International Journal of Social Robotics (2023) 15:165–183
https://doi.org/10.1007/s12369-022-00946-2

Utilizing an Emotional Robot Capable of Lip-Syncing in Robot-Assisted Speech Therapy Sessions for Children with Language Disorders

Alireza Esfandbod[1] · Zeynab Rokhi[1] · Ali F. Meghdari[1,2] · Alireza Taheri[1] · Minoo Alemi[1,3] · Mahdieh Karimi[4]

Accepted: 4 November 2022 / Published online: 23 November 2022

Abstract

This study scrutinizes the impacts of utilizing a socially assistive robot, the RASA robot, during speech therapy sessions for children with language disorders. Two capabilities were developed for the robotic platform to enhance children-robot interactions during speech therapy interventions: facial expression communication (containing recognition and expression) and lip-syncing. Facial expression recognition was conducted by training several well-known CNN architectures on one of the most extensive facial expressions databases, the AffectNet database, and then modifying them using the transfer learning strategy performed on the CK+ dataset. The robot's lip-syncing capability was designed in two steps. The first step was concerned with designing precise schemes of the articulatory elements needed during the pronunciation of the Persian phonemes (i.e., consonants and vowels). The second step included developing an algorithm to pronounce words by disassembling them into their components (including consonants and vowels) and then morphing them into each other successively. To pursue the study's primary goal, two comparable groups of children with language disorders were considered, the intervention and control groups. The intervention group attended therapy sessions in which the robot acted as the therapist's assistant, while the control group only communicated with the human therapist. The study's first purpose was to compare the children's engagement while playing a mimic game with the affective robot and the therapist, conducted via video coding. The second objective was to assess the efficacy of the robot's presence in the speech therapy sessions alongside the therapist, accomplished by administering the Persian Test of Language Development, Persian TOLD. According to the first scenario, playing with the affective robot is more engaging than playing with the therapist. Furthermore, the statistical analysis of the study's results indicates that participating in robot-assisted speech therapy (RAST) sessions enhances children with language disorders' achievements in comparison with taking part in conventional speech therapy interventions.

Keywords Social robots · Child-robot interaction (CRI) · Children with language disorders · Facial expression recognition · Convolutional neural network (CNN) · Lip-syncing · Speech therapy

1 Introduction

Inclusive education is based on treating individuals with diverse capabilities attentively through a perfect tutoring procedure [1, 2]. By way of explanation, inclusive education provides an opportunity for learners with differentiating characteristics to be educated in an equitable context and acquire further training achievements. Several studies recently conducted in the field of education revealed that the educational process benefits from the relationship between educational assistive tools and learners much more than the relationship between tutors and learners [3]. Hence, an important aspect of teaching is utilizing appropriate educational tools to improve the students' acquisition and engagement. Social robots are

✉ Ali F. Meghdari
meghdari@sharif.edu

[1] Social and Cognitive Robotics Laboratory, Center of Excellence in Design, Robotics, and Automation (CEDRA), Sharif University of Technology, Tehran, Iran

[2] Fereshtegaan International Branch, Islamic Azad University, Tehran, Iran

[3] Department of Humanities, West Tehran Branch, Islamic Azad University, Tehran, Iran

[4] Mahan Learning Disorders Center, Tehran, Iran

novel educational assistive technologies that assist educators by promoting learning efficiency [4].

Delayed Speech Development (DSD) is an incapacity to deploy communicational skills in infancy [5]. This disruptive delay is frequently accompanied by mental retardation, which can influence toddlers' socialization [6, 7]. Accordingly, these children face tremendous challenges in expressing their thoughts and comprehending information from their environment, leading them to be classified as socially vulnerable children. Statistics indicated that 8–10% of all preschoolers face DSD problems [8]. Ordinarily, the disorder's initial symptoms appear around 18 months when a child does not make any effort to repeat the words they hear. At 24 months, their vocabulary is restricted to single words, and at 30–36 months, an apparent lack of skill in making sentences can be observed. Generally, these children are only able to use memorized phrases gathered from games or animations [9]. Speech therapy sessions mitigate some of the problems related to language disorders and hone special needs children's communicational and verbal skills. In recent decades, employing audio and visual content during speech therapy sessions has become very popular among speech therapists due to their potential benefits in increasing the efficiency of the interventions [10–12]. However, these tools only lead to one-way interaction; in other words, children's responses are not reciprocated in these approaches; therefore, a conversation, which is a prerequisite to communication, cannot be formed [13].

Among various assistive technologies utilized in therapy sessions, social robots have received growing attention in recent years due to their potential role as mediators between therapists and children [4, 14, 15]. Involving social robots in clinical settings increases participants' attention, improves individuals' social behaviors, and sustains their level of engagement during therapeutic interventions. The encouraging implications of employing social robots in therapy interventions for individuals with various impairments, such as Autism Spectrum Disorder (ASD) [16–18], Down syndrome [19, 20], and Hearing impairments [21], have underscored the encouraging prospects of these promising assistive tools in terms of providing equal educational opportunities for special needs children [22]. A noteworthy characteristic of employing a social robot as an assistive tool in therapy sessions is the two-way interaction formed between the robot and the child, which encourages different aspects of the children's behaviors, such as attention span and willingness to learn.

This research endeavors to explore the potential benefits of employing the RASA robot [23, 24] in speech therapy sessions via quantitative analysis. In this regard, two scenarios were carried out: the first was associated with comparing the children's engagement level in an imitation game played with the therapist and the robot, and the second was concerned with investigating the efficacy of the robot's presence in speech therapy sessions. The children's awareness of the robot's capabilities to recognize its users' facial expressions and express several emotional states forms a positive preconception about the robot's intelligence level and the complexity of its behaviors, which helps to sustain the children's engagement through long-term interaction with the robot [25]. Additionally, when children assess a social robot, the delight they have experienced through their interactions affects the acceptance of the robot. Enjoyment is a crucial element in the investigation of social robots' acceptance; it diminishes the individuals' anxiety and makes them feel more confident about their ability to communicate with this technology [26–28]. Thus, the first scenario could benefit the second by increasing the children's willingness to approve the social robot as an educational assistive tool in speech therapy sessions. To scrutinize the impacts of the robot's presence in the two scenarios, two groups of children were recruited to participate in our examinations: the first one (the intervention group) participated in robot-assisted speech therapy (RAST) sessions and the second one (the control group) participated in traditional therapy sessions. Each group of participants was comprised of six children with language disorders (four males and two females) with a mean age of 6.4 years. To accomplish the robot's objective in terms of interacting with children in the RAST sessions, a facial expression recognition system and an accurate lip-syncing system were developed and implemented on the RASA robot. To do this, several well-established Convolutional Neural Network (CNN) architectures were trained on the AffectNet emotional database [29] and modified via the transfer learning technique performed on the CK+ database to become more suitable for the RAST sessions. The other significant aspect of this study that distinguishes it from previous works [3, 4, 30] is the design of the robot's mouth, which can precisely synchronize lip movement with the robot's speech. In this way, children with language disorders can better learn the exact pronunciation of each word by concentrating on the robot's lips.

The rest of the paper is organized as follows: Sect. 2 is devoted to elucidating related works. Section 3 explains the design of the affective interaction system composed of three main phases: recognizing facial expressions, expressing various emotions, and implementing the system on the RASA robot. Section 4 describes the development of an appropriate human-like lip-syncing system for the robot. This section explains the design of the robot's visual articulatory elements for each Persian phoneme in detail, along with the algorithm utilized to attain human-like lip-syncing. Section 5 gives details about the experimental procedure and discusses the two scenarios carried out in the study. The first scenario sought to answer a primarily exploratory question: Are

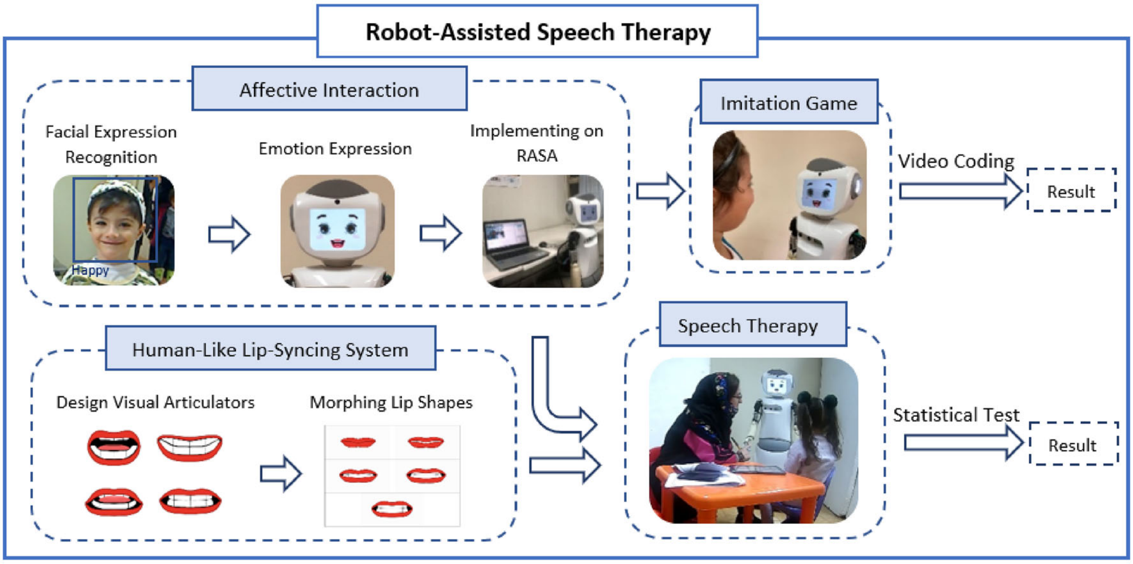


Fig. 1 An overview of the study

children with language disorders more inclined to play a collaborative emotional imitation game with a social robot or a therapist? The second scenario aimed to explore the effects of the robot's presence in speech therapy interventions on the individuals' language skills development and was assessed by comparing the progress of two groups of children with language disorders, those who participated in robot-assisted therapy sessions with those who took part in conventional sessions. Section 6 analyzes the results and discusses the outcomes via conducted statistical tests. The assessment tools used in this analysis include video coding for the first scenario and the Persian version [31] of the Test of Language Development (TOLD) [32] for the second scenario. The following section discusses the limitations of this study and future works. Finally, the conclusion is drawn in Sect. 8. An overview of the study is shown in Fig. 1.

sessions. They conducted their experiments for two weeks (three sessions per week) with four DSD children between four and six years old. In this study, the robot played the role of entertainer by performing dances, playing games, and telling several fairy tales. The study results suggested that RAT could be regarded as a practical approach to encourage DSD children and facilitate their development in pronouncing simple sentences and singing well-known songs with the robot. However, the robotic platforms utilized in references [33–38] did not possess precise visual articulatory systems; consequently, the RAT scenarios conducted in these studies were based on auditory-verbal therapy (concerned with developing auditory and verbal skills) and were ineffective in enhancing children's capabilities with regard to lip-reading and perceiving other non-verbal cues [37].

2 Related Work

2.1 Robot-Assisted Speech Therapy

According to studies investigating the potential of social robots in speech therapy interventions for children suffering from different impairments such as ASD [33], Cleft Lip and Palate (CL/P) [34], Cerebral Palsy (CP) [35], Hearing impairments [36, 37], and DSD [38], the presence of a robotic assistant is beneficial in terms of providing incentives for children to participate in therapy sessions and improving their verbal and communication skills. For example, in Ref. [38], Zhanatkyzy and Turarova used the NAO robot to investigate the effectiveness of robot-assisted therapy (RAT)

2.2 Facial Communication Channels in HRI

By and large, blurring the distinctions between therapists and socially assistive robots in terms of communication methods used to interact with children could lead to progress in human–robot interaction (HRI). Moreover, real-time interaction between children and robots can positively affect both the learning process and social development [39, 40]. Thus, augmenting human-like features to a socially assistive robot, such as real-time recognition and expression of emotional states, body gestures, and lip-syncing, makes the robot more socially acceptable [41].

2.2.1 Facial Expression Recognition in HRI

Following advances in computer vision technologies, developing emotional facial expression recognition systems for social robots via various machine learning algorithms and promoting the robots' emotional intelligence have been trending upward [42–45]. Since facial cues are essential elements in an affective interaction, their recognition and expression lead to more in-depth communications [46–51]. Furthermore, the more extravagantly the social robot behaves, the more it encourages children to remain engaged through long-term interactions with the robot [25]. In social robots' acceptance, sociability is a primary factor attributed to the users' opinions about the robot's social, emotional, and cognitive skills [28]. Hence, the robot's capabilities in terms of recognizing and expressing various emotional states influence the individuals' evaluation of the robot's intelligence level and heighten the robot's acceptance. Different machine learning methods, e.g., deep learning, have been extensively used in the literature to promote social robots' emotional intelligence [52]. Ref. [53] trained the Xception architecture [54] on the FER-2013 database [55] and implemented the trained model on the Pepper humanoid robot. In that study, the robot was able to recognize pedestrians' emotions (neutral, happiness, sadness, and anger) and consider their emotional states to perform emotion-aware navigation while moving among them. In Ref. [47], the VGG16 Network [56] was trained on the FERW database to develop a model capable of recognizing seven basic emotions; the trained model was then implemented on the XiaoBao robotic platform to improve the quality of the robot's interactions.

2.2.2 Lip-Syncing in HRI

Lip-syncing is a key factor in human–human interactions, and its precise presentation could result in a better perception of the communicators' purposes [57]. The visual components of human articulatory systems (lips, tongue, teeth, and jaw) and their motions convey the sounds generated by the vocal tract [58]. Due to the importance of multimodal communication in social robotics, many studies have focused on synchronizing lip movements with speech to take advantage of audio-visual information [59]. Cid and Manso [60] concluded that a robot's verbal articulation could be improved by compounding two sources of signals, auditory cues (pitch, pause, and emphasize) and visual cues (lip motions, facial expressions, and body gestures). Ref. [61] found that possessing a dynamic and human-like mouth could increase the acceptance of the robot. The significance of this type of mouth for a socially assistive robot is much more critical in speech therapy interventions where the ultimate goal is emulating natural speech.

3 Emotional Interaction System Design

3.1 Facial Expression Recognition System

As previously mentioned, social robots with the capacity to interact with children emotionally can substantially attract their attention [62]. Generally, emotional interaction is comprised of emotion recognition and expression that can be conveyed through assorted audio and visual channels, including facial cues, which are a primary way of displaying feelings in human–human interactions. End-to-end neural networks are ubiquitously utilized among different machine learning algorithms for facial expression recognition tasks. The two principal aspects of developing a well-trained model for a recognition task are adopting proper databases and suitable architectures.

Developed by Mohammad H. Mahoor and his colleagues in 2017, AffectNet is one of the most comprehensive wild emotional datasets comprising approximately 1 M web images [29]. This dataset consists of two main parts, manually and automated annotated images. Manually labeled images, the focus of this study, are classified into eight expressions and three invalid emotion categories: neutral, sad, happy, surprise, fear, anger, disgust, contempt, none, uncertain, and non-face. It should be noted that invalid emotion categories (none, uncertain, and non-face) were not considered in the training process of the current study. Due to the copious number of annotated images and wild hallmarks of the dataset, training an appropriate CNN architecture on this dataset will yield a well-trained model with superior generalization capability, which can be used in real-world applications.

The extended Cohn-Kanade (CK+) is another standard facial expression dataset developed by Patrick Lucey [63]. With only a tiny number of samples consisting of 327 sequences across 123 subjects gathered in a controlled condition, it superficially resembles the sequences captured by the RASA robot's head-mounted camera in the laboratory. In this paper, similar to [64, 65], the last three frames of each labeled sequence were categorized as one of the basic emotions, and the sequences' first frames were extracted as neutral. Table 1 summarizes the total number of images per expression for each dataset.

Sample images of the AffectNet and CK datasets and an image captured by the robot's camera in the lab environment are shown in Fig. 2.

As the figure shows, the images captured by the robot's camera and the CK+ images have two conspicuous similarities; both were captured in a straight-ahead position and a standard lab environment.

In this study, the facial expression recognition system was designed and implemented on the robotic platform in three steps. In the first step, several noted architectures were

Table 1 The total number of images in the AffectNet and CK+ datasets per each expression [29, 65]

	Neutral	Sad	Happy	Surprise	Fear	Anger	Disgust	Contempt
CK+	593	84	207	249	75	135	177	54
AffectNet	75,374	25,959	134,915	14,590	6878	25,382	4303	4250

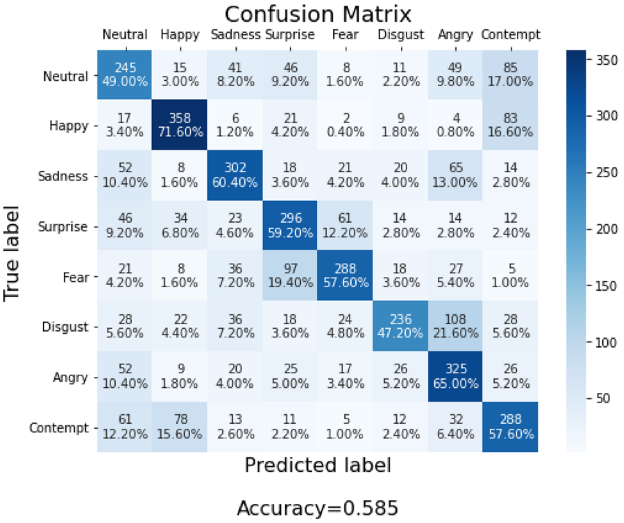
Fig. 2 **a** An image captured by the robot's camera in the experimental setup, **b** a sample image of the AffectNet [29], and **c** a sample image of the CK+ datasets [63]

trained on the AffectNet dataset, and the results were compared through various evaluation metrics, such as accuracy, F1-score [66], Cohen's kappa [67], and area under the ROC curve (AUC) [68], to achieve an accurate model. Afterward, to enhance the system's performance in interactions with the robot's users in the laboratory, the model (selected according to its performance on the AffectNet test set) was then adapted via the transfer learning technique conducted on the CK+ dataset. Finally, the modified model was implemented on the RASA robot.

3.1.1 Step One: Model Training

Several well-known CNN architectures, including MobileNet [69], MobileNet v2 [70], NASNET [71], DenseNet169 [72], DenseNet121 [72], Xception [54], Inception v3 [73], and VGG16 [56], with satisfactory performance on the ImageNet dataset [74], were trained on the AffectNet dataset. According to the dataset's instruction manual, faces were cropped and resized to 224 × 224. Then, the corresponding preprocesses were applied to the images for each network. In order to achieve a better-generalized model, data augmentation was performed via three standard techniques: rotation (from -10 to 10 degrees), translation (up to 10% in both x and y directions), and horizontal flipping. The Adam optimizer was utilized with a learning rate of 1e-5 and a momentum of 0.9. The weighted-loss function was also used to compensate for the adverse effects of the imbalanced training set. The mentioned networks were trained over ten epochs. For each network, the maximum batch size was limited by the available memory of the hardware: 64 for MobileNet, 64 for MobileNet v2, 64 for NASNET, 32 for DenseNet169, 32 for DenseNet121, 16 for Xception, 16 for Inception v3, and 8 for VGG16. All the networks were trained on an NVIDIA GeForce GTX 1080Ti GPU using Keras framework. Table 9, presented in the "Appendix" section, summarizes the accuracy of the trained models. A comparison of the various networks' accuracies led us to adopt the MobileNet architecture for the facial expression recognition task due to the number of parameters and superior performance on the AffectNet test set. The confusion matrix of the MobileNet model is shown in Table 2.

Table 2 Confusion matrix of the trained MobileNet architecture on the AffectNet test set



	Neutral	Happy	Sadness	Surprise	Fear	Disgust	Angry	Contempt
Neutral	245 49.00%	15 3.00%	41 8.20%	46 9.20%	8 1.60%	11 2.20%	49 9.80%	85 17.00%
Happy	17 3.40%	358 71.60%	6 1.20%	21 4.20%	2 0.40%	9 1.80%	4 0.80%	83 16.60%
Sadness	52 10.40%	8 1.60%	302 60.40%	18 3.60%	21 4.20%	20 4.00%	65 13.00%	14 2.80%
Surprise	46 9.20%	34 6.80%	23 4.60%	296 59.20%	61 12.20%	14 2.80%	14 2.80%	12 2.40%
Fear	21 4.20%	8 1.60%	36 7.20%	97 19.40%	288 57.60%	18 3.60%	27 5.40%	5 1.00%
Disgust	28 5.60%	22 4.40%	36 7.20%	18 3.60%	24 4.80%	236 47.20%	108 21.60%	28 5.60%
Angry	52 10.40%	9 1.80%	20 4.00%	25 5.00%	17 3.40%	26 5.20%	325 65.00%	26 5.20%
Contempt	61 12.20%	78 15.60%	13 2.60%	11 2.20%	5 1.00%	12 2.40%	32 6.40%	288 57.60%

Other evaluation metrics for the CNNs mentioned above are also concisely presented in Table 10 in the "Appendix" section. It is worth noting that the AffectNet dataset's annotators concurred with each other on 60.7% of the images [29].

Table 3 A comparison of the MobileNet evaluation metrics on the CK+ test set, before and after transfer learning

	Before transfer learning	After transfer learning
Accuracy	0.85	0.95
F1 score	0.77	0.92
Recall	0.85	0.91
Precision	0.75	0.97
Cohen Kappa	0.76	0.94
AUC	0.86	0.98

3.1.2 Step Two: Model Adaptation

In this step, the MobileNet model, chosen in the previous step, was evaluated on the CK+ dataset and tuned by the transfer learning technique. The dataset was split into train and test sets to assess the network's performance on the CK+ . The splitting procedure was subject-based, so 10% of the subjects, randomly selected, formed the test set. The face detection for this dataset was done by the Viola-Johns method [75], and the previous preprocesses and augmentation techniques were applied, as explained above. While the features extracted in the early layers of a CNN model are more generic, the ones extracted from later layers are more dataset-specific. Hence, to optimize the model on the CK+, the 20 earliest layers' parameters were frozen, and the other layers' parameters were tuned over ten epochs. Table 3 represents the accuracy and the evaluation metrics of the MobileNet model on the CK+ test set before and after performing transfer learning on the CK+ train set.

As Table 3 presents, transfer learning improved the model's performance on the CK+ test set. Due to the similarity between the study's experimental environment and the CK+ , we could reasonably expect to acquire a more precise facial expression recognition system after the tuning.

3.1.3 Step Three: Implementing the Facial Expression Recognition System on the RASA Robot

The humanoid robotic platform utilized in the study was RASA, designed and manufactured at CEDRA (Center of Excellence in Design, Robotics, and Automation) at the Sharif University of Technology [23, 24]. This socially assistive robot aims to interact with special needs children. Figure 3 displays the employed robotic platform.

The robot's abilities to perform real-time recognition and react authentically are critical factors in providing a natural interaction. Hence, due to the limited power of the graphics processing unit of the robot's onboard computer, it would be beneficial to use an external graphics processing unit to execute the facial expression recognition task's computational cost. Accordingly, an external NVIDIA GeForce RTX 2070 GPU was deployed to do graphical computations. To implement the developed emotional system on the RASA, the robot's onboard camera first captured the user's image. Next, a ROS node was used to stream the image topic. Then, a python code was developed to capture live stream video from the robot's IP and apply Viola-John's face detector algorithm [75] to the received data. Following the face detection, the CNN model was used to predict the user's facial expression. By way of response, the proper reaction, according to the HRI scenario, was selected and published on a ROS topic. Ultimately, the robot reacted according to the subscribed message. In this scheme, only the tasks of streaming the video and subscribing to messages were loaded onto the robot's onboard computer.

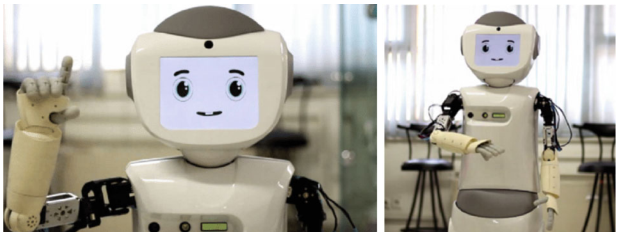


Fig. 3 The RASA socially assistive robot

3.2 Facial Expressions

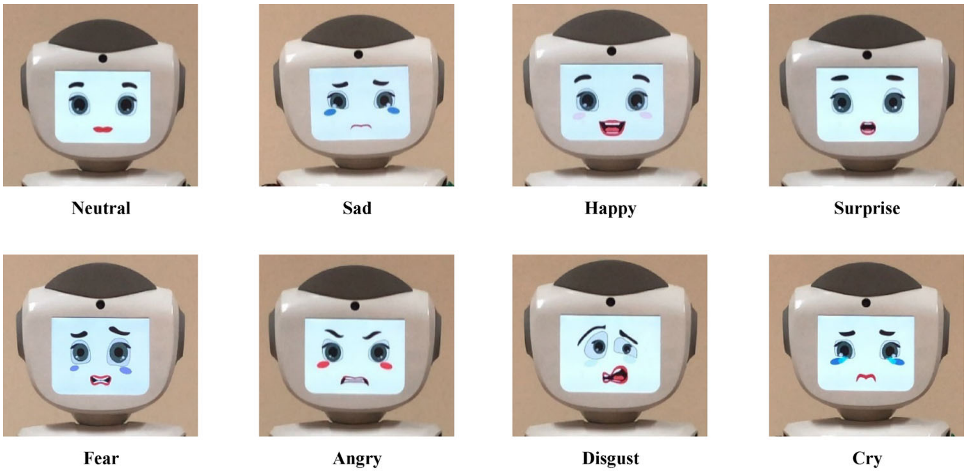
To achieve a two-way interaction between the robot and a child, not only is it essential to recognize the child's emotional state, but the robot must also depict a justifiable expression. Therefore, designing appealing facial expressions for the robot is crucial. In the current study's speech therapy scenarios, the robot should be able to convey emotional messages and enunciate letters and words simultaneously. Thus, the robot's emotional expressions should not depend only on articulatory visual elements. Hence, several other components, such as eyes, eyebrows, and cheeks, were also considered in the design of the robot's emotional states. In this way, the robot will be able to express emotions and lip-sync concurrently.

Figure 4 depicts the eight emotional states designed for the robot's face.

4 Lip-Syncing System

4.1 Graphic Design

Developing a lip-syncing system with realistic articulators could boost the robot's efficacy in the RAST sessions. To



Neutral

Sad

Happy

Surprise

Fear

Angry

Disgust

Cry

Fig. 4 Designed emotional states of the robot

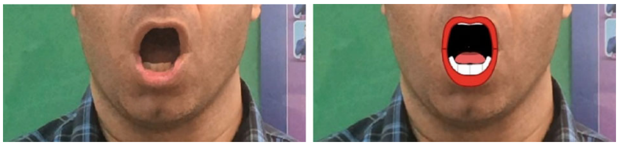


Fig. 5 The procedure of designing the robot's articulators for a particular phoneme

achieve a perceptible visual articulatory system, an Iranian sign language interpreter was hired to pronounce Persian phonemes (including vowels and consonants), and the articulators were thoroughly sketched based on the images captured from him in a straight-ahead position. Figure 5 illustrates the procedure of sketching the robot's visual articulatory elements for a particular phoneme.

Figure 6 shows the individual shapes sketched for Persian phonemes, including twenty-two consonants and six vowels.

4.2 Morphing Algorithm

The algorithm proposed for the lip-syncing system executes according to a three-step process, including receiving a word, disassembling it into basic phonemes, and morphing them into each other smoothly. In the procedure of morphing a phoneme into its subsequent one, the deformation of the articulators should be minimized to achieve a natural visual articulation. Furthermore, although minor defects in the sketches could be ignored by spectators, discrete and unnatural transitions of elements are not permissible. Following a path between the initial and final points with a constant velocity throughout the transition procedure adversely affects the fluidity of movement and leads to unnatural motions [76]; this problem could be addressed by adding acceleration terms [77].

In animation jargon, an easing function describes the way that the transition from an initial point to a final point occurs by determining the velocity and acceleration terms. Figure 7 demonstrates some well-established easing functions.

After examining the easing functions presented in Fig. 7, the "InOutExpo" function was selected due to its capability to provide a natural and smooth transformation. The equation of this function is given by [76]:

$$y = \begin{cases} 0 & x = 0 \\ 2^{20x-11} & x \in \left(0,\ \frac{1}{2}\right] \\ 1 - 2^{-20x+9} & x \in \left(\frac{1}{2},\ 1\right) \\ 1 & x = 1 \end{cases} \tag{1}$$

The smooth transition from a particular articulatory element of a phoneme to its corresponding component in another phoneme within successive frames could be accomplished by dividing each shape into numerous points and employing an easing function to determine the points' transition characteristics. The two shapes' corresponding points are determined by minimizing the deformation according to the following penalty function:

$$J = \sqrt{\frac{\sum_{i=1}^{N}\left(x_i - \hat{x}_i\right)^2}{N}} \tag{2}$$

Thus, the transition problem is simplified to the optimization of the least square penalty function presented in Eq. (2). Increasing the number of points makes the transition smoother; however, it accrues more computational cost. In this study, the KUTE library was deployed for morphing the articulators' vector shapes sketched by Adobe Illustrator. The sign language interpreter and the speech therapist assessed the caliber of the proposed articulatory system by visual examinations. Figure 8 demonstrates the procedure of lip-syncing a Persian word.

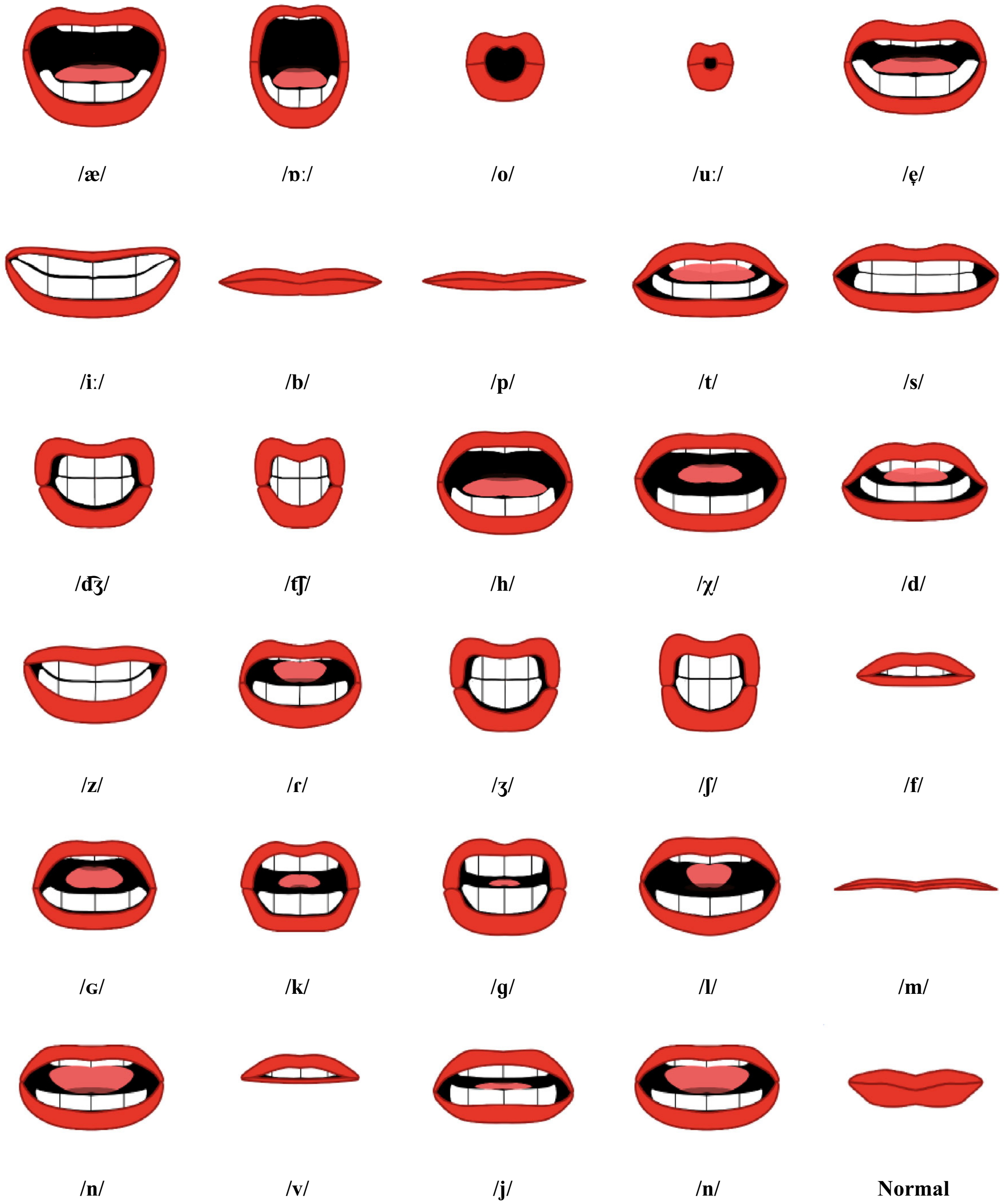


Fig. 6 The sketches of the articulatory elements for the Persian phonemes and the normal state

5 Methodology

In this study, two groups of children with language disorders were investigated to assess the efficacy of utilizing social robots as assistive tools in speech therapy. The first group (the intervention group) was enrolled in the RAST interventions, while the second group (the control group) participated in conventional speech therapy sessions. The under-investigation groups underwent two scenarios: a ten-minute imitation game and a set of thirty-minute speech therapy sessions.

The first scenario was designed to scrutinize the children's engagement level through a mimic game played with the robot (for the intervention group) and the therapist (for the control group). This examination endeavored to compare the

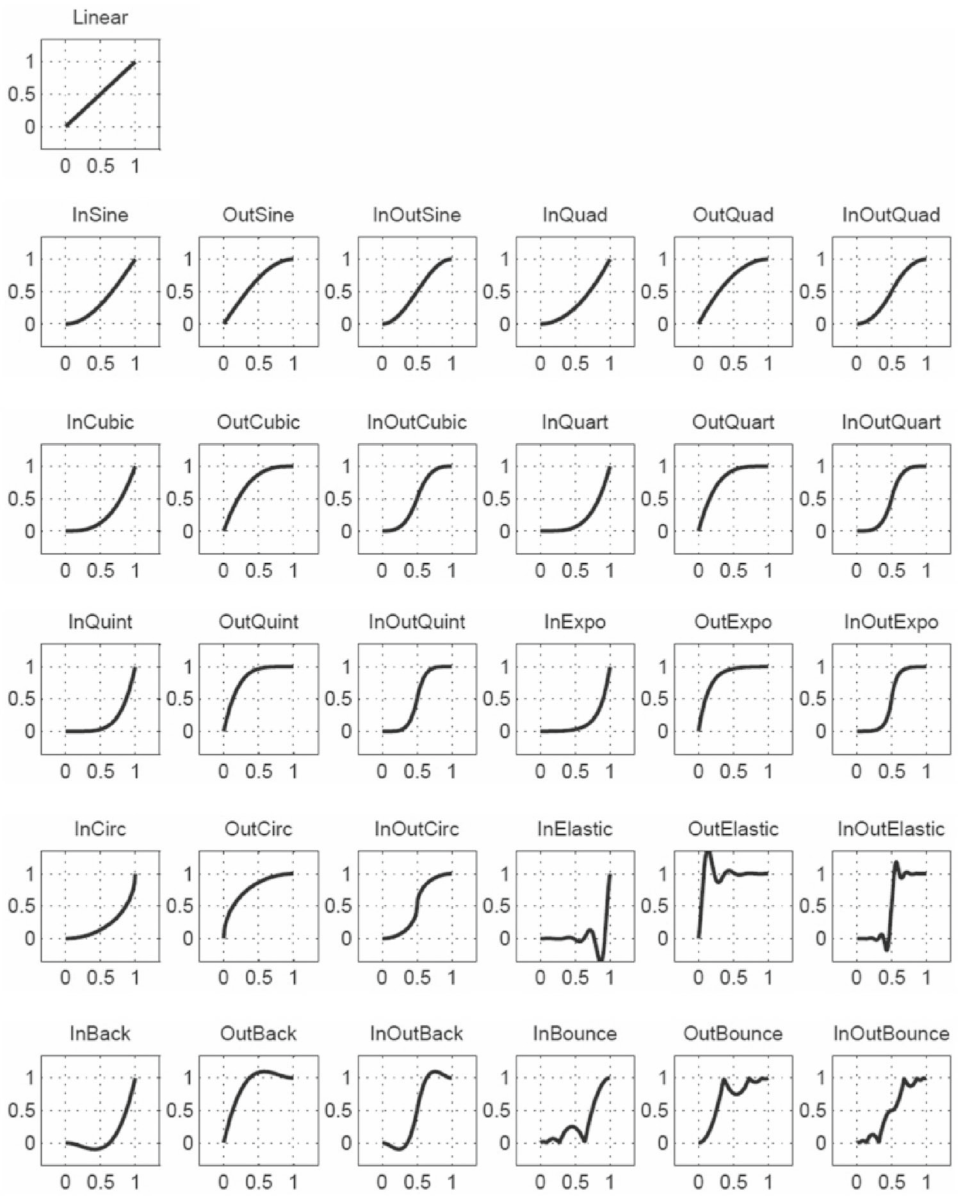


Fig. 7 Penner’s easing functions [76]

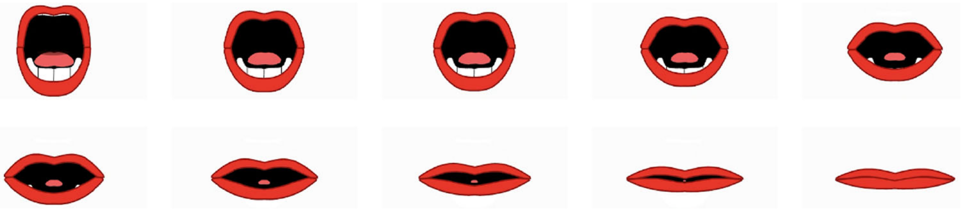


Fig. 8 The procedure of lip-syncing (/ɒːb/), which means water in Persian

duration participants looked at the therapist/robot employing the manual video coding technique. The second scenario was formulated to examine the potential of deploying social robots in speech therapy sessions for children with language disorders. This inspection investigated the language skills progression of individuals who participated in the RAST sessions and traditional therapeutic interventions. To accomplish this objective, the Persian TOLD was taken from the intervention and control groups in two phases: a pre-test and a post-test. The pre-intervention scores were utilized to assess the comparability of the two groups’ initial language levels, and the post-intervention scores were used to investigate the efficacy of the RAST sessions.

Each group participated in ten one-to-one therapy sessions, one session per week. The first session (Week #1) was devoted to familiarizing the children with the experimental setup (to negate the novelty factor effects) and administrating the pre-test. The second session (Week #2) was dedicated to performing the first scenario, the imitation game, and the other sessions were held to explore the social robot's potential in speech therapy. One week after the final speech therapy session, the post-test was given to both groups of children.

5.1 Participants

In this exploratory study, the intervention group was made up of six native Persian-speaking children with language disorders (two female, four male) with an average age of 6.4 years and a standard deviation of 2.2 months. The control group consisted of the same number of native Persian-speaking children, the same distribution of genders, and an average age of 6.4 years with a standard deviation of 1.9 months. In order to make a precise evaluation, both groups of children were selected from individuals who attended weekly traditional speech therapy sessions at the Mahan Language Disorders Center. Furthermore, they were asked not to participate in any other therapy sessions from two weeks before the start of our investigation until the end of it. According to a post-hoc power analysis conducted via G*Power 3.1 Software [78], for the sample size of N = 6 per group, the power of this pilot study is 12% considering a medium effect size of 0.5 and a significance level of 0.05%.

5.2 Experimental Setup

The RAST sessions were conducted at the Social and Cognitive Robotics Lab at the Sharif University of Technology. Three cameras, two located in the room's corners and one mounted on the robot's head, recorded all interventions. In all sessions, the speech therapist was present beside the robot. In another room (control room), the robot's operators controlled and narrated the robot's dialogues, synchronizing their voice with the lip-syncing of the robot through videos they received from the RAST sessions. Hence, a real-time human voice was synchronously compounded with the robot's articulation to communicate with the children. Two speakers were also placed in the room to play the filtered operator's voice, which was made to sound like a child by changing its pitch. The schematic of the experimental setup, including the intervention and control room, is shown in Fig. 9.

The conventional speech therapy sessions of the second group were also held in the aforementioned experimental setup without the social robot's presence to eliminate the environmental conditions' impact.

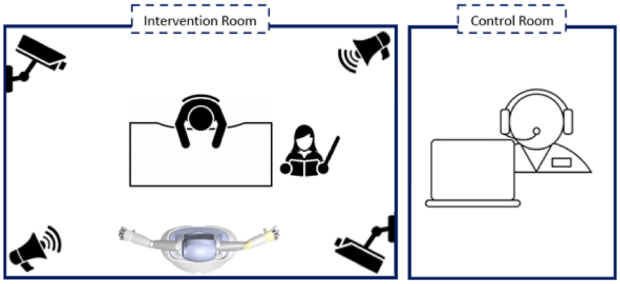


Fig. 9 The schematic of the intervention and control rooms

5.3 Intervention Scenarios

Speech therapy aims to hone individuals' communication skills and enhance the participants' abilities to grasp and express thoughts, ideas, and feelings. Therefore, engaging children and boosting their achievements from therapy sessions are two pre-eminent factors that should be considered in the interventions.

5.3.1 Scenario One: The Investigation of the Children's Engagement Level via an Imitation Game

The current scenario was designed to explore whether or not children with language disorders are more engaged during interaction with the social robot than with the therapist. Scenario one was a facial expression mimicry game that required children to stand in front of a playmate (robot/therapist) and imitate its facial expressions. The scenario was conducted in a ten-minute intervention where the child's playmate revealed a random facial expression and waited until the child emulated the same emotional state. The playmate would express the next emotion when the imitation was performed correctly. The intervention group played with the robot throughout the scenario, while the control group played with the therapist.

5.3.2 Scenario Two: Utilizing a Social Robot for the Therapy of Children with Language Disorders

In this scenario, two sets of thirty-minute speech therapy sessions were carried out for the two groups of participants. The intervention group attended RAST sessions, while the control group participated in conventional speech therapy interventions. During the RAST sessions, the RASA robot interacted with children in various ways, i.e., teaching the correct pronunciation of words via lip-syncing, providing a system of reward and punishment by expressing different emotional states, asking multiple questions, and guiding children to answer the therapist's questions nicely. Five frequent tasks (extracted from relevant studies [79–83]) were performed in each speech therapy session for both groups of

Table 4 The list of activities performed in the speech therapy sessions

Activity	Description
Picture description	Ask children to describe the components in the picture accurately [79]
Oral narrative	Ask children to generate a narrative in response to the wordless picture book [80]
Syntactic understanding	Ask children to re-tell the targeting story, using full sentences as well as asking and answering a wide range of "wh" questions [81]
Picture identification	Ask children to respond by pointing to one of twelve pictures following each stimulus presentation [82]
Oral imitation	ask children to imitate specific words that are in different syllable categories [83]

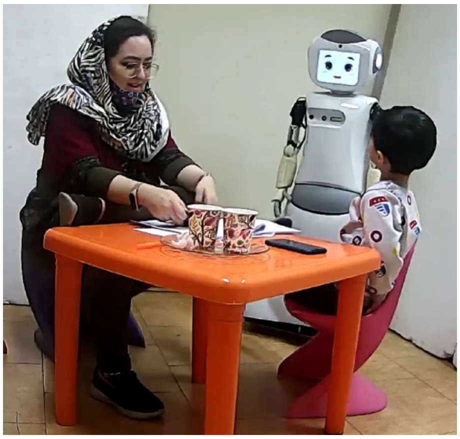


Fig. 10 A robot-assisted speech therapy session

participants to facilitate the children's oral language development. Table 4 describes the list of the activities conducted in the therapeutic interventions.

Figure 10 displays the therapist and the robot in a RAST session.

5.4 Assessment Tools

5.4.1 Scenario One: Assessment of Children's Engagement Level via an Imitation Game

In the context of HRI, content analysis of the interventions' recorded videos was extensively employed to probe individuals' behavioral patterns [84–86]. Meanwhile, analyzing the gaze data (frequencies and durations of gazes) provides metrics quantifying individuals' engagement throughout human–human and human–robot interactions [87–89]. In this regard, in the first scenario, the evaluation of the participants' engagement was conducted via deploying the manual video coding technique to elicit the children's gaze information from the videos of the therapy sessions. The video coding was performed by two raters separately according to the following procedure.

First, due to the oscillating attribute of the participants' attention and distraction during interventions, the game's duration was segmented into specific equal spans ($\Delta t = 20s$). Secondly, in each span, the interval's raw score was defined as the portion of the time children spent gazing at their playmates (either the robot or the therapist). Afterward, the mean score of each span was calculated by averaging the coders' raw scores. Finally, the individuals' engagement scores were computed by taking the integral of the participants' mean scores over the interventions' duration and dividing it by the length of the sessions. The Pearson correlation coefficient between the two raters' raw scores was calculated to determine the inter-rater reliability of the results.

5.4.2 Scenario Two: Assessment of the RASA Social Robot's Utility in Speech Therapy for Children with Language Disorders

In the second scenario, the Persian version of the TOLD was used to evaluate the impacts of the robot on children's language development. This questionnaire is a certified tool for evaluating preschooler language abilities in six core and three supplemental subsets. The test's subsets are summarized in Table 5.

A speech therapist was hired to hold the therapy interventions and score the Persian TOLD questionnaire. Following the test scoring instructions, the therapist asked each child several items to rate the test subsets. If the participant correctly answered the therapist's question, they would have received a score of one; otherwise, they would have been given zero. Thus, the number of items the children correctly answered in each subset's examination determined their respective raw scores. To eliminate the potential impact of the children's age in assessing their language development, the TOLD proposed tables regarding the participants' ages to convert their raw scores into scaled scores varying between 0 and 20. In the current study, the normalized scaled scores (the scaled scores divided by 20) were adopted as metrics to compare the participants' oral language enhancement in the speech therapy scenarios. By integrating the subsets' scores, composite scores were calculated that disclose the children's development concerning primary facets of language, including listening, organizing, speaking, semantics, grammar, phonology, and overall language ability. The corresponding score of each language dimension was evaluated by summing the scaled scores of the subsets associated with

Table 5 The TOLD subsets' descriptions [31, 32]

Subsets	Item	Description
Picture vocabulary	30	Measures a child's understanding of the meaning of spoken Persian words
Relational vocabulary	30	Measures a child's understanding and ability to orally express the relationships between two spoken stimulus words
Oral vocabulary	28	Measures a child's ability to give oral definitions to common Persian words that the examiner speaks
Syntactic understanding	25	Measures a child's ability to comprehend the meaning of sentences
Sentence imitation	30	Measures a child's ability to imitate Persian sentences
Morphological completion	28	Measures a child's ability to recognize, understand, and use common Persian morphological forms
Word discrimination	20	Measures a child's ability to recognize the differences in significant speech sounds
Word analysis	14	Measures a child's ability to segment words into smaller phonemic units
Word articulation	20	Measures a child's ability to utter important Persian speech sounds

the under-investigation skill and normalizing the calculated score according to the number of subsets involved in the skill leading to a score between zero and one [31]. Table 6 demonstrates the association of the TOLD subsets with primary language skills.

6 Results and Discussion

In the explained scenarios, the scores of both groups (intervention and control groups) were separately evaluated by the proposed assessment tools and then examined via statistical analysis using Minitab software. The *p*-values of the tests were employed to identify any significant differences between the intervention and control groups.

Table 6 The association between the TOLD subsets and primary language skills [32]

Language skills	Picture vocabulary	Relational vocabulary	Oral vocabulary	Syntactic understanding	Sentence imitation	Morphological completion	Word discrimination	Word analysis	Word articulation
Listening	✓			✓			✓		
Organizing		✓			✓			✓	
Speaking			✓			✓			✓
Semantics	✓	✓	✓						
Grammar				✓	✓	✓			
Phonology							✓	✓	✓
Overall language ability	✓	✓	✓	✓	✓	✓	✓	✓	✓

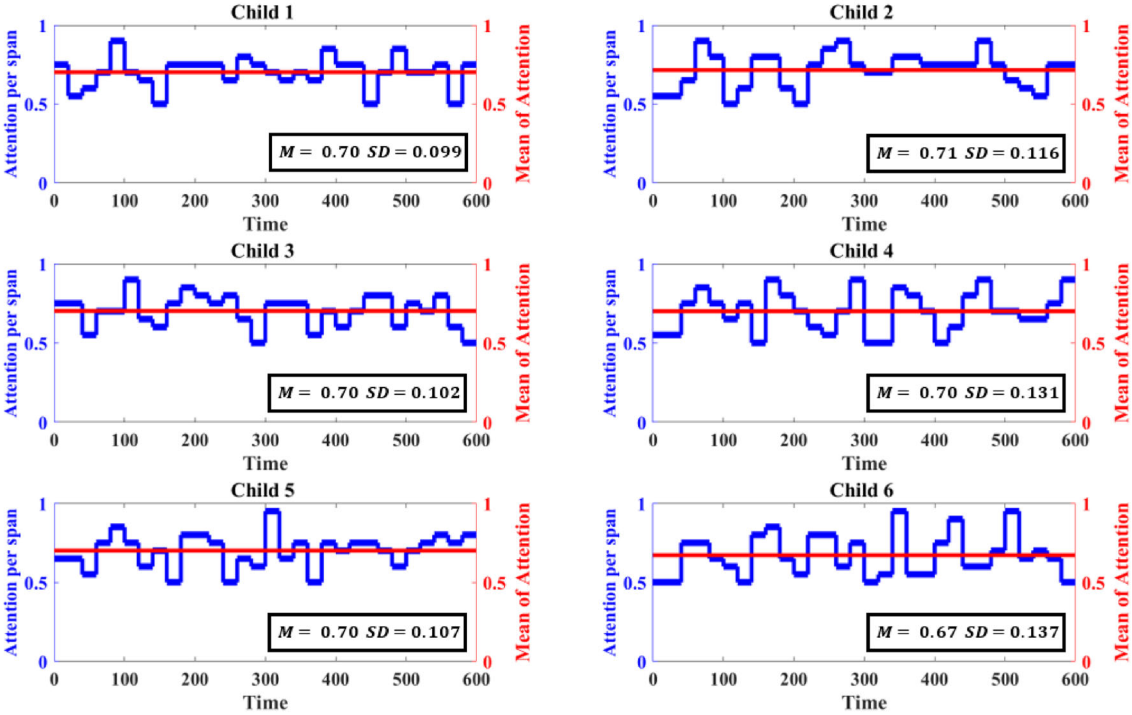


Fig. 11 The scores of the children's engagement in the intervention group

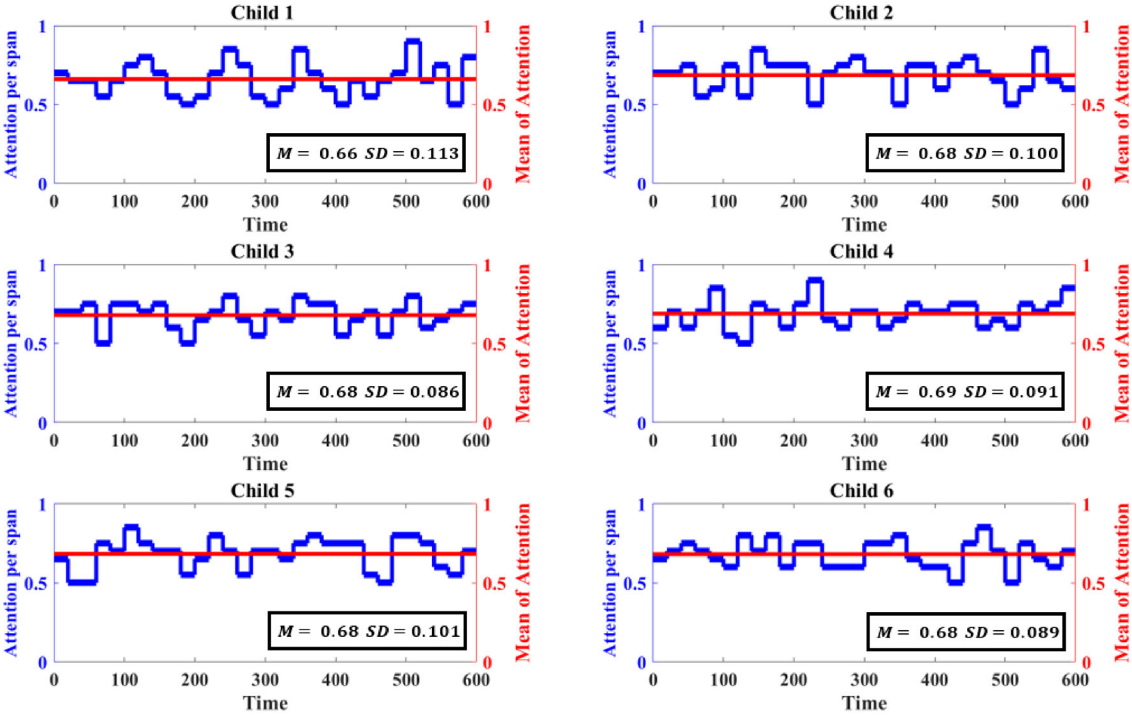


Fig. 12 The scores of the children's engagement in the control group

Table 7 The engagement scores of the intervention and control groups

	Intervention group (N = 6)		Control group (N = 6)		p-value	Cohen's d
	M	SD	M	SD		
Score	0.70	0.014	0.68	0.009	0.025	1.66

6.1 Content Analysis of the Recorded Videos

Figures 11 and 12 present the scores of each child's engagement for the intervention and control groups, respectively. Table 7 also encapsulates the average and standard deviation of the two groups' engagement scores and the statistical analysis results.

As Table 7 demonstrates, the results of the t-test indicate that the engagement scores of the intervention group are significantly higher than the control group ($p = 0.025 < 0.05$). Statistic measures show that the intervention group members, on average, gazed at the robot for 11.7 s more than the control group participants looked at the therapist. Furthermore, the large Cohen's d effect size ($1.66 > 0.8$) indicates the children disclosed a higher intention to play with the social robot than the therapist. According to the Pierson correlation coefficient, the two raters' raw scores were strongly correlated ($r = 0.72 > 0.7$).

6.2 TOLD Analysis

This study quantified the participants' language skills and overall language abilities through the scoring system proposed by the Persian TOLD questionnaire. The groups' pre-test scores were statistically analyzed to explore the comparability of the intervention and the control groups in terms of their initial language levels. Furthermore, the implications of the robot's presence on the participants' language skills improvement were assessed by computing their progress scores, defined by subtracting the children's pre-test scores from their post-test scores. The normalized results of these measures and their corresponding statistical analysis are summarized in Table 8. A Bonferroni correction ($\alpha' = \alpha/k$, where k defines the number of tests; in this study, the k is equal to nine, which is the number of the TOLD subsets) was utilized for the pairwise comparisons to avoid a Type I error.

According to Table 8, p-values related to the administered pre-test highlight no significant differences ($p > 0.05$) between the intervention and control groups regarding the initial levels of language development metrics; thus, these groups can be considered comparable. The Bonferroni post-hoc tests indicate that only the scores of the "Word Discrimination" subset were significantly different between the intervention and control groups ($p < 0.005$).

As previously mentioned, the scores of the primary language skills could be evaluated by summing and normalizing the scores of the associated subsets, as explained in Table 6. Table 8 reveals that the two groups' primary language skills scores in the pre-test were not significantly different, which means that the initial states of the groups were comparable. According to the Bonferroni post-hoc tests, the intervention group made significantly more progress in primary language skills than the control group. Furthermore, the overall language ability score, calculated by summing and normalizing the nine subsets' scores of the TOLD, is a measure that represents the total language development of the children. The analysis of this metric shows that the overall language ability of the children who interacted with the robot improved significantly more than those who took part in conventional speech therapy sessions. The results of this preliminary exploratory investigation shed light on the encouraging implications of utilizing social robots in speech therapy sessions which are in agreement with the results of Ref. [4]. However, the limited number of study participants prevents us from making a generalized claim about the robot's efficacy through interaction with other children with language disorders.

7 Limitations and Future Work

The COVID-19 pandemic restricted the families willing to collaborate with our research group, which resulted in the small number of study participants. This issue was a serious limitation of this examination, which underpowered the study, as proven by the power analysis. The fact that the authors had no control over the families of the special needs children and could not in good conscience deprive them of therapies for a longer span before the beginning of the examination was another study limitation. Consequently, a thorough separation of the possible influences of the previously experienced therapy sessions on the current therapeutic interventions was impossible. The temporary displeasure of a few children in a limited number of training sessions was another limitation of the current study. Although the training sessions were based on one-to-one interactions, a few children would have refused to participate in some sessions if one of their families or companions had not been present at the beginning of the training sessions. It should be noted that this could lead to possible social bias in the study's results, which was inevitable. Due to the lack of similar studies about the

Table 8 Comparison between the two groups of children's language development metrics

Language development metrics	Pre-test scores			Progress scores		
	Intervention group mean (SD)	Control group mean (SD)	*P* value	Intervention group mean (SD)	Control group mean (SD)	*P* value
Picture vocabulary	0.33 (0.103)	0.32 (0.075)	0.757	0.27 (0.082)	0.20 (0.063)	0.148
Relational vocabulary	0.30 (0.167)	0.33 (0.103)	0.689	0.35 (0.105)	0.18 (0.075)	0.012
Oral vocabulary	0.40 (0.126)	0.42 (0.117)	0.818	0.37 (0.082)	0.15 (0.122)	0.007
Syntactic understanding	0.28 (0.172)	0.27 (0.121)	0.851	0.38 (0.133)	0.15 (0.084)	0.007
Sentence imitation	0.28 (0.117)	0.27 (0.121)	0.814	0.38 (0.133)	0.15 (0.055)	0.007
Morphological completion	0.30 (0.179)	0.28 (0.098)	0.847	0.30 (0.110)	0.17 (0.052)	0.031
Word discrimination	0.47 (0.197)	0.43 (0.082)	0.715	0.30 (0.141)	0.18 (0.041)	0.110
Word analysis	**0.40 (0.141)**	**0.38 (0.172)**	**0.859**	**0.50 (0.089)**	**0.22 (0.075)**	**0.000**
Word articulation	0.65 (0.259)	0.63 (0.151)	0.895	0.28 (0.194)	0.13 (0.103)	0.139
Listening	**0.36 (0.136)**	**0.34 (0.049)**	**0.719**	**0.32 (0.069)**	**0.18 (0.045)**	**0.003**
Organizing	**0.33 (0.127)**	**0.33 (0.083)**	**1.000**	**0.41 (0.045)**	**0.18 (0.018)**	**0.000**
Speaking	**0.45 (0.146)**	**0.44 (0.083)**	**0.938**	**0.32 (0.046)**	**0.15 (0.046)**	**0.000**
Semantics	**0.34 (0.128)**	**0.35 (0.072)**	**0.858**	**0.33 (0.061)**	**0.18 (0.050)**	**0.001**
Grammar	**0.29 (0.141)**	**0.27 (0.098)**	**0.818**	**0.35 (0.045)**	**0.15 (0.034)**	**0.000**
Phonology	**0.51 (0.168)**	**0.48 (0.105)**	**0.790**	**0.36 (0.083)**	**0.18 (0.034)**	**0.002**
Overall language ability	**0.38 (0.132)**	**0.37 (0.045)**	**0.876**	**0.35 (0.035)**	**0.17 (0.022)**	**0.000**

Significant values are reported in bold

utility of social robots in speech therapy interventions, it was hard for our team to compare the outcomes of this research with others comprehensively.

In our future work, we will increase the number of participants and consider the subjects' gender as an independent variable to see whether the current findings can be generalized to RAST sessions for children with language disorders. Moreover, to encourage children to participate in the RAST sessions, they were initially engaged with the robot via an imitation game. However, the influences of the gaming scenario and the augmented robot's features, including facial expression recognition and lip-syncing capabilities, on the therapeutic interventions were not explicitly investigated. Thus, further inspections would be required to rigorously assess whether the children's language progress is attributed to only the robot's presence or the augmented capabilities implemented on the robot. Additionally, the novelty of the robot could have repercussions on the outcomes of the two scenarios. Although the first week of the examination was dedicated to introducing the robot to children, quantitative investigations would be beneficial in negating the novelty factor's impacts.

8 Conclusion

This paper addressed the potential benefits of employing a socially assistive robot in speech therapy interventions. The main focus of the study was to evaluate the robot's capacity to engage children with language disorders and enhance their learning achievements. To attain the interventions' objectives, two capabilities, facial expression recognition and lip-syncing, were developed for the employed robotic platform, the RASA social robot. The facial expression recognition model was achieved by training various well-known CNNs on the AffectNet database and modifying via the transfer learning technique to enhance the system's performance in the robot's environment. The lip-syncing capability was developed by designing and implementing an articulatory system on the robot, which endeavored to imitate human articulation. The study's results, acquired by video coding, the Persian TOLD, and statistical analysis, revealed the prospects of using the RASA robot in speech therapy sessions for children with language disorders. However, one should avoid expecting considerable improvements and consider this study's reported findings as preliminary exploratory results that must be interpreted with caution since the small number of subjects limits the investigation, as proven by the power analysis.

Acknowledgements This study was funded by the "Dr. Ali Akbar Siassi Memorial Research Grant Award" and The Sharif University of Technology (Grant No. G980517). We also thank Mrs. Shari Holderread for the English editing of the final manuscript.

Author Contributions All authors contributed to the study's conception and design. Material preparation, data collection, and analysis were performed by AE and ZR. The first draft of the manuscript was written by AE and ZR, and all authors commented on the manuscript. All authors read and approved the final manuscript.

Data Availability All data from this project (videos of the sessions, results, scores of the performances, etc.) are available in the archive of the Social & Cognitive Robotics Laboratory.

Code Availability All of the codes are available in the archive of the Social & Cognitive Robotics Laboratory. If the readers need the codes, they may contact the corresponding author.

Declarations

Conflict of interest Authors Alireza Taheri received a research grant from the "Sharif University of Technology" (Grant No. G980517). The authors Alireza Esfandbod, Zeynab Rokhi, Ali F. Meghdari, Minoo Alemi, and Mahdieh Karimi declared that they had no conflict of interest.

Ethical Approval Ethical approval for the protocol of this study was provided by the Iran University of Medical Sciences (#IR.IUMS.REC.1395.95301469).

Consent to Participate Informed consent was obtained from all individual participants included in the study.

Consent for Publication The authors affirm that human research participants provided informed consent for the publication of all participants' images. All of the participants have consented to the submission of the results of this study to the journal.

Appendix

In the Table 9, the Top 1 value describes the proportion of test samples for which the model prediction results (predicted labels with the highest probability) are in harmony with their correct labels. The Top 2 value indicates the proportion of the test samples in which the predicted classes with the two highest probabilities match their corresponding real labels.

Table 9 The accuracy of various networks trained on the AffectNet train set and tested on the AffectNet test set

CNN	MobileNet		MobileNet V2		NASNET		DenseNet 121		DenseNet 169		Xception		Inception V3		VGG 16	
	Top 1	Top 2	Top 1	Top 2	Top 1	Top 2	Top 1	Top 2	Top 1	Top 2	Top 1	Top 2	Top 1	Top 2	Top 1	Top 2
Accuracy	0.58	0.90	0.56	0.78	0.56	0.76	0.58	0.80	0.59	0.79	0.56	0.77	0.58	0.77	0.59	0.79

Table 10 The evaluation metrics of various networks trained on the AffectNet train set and tested on the AffectNet test set

	MobileNet	MobileNet V2	NASNET	DenseNet 121	DenseNet 169	Xception	Inception V3	VGG 16
Accuracy	0.58	0.56	0.56	0.58	0.59	0.56	0.58	0.59
F1 score	0.58	0.56	0.55	0.58	0.59	0.56	0.58	0.59
Recall	0.58	0.56	0.56	0.58	0.59	0.56	0.58	0.59
Precision	0.59	0.57	0.56	0.58	0.60	0.58	0.58	0.60
Cohen Kappa	0.52	0.50	0.49	0.52	0.53	0.50	0.52	0.53
AUC	0.76	0.75	0.75	0.76	0.77	0.76	0.76	0.73
#Parameters	4 M	3 M	5 M	8 M	14 M	23 M	24 M	138 M

References

1. Thomas G, Vaughan M (2004) Inclusive education: readings and reflections. Inclusive Education, ERIC
2. Meghdari A, Alemi M (2020) STEM teaching-learning communication strategies for deaf students
3. Lee H, Hyun E (2015) The intelligent robot contents for children with speech-language disorder. J Educ Technol Soc 18(3):100–113
4. Estévez D, Terrón-López M-J, Velasco-Quintana PJ, Rodríguez-Jiménez R-M, Álvarez-Manzano V (2021) A case study of a robot-assisted speech therapy for children with language disorders. Sustainability 13(5):2771
5. Pinborough-Zimmerman J, Satterfield R, Miller D, Bilder S (2007) Hossain, and W. McMahon, Communication disorders: prevalence and comorbid intellectual disability, autism, and emotional/behavioral disorders. Am J Speech Lang Pathol
6. Daily DK, Ardinger HH, Holmes GE (2000) Identification and evaluation of mental retardation. Am Fam Phys 61(4):1059–1067
7. Stevenson J, Richman N (1976) The prevalence of language delay in a population of three-year-old children and its association with general retardation. Dev Med Child Neurol 18(4):431–441
8. Bobylova MY, Braudo T, Kazakova M, Vinyarskaya I (2017) Delayed speech development in children: introduction to terminology. Russ J Child Neurol 12(1):56–62
9. Shriberg LD, Kwiatkowski J, Mabie HL (2019) Estimates of the prevalence of motor speech disorders in children with idiopathic speech delay. Clin Linguist Phon 33(8):679–706
10. Beukelman DR, Light,JC (2020) Augmentative and alternative communication: Supporting children and adults with complex communication needs, Paul H. Brookes Publishing Company
11. Buggey T (1995) An examination of the effectiveness of videotaped self-modeling in teaching specific linguistic structures to preschoolers. Topics Early Childh Spec Educ 15(4):434–458
12. van Balkom H, Verhoeven L, van Weerdenburg M, Stoep J (2010) Effects of parent-based video home training in children with developmental language delay. Child Lang Teach Ther 26(3):221–237
13. Han J, Jo M, Park S, Kim S (2005) The educational use of home robots for children. In: ROMAN 2005. IEEE international workshop on robot and human interactive communication, IEEE, pp 378–383
14. Jeon KH, Yeon SH, Kim YT, Song S, Kim J (2014) Robot-based augmentative and alternative communication for nonverbal children with communication disorders. In: Proceedings of the 2014 ACM international joint conference on pervasive and ubiquitous computing, pp 853–859
15. Malik NA, Yussof H, Hanapiah FA (2014) Development of imitation learning through physical therapy using a humanoid robot. Procedia Comput Sci 42:191–197
16. Cao H-L et al (2019) Robot-enhanced therapy: development and validation of supervised autonomous robotic system for autism spectrum disorders therapy. IEEE Robot Autom Mag 26(2):49–58
17. Kumazaki H et al (2018) The impact of robotic intervention on joint attention in children with autism spectrum disorders. Mol Autism 9(1):1–10
18. Taheri A, Meghdari A, Alemi M, Pouretemad H (2019) Teaching music to children with autism: a social robotics challenge. Scientia Iranica 26:40–58
19. Lehmann H, Iacono I, Dautenhahn K, Marti P, Robins B (2014) Robot companions for children with down syndrome: a case study. Interact Stud 15(1):99–112
20. Aslam S, Standen PJ, Shopland N, Burton A, Brown D (2016) A comparison of humanoid and non-humanoid robots in supporting the learning of pupils with severe intellectual disabilities. In: 2016 International conference on interactive technologies and games (ITAG), IEEE, pp 7–12
21. Özkul A, Köse H, Yorganci R, Ince G (2014) Robostar: an interaction game with humanoid robots for learning sign language. In: 2014 IEEE international conference on robotics and biomimetics (ROBIO 2014), IEEE, pp 522–527
22. Daniela L, Lytras MD (2019) Educational robotics for inclusive education, Springer
23. Meghdari A, Alemi M, Zakipour M, Kashanian SA (2019) Design and realization of a sign language educational humanoid robot. J Intell Rob Syst 95(1):3–17
24. Zakipour M, Meghdari A, Alemi M (2016) RASA: a low-cost upper-torso social robot acting as a sign language teaching assistant. In: International conference on social robotics, Springer, pp 630–639
25. Leite I, Martinho C, Paiva A (2013) Social robots for long-term interaction: a survey. Int J Soc Robot 5(2):291–308
26. Shin D-H, Choo H (2011) Modeling the acceptance of socially interactive robotics: social presence in human–robot interaction. Interact Stud 12(3):430–460
27. Heerink M, Kröse B, Evers V, Wielinga B (2010) Assessing acceptance of assistive social agent technology by older adults: the almere model. Int J Soc Robot 2(4):361–375
28. De Graaf MM, Allouch SB (2013) Exploring influencing variables for the acceptance of social robots. Robot Auton Syst 61(12):1476–1486
29. Mollahosseini A, Hasani B, Mahoor MH (2017) Affectnet: adatabase for facial expression, valence, and arousal computing in the wild. IEEE Trans Affect Comput 10(1):18–31
30. Egido-García V, Estévez D, Corrales-Paredes A, Terrón-López M-J, Velasco-Quintana P-J (2020) Integration of a social robot in a pedagogical and logopedic intervention with children: a case study. Sensors 20(22):6483
31. Hassanzade S, Minayi A (2002) Test of language development (TOLD-P: 3), normalized in Persian. Research institute of exceptional children publishers, Tehran
32. Newcomer PL, Hammill DD (2008) Told-p: 4: test of language development. Primary. Pro-Ed Austin, TX
33. Boccanfuso L, Scarborough S, Abramson RK, Hall AV, Wright HH, O'Kane JM (2017) A low-cost socially assistive robot and robot-assisted intervention for children with autism spectrum disorder: field trials and lessons learned. Auton Robot 41(3):637–655
34. Ramamurthy P, Li T (2018) Buddy: a speech therapy robot companion for children with cleft lip and palate (cl/p) disorder. In: Companion of the 2018 ACM/IEEE international conference on human–robot interaction, pp 359–360
35. Robles-Bykbaev V et al (2016) Robotic assistant for support in speech therapy for children with cerebral palsy. In: 2016 IEEE international autumn meeting on power, electronics and computing (ROPEC), IEEE, pp 1–6
36. Ioannou A, Andreva A (2019) Play and learn with an intelligent robot: enhancing the therapy of hearing-impaired children. In: IFIP conference on human–computer interaction, Springer, pp 436–452
37. Андреева А, Йоану А (2020) Robot-assisted speech and language therapy for children with hearing impairment. СПЕЦИАЛНА ПЕДАГОГИКА И ЛОГОПЕДИЯ 1(1):75–91
38. Zhanatkyzy A, Turarova A, Telisheva Z, Abylkasymova G, Sandygulova A (2019) Robot-assisted therapy for children with delayed speech development: a pilot study. In: 2019 28th IEEE international conference on robot and human interactive communication (RO-MAN), IEEE, pp 1–5
39. Belpaeme T, Kennedy J, Ramachandran A, Scassellati B, Tanaka F (2018) Social robots for education: a review. Sci Robot 3:21
40. Istenic Starcic I, Bagon S (2014) ICT-supported learning for inclusion of people with special needs: review of seven educational technology journals, 1970–2011. Br J Educ Technol 45(2):202–230
41. Fong T, Nourbakhsh I, Dautenhahn K (2003) A survey of socially interactive robots. Robot Auton Syst 42(3–4):143–166

42. Barros P, Weber C, Wermter S (2015) Emotional expression recognition with a cross-channel convolutional neural network for human–robot interaction. In: 2015 IEEE-RAS 15th international conference on humanoid robots (humanoids), IEEE, pp 582–587
43. Li T-HS, Kuo P-H, Tsai T-N, Luan P-C (2019) CNN and LSTM based facial expression analysis model for a humanoid robot. IEEE Access 7:93998–94011
44. Lopez-Rincon A (2019) Emotion recognition using facial expressions in children using the NAO Robot. In: 2019 International conference on electronics, communications and computers (CONIELECOMP), IEEE, pp 146–153
45. Webb N, Ruiz-Garcia A, Elshaw M, Palade V (2020) Emotion recognition from face images in an unconstrained environment for usage on social robots. In: 2020 International joint conference on neural networks (IJCNN), IEEE, pp 1–8
46. Meghdari A, Shouraki SB, Siamy A, Shariati A (2016) The real-time facial imitation by a social humanoid robot. In: 2016 4th International conference on robotics and mechatronics (ICROM), IEEE, pp 524–529
47. Chen H, Gu Y, Wang F, Sheng W (2018) Facial expression recognition and positive emotion incentive system for human–robot interaction. In: 2018 13th World congress on intelligent control and automation (WCICA), IEEE, pp 407–412
48. Ruiz-Garcia A, Elshaw M, Altahhan A, Palade V (2018) A hybrid deep learning neural approach for emotion recognition from facial expressions for socially assistive robots. Neural Comput Appl 29(7):359–373
49. Esfandbod A, Rokhi Z, Taheri A, Alemi M, Meghdari A (2019) Human–robot interaction based on facial expression imitation. In: 2019 7th international conference on robotics and mechatronics (ICRoM), IEEE, pp 69–73
50. Liu Z et al (2017) A facial expression emotion recognition based human–robot interaction system
51. Nijssen SR, Müller BC, Bosse T, Paulus M (2021) You, robot? The role of anthropomorphic emotion attributions in children's sharing with a robot. Int J Child Comput Interact 30:100319
52. McColl D, Hong A, Hatakeyama N, Nejat G, Benhabib B (2016) A survey of autonomous human affect detection methods for social robots engaged in natural HRI. J Intell Rob Syst 82(1):101–133
53. Bera A et al (2019) The emotionally intelligent robot: improving social navigation in crowded environments. arXiv preprint arXiv:1903.03217
54. Chollet F (2017) Xception: deep learning with depthwise separable convolutions. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 1251–1258
55. .Goodfellow J et al (2013) Challenges in representation learning: a report on three machine learning contests. In: International conference on neural information processing, Springer, pp 117–124
56. Simonyan K, Zisserman A (2014) Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556
57. McGurk H, MacDonald J (1976) Hearing lips and seeing voices. Nature 264(5588):746–748
58. Hyung HJ, Ahn BK, Choi D, Lee D, Lee DW (2016) Evaluation of a Korean Lip-sync system for an android robot. In: 2016 13th International conference on ubiquitous robots and ambient intelligence (URAI), IEEE, pp 78–82
59. Moubayed SA, Skantze G, Beskow J (2013) The furhat back-projected humanoid head–lip reading, gaze and multi-party interaction. Int J Hum Rob 10(01):1350005
60. Cid FA, Manso LJ, Calderita LV, Sánchez A, Nuñez P (2012) Engaging human-to-robot attention using conversational gestures and lip-synchronization. J Phys Agents 6(1):3–10
61. Castro-Gonzalez A, Alcocer-Luna J, Malfaz M, Alonso-Martin F, Salichs MA (2018) Evaluation of artificial mouths in social robots. IEEE Trans Hum Mach Syst 48(4):369–379
62. Castellano G, Leite I, Pereira A, Martinho C, Paiva A, Mcowan PW (2013) Multimodal affect modeling and recognition for empathic robot companions. Int J Hum Rob 10(01):1350010
63. Lucey P, Cohn JF, Kanade T, Saragih J, Ambadar Z, Matthews I (2010) The extended cohn-kanade dataset (ck+): a complete dataset for action unit and emotion-specified expression. In: 2010 IEEE computer society conference on computer vision and pattern recognition-workshops, IEEE, pp 94–101
64. Mohammadi MR, Fatemizadeh E, Mahoor MH (2014) PCA-based dictionary building for accurate facial expression recognition via sparse representation. J Vis Commun Image Represent 25(5):1082–1092
65. Cugu I, Sener E, Akbas E (2019) Microexpnet: an extremely small and fast model for expression recognition from face images. In: 2019 Ninth international conference on image processing theory, tools and applications (IPTA), IEEE, pp 1–6
66. Sokolova M, Japkowicz N, Szpakowicz, S (2006) Beyond accuracy, F-score and ROC: a family of discriminant measures for performance evaluation. In: Australasian joint conference on artificial intelligence, Springer, pp 1015–1021
67. Cohen J (1960) A coefficient of agreement for nominal scales. Educ Psychol Meas 20(1):37–46
68. Jeni LA, Cohn JF, De La Torre F (2013) Facing imbalanced data–recommendations for the use of performance metrics. In: 2013 Humaine association conference on affective computing and intelligent interaction, IEEE, pp 245–251
69. Howard AG et al (2017) Mobilenets: efficient convolutional neural networks for mobile vision applications. arXiv preprint arXiv:1704.04861
70. Sandler M, Howard A, Zhu M, Zhmoginov A, Chen LC (2018) Mobilenetv2: inverted residuals and linear bottlenecks. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 4510–4520
71. Zoph B, Vasudevan V, Shlens J, Le QV (2018) Learning transferable architectures for scalable image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 8697–8710
72. Huang G, Liu Z, Van Der Maaten L, Weinberger KQ (2017) Densely connected convolutional networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 4700–4708
73. Szegedy C, Vanhoucke V, Ioffe S, Shlens J, Wojna Z (2016) Rethinking the inception architecture for computer vision. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 2818–2826
74. Deng J, Dong W, Socher R, Li LJ, Li K, Fei-Fei L (2009) Imagenet: a large-scale hierarchical image database. In: 2009 IEEE conference on computer vision and pattern recognition, IEEE, pp 248–255
75. Viola P, Jones M (2001) Rapid object detection using a boosted cascade of simple features. In: Proceedings of the 2001 IEEE computer society conference on computer vision and pattern recognition. CVPR 2001, vol 1, IEEE, pp I-I
76. Izdebski Ł , Sawicki D (2016) Easing functions in the new form based on bézier curves. In: International conference on computer vision and graphics, Springer, pp 37–48
77. Penner R (2002) Motion, tweening, and easing. Programming macromedia flash MX, pp 191–240
78. Faul F, Erdfelder E, Buchner A, Lang A-G (2009) Statistical power analyses using G* Power 3.1: tests for correlation and regression analyses. Behav Res Methods 41(4):1149–1160
79. Patel R, Connaghan K (2014) Park play: a picture description task for assessing childhood motor speech disorders. Int J Speech Lang Pathol 16(4):337–343

80. Epstein S-A, Phillips J (2009) Storytelling skills of children with specific language impairment. Child Lang Teach Ther 25(3):285–300
81. Snow PC, Eadie PA, Connell J, Dalheim B, McCusker HJ, Munro JK (2014) Oral language supports early literacy: a pilot cluster randomized trial in disadvantaged schools. Int J Speech Lang Pathol 16(5):495–506
82. Geers A, Brenner C, Davidson L (2003) Factors associated with development of speech perception skills in children implanted by age five. Ear Hear 24(1):24S-35S
83. Newmeyer AJ et al (2007) Fine motor function and oral-motor imitation skills in preschool-age children with speech-sound disorders. Clin Pediatr 46(7):604–611
84. Pettinati MJ, Arkin RC, Shim J (2016) The influence of a peripheral social robot on self-disclosure. In: 2016 25th IEEE international symposium on robot and human interactive communication (RO-MAN), IEEE, pp 1063–1070
85. Taheri A, Meghdari A, Alemi M, Pouretemad H (2018) Human–robot interaction in autism treatment: a case study on three pairs of autistic children as twins, siblings, and classmates. Int J Soc Robot 10(1):93–113
86. Jones C, Sung B, Moyle W (2015) Assessing engagement in people with dementia: a new approach to assessment using video analysis. Arch Psychiatr Nurs 29(6):377–382
87. Anzalone SM, Boucenna S, Ivaldi S, Chetouani M (2015) Evaluating the engagement with social robots. Int J Soc Robot 7(4):465–478
88. Argyle M, Dean J (1965) Eye-contact, distance and affiliation. Sociometry, pp 289–304
89. Ivaldi S, Lefort S, Peters J, Chetouani M, Provasi J, Zibetti E (2017) Towards engagement models that consider individual factors in HRI: on the relation of extroversion and negative attitude towards robots to gaze and speech during a human–robot assembly task. Int J Soc Robot 9(1):63–86

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Alireza Esfandbod is a Ph.D. candidate at the Mechanical Engineering Department of the Sharif University of Technology, Tehran, Iran. His research interests include Mechatronics, Robotics, Social Robotics, Pattern Recognition, Computer Vision, Artificial Intelligence, Machine Learning, and their applications in Human-Robot Interaction.

Zeynab Rokhi received an M.Sc. degree in Mechanical Engineering from the Sharif University of Technology. Her research interests are Social Robotics, Computer Vision, Deep learning, Machine learning, and Human-Robot Interaction. She is curreny a Ph.D. student at the McMaster University in Canada.

Ali F. Meghdari is a Professor Emeritus of Mechanical Engineering and Robotics at Sharif University of Technology (SUT) in Tehran. Professor Meghdari has performed extensive research in various areas of robotics; social and cognitive robotics, mechatronics, and modeling of biomechanical systems. He has been the recipient of various scholarships and awards, including: the 2012 Allameh Tabatabaei distinguished professorship award by the National Elites Foundation of Iran (BMN), the 2001 Mechanical Engineering Distinguished Professorship Award from the Ministry of Science, Research & Technology (MSRT) in Iran, and the 1997 ISESCO Award in Technology from Morocco. He is the founder of the Centre of Excellence in Design, Robotics, and Automation (CEDRA), an affiliate member of the Iranian Academy of Sciences (IAS), a Fellow of the American Society of Mechanical Engineers (ASME), and the Founder and Chancellor of Islamic Azad University- Fereshtegaan International Branch (for students with special needs; primarily the Deaf).

Alireza Taheri is an Assistant Professor of Mechanical Engineering with an emphasis on Social and Cognitive Robotics at Sharif University of Technology, Tehran, Iran. He is the Head of the Social and Cognitive Robotics Lab at Sharif University of Technology, and Vice-Chairman of Department of Mechanical Engineering.

Minoo Alemi received her Ph.D. in Applied Linguistics from Allameh Tabataba'i University in 2011. She is currently an Associate Professor and Division Head of Applied Linguistics at the Islamic Azad University, West-Tehran Branch. She is the cofounder of Social Robotics in Iran, a title she achieved as a Post-Doctoral research associate at the Social Robotics Laboratory of the Sharif University of Technology. Her areas of interest include discourse analysis, interlanguage pragmatics, materials development, and RALL. Dr. Alemi has been the recipient of various teaching and research awards from Sharif University of Technology, Allameh Tabataba'i University, Islamic Azad University, and Int. Conf. on Social Robotics (ICSR-2014).

Mahdieh Karimi received an M.Sc. degree in the psychology of exceptional children from the Islamic Azad University, Science and Research Branch. She is an expert in speech and language pathology, with fifteen years of experience in this field, and the founder of the Mahan Rehabilitation Clinic.