Evaluating a multi-avatar game for speech therapy applications

Conference Paper · November 2018

DOI: 10.1145/3284869.3284913

7 authors, including:

Antonio Origlia
University of Naples Federico II
45 PUBLICATIONS **148** CITATIONS
SEE PROFILE

Federico Altieri
University of Padova
3 PUBLICATIONS **0** CITATIONS
SEE PROFILE

Claudio Zmarich
Italian National Research Council
90 PUBLICATIONS **346** CITATIONS
SEE PROFILE

Antonio Roda
University of Padova
72 PUBLICATIONS **455** CITATIONS
SEE PROFILE

Evaluating a multi-avatar game for speech therapy applications

Antonio Origlia
University of Naples "Federico II"
antonio.origlia@unina.it

Federico Altieri
University of Padua
altieri@dei.unipd.it

Giorgia Buscato
University of Padua
giorgia.buscato@studenti.unipd.it

Alice Morotti
University of Padua
alice.morotti@studenti.unipd.it

Claudio Zmarich
Institute of Cognitive Sciences and Technologies - CNR
claudio.zmarich@cnr.it

Antonio Rodá
University of Padua
roda@dei.unipd.it

Piero Cosi
Institute of Cognitive Sciences and Technologies - CNR
piero.cosi@cnr.it

ABSTRACT

In this paper, we propose a new set of experiments to further evaluate the performance of a previously presented system based on an adaptive strategy for stimuli selection masked behind a gamified activity. This involves two virtual agents creating a social setting designed to support a narrative to engage young children. With respect to previously obtained results, we further evaluate the quality of the support for diagnosis purposes and we present a first investigation concerning the applicability of the system for therapeutic goals.

CCS CONCEPTS

• **Applied computing → Health care information systems**; **Health informatics**;

KEYWORDS

Gamification, Dialogue systems, speech therapy

ACM Reference Format:
Antonio Origlia, Federico Altieri, Giorgia Buscato, Alice Morotti, Claudio Zmarich, Antonio Rodá, and Piero Cosi. 2018. Evaluating a multi-avatar game for speech therapy applications. In *International Conference on Smart Objects and Technologies for Social Good (Goodtechs '18), November 28–30, 2018, Bologna, Italy*. ACM, New York, NY, USA, 6 pages. https://doi.org/10.1145/3284869.3284913

1 INTRODUCTION

Administering linguistic tests to young children is a difficult task as the speech therapist must deal with attentional issues linked to the intrinsic interest children must have towards the proposed activity. Moreover, limited time and the extreme variability that can be observed among children represent further issues that are difficult to overcome. Engaging children in activities that do not possess intrinsic value for them causes lower performances and unwillingness to engage spontaneously [4]. Games have been repeatedly used, in the past, to administer tests to young children as playful activities represent an added value to this kind of activity [16] and the power of social, playful activities has been linked to biological characteristics of mammals [12]. Involving artificial agents like robots in such tasks has been done in previous works for a number of tasks, for example [14, 5], among which the treatment of speech language disorders [7, 19].

Our previous work on the subject [10] has introduced two complementary agents (a virtual avatar and a physical robot) to administer a linguistic discrimination tests to young children (5 years old). The system narrative supports a social situation in which the child helps the virtual avatar to teach the robot how to speak. This setting relies on the *learning by teaching* approach to reduce the stress for the children, who are not explicitly evaluated. Also, the system relies on graph-based knowledge of language (Italian) to adaptively select the most informative stimulus to present depending on the observed performance. This differs from traditional, scripted approaches as the test is built on the fly as the child interacts with the system.

In previous work we presented the system architecture [10] and a first set of experiments to evaluate the applicability of the approach to diagnosis tasks [11]. In this work we present a new set of experiments to extend the evaluation of the diagnostic capabilities of the system and we also introduce an investigation of the potential impact the approach has as a therapeutic tool. This ongoing work builds upon the experience of the Colorado Literacy Tutor [2] and of [3].

In the following sections, we summarise the system architecture and we provide the details of the adaptive approach we developed for online stimuli selection.

2 SYSTEM ARCHITECTURE

Among several types of discrimination tests, we choose the standard AX or "same-different" procedure. Traditionally, AX tests to

Goodtechs '18, November 28–30, 2018, Bologna, Italy

ACM ISBN 978-1-4503-6581-9/18/11…$15.00
https://doi.org/10.1145/3284869.3284913

evaluate the phonemes discrimination capability of young children are designed as scripts and software traditionally used to administer this kind of test also follows scripts (e.g. [1]). These contain a series of (non-) word pairs presenting phoneme oppositions (i.e. *'pepi* / *'pebi*) in different syllabic structures (i.e. CV-CV is a disyllabic structure where each syllable has a single heading consonant). The child is given the task to indicate, after listening to the experimenter reading the stimuli, whether the two (non-)words are the same or if they are different. These tests are designed in such a way that consonants presenting a single distinctive feature [6] are opposed at each time (e.g. voiced/unvoiced, sonorant/non-sonorant). Control stimuli are present in such tests as pairs composed by the same (non-)word repeated twice and by pairs composed by completely different (non-)words. This approach is necessary as it is impossible for a human expert to dynamically select stimuli pairs that comply to a set of very strict constraints. Specifically, each stimulus must:

- present opposed consonants that differ in exactly one feature
- syllabic structure must be the same in the two (non-)words
- present the opposition in a precise position in the syllabic structure (e.g. the onset consonant of the second syllable)
- the stress must be in the same place in the two (non-)words

The system architecture presented in [10] has two main purposes: a) dynamically adapt the test to the child's performance and b)support groups of virtual agents to establish social setups. The first goal is accomplished by introducing a probabilistic dialogue manager (Opendial [8]) and a graph based representation of knowledge of Italian language sounds implemented using the Neo4J database [18]. The probabilistic framework uses utility functions to estimate how informative each available stimulus is about the observed child to dynamically select the most informative one depending on previous performance. This approach lets the system compile a probabilistic summary for each linguistic trait (as defined in [13], with the exception of the introduction we made of the *Length* feature) describing the estimated probability that the child will be able to discriminate specific traits depending on the observed performance. The interface of the system consists of a Nao robot and of a virtual avatar animated in the Unreal Engine 4[1]. Differently from previous tests, where we used Nao's synthetic voice engine for the robot, both actors are now provided with a synthetic voice provided by the Mivoq Voice Synthesis Engine[2]. This improves the quality of the system as accidental mismatches caused by the different synthesis engines are now eliminated so that the test is more consistent. Since the ability to adequately use a tablet interface appears to be reliable for 5 years old and older children [17], this is the chosen method for feedback input from the children we selected. The updated system architecture is shown in Figure 1. The reader is referred to [10] for the technological details.

3 ADAPTIVE GAMIFIED TESTS

In order to dynamically select the stimuli to present through the two agents, the underlying dialogue system uses a probabilistic model designed to take into account the observe performance and select the next stimulus accordingly to maximise the informative utility. In this section we detail the probabilistic model driving the system.

The probability of a subject to assign a label to the presented opposition is a binomial distribution (Equal/Different). Therefore, to represent a priori probabilities built using previous feedback, the conjugate prior of the binomial distribution, the Beta distribution, is used. Following the Opendial implementation, a two dimensional Dirichlet probability density function with parameters (α_1, α_2) is used to model the conjugate prior. The a priori probabilities of the labels a subject can assign to a stimulus are represented as

$$\theta_{t_m}(Hit, Miss; (\alpha_1, \alpha_2)) \quad | \quad t_m \in \mathbf{T} \tag{1}$$

In order to evaluate which stimulus represents the most informative choice given the available probability density functions, we consider the entropy as a first factor. Since the densities are symmetrical among 0.5 on the two considered dimensions, we define the entropy of the m-th tract t_m as

$$H(\theta_{t_m}(Hit, Miss; (\alpha_1, \alpha_2))) = -\sum_{i=1}^{n} p_{t_m}(x_i; \alpha_1) log(p_{t_m}(x_i; \alpha_1)) \tag{2}$$

where $p_{t_m}(x_i; \alpha_1)$ represents the sampled probability density function at point x_i on the first dimension. The maximally entropic feature is therefore defined as

$$t_H = max(H(\theta_{t_m}(Hit, Miss; (\alpha_1, \alpha_2)))) \quad | \quad t_m \in O_{ij} \tag{3}$$

The entropy-based utility function for the opposition $(s_i, s_j$ is computed by subtracting t_H to the maximum utility value normalised by the maximally entropic feature in $\mathbf{T}$ as

$$U_H(s_i, s_j) = 1 - \frac{t_H}{min(H(\theta_{t_m}(Hit, Miss; (\alpha_1, \alpha_2))))} \quad | \quad t_m \in \mathbf{T} \tag{4}$$

This function assigns higher utility values to stimuli presenting the opposing features for which the associated probability density functions present the higher uncertainty.

An opposition presenting more than one highly entropic feature is not an optimal choice as it is not possible to evaluate which feature influenced the outcome. This is the reason why, for scripted tests, it is not possible to use phoneme pairs opposing more than one feature, which becomes a problem in tests opposing words as there may not be phonological neighbours with the specified structure opposing exactly the phonemes involved in the feature of interest. For a dynamically constructed test, instead, given a generic pair (s_i, s_j), let's consider the simple case of two features found in opposition so that $O_{ij} = t_1, t_2$, if t_1 is likely not to be discriminated by the subject, it is possible to use the opposition (s_i, s_j) to investigate t_2. Following the approach adopted by Opendial, Bayesian inference is used to adequately update all the involved probability distributions. The usefulness of the opposition becomes greater as the probability t_1 not to be discriminated becomes higher. On the other hand, if t_1 is likely to be discriminated by the subject, the opposition becomes less useful to investigate t_2. To include this evaluation, we first define the mean probability for an opposition on the m-th feature to be missed as

[1] www.unrealengine.com
[2] www.mivoq.it

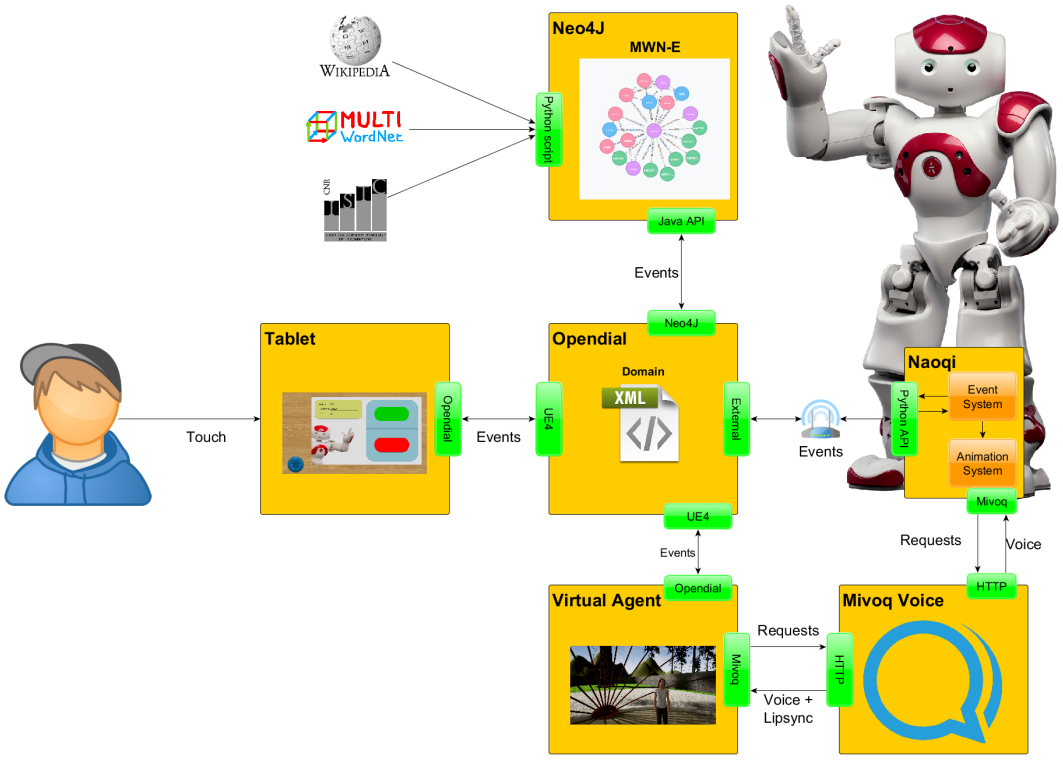

Figure 1: System Architecture.

$$\mu(\theta_{t_m}(Miss;\alpha_2)) = \frac{\sum_{i=1}^{n} p_{t_m}(x_i;\alpha_2)}{n} \tag{5}$$

The mean probability of the most entropic feature in O_{ij} is taken as reference and defined as

$$\mu_H = \mu(\theta_{t_H}(Miss;\alpha_2)) \tag{6}$$

The mean-based utility function is, then, defined as the minimum difference between μ_H and $\mu(P_{t_m}(Miss;\alpha_2))$ computed for all other features in O_{ij}. The minimum difference can be negative, indicating that in O_{ij} there is a feature that is likely to be discriminated by the subject. To normalise the score in the range $[0,1]$, we define the mean-based utility function as

$$U_\mu(s_i,s_j) = \frac{min(\mu_H - \mu(\theta_{t_m}(Miss;\alpha_2))) + 1}{2} \quad | \quad t_m \in O_{ij} \backslash \{t_H\} \tag{7}$$

This function assigns a higher utility value to oppositions presenting a single, highly entropic, feature together with features that have been found not to be discriminated. The higher the likelihood of other features not to be discriminated, the higher the utility.

Since the task complexity can be influenced by the age acquisition difference in the involved phonemes, we model a substitution-based utility function as

$$U_S(s_i,s_j) = 1 - \frac{j-i}{n} \tag{8}$$

This function assigns a higher value to phoneme oppositions that are closer to each other in **S**. As this is a relative measure of phoneme-based complexity for the (s_i,s_j) opposition, we also need an absolute measure to prefer phonemes acquired earlier. We therefore define an acquisition-based utility function as

$$U_A(s_i,s_j) = 1 - \frac{i+j}{2n} \tag{9}$$

Since $U_H(s_i,s_j), U_\mu(s_i,s_j), U_S(s_i,s_j)$ and $U_A(s_i,s_j)$ are different measures of the same object (s_i,s_j) sharing the same range $[0,1]$, the final utility function for the opposition (s_i,s_j) is computed as the harmonic mean of these four measures and is therefore defined as

$$U(s_i,s_j) = \frac{4}{\frac{1}{U_H(s_i,s_j)} + \frac{1}{U_\mu(s_i,s_j)} + \frac{1}{U_S(s_i,s_j)} + \frac{1}{U_A(s_i,s_j)}} \tag{10}$$

This function lets the dialogue manager select the optimal stimulus for the next turn. The algorithm for dialogue management, implemented in Opendial and exploiting the MWN-E data, proposes a stimulus at each step and updates the probability distributions according to the feedback given by the subject. The algorithm to administer the test can be summarised as shown in Algorithm 1.

4 EXPERIMENTAL SETTING

The interface proposed to the child to mask the discrimination test supports a narrative in which the Nao robot wants to learn how

Algorithm 1 The dialogue management algorithm

procedure $wordsExist(s_i, s_j)$
if $isEmpty(queryMWNE(s_i, s_j))$ **then**
  return $false$
return $true$
end procedure

procedure $main$
for all t_m **do**
  $\theta_{t_m} = Dirichlet(1, 1)$
  $\mu(\theta_{t_m}) = 0.5$
for $k = 1 \quad to \quad MAXITEMS$ **do**
  for all (s_i, s_j) **do**
    if $wordsExist(s_i, s_j)$ **then**
      $U[i, j] = U(s_i, s_j)$
      $W[i, j] = queryMWNE(s_i, s_j)$
      $O[i, j] = opposed features(s_i, s_j)$
    else
      $U[i, j] = 0$
  $feedback = presentStimulus(W[argmax(U)])$
  for all $t_m \quad in \quad O[argmax(U)]$ **do**
    $update(\theta_{t_m}, feedback)$
    $update(\mu(\theta_{t_m}), feedback)$
end procedure

to speak and the 3D character needs the child's help to teach it. Through this approach, the child is given an authoritative role to avoid making him feel threatened or evaluated. When the system starts, an introductory scenario is presented and the 3D character introduces itself. The scenario ends with the 3D character asking the child to caress Nao in order to wake it up. This has both the goal of providing the invitation to play and to establish physical contact between Nao and the child. Whether the physical attributes of robots constitute an advantage for acceptability *per se* is still a debated issue. In our work, we attempt to fully exploit the physical presence of the robot by presenting tasks that require the child to physically interact with it. By proposing activities that a 3D character simply cannot be involved into, we capitalise on the robot's potential to provide a more engaging multisensorial experience. Caressing, in particular, is a powerful social mean to build attachment. On the other hand, the high level of control over the 3D character movements allows to efficiently represent its higher competence in the considered setup: differently from Nao, the avatar can move the lips and change its facial expressions, providing effective indications on how to continue playing. An advantage of the presented architecture is that different virtual agents can be combined to build the test upon the various advantages they offer. As a final step, the child is required to provide a same/different feedback using an evaluation card that appears on the tablet.

Concerning the recruitment phase, two groups of children were selected from two different schools in Padua (Italy), in collaboration with the schools' personnel and after presenting the experimental setup to the parents, who gave their consent to the participation of the children to the experiments. For the experimental group, the initially set of candidates comprised 14 children, which were reduced to 8 because of age incompatibility (3), unwillingness to participate (1) and subsequent absence for holidays (2). For the control group, the initial set of 5 candidate children was reduced to 4 because one child was absent for holidays.

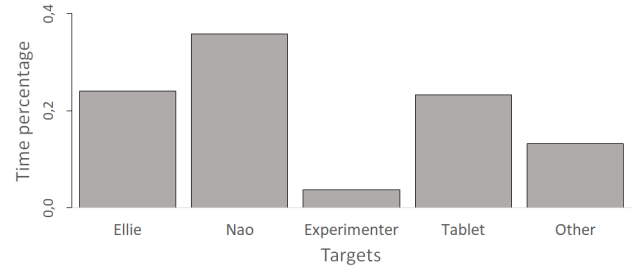

Figure 2: Overall gaze distribution.

To obtain a standard reference for the applicability of the system to diagnosis tasks, the children of the experimental group were administered a) the non-word phonological discrimination subtest of BVN 5-12 [15] and the b) word phonological discrimination and c) word and d) non-word repetition subtests of BVL 4-12 [9] both before and after being exposed to the experience with the presented system. In order to evaluate the presence of a potential impact of the system on the performance of the children by removing the training effect of the first standard test, subjects from the control group were administered the reference tests at a distance of 7 days. For logistics reasons, the sessions with the experimental group were organised in subsequent school days. It was also necessary to administer the test in rooms adjacent to playground areas, thus making the setting more difficult for the system because of some background noise.

5 RESULTS

First, we estimate the level of engagement the children showed towards the system. Using the ELAN software [20], we annotated the gaze targets of 4 randomly chosen children in the recorded videos of the four sessions, specifying the time intervals in which the child was looking at: a) the experimenter, b) the tablet, c) the avatar (Ellie), d) the Nao robot or e) other objects. The overall gaze distribution, shown in Figure 2, shows that the children were paying attention to the system elements for more than 70% of the time.

More in detail, the gaze distribution computed over the different sessions shows that the percentage of attention towards Nao remains unchanged over the four sessions. Attention towards the experimenter tends to decrease as the children get accustomed to the system and attention towards the avatar lowers mostly in favour of the tablet. This is because the children understand that it is sufficient to listen to the avatar rather than look at it in order to complete the task. These results are compatible with the ones presented in our previous work.

In order to evaluate the performances, the posterior probability density functions associated with each tract are considered. First of all, the distribution associated with the control stimuli is used to validate the session. This is marked as *Reliable* if the difference between the estimated probability of the child to provide a correct

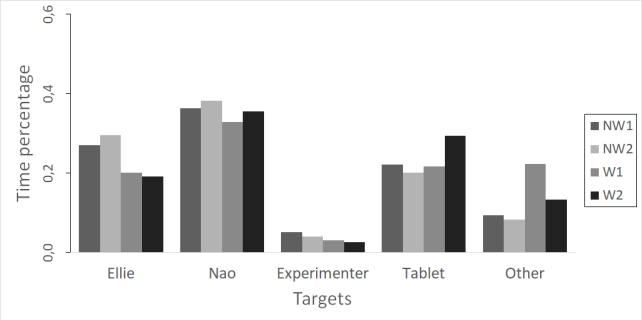

Figure 3: Gaze distribution over the targets involved in the experimental setting (details per session)

Table 1: Example summary (F3/W2)

Feature	$\mu(Ok)$	$\mu(Wrong)$	σ	N	Evaluation
Control	0.888	0.112	0.06	28	Reliable
Sonorant	0.75	0.25	0.202	2	Weak Ok
Continuous	0.779	0.221	0.114	12	Strong Ok
D.Solution	0.775	0.225	0.168	3	Strong Ok
Voiced	0.622	0.378	0.129	14	Strong Ok
Nasal	0.5	0.5	0.289	0	Unknown
Lateral	0.760	0.24	0.193	2	Weak Ok
Coronal	0.807	0.193	0.109	12	Strong Ok
Anterior	0.81	0.189	0.129	8	Strong Ok
Posterior	0.5	0.5	0.289	0	Unknown
Length	0.827	0.173	0.146	4	Strong Ok

Table 2: Entry test results

	M1	M2	F1	F2	F3	F4	M3	F5
Word disc	0.90	1.00	1.00	1.00	1.00	0.90	0.90	1.00
Word rep	0.60	1.00	0.80	0.93	1.00	0.80	0.87	0.93
Nonce disc	0.86	0.97	0.86	0.92	1.00	0.84	0.81	0.95
Nonce rep	0.60	0.67	0.67	0.87	0.87	0.73	0.53	0.80
HMean	0.71	0.88	0.82	0.93	0.96	0.81	0.74	0.91

Table 3: Exit test results

	M1	M2	F1	F2	F3	F4	M3	F5
Word disc	1,00	1,00	1,00	0,93	1,00	0,80	0,90	1,00
Word rep	0,93	0,87	1,00	1,00	0,93	0,67	1,00	0,93
Nonce disc	0,97	0,97	0,92	1,00	0,97	0,89	0,95	0,97
Nonce rep	0,87	0,93	1.00	0.87	1.00	0.60	0.87	1.00
HMean	0.94	0.94	0.98	0.95	0.98	0.72	0.93	0.98

answer is higher than the probability of the child to provide a wrong answer for the control stimuli in a statistically significant way using one tailed t-tests ($\alpha = 0.01$). All tests were marked as *Reliable* in this set of experiments.

Since the means of the a posteriori probability density functions associated with each feature must be evaluated depending on the significance of the result, a simple scoring system was adopted to obtain the ranking. The considered categories, together with the conditions they describe, are defined as follows:

- **Reliable/Unrealiable**: whether the probability of the child correctly identifying a control stimulus is significantly higher than the opposite case (and viceversa);
- **Strong problem**: the probability of the child giving a wrong answer is significantly higher than the probability of obtaining a correct answer;
- **Weak problem**: the probability of the child giving a wrong answer is higher than the probability of obtaining a correct answer but not in a statistically significant way;
- **Unknown**: the system did not present a sufficient number of stimuli to compute the statistical significance of the difference ($n < 2$);
- **Weak Ok**: the probability of the child giving a wrong answer is lower than the probability of obtaining a correct answer but not in a statistically significant way;
- **Strong Ok**: the probability of the child giving a wrong answer is significantly lower than the probability of obtaining a correct answer.

Table 4: Scoring table for features based evaluation

Result	Score
Strong problem	-2
Weak problem	-1
Unknown/Unreliable	0
Weak Ok	1
Strong Ok	2

Table 5: Subjects ranking

T1	T2	System
F3	F1	F3
F2	F3	F4
F5	F5	M2
M2	F2	F2
F1	M2	M1
F4	M1	F1
M3	M3	F5
M1	F4	M3

The scoring system we adopted is presented in Table 4 and the rankings obtained with the entry test (T1), the exit test (T2) and from the system evaluation (System) are reported in Table 5.

To compare the different rankings, we consider the average quadratic distance between the positions of the T1 ranking and the positions of the T2 ranking as a reference to estimate variations in the subjects' performance between the two tests. The ranking difference between the entry and exit test was 3.75. The average quadratic distance between the positions of the T1 ranking and the positions of the ranking obtained by the system was 6. A two tailed t-test to compare the two averages indicated that no statistically significant difference could be found between the two rankings ($\rho > 0.45$), suggesting that the ranking obtained by the system

Table 6: Entry test results (Control group)

	CGM1	CGF1	CGF2	CGF3
Word disc	0.80	0.97 0	0.77	1
Word rep	0.80 0	0.87	0.80	0.73
Nonce disc	0.81	0.97	0.86	0.81
Nonce rep	0.80	0.87	0.80	0.33
HMean	0.80	0.92	0.81	0.61

Table 7: Exit test results (Control group)

	CGM1	CGF1	CGF2	CGF3
Word disc	0.87	0.97	0.87	0.90
Word rep	1	0.93	0.93	0.73
Nonce disc	0.89	0.89	0.89	0.84
Nonce rep	0.80	1	0.80	0.47
HMean	0.88	0.95	0.87	0.69

Table 8: Subjects ranking (Control group)

T1	T2
CGF1	CGF1
CGF2	CGM1
CGM1	CGF2
CGF3	CGF3

does not differ significantly from the one obtained by administering the T1 test. As for other measures in this work, however, the reduced size of the considered groups does not allow us to draw definitive conclusions about the system being able to replicate the performance of static tests. In this work, we also investigate the possibility that the proposed system may be used for therapeutic goals. A control group composed of four subjects (three females and one male) was administered the entry and exit tests without being exposed to the proposed system. The registered variations represent a reference for natural variations in the capabilities of the children induced by both spontaneous improvement due to more time for natural learning process and the training effect caused by the administration of the entry test. The results of the control group in both the entry and exit tests are reported in Tables 6 and 7.

The obtained rankings are represented in Table 8. The average quadratic distance between the positions of the subjects in the two rankings is 0.5. This result would suggest that the subjects that were not exposed to the system improved their performance in a more uniform way with respect to the group that was exposed to the system, suggesting an impact at least on some of the subjects of the experimental group. However, as the difference between this measure and the average quadratic distance obtained from the experimental group is not statistically significant ($\rho > 0.25$), a new experiment involving a larger group of subjects is needed to confirm this indication.

6 CONCLUSIONS

We have presented a new set of experiments to investigate the capabilities of a gamified, adaptive approach to speech therapy to support both evaluation and therapy. The obtained results, while still preliminary, suggest that the system is able to provide information about the performance of the subjects that is comparable to the one obtained with traditional tests. Also, more detailed reports can be obtained through the proposed system. Also, we investigated the impact the system may have on the performance of children in this kind of tasks and found indications that, with respect to children who were not exposed to the gamified experience, at least some of the children who interacted with the system may have obtained beneficial effects. As the considered groups were small, however, further testing is needed to confirm these indications.

REFERENCES

[1] Carine André, Alain Ghio, Christian Cavé, and Bernard Teston. 2007. PERCEVAL: a computer-driven system for experimentation on auditory and visual perception. *CoRR*, abs/0705.4415. http://arxiv.org/abs/0705.4415.

[2] Ronald A Cole. 2003. Roadmaps, journeys and destinations speculations on the future of speech technology research. In *Eighth European Conference on Speech Communication and Technology*.

[3] P Cosi, R Delmonte, S Biscetti, RA Cole, B Pellom, and S van Vuren. 2004. Italian literacy tutor tools and technologies for individuals with cognitive disabilities. In *Proceedings of InSTIL/ICALL2004-NLP and Speech Technologies in Advanced Language Learning Systems-Venice*. Vol. 17, 19.

[4] Edward L Deci, Richard Koestner, and Richard M Ryan. 1999. A meta-analytic review of experiments examining the effects of extrinsic rewards on intrinsic motivation. *Psychological bulletin*, 125, 6, 627.

[5] Deanna Hood, Severin Lemaignan, and Pierre Dillenbourg. 2015. When children teach a robot to write: an autonomous teachable humanoid which uses simulated handwriting. In *Proceedings of the Tenth Annual ACM/IEEE International Conference on Human-Robot Interaction*. ACM, 83–90.

[6] Roman Jakobson, C Gunnar Fant, and Morris Halle. 1951. *Preliminaries to speech analysis: The distinctive features and their correlates*. MIT press.

[7] Hawon Lee and Eunja Hyun. 2015. The intelligent robot contents for children with speech-language disorder. *Educational Technology and Society*, 18, 3, 100–113.

[8] Pierre Lison and Casey Kennington. 2016. Opendial: a toolkit for developing spoken dialogue systems with probabilistic rules. *ACL 2016*, 67.

[9] Andrea Marini, Luigi Marotta, Sara Bulgheroni, and Franco Fabbro. 2015. Batteria per la valutazione del linguaggio in bambini dai 4 ai 12 anni (bvl_4-12). *Firenze: Giunti OS*.

[10] Antonio Origlia, Piero Cosi, Antonio Rodà, and Claudio Zmarich. 2017. A dialogue-based software architecture for gamified discrimination tests. In *Proceedings of GHItaly*.

[11] Antonio Origlia, Antonio Rodá, Claudio Zmarich, Piero Cosi, Stefania Nigris, Benedetta Colavolpe, and Ilaria Brai. 2018. Gamified discrimination tests for speech therapy applications. In *Proceedings of AISV (to appear)*.

[12] Stephen W Porges. 2007. The polyvagal perspective. *Biological psychology*, 74, 2, 116–143.

[13] Stephan Schmid. 1999. *Fonetica e fonologia dell'italiano*. Paravia scriptorium.

[14] Fumihide Tanaka and Shizuko Matsuzoe. 2012. Children teach a care-receiving robot to promote their learning: field experiments in a classroom for vocabulary learning. *Journal of Human-Robot Interaction*, 1, 1.

[15] P. Tressoldi, M. Vio, M. Gugliotta, P.S. Bisiacchi, and M. Cendron. 2005. Batteria di valutazione neuropsicologica per l'etá evolutiva (bvn 5-11). *Erickson: Trento*.

[16] Colwyn Trevarthen. 2009. The functions of emotion in infancy. In *The healing power of emotion: Affective neuroscience, development & clinical practice (Norton Series on Interpersonal Neurobiology)*. D. Fosha, D. J. Siegel, and M. F. Solomon, (Eds.) WW Norton & Company, 55–85.

[17] Radu-Daniel Vatavu, Gabriel Cramariuc, and Doina Maria Schipor. 2015. Touch interaction for children aged 3 to 6 years: experimental findings and relationship to motor skills. *International Journal of Human-Computer Studies*, 74, 54–76.

[18] Jim Webber. 2012. A programmatic introduction to neo4j. In *Proceedings of the 3rd annual conference on Systems, programming, and applications: software for humanity*. ACM, 217–218.

[19] Jacqueline Kory Westlund and Cynthia Breazeal. 2015. The interplay of robot language level with children's language learning during storytelling. In *Proceedings of the Tenth Annual ACM/IEEE International Conference on Human-Robot Interaction Extended Abstracts*. ACM, 65–66.

[20] Peter Wittenburg, Hennie Brugman, Albert Russel, Alex Klassmann, and Han Sloetjes. 2006. Elan: a professional framework for multimodality research. In *Proceedings of LREC*. Vol. 2006, 5th.