). P15 could not read the sentence and, at the same time, express an emotional tone: he modulated his voice to express the emotion by pronouncing non-words. Session after session, participants began using tricks to express emotions with their voices. For example, P3 and P10 added "Ah" and "Oh" to the given sentences. P8 and P18 acted with their voices and helped themselves with gestures and facial expressions. P8 said that at the beginning, she was shy about acting. Then, she imagined being in a beautiful place with her boyfriend and managed to act out a sentence in a joyful way. P10 said she liked acting with Emoty because she felt safe; if she had to do it in front of other people, she would be afraid of making mistakes. Before acting out the sentence for Emoty, she thought on her own about situations that reminded her of the required emotion. In addition, she said that since she met Emoty, she has been more careful about people's emotions and the emoji she uses in her messages on WhatsApp. P17 said that acting was challenging for him, and he was pleased when he was able to perform the asked emotion appropriately.

The facilitators suggested that for some users may be necessary to work first on emotion recognition and then on production. Therefore, they proposed adding a new activity in which Emoty plays some emotional speech recordings by some actors and asks the user to recognize the emotion they express with the voice. The facilitators also pointed out that it might be necessary for some individuals to work on specific emotions. Therefore, they wished to customize the user experience with Emoty for each user.

Conversational experience The facilitators positively commented on the choice to structure the experience as a gamified conversation because participants were generally engaged during the sessions. They also liked Emoty as a stimulus for a broader range of communication skills beyond emotional ones. F2 said: "Although I guess that in the same amount of time of a session with Emoty, many more emotional tasks may be performed by users with a different interaction paradigm (for example, with a touch-based application), I appreciated Emoty because I believe it has the potential to help users improving many communication and social skills in a safe and controlled context: respecting turn-taking, rephrasing concepts to be better understood, and articulating and pronouncing words clearly". In addition, facilitators liked the style of the dialogs with Emoty because they included many repetitions of the main concepts, and the user could always request to repeat the last sentence said by the agent.

Kolb's cycle The repetitive structure of the dialogs supported by Emoty was partially a consequence of the underlying structured stages of Kolb's cycle. The facilitators generally appreciated the four steps of the empirical learning process. F2 liked the *application* phase of the cycle because it allowed participants to connect what they learned during the acting activity to real-life scenarios. For example, when Emoty reported that "joy is the emotion you get when mom bakes your favorite dessert", P17 commented aloud, "Yes, it is true!" F1 stated: "The *reflection* phase of Kolb's cycle was useful to stimulate the self-critical capacity of some participants". For example, P9 and P2 responded several times that they thought they could have acted better in the proposed sentence. As a drawback, F1 also pointed out: "For some individuals with NDD *reflection* stage might be too difficult."

Consequently, F1 suggested implementing a way for facilitators to skip the *reflection* stage and directly move from experience to *conceptualization* when necessary.

5.2 Quantitative results

On average, participants played for 660 seconds per session (*SD* = 228), including 160 seconds to perform onboarding (*SD* = 122). Participants uttered 25.153 sentences (*SD* = 11.620), including 8.224 sentences during onboarding (*SD* = 7.730). They were misunderstood or were not recognized as speech 12.871 times (*SD* = 11.732), including 5.259 times during onboarding (*SD* = 7.815). They performed 2.035 acting tasks (*SD* = 1.052) and 4.553 acting attempts per session (*SD* = 2.228).

From the data automatically gathered by Emoty, we estimated the values of the variables that we used for analysis.

5.2.1 Session 1 vs session 5

Eight Wilcoxon signed-rank tests were used to compare our variables in sessions 1 and 5. The following variables were used: *total utterances over total time*, *tasks over total time*, *attempts over total time*, *onboarding low quality utterances over onboarding utterances*, *total low quality utterances over total utterances*, *onboarding utterances over total utterances*, *onboarding time over total time*, and *acting performance*.

The results are reported in Table 3 and are shown in Fig. 4. For each variable, there was a significant difference between the scores in Session 1 and Session 5.

Overall, the rate of utterances, acting tasks, and acting attempts over time by participants with NDD increased after five sessions with Emoty. Also, the interaction quality between the participants and the agent improved after five sessions, as they were understood more often than in the first session (both specifically during onboarding and in general). Both the number of utterances produced by the participants during onboarding over the total number of utterances and the amount of time spent for onboarding over the total amount of time decreased after five sessions with our conversational agent. This means that after five sessions with Emoty users proportionally dedicated more utterances and time to the acting activity. Finally, after five sessions with Emoty the participants with NDD learned how to perform better in the acting activity.

5.2.2 Onboarding vs acting activity

We ran five Wilcoxon signed rank tests to compare session by session the rate of sentences over time between the on-boarding and the acting activity. The variables we used were *onboarding utterances over onboarding time* and *activity utterances over activity time*.

For each session, there was a significant difference between the scores given for the utterances over time in the onboarding phase compared with the same rate during the acting activity (as reported in Table 4). In conclusion, *the results show that the rate of interactions during onboarding was higher than that during the acting activity* (see Fig. 5).

Table 3 Comparison of the variables in session 1 and in session 5

	Session 1		Session 5					
Variable	*M*	*SD*	*M*	*SD*	n	Z	p	
total utterances over total time	0.034	0.011	0.043	0.014	17	–2.675	0.007	**
tasks over total time	0.002	0.001	0.004	0.002	17	–2.947	0.003	**
attempts over total time	0.002	0.001	0.008	0.004	17	–2.223	0.026	*
onboarding low quality utterances over onboarding utterances	0.559	0.240	0.408	0.308	17	–2.718	0.007	**
total low quality utterances over total utterances	0.506	0.251	0.400	0.275	17	–2.296	0.022	*
onboarding utterances over total utterances	0.464	0.271	0.335	0.269	17	–2.223	0.026	*
onboarding time over total time	0.409	0.268	0.291	0.276	17	–2.430	0.015	*
acting performance	0.042	0.074	0.299	0.264	17	–3.112	0.002	**

*p<0.05, **p<0.01

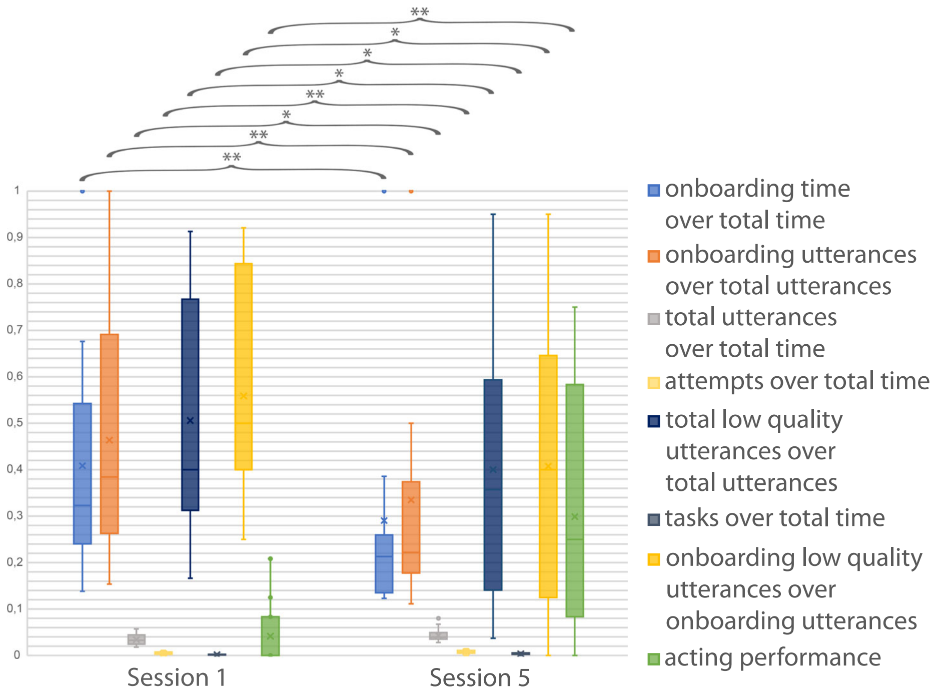

Fig. 4 Comparison of the variables in session 1 and in session 5. The bottom and top of each box represent respectively the first and third quartile of the data. The horizontal solid line represents the median value. The cross marker represents the mean value. Full dots represent outliers. The whiskers represent the maximum and minimum values

5.2.3 Participants with NDD vs control user

The control user interacted with Emoty for 219 seconds (including 51 seconds to perform onboarding), pronouncing 11 total sentences (including 3 sentences for onboarding). The automatic speech recognition system always comprehended the user's sentences without misunderstanding. During the acting activity, the user performed four sentences expressing

Table 4 The comparison of the mean (*M*) and standard deviation (*SD*) of the utterances over time produced by the users during the onboarding phase and the acting activity for the five sessions (S1-S5)

	onboarding utterances over onboarding time		activity utterances over activity time				
Session	*M*	*SD*	*M*	*SD*	n	Z	p
S1	0.039	0.012	0.026	0.014	17	–2.817	0.005 **
S2	0.047	0.024	0.030	0.013	17	–2.296	0.022 *
S3	0.059	0.013	0.031	.010	17	–3.621	0.000 ***
S4	0.052	0.016	0.037	.013	17	–3.154	0.002 **
S5	0.053	0.018	0.033	0.015	17	–2.580	0.010 *

*p<0.05, **p<0.01, ***p<0.001

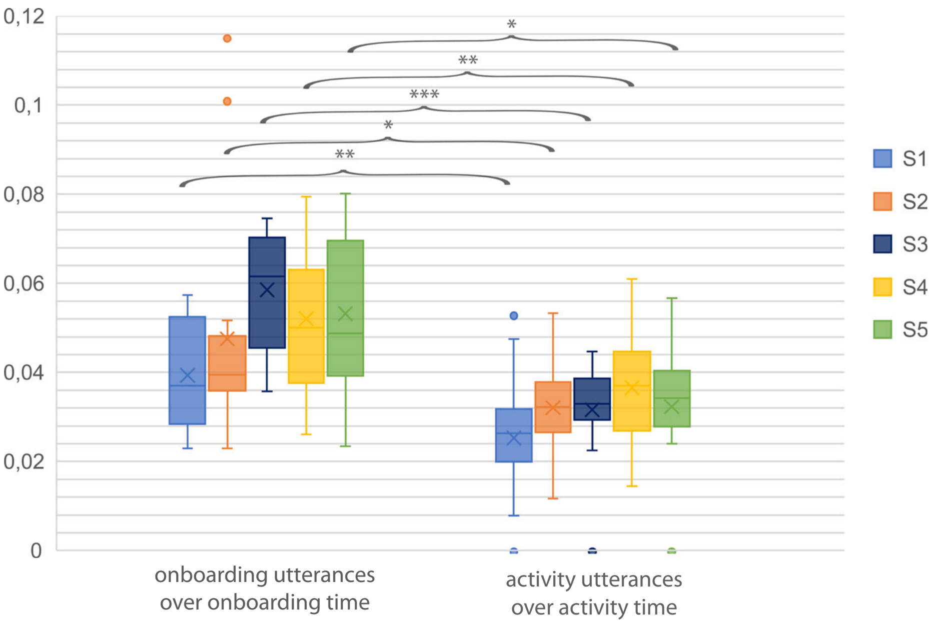

Fig. 5 Comparison of the rate of sentences over time between onboarding and the acting activity. *p < 0.05, **p < 0.01, ***p < 0.001. The bottom and top of each box represent respectively the first and third quartile of the data. The horizontal solid line represents the median value. The cross marker represents the mean value. Full dots represent outliers. The whiskers represent the maximum and minimum values

anger, surprise, fear, and sadness. The system always recognized the correct emotion in the first attempt.

Ten one-sample Wilcoxon signed-rank tests were used to compare some variables describing the sessions by participants with those describing the session by the control user. The variables we used were *total utterances over total time*, *attempts over total time*, *tasks over total time*, *total low quality utterances over total utterances*, and *acting performance*.

As reported in Table 5, for each variable, there was a significant difference between the scores referring to the participants in session 1 and session 5 and those referring to the control user. As depicted respectively in Fig. 6A, B, and C, *the rates of utterances, acting attempts, and acting tasks by participants with NDD were lower than those of the neurotypical control user in both the first and last sessions with Emoty.* Also, *participants with NDD had more difficulty than the control user in being understood by the automatic speech recognition system (see* Fig. 6D*).* Finally, as shown in Fig. 6E, *the acting performance of the control user was higher than that of the population.*

6 Discussion

We discuss the findings of the study with respect to our research questions (RQ1 and RQ2) and research hypotheses (H1a, H1b, H1c, H1d, and H2).

Table 5 Comparison of the variables to the baseline by the control user

	Control user	Participants with NDD						
Variable	Score	Session	*M*	*SD*	n	Z	p	
total utterances over total time	0.050	1	0.034	0.011	17	–3.385	0.001	***
		5	0.043	0.014	17	–1.965	0.049	*
tasks over total time	0.020	1	0.002	0.001	17	–3.622	0.000	***
		5	0.004	0.002	17	–3.622	0.000	***
attempts over total time	0.020	1	0.002	0.001	17	–3.622	0.000	***
		5	0.008	0.004	17	–3.622	0.000	***
total low quality utterances over total utterances	0.000	1	0.506	0.251	17	3.622	0.000	***
		5	0.400	0.275	17	3.622	0.000	***
acting performance	1.000	1	0.042	0.074	17	–3.628	0.000	***
		5	0.299	0.264	17	–3.628	0.000	***

*p<0.05, ***p<0.001

6.1 [RQ1] Is Emoty usable by people with NDD?

[H1a - Participants can easily interact with Emoty since their first usage] Rates of sentences, tasks, and attempts over time by participants substantially differed from those of the control user, indicating that, at first glance, people with NDD found it problematic to interact with Emoty.

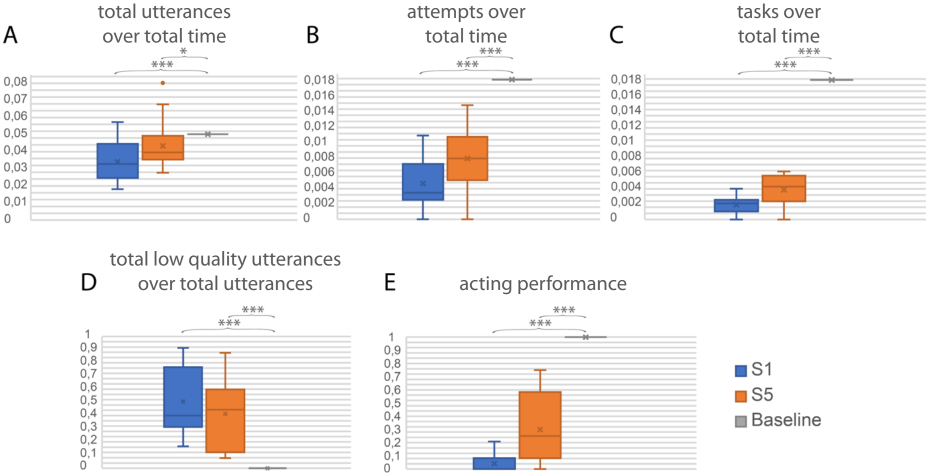

Fig. 6 Comparison of the variables to the baseline by the control user. *p<0.05, ***p<0.001. The bottom and top of each box represent respectively the first and third quartile of the data. The horizontal solid line represents the median value. The cross marker represents the mean value. Full dots represent outliers. The whiskers represent the maximum and minimum values

The first obstacle to using Emoty in autonomy was the wake action: most participants could not point and click the button with the mouse before speaking. For this reason, facilitators had to assist them in the interaction with Emoty, clicking on the button for them. Interaction designers for NDD might consider using different wake actions based on the special needs of each specific population of users. Alternatives could be of a different nature (vocal, tactile, visual, event-based, and motion-based). According to the psychologists and therapists involved in this study, the ideal solution would be to interact using free-form speech (no wake action). However, this interaction paradigm is technically complex to develop because it involves message filtering and turn-taking capabilities by the machine [17]. Another paper [17] proposed the use of identical actions both to wake up and put to sleep the agent and provided a theoretical argument based on the theory of *partner-perceived communication* [24, 56]. This theory states that predictability and repetitiveness facilitate giving meaning to sentences even for people with complex communication needs. In the literature, there is no other material about the best wake and sleep actions for people with NDD, and a study urges to be performed. Previous research on neurotypical children [17] suggested the use of a physical button to wake up and put to sleep the conversational agent. This finding might be the starting point for an exploratory study on NDD. To ensure completeness, it must be considered that the use of a physical button has the significant disadvantage of making web-based conversational agents such as Emoty less scalable because it requires specific hardware to work.

[H1b - Interaction with Emoty becomes easier over time] Despite the initial challenges, the interaction with Emoty became smoother over time. The sentences, tasks, and attempts over time increased considerably from session 1 to session 5. Some participants learned to use the required wake action autonomously. In addition, the relative number of sentences and the relative amount of time users dedicated to onboarding generally decreased after five sessions with Emoty. This means that proportionally more time and sentences were dedicated to the most crucial part of the session: the acting activity.

[H1c - Users with NDD are not always understood by Emoty] The interaction with Emoty was generally limited by the automatic speech recognition system integrated into the agent, which was not sufficiently accurate, especially for users with some speech impairments. Psychologists and therapists did not consider misunderstandings as totally negative because they pushed participants to pronounce sentences better. The same optimistic vision was shared in a previous study on Google Home and children with autism [18]. However, we understand from another study with Amazon Echo and children with autism [1] that continuous misunderstanding could lead to frustration for the user, making the use of conversational agents ineffective. Consequently, beyond improving the conversational technology's automatic transcription and comprehension capabilities, agents for NDD might also offer therapists and caregivers the possibility to be controlled in real-time whenever they want during the session. Emoty might have a side-application for facilitators to prevent any possible misunderstanding by manually recognizing the sentences uttered by the user without even running any automatic transcription or natural language understanding analysis. In addition, we agree with the authors of [8] that a conversational agent for NDD should provide more specific communication feedback to collaborate on communication repair (the ability to persist in communication and to modify, repeat, or revise a signal when the initial communication attempt fails). Indeed, as already pinpointed in [30], it would be helpful for the user and the facilitator to receive a clear explanation of the reason why the agent is unable to answer appropriately. This way, if the issue becomes clear, the facilitator can help

the user employ different communication repair strategies depending on the context, such as by defining terms [8].

In our study, the web application screen played a crucial role in supporting verbal communication. It effectively provided visual feedback to the user about how the agent was functioning, presenting the tasks to be completed, and reinforcing concepts conveyed by voice. Our findings are in line with the literature, which reports that various stimuli can support people in enhancing verbal communication abilities [21]. The VAK model [6] identifies three learning modalities:

- Visual learning, which exploits graphs, charts, maps, diagrams, pictures, paintings, and other kinds of visual stimulation;
- Auditory learning, which depends on listening and speaking;
- Kinesthetic learning, which requires gestures, body movements, and object manipulation to process new information.

The authors of [5] stated that learners generally appear to benefit from mixed modality presentations, for instance, using both auditory and visual techniques as in the case of Emoty.

From our study, conversation designers can gain insights into the communicative effectiveness of emojis and colors associated with emotions in people with NDD. In addition, therapists and psychologists have stressed the importance of the screen to maintain the attention level of users with NDD. A previous study [81] suggested the use of a virtual character as the best embodiment for a conversational agent interacting with people with NDD. Therefore, in the future, we intend to increase the animated content on the screen using a virtual character anthropomorphizing Emoty.

When we compared the behavior of the users during the two phases of each session with Emoty (the onboarding and the acting activity), we noticed that the rate of interactions during the first phase was generally higher than during the latter. We hypothesize that the task demanded during onboarding (answering three questions for welcoming and introducing the session) is more accessible than the acting task required in the second phase. About the acting activity, facilitators appreciated that the experience with Emoty was structured as gamified and that it followed the principles of empirical learning. Some participants found the reflection phase of Kolb's cycle challenging because they did not have the required self-critical skills to evaluate their performance or cognitive skills to understand what they were asked to do. Although reflection is a fundamental step in the empirical learning process, psychologists and therapists have suggested that it could be skipped in some cases or with some particular users. Consequently, to make the experience with Emoty more dynamic and customizable, the next version of our agent will allow facilitators to skip the reflection phase of Kolb's cycle. Without Kolb's reflection step, dialogs still have a basis in the literature and follow the guidelines in [85], who proposed activity cycles for people with NND including progress, feedback, and generalization of the skills learned during the session to daily life.

[H1d - Misunderstandings decrease over time] Misunderstandings decreased over time. This means that the quality of interaction between the participants and the agent generally improved after five sessions. Some participants developed communication strategies to make themselves understood by the conversational agent, such as using certain words that were more likely to be heard. Others realized that they were better understood by pronouncing and articulating words better. Improvements were most evident during the onboarding phase, probably because participants were focused only on making themselves understood and not on other aspects of communication such as voice intonation.

6.2 [RQ2]Is Emoty effective in helping people with NDD improve their expression of emotions with the voice?

[H2 - Users improve over time their ability to express emotions with the voice] This study confirmed the weakness of people with NDD in terms of emotional skills [2] and highlighted the need to train them to express emotions with the voice. The acting performance of the participants was largely lower than that of the control user, and some participants acknowledged that they experienced difficulty while acting.

Nevertheless, participants' acting performance substantially increased from the first to the last session with Emoty, demonstrating the potential of the agent to stimulate the emotion expression skills of people with NDD. Our agent created a safe environment for participants who interacted without fear of making mistakes or being judged. However, we are aware that the results obtained with the conversational agent do not necessarily reflect an improvement in expressing emotions when communicating with other people or even an increased awareness of one's emotional sphere. Indeed, our findings may reveal just an upgrade of the performance in the acting activity within the sessions with the technology [40]. One participant spontaneously reported that since she met Emoty, she pays more attention to other people's emotions and the emojis she puts in her WhatsApp messages. However, we did not collect any comments from psychologists or therapists at daycare centers that support this progress in daily life in her or other participants. To further investigate the effect of the usage of Emoty on the emotional sphere of the users, we plan a follow-up experiment; we want to measure participants' emotional capabilities before and after an extended period of usage of our agent. The gold standard measures of emotional skills, which are generally referred to as *emotional intelligence* [62], are the tests called MSCEIT [61] and MSCEIT-YV [74]. Unfortunately, these methods are limited to assessing facial emotion recognition and do not rely on speech emotion recognition or expression. In addition, we know that generally, for emotion recognition, there are differences according to the cue type, with facial expression better recognized than speech [32]. For this reason, we will need to integrate test results with observations by people close to participants (psychologists, therapists, teachers, and parents) to address our research question comprehensively. Moreover, we plan to administer the TAS-20 [4] questionnaire to participants in the following study to assess their level of alexithymia with multiple measures as suggested in [58].

6.3 Limitations of the study

Our research brings up a few limitations.

The sample we were able to include in the study (19 people with NDD) was small and heterogeneous in terms of age and cognitive impairment level, potentially impacting the generalization of the findings for the whole population. Still, it should be noticed that the sample size of our study is wider than previous ones involving people with NDD and conversational agents for emotional training (M = 11.14, SD = 10.04 as reported in Section 2). As a consequence of the tiny sample size, the results of our statistical tests are to be considered preliminary and require further investigation to validate them.

In addition, the fact that the current study participants did not have an official diagnosis of alexithymia but were only observed to have traits associated with it might be considered a limitation. However, we know that it is common for people with NDD not to be adequately diagnosed with alexithymia because of the difficulty of performing a proper assessment or because other comorbid conditions overshadow it [11, 58]. To overcome this limitation, in the future, we will extend the research to a broader population, and a precise diagnosis of alexithymia will be a strict inclusion criterion.

The study took place in a limited period (five sessions in two and a half months), and, consequently, only preliminary insights emerged and not a complete overview of the valuable features for the adoption of Emoty in a therapeutic context. For example, we are aware that learning how to interact with technology might be a significant challenge for people with special needs, and we cannot exclude the possibility that some of the participants could interact with Emoty more easily if they had more time. In addition, the positive trends found let us wonder if they would persist in longer experimentation. Although five sessions were in line with other existing studies with persons with NDD [70, 82], a longer study could be performed to obtain more reliable results and to analyze post-novelty effects and consolidation of learning.

Regarding the experimental procedure, we are aware that the presence of the test observer in the experimental room could have affected the participants' experience during the sessions, which could be considered a limitation of our study. However, we feel that we took all the proper measures to mitigate this issue (see Section 4.3). In addition, one might criticize the set of emotions we chose for the sessions with Emoty. For future studies, we want to give the facilitator the possibility of customizing the experience for each participant so that they can work on a subset of the Big Six emotions by Ekman or even on a set of more complex emotions (e.g., the wheel of emotions in [68]) depending on the special needs of every single user. Moreover, we might think about variations of the acting activity protocol for some future studies. For instance, we may include the possibility of accessing examples of emotion expressions or imitation-based help from the very beginning of each acting task.

Finally, we tested a semi-automated version of Emoty by exploiting a manually controlled speech emotion recognition module. We plan to assess the accuracy of our automatic emotion recognizer on the audio recordings we collected during this study from people with NDD. In this way, we can check whether the predictions by the system are in line with the feedback that the facilitators gave on the same audio recordings, and we can quantitatively evaluate how ready our model is to enable a fully automated version of Emoty.

7 Conclusion

Emoty exploits conversational technology to mitigate the main effects of alexithymia, i.e., severe impairments in expressing and recognizing emotions, which are very frequent in persons with NDD. The system provides an emotional trainer that exploits the communication potential of the voice and helps people learn to use the vocal tone for emotion expression.

To gain insights into the usability of Emoty among people with NDD and its effectiveness as emotional trainer, we performed an empirical study involving 19 people with mild, moderate, and severe levels of cognitive disability and traits associated with alexithymia. Participants were asked to interact with Emoty during five sessions and verbalize the assigned sentences expressing the given emotions with the voice. They received real-time feedback on their acting performance and were guided into an empirical learning process according to the steps of Kolb's cycle.

Concerning usability, our findings indicate that it was not easy for persons with NDD to interact with Emoty since the very first usage, but the interaction with the system became more straightforward as users acquired familiarity: 17 of the 19 participants could perform all Emoty activities with progressively decreasing help from the facilitator. Caregivers ascribed the initial difficulty of using Emoty to two main factors: (i) the persons with NDD were often misunderstood by the automatic speech recognition system; (ii) they found it unnatural to click on a button every time they wanted to speak.

Concerning effectiveness, the study results suggest that Emoty has the potential of bringing some benefits and complementing traditional interventions for people with NDD. Participants initially had poor acting performance, manifesting their weakness in emotion expression, but, in general, they progressively improved while interacting with Emoty. Our study also indicate an improvement of other capabilities during the experience with Emoty, regarding the ability of respecting turn-taking, rephrasing concepts to be better understood, pronouncing words, and articulating sentences clearly. Still, we have not measured any improvement in emotion expression and other communication skills generalized in contexts outside the study sessions, e.g., while communicating with other people in every day life contexts. We plan a follow-up study to address this issue, assessing participants' communication capabilities concern emotion expression, and beyond, in multiple ways: with a standardized, validated test before and after an extended period of usage of Emoty, and questionnaires with participants' parents and their caregivers working with them day by day.

Despite its weaknesses, the empirical study reported in the paper advances our knowledge of conversational agents for people with NDD, and paves the way for a better understanding of the cognitive and emotional mechanisms that come into play when this population interacts with natural language-based interfaces. Indeed, a number of lessons emerged from our research that can benefit designers, developers, and researchers in the field.

First, conversational agents for persons with NDD should provide new and more intuitive wake actions for conversational agents. Further research is required to explore different design solutions for notifying the user's intention to speak, to better address the characteristics of users with special needs, and also to investigate different interaction paradigms beyond digital buttons (for example, tangible interaction with physical buttons [17]).

The second lesson is the need for conversational agents for NDD to provide user-specific communication feedback to collaborate on communication repair when the initial communication attempt fails. This might be addressed by enhancing the conversational agent with knowledge of the communication repair approaches used in speech therapy.

The third and perhaps most challenging lesson is the need to improve conversational agents' transcription and comprehension capability to take into account the linguistic limitations of persons with NDD. This would require the creation of larger, emotionally tagged speech datasets generated by neurotypical persons and persons with NDD to train a machine learning model for speech classification. To address this issue, we are currently collecting tagged speech-based sentences via crowdsourcing (see https://emozionalmente.i3lab.group/).

Acknowledgments The authors would like to thank all therapists from the center "Fraternità e Amizicia" who contributed to the co-design process of Emoty, and all persons from the centers "L'Impronta" and "Collage" who participated in the empirical study.

Author Contributions F. C. and F. G. conceived and designed research. Material preparation, data collection and analysis were performed by F.C. The first draft of the manuscript was written by F. C. and all authors commented on previous versions of the manuscript. All authors read and approved the final manuscript.

Funding Open access funding provided by Politecnico di Milano within the CRUI-CARE Agreement. This research work was not funded by any organization or supported by funding schemes.

Data Availability The datasets generated and analyzed during the current study are available from the corresponding authors on reasonable request.

Declarations

Ethics approval The study protocol was approved by the Ethics Committee of Politecnico di Milano.

Conflict of Interests The authors declare no conflict of interest, financial or otherwise.

References

1. Allen AA, Shane HC, Schlosser RW (2018) The echo™as a speaker-independent speech recognition device to support children with autism: an exploratory study. Adv Neurodev Disord 2(1):69–74
2. American Psychiatric Association et al (2013) Diagnostic and statistical manual of mental disorders (dsm-5®). American Psychiatric Pub
3. Askari F (2018) Studying facial expression recognition and imitation ability of children with autism spectrum disorder in interaction with a social robot. University of Denver
4. Bagby RM et al (1994) The twenty-item Toronto Alexithymia Scale-II. Convergent, discriminant, and concurrent validity. J Psychosomatic Res 38(1):33–40
5. Barbe WB, Milone MN Jr (1981) What we know about modality strengths. Educ Leadersh 38(5):378–80
6. Barbe WB, Milone MN, Swassing RH (1988) Teaching through modality strengths: Concepts and practices. Zaner-Bloser
7. Bekele E, Wade J, Bian D, Fan J, Swanson A, Warren Z, Sarkar N (2016) Multimodal adaptive social interaction in virtual environment (masi-vr) for children with autism spectrum disorders (asd). In: 2016 IEEE virtual reality (VR). IEEE, pp 121–130
8. Beneteau E, Richards OK, Zhang M, Kientz JA, Yip J, Hiniker A (2019) Communication breakdowns between families and alexa. In: Proceedings of the 2019 CHI conference on human factors in computing systems CHI '19. Association for Computing Machinery, New York, pp 1–13
9. Berlo DK (1987) El proceso de la comunicación: introducción a la teoría ya la práctica. In: El proceso de la comunicación: introducción a la teoría ya la práctica. Editorial El Ateneo, pp 173–173
10. Bernardini S, Porayska-Pomsta K, Smith TJ (2014) Echoes: an intelligent serious game for fostering social communication in children with autism. Inf Sci 264:41–60
11. Berthoz S, Hill EL (2005) The validity of using self-reports to assess emotion regulation abilities in adults with autism spectrum disorder. European Psychiatry 20(3):291–298
12. Boyd-Graber JL, Nikolova SS, Moffatt KA, Kin KC, Lee JY, Mackey LW, Tremaine MM, Klawe MM (2006) Participatory design with proxies: developing a desktop-pda system to support people with aphasia. In: Proceedings of the SIGCHI conference on human factors in computing systems, pp 151–160
13. Braun V, Clarke V (2006) Using thematic analysis in psychology. Qualitative Research in Psychology 3(2):77–101
14. Car LT, Dhinagaran DA, Kyaw BM, Kowatsch T, Rayhan JS, Theng YL, Atun R (2020) Conversational agents in health care: scoping review and conceptual analysis. J Med Int Res 22:8
15. Catania F, Beccaluva E, Garzotto F (2019) The conversational agent "emoty"? perceived by people with neurodevelopmental disorders: Is it a human or a machine? In: International workshop on chatbot research and design. Springer, pp 65–78
16. Catania F, Di Nardo N, Garzotto F, Occhiuto D (2019) Emoty: an emotionally sensitive conversational agent for people with neurodevelopmental disorders. In: Proceedings of the 52nd Hawaii international conference on system sciences, pp 2014–2023

17. Catania F, Spitale M, Cosentino G, Garzotto F (2020) What is the best action for children to “wake up” and “put to sleep” a conversational agent? a multi-criteria decision analysis approach. In: Proceedings of the 2nd conference on conversational user interfaces, pp 1–10
18. Catania F, Spitale M, Garzotto F (2021) Toward the introduction of google assistant in therapy for children with neurodevelopmental disorders: An exploratory study. In: Extended abstracts of the 2021 CHI conference on human factors in computing systems, pp 1–7
19. Catania F, Spitale M, Garzotto F (2022) Conversational agents in therapeutic interventions for neurodevelopmental disorders: a survey. ACM Comput Surv. https://doi.org/10.1145/3564269, Just Accepted
20. Chevalier P, Li JJ, Ainger E, Alcorn AM, Babovic S, Charisi V, Petrovic S, Schadenberg BR, Pellicano E, Evers V (2017) Dialogue design for a robot-based face-mirroring game to engage autistic children with emotional expressions. In: International conference on social robotics. Springer, pp 546–555
21. Coffield EK, Moseley HE, Hall E, Ecclestone K (2004) Learning styles and pedagogy in post 16 education: a critical and systematic review. Learning and Skills Research Centre, London
22. Cooper HE, Camic PM, Long DL, Panter AT, Rindskopf DE, Sher KJ (2012) Apa handbook of research methods in psychology, vol 2: Research designs: Quantitative, qualitative, neuropsychological, and biological. American Psychological Association
23. Costantini G, Iaderola I, Paoloni A, Todisco M (2014) Emovo corpus: an italian emotional speech database. In: International conference on language resources and evaluation (LREC 2014). European Language Resources Association (ELRA), pp 3501–3504
24. Costantino MA (2011) Costruire libri e storie con la caa: gli in-books per l’intervento precoce e l’inclusione. Erickson
25. Da Silva ACN, Vasco AB, Watson JC (2018) Alexithymia and therapeutic alliance: a multiple case study comparing good and poor outcome cases. Research in Psychotherapy: Psychopathology, Process, and Outcome 21:2
26. Derks D, Bos ArjanER, Von Grumbkow J (2008) Emoticons and online message interpretation. Soc Sci Comput Rev 26(3):379–388
27. Deterding S, Dixon D, Khaled R, Nacke L (2011) From game design elements to gamefulness: defining “gamification”. In: Proceedings of the 15th international academic MindTrek conference: envisioning future media environments, pp 9–15
28. Dickerson P, Robins B, Dautenhahn K (2013) Where the action is: a conversation analytic perspective on interaction between a humanoid robot, a co-present adult and a child with an asd. Interact Stud 14(2):297–316
29. Diederich S, Brendel AB, Kolbe LM (2019) On conversational agents in information systems research: Analyzing the past to guide future work. In: Wirtschaftsinformatik, pp 1550–1564
30. Druga S, Williams R, Breazeal C, Resnick M (2017) Hey google, is it ok if i eat you?: initial explorations in child-agent interaction. In: Proceedings of the 2017 conference on interaction design and children, pp 595–600
31. Ekman P (1999) Basic emotions. Handb Cogn Emot 98(45–60):16
32. Elfenbein HA, Jang D, Sharma S, Sanchez-Burks J (2017) Validating emotional attention regulation as a component of emotional intelligence: a stroop approach to individual differences in tuning in to and out of nonverbal cues. Emotion 17(2):348
33. Epifânio JC, Da Silva LF (2020) Scrutinizing reviews on computer science technologies for autism: issues and challenges. IEEE Access 8:32802–32815
34. Fachantidis N, Syriopoulou-Delli CK, Zygopoulou M (2020) The effectiveness of socially assistive robotics in children with autism spectrum disorder. Int J Develop Disab 66(2):113–121
35. Følstad A, Brandtzaeg PB (2020) Users’ experiences with chatbots: findings from a questionnaire study. Quality and User Experience 5(1):1–14
36. Google Google dialogflow. Accessed 13 June 2022, https://dialogflow.cloud.google.com/
37. Google Google speech to text. Accessed 13 June 2022, https://cloud.google.com/speech-to-text
38. Google Google text to speech. Accessed 13 June 2022, https://cloud.google.com/text-to-speech
39. Granic I, Lobel A, Engels RCME (2014) The benefits of playing video games. American Psychologist 69(1):66
40. Grossard C, Grynspan O, Serret S, Jouen A-L, Bailly K, Cohen D (2017) Serious games to teach social interactions and emotions to individuals with autism spectrum disorders (asd). Comput Educ 113:195–211
41. Hamzah MSJ, Shamsuddin S, Miskam MA, Yussof H, Hashim KS (2014) Development of interaction scenarios based on pre-school curriculum in robotic intervention for children with autism. Procedia Comput Sci 42:214–221
42. Hart SG (2006) Nasa-task load index (nasa-tlx); 20 years later. In: Proceedings of the human factors and ergonomics society annual meeting, vol 50. Sage publications Sage CA, Los Angeles, pp 904–908

43. Hill E et al (2005) The validity of using self-reports to assess emotion regulation abilities in adults with autism spectrum disorder. European Psychiatry: the Journal of the Association of European Psychiatrists 20:291–8
44. Hobert S, von Wolff RM (2019) Say hello to your new automated tutor - a structured literature review on pedagogical conversational agents. In: Wirtschaftsinformatik, pp 1–14
45. Hussain S, Sianaki OA, Ababneh N (2019) A survey on conversational agents/chatbots classification and design techniques. In: Workshops of the International conference on advanced information networking and applications. Springer, pp 946–956
46. Jordan PW, Thomas B, McClelland IL, Weerdmeester B (1996) Usability evaluation in industry. CRC Press
47. Jordan PW, Thomas B, McClelland IL, Weerdmeester B (1996) Usability evaluation in industry. CRC Press
48. Karanchery S, Palaniswamy S (2021) Emotion recognition using one-shot learning for human-computer interactions. In: 2021 International conference on communication, control and information sciences (ICCISc), vol 1. IEEE, pp 1–8
49. Kim Y-D, Hong J-W, Kang W-S, Baek S-S, Lee H-S, An J (2010) Design of robot-assisted observation system for therapy and education of children with autism. In: International conference on social robotics. Springer, pp 222–231
50. Kinnaird E, Stewart C, Tchanturia K (2019) Investigating alexithymia in autism: a systematic review and meta-analysis. Eur Psychiatry 55:80–89
51. Kolb DA (2014) Experiential learning: experience as the source of learning and development. FT Press
52. Kremelberg D (2010) Practical statistics: a quick and easy guide to ibm® spss® statistics, stata, and other statistical software. SAGE publications
53. Laux LF, McNally PR, Paciello MG, Vanderheiden GC (1996) Designing the world wide web for people with disabilities: a user-centered design approach. In: Proceedings of the second annual ACM conference on Assistive technologies, pp 94–101
54. Lewis C, Wharton C (1997) Chapter 30 - cognitive walkthroughs. In: Helander MG, Landauer TK, Prabhu PV (eds) Handbook of human-computer interaction. 2nd edn., North-Holland, pp 717–732
55. Li J, Davison D, Alcorn A, Williams A, Dimitrijevic SB, Petrovic S, Chevalier P, Schadenberg B, Ainger E, Pellicano L et al (2020) Non-participatory user-centered design of accessible teacher-teleoperated robot and tablets for minimally verbal autistic children. In: Proceedings of the 13th ACM international conference on pervasive technologies related to assistive environments, pp 1–9
56. Light J, Drager K (2007) Aac technologies for young children with complex communication needs: State of the science and future research directions. Augmentative and Alternative Communication 23(3):204–216
57. Liu X, Wu Q, Zhao W, Luo X (2017) Technology-facilitated diagnosis and treatment of individuals with autism spectrum disorder: An engineering perspective. Appl Sci 7(10):1051
58. Lumley MA, Neely LC, Burger AJ (2007) The assessment of alexithymia in medical settings: implications for understanding and treating health problems. J Person Assess 89(3):230–246
59. Marchi E, Schuller B, Batliner A, Fridenzon S, Tal S, Golan O (2012) Emotion in the speech of children with autism spectrum conditions: Prosody and everything else. In: Proceedings 3rd Workshop on Child, Computer and Interaction (WOCCI), Satellite Event of INTERSPEECH, pp 1–8
60. Marino F, Chilà P, Sfrazzetto ST, Carrozza C, Crimi I, Failla C, Busà M, Bernava G, Tartarisco G, Vagni D et al (2020) Outcomes of a robot-assisted social-emotional understanding intervention for young children with autism spectrum disorders. Journal of Autism and Developmental Disorders 50(6):1973–1987
61. Mayer JD (2002) Msceit: Mayer - salovey - caruso emotional intelligence test. Toronto, Canada: Multi-Health Systems
62. Mayer JD, Salovey P (1993) The intelligence of emotional intelligence. Intelligence 17(4):433–442
63. Mehrabian A (2017) Communication without words. In: Communication theory. Routledge, pp 193–200
64. Mower E, Black MP, Flores E, Williams M, Narayanan S (2011) Rachel: design of an emotionally targeted interactive agent for children with autism. In: 2011 IEEE International conference on multimedia and expo. IEEE, pp 1–6
65. National Institute of Mental Health (2018) Autism spectrum disorder. Accessed 13 June 2022, https://nimh.nih.gov/health/topics/autism-spectrum-disorders-asd/index.shtml
66. Palestra G, Varni G, Chetouani M, Esposito F (2016) A multimodal and multilevel system for robotics treatment of autism in children. In: Proceedings of the international workshop on social learning and multimodal interaction for designing artificial agents, pp 1–6
67. Pennisi P, Tonacci A, Tartarisco G, Billeci L, Ruta L, Gangemi S, Pioggia G (2016) Autism and social robotics: a systematic review. Autism Res 9(2):165–183

68. Plutchik R (2001) The nature of emotions: human emotions have deep evolutionary roots, a fact that may explain their complexity and provide tools for clinical practice. American Scientist 89(4):344–350
69. Provoost S, Lau HM, Ruwaard J, Riper H (2017) Embodied conversational agents in clinical psychology: a scoping review. Journal of Medical Internet Research 19(5):e151
70. Razavi SZ, Ali MR, Smith TH, Schubert LK, Hoque ME (2016) The LISSA virtual human and ASD teens: an overview of initial experiments. In: International conference on intelligent virtual agents, pp 460–463
71. Reeves B, Read JL (2009) Total engagement: how games and virtual worlds are changing the way people work and businesses compete. Harvard Business Press
72. Ribeiro PC, Raposo AB (2014) Comfim: a game for multitouch devices to encourage communication between people with autism. In: 2014 IEEE 3nd International conference on serious games and applications for health (SeGAH). IEEE, pp 1–8
73. Ricciardi L et al (2015) Alexithymia in neurological disease: a review. The Journal of Neuropsychiatry and Clinical Neurosciences 27(3):179–187
74. Rivers SE, Brackett MA, Reyes MR, Mayer JD, Caruso DR, Salovey P (2012) Measuring emotional intelligence in early adolescence with the msceit-yv: psychometric properties and relationship with academic performance and psychosocial functioning. J Psychoeduc Assess 30(4):344–366
75. Russo A, D'Onofrio G, Gangemi A, Giuliani F, Mongiovi M, Ricciardi F, Greco F, Cavallo F, Dario P, Sancarlo D et al (2019) Dialogue systems and conversational agents for patients with dementia: the human-robot interaction. Rejuvenation Research 22(2):109–120
76. Saleh MA, Hashim H, Mohamed NN, Abd Almisreb A, Durakovic B (2020) Robots and autistic children: a review. Periodicals of Engineering and Natural Sciences 8(3):1247–1262
77. Sampath H, Agarwal R, Indurkhya B (2013) Assistive technology for children with autism-lessons for interaction design. In: Proceedings of the 11th Asia Pacific conference on computer human interaction, pp 325–333
78. Scassellati B, Boccanfuso L, Huang C-M, Mademtzi M, Qin M, Salomons N, Ventola P, Shic F (2018) Improving social skills in children with asd using a long-term, in-home social robot. Sci Robot 3:21
79. Scherer KR (1987) Toward a dynamic theory of emotion: the component process model of affective states. Geneva studies in Emotion and Communication 1:1–98
80. Shen S, Doyle-Thomas KrissyAR, Beesley L, Karmali A, Williams L, Tanel N, McPherson AC (2017) How and why should we engage parents as co-researchers in health research? A scoping review of current practices. Health Expect 20(4):543–554
81. Spitale M, Catania F, Crovari P, Garzotto F (2020) Multicriteria decision analysis and conversational agents for children with autism. In: Proceedings of the 53rd Hawaii international conference on system sciences, pp 1–10
82. Spitale M, Silleresi S, Cosentino G, Panzeri F, Garzotto F (2020) Whom would you like to talk with? Exploring conversational agents for children's linguistic assessment. In: Proceedings of the interaction design and children conference, pp 262–272
83. Tanaka H, Negoro H, Iwasaka H, Nakamura S (2017) Embodied conversational agents for multimodal automated social skills training in people with autism spectrum disorders. PloS one 12:8
84. Trevisan DA, Bowering M, Birmingham E (2016) Alexithymia, but not autism spectrum disorder, may be related to the production of emotional facial expressions. Molecular autism 7(1):1–12
85. Valencia K, Rusu VZ, Jamet E, Zúñiga C, Garrido E, Rusu C, Quiñones D (2020) Technology-based social skills learning for people with autism spectrum disorder. In: International conference on human-computer interaction. Springer, pp 598–615
86. Valentine AZ, Brown BJ, Groom MJ, Young E, Hollis C, Hall CL (2020) A systematic review evaluating the implementation of technologies to assess, monitor and treat neurodevelopmental disorders: a map of the current evidence. Clin Psychol Rev, 101870
87. Villano M, Crowell CR, Wier K, Tang K, Thomas B, Shea N, Schmitt LM, Diehl JJ (2011) Domer: a wizard of oz interface for using interactive robots to scaffold social skills for children with autism spectrum disorders. In: 2011 6th ACM/IEEE International Conference on Human-Robot Interaction (HRI). IEEE, pp 279–280
88. Yao Q (2014) Multi-sensory emotion recognition with speech and facial expression. The University of Southern Mississippi