IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received June 1, 2021, accepted June 15, 2021, date of publication June 18, 2021, date of current version July 7, 2021.

Digital Object Identifier 10.1109/ACCESS.2021.3090516

Evaluating the Impact of an Autonomous Playing Mode in a Learning Game to Train Oral Skills of Users With Down Syndrome

DAVID ESCUDERO-MANCEBO, MARIO CORRALES-ASTORGANO, VALENTÍN CARDEÑOSO-PAYO, (Member, IEEE), AND CÉSAR GONZÁLEZ-FERRERAS
Departamento de Informática, Universidad de Valladolid, 47011 Valladolid, Spain

Corresponding author: David Escudero-Mancebo (descuder@infor.uva.es)

This work was supported in part by the Ministerio de Ciencia, Innovación y Universidades and the European Regional Development Fund (FEDER) under Grant TIN2017-88858-C2-1-R, and in part by the Consejería de Educación de la Junta de Castilla y León under Grant VA050G18.

This work involved human subjects or animals in its research. Approval of all ethical and experimental procedures and protocols was granted by the Research Ethics Committee of the University of Valladolid under Approval No. PI 20-1639 NO HCUV, and was aligned with the Spanish legislation.

ABSTRACT The use of ICT tools is broadly extended among people with intellectual disabilities and also, to a lesser degree, the use of learning tools including learning games. Although the use of learning games is widely accepted due to its high engagement capacity, there are few studies that analyze its usability for people with intellectual disabilities. This work presents an evaluation of the impact of adding an autonomous playing mode on the usability of a learning game designed to aid the training of oral skills in people with Down syndrome. A study in which the effectiveness, efficiency and user satisfaction of a learning game are compared when the system is used with different degrees of teacher supervision is carried out. A learning game originally designed to train oral competencies for people with Down syndrome in a teacher supervised scenario is adapted to allow its autonomous use, by including a module that provides the automatic assessment of oral productions. The use of the tool is thus compared in three different scenarios: a supervised environment, autonomous use, and laboratory use with multiple users working in parallel. The different usability evaluation instruments used reveal that, although there are no differences in the degree of engagement, there may be important differences regarding session performance: the quality of the audios is lower in the laboratory sessions and the number of errors increases in the autonomous sessions. We conclude that, although the autonomous use of learning games by users with intellectual disabilities is possible, and this can lead to considerable savings in human resources, if the feedback provided by the game is not comparable with that provided by the teacher, performance may drop considerably although the degree of engagement is maintained.

INDEX TERMS Computer assisted pronunciation training, intelligent tutoring systems, down syndrome speech.

I. INTRODUCTION

Information and communication technologies (ICT) have entered our daily activities, and this is also true for people with intellectual disabilities [1], [2]. For example, social networks, one of the most used ICT tools, are reported to be frequently used by people with intellectual disabilities [3].

The associate editor coordinating the review of this manuscript and approving it for publication was Francisco J. Garcia-Penalvo.

One of the areas in which ICTs have been most ambitiously introduced is the field of education, with specific learning and skills training tools. Among these applications, we can find proposals for software tools and learning games specially designed for people with intellectual disabilities [4]–[6].

One of the competencies in which certain people with intellectual disabilities need specific training is in communication skills. For example, people with Down syndrome often have important social and interaction limitations derived from poor

communication management of oral production [7]. There are tools for the training of oral communication that focus on pronunciation (see section II-A), and many of them are gamified (Section II-B).

We presented a gamified tool for training oral competencies related to prosody in [8] aimed at users with Down syndrome. The use of the tool has allowed us to learn more about the speech of this type of users [9]. One of the needs identified during these years of learning game testing is to offer users the possibility of playing autonomously. The current version is designed for the player to practice with an adult (typically the teacher or the therapist) who assists the user and makes decisions about the quality of the audio and whether or not it is necessary to repeat activities. Training sessions have made it possible to compile a wide corpus [10] that has enabled the training of an automatic system (described in section IV) that plays the role of an intelligent tutor, allowing the user to play autonomously.

In this article, we analyze how the incorporation of a module for the autonomous use of the learning game affects usability. To do so, we ask the following research questions:

RQ1: Is it possible to drive the training of speech pronunciation in a learning game with an automatic component?

RQ2: How does the introduction of the automatic module affect usability and enable autonomous training?

To answer these questions, we compare the results of using the tool in three different scenarios: supervised use by the therapist, autonomous use by the student and semi-supervised use in a laboratory where several students practice simultaneously with a single teacher. Comparing different versions of a video game is a challenging task because evaluating usability in educational video games requires taking into account the application domain and both the player characteristics and the therapist or teacher perspective. In [11], authors present a systematic review of methods for evaluating usability in learning games. The results reveal the diversity of approaches, dimensions evaluated and instruments used in the different state of the art works. One of the reasons for this diversity is the multiple applications of the games, but another one is that the concept of usability is merged with those of player experience and playability every time the interaction is gamified and there is still some controversy related with the definition of such terms [12]. In this work we use the well known framework of Quesenbery [13] for presenting the concepts of effectiveness, efficiency, easiness to learn, error tolerance and engagement and relating our evaluation instruments with them. To avoid misinterpretations, we take care of presenting the definition of these dimensions and declaring what are the particular aspects we are interested in comparing, taking into account our research goal of including an automatic module for evaluating the oral production activities of the learning game. Section III presents the tools used to analyze the usability dimensions relative to the effectiveness, efficiency, easy to learn, tolerance to errors and engagement. The results presented in section VI show important differences in performance that do not affect engagement. We end the article with the discussion and conclusions in sections VII and VIII.

II. RELATED WORK

People with speech disorders need treatment, but there is a limited accessibility to speech and language pathologists. The lack of appropriate treatment can limit speakers integration in society. People with speech disorders experience difficulties to communicate in public environments, where they feel less comfortable [14]. Moreover, speech impairment has an adverse impact on such life activities as learning, applying knowledge, focusing attention, thinking, communication, interpersonal relationships with friends and family, or acquiring and maintaining a job [15]. Therefore, the use of Computer-Aided Speech Therapy tools is a reasonable alternative. In fact, virtual speech therapists are becoming increasingly popular because of their availability, versatility and portability. They result in more affordable services as well. Moreover, tools have the potential of obtaining objective measures of the speech produced by the patient. These software tools may supplement face-to-face speech and language therapy and facilitate higher practice intensity. Finally, speech therapy software may also increase engagement and motivation with learning tasks.

Computer programs are especially useful for speech therapy as they allow the incorporation of audio files that capture the differences between pronunciations, in order to work in the perception domain, and to record their own productions, in order to work in the production domain. Perception activities focus on the discrimination between particular speech elements, while production activities focus on acquiring utterances pronounced by the student until a correspondence with the model is achieved. Both perception and production activities can be automated using software. Furthermore, the different linguistic activities can be contextualized, which also helps in the learning process.

A. SPEECH TECHNOLOGIES FOR PRONUNCIATION TRAINING

There are several techniques for speech therapy whose objective is to train and improve the different communicative skills in individuals who have a speech disorder [16]. Speech technologies can be used to develop software tools to assist therapists and patients in the diagnosis and treatment of the different disorders. There are also some tools that can be used autonomously by the patients. The tools that use speech technologies must be robust with respect to the sources of speech variability that are characteristic of individuals with voice disorders.

Tools need to include a module of automatic evaluation of the speech quality. This automatic evaluation allows feedback feedback to be provided the patients. Several experiments carried out to develop and evaluate such assessment modules are found in the literature. The experiments are

focused mostly on specific aspects or reduced populations. Some works focus on the speech intelligibility of people with aphasia [17]–[19]. Other experiments are reported with patients who had their larynx removed due to cancer and with children with cleft lip and palate [20]. The evaluation of speech intelligibility in individuals with advanced head and neck cancer, cerebral palsy and amyotrophic lateral sclerosis is described in [21]. Others try to predict the intelligibility in Parkinson's disease [22] or for dysarthric speech [23]–[29]. Finally, the assessment of atypical prosody in children with autism spectrum disorder is presented in [30]. In all these experiments, a speech corpus is used to train and evaluate the classification systems, and a subjective evaluation carried out by experts is used as a gold standard.

There are tools to train phonetic articulation [31] and to train basic speech production skills (intensity, blow, vocal onset, phonation time, tone, and vocalic articulation) [32]. Other tools incorporate automatic speech recognition technology to allow aphasia patients to perform word naming training exercises [33]. A system for the stuttering problem using speech recognition, text-to-speech and a talking head is described in [34]. A system for automated speech therapy in childhood apraxia of speech (CAS) is presented in [35]. The system is able to identify the three main types of error commonly associated with CAS: groping errors, articulation errors and prosodic errors.

In the case of prosody learning, a software that augments text with visual prosodic cues to improve expressive reading is presented in [36]. An experiment with children suggests that beginning readers benefit from explicit visual prosodic cues and this improves oral reading expressiveness. There are systems for teaching prosody for children with a hearing impairment or with a cochlear implant [37]. Intensity, intonation and rhythm are presented visually to the students as visual feedback and automatic assessment scores are also given.

Another tool to assess prosody in clinical and developmental populations with atypical speech motor control is described in [38]. The tool evaluates the user's ability to reconstruct prosodic contrasts in a visual-spatial domain. Speech features of pitch and duration are represented on a 2D graphical display and the user can arrange linguistic elements in order to alter the prosodic dimensions. The resulting configuration is synthesized to produce novel prosody.

Talking heads and audiovisual speech synthesis have been used to show patients how to make a correct articulation. For instance, a virtual talking head was adapted to show articulatory movements of the lips, jaw, tongue, and velum [39]. This tool has been developed for speech therapists for use in speech therapy to explain the correct pronunciation of different phonemes. A talking head is also used in a system for hard of hearing children [40]. The talking head provides the audiovisual representation of the speech process with the visual representation of articulation. The system includes automated speech production assessment, which allows feedback to be provided about the pronunciation quality of words and sentences uttered during unsupervised practice.

Some of the systems allow the pathologist to assign exercises to each patient. There are systems that include an authoring tool that allows the development of exercises, assigning them to their clients, and monitoring their performance [41]. The system uses the visual animation of an animated agent to deliver speech and language therapy to persons with aphasia. Other systems allow a speech language pathologist to assign speech production exercises to each child through a web interface [42]. Then, the child practices these exercises in the mobile app and the pathologist can review the individual recordings and the automated scores assigned by the system through a web interface. The pathologist can then provide feedback to the child and adapt the training program as needed.

B. DIGITAL LEARNING GAMES AND GAMIFICATION ON COMPUTER-ASSISTED PRONUNCIATION TRAINING

The potential of games to improve motivation and engagement in education has been examined. The idea is to apply typical elements of computer games, such as feedback, guidance, time pressure, and rewards [43]. These elements can also be included in non-gaming systems to improve user experience and engagement [44]. Games allow learners the chance to practice in environments close to real world situations. Rather than applying learning in a vacuum, games provide a context for learning, allowing players to apply what they have learned to solve real life problems [45]. Modern theories of effective learning suggest that learning is most effective when it is active, experiential, situated, problem-based and provides immediate feedback [46]. Video games can include activities which have these features.

We have found few examples of the use of games for pronunciation training for individuals with speech disorders in the state of the art. A within-subjects experiment using a game for providing speech training in elderly patients with dysarthria due to Parkinson's disease or a stroke is described in [47]. Treasure Hunters is a two-player cooperative game in which players navigate a virtual map and need to help each other to find the treasure. Players receive automatic feedback on voice loudness and pitch. Another example is described in [48] for children with childhood apraxia of speech. The game integrates automatic speech recognition. Children and pathologists found the speech-controlled games interesting and fun. This results in increased engagement leading to more intense practice. Apraxia World is another speech therapy game for children with CAS [49]. The game integrates speech exercises into an engaging platform-style game. The exercises must be completed in order to advance to the next level. A video game for speech therapy in children with early diagnosed hearing disability is presented in [50]. The games integrates spoken user interaction by means of automatic speech recognition and benefits from visual feedback. The results show that the use of a serious game in therapy can be entertaining, stimulating enjoyment and engagement, which allows attention and

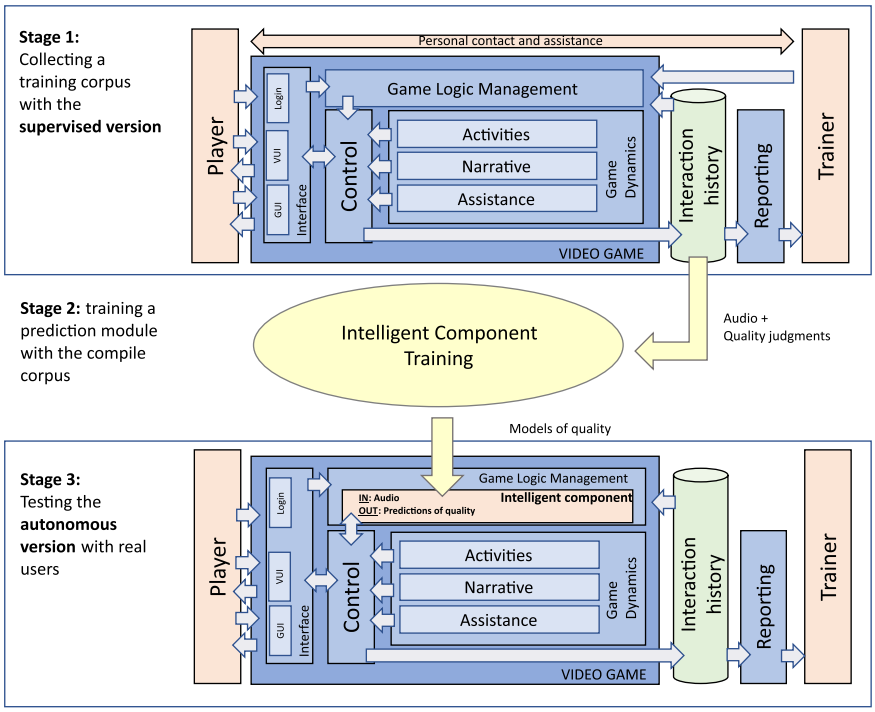

FIGURE 1. **Stages of the system development. Stage 1: Collecting a training corpus with the supervised mode of the video game. Stage 2: Training a module for the automatic prediction of quality of the users' oral productions. Stage 3: Testing the autonomous mode with real users.**

enthusiasm to be maintained in routine therapy tasks. Finally, a speech therapy game for children with speech disorders in a 3D environment is described in [51]. The game is developed to support the traditional therapy sessions.

III. THE LEARNING VIDEO GAME

Fig. 1 shows the software architecture and its evolution for including a component that permits autonomous playing. We focus in this section on presenting the aspects of the tool that are common in both versions and section IV reports details about the automatic assessment module.

The current version of the video game is an evolution of the tool named "La Piedra Mágica" (The Magic Stone) [52]. Fig. 2 shows screenshots of the video game. It is a graphical adventure video game where players have to solve training activities in order to go through the different scenarios and complete the whole story. Both the player and the trainer interact with the tool, as shown in the upper part of Fig. 1. The player uses GUI (Graphic User Interface) resources like mouse clicks and interactive icons to navigate through the stages, and VUI (Voice User Interface) resources as the microphone and the speakers to introduce oral utterances and to receive audio messages. The trainer also interacts with the system as he/she assists the students during playing. In particular, the trainer is responsible for deciding whether the quality of the player's performance in the oral production activities is good enough to allow the user to continue playing or, if not, to require him/her to keep on trying the same activity. The

FIGURE 2. **Screenshots of the learning game: comprehension activities (top left) production activities (top right) and visual activities (bottom).**

control component is responsible for presenting the activities to the player following the scripted narrative of the graphic adventure. It also detects problematic situations and sends assistance messages to the user, for example, when there is a long delay in the interaction (see section III-B).

PRADIA is an evolved version of "La Piedra Mágica" that includes a module for reporting information about player performance [53]. All the information about the user interaction (timing, number of attempts to complete a task, number of mouse clicks, number of helps showed to the user, ...) is registered and the recordings of the production activities are stored. The speech therapist can use this information

to analyze the players' performance along different game sessions. In addition, the audio recordings increase the speech corpus. The audio recordings stored during training sessions were analyzed to identify the differences in some acoustic features related to prosody (fundamental frequency, energy and temporal) between people with Down syndrome and people without intellectual disabilities, as well as the impact of prosody for identifying a voice as typical or atypical [9].

A. TRAINING ACTIVITIES OF ORAL COMPETENCIES

Activities are oriented to fulfilling the final goal of the learning tool, which is the training of oral competencies of players as far as managing pragmatics and prosody are concerned. During the graphic adventure, significant scenarios are presented that simulate current situations from daily life like taking a bus or shopping. While playing, the users of the video game are evaluated as to their verbal competencies concerning their linguistic and prosodic functions. Both prosodic and linguistic functions are a matter of interest in the study of DS language development: it has been shown that DS speakers present deficits in prosodic production affecting focus, chunking and turn-end [54], [55]; as for linguistic functions, it has been reported how DS speakers display areas of substantial pragmatic weakness, such as in rendering descriptions and interacting [56], [57]; finally, it has been highlighted that expressive language skills are more impaired than receptive skills in young individuals with DS when referring to pragmatics [58].

Players are encouraged by the game to train their speech, keeping in mind such prosodic aspects as intonation, expression of emotions or syllabic emphasis. In the activities, the players are introduced by the game into different conversations with game characters, where they have to choose between different options to continue the dialogue (comprehension activities) or to record some sentences related to the dialogue context (production activities), depending on the activity. Finally, there are other activities that were included to add variety to the game and to train other skills not directly related to speech training (visual activities).

B. ENGAGING USERS WITH INTELLECTUAL DISABILITIES

Engagement and immersion are important success factors for educational video games, specially in people with intellectual disabilities. The PRADIA video game was developed following some design guidelines that have proved to be useful in video games focused on people with intellectual disabilities [59].

As for the interface design, the development of the scenarios, items and characters had a uniform design, close to cartoons, but without making them too childish. Bright colors were used in accordance with the scenarios represented in the game. A simple text font is used, with a larger size than usual to make it easier to read. The sound instructions use simple and brief language to improve understanding. The inclusion of visual clues on the game scenarios to help players identify the next goal is a necessity, with the aim of avoiding players getting stuck. At the beginning of the game, players can choose an avatar from among four options. It represents the players' image in the game, allowing players to identify themselves with the character of the story. The video game included an assistant to guide the players during the game and to help them in the training activities. It also reminds players of the current goal. The scenes, characters and items that the players find in the game are representations of real world elements, with some imaginary elements included to motivate the player.

Each activity offers the users feedback according to the results obtained. However, due to the difficulties presented by players with Down syndrome, it is important not to cause frustration that can lead to an abandonment of the game. After a limited number of trials (3), they are allowed to progress regardless of the results, but the feedback is different depending on the results. A positive feedback is shown when they are right and a negative feedback is shown when they go wrong, but the negative feedback is complemented by a positive message that helps them to keep playing and not get demoralized. In order not to stress the users, no scores of any kind are used.

IV. THE MODULE FOR AUTOMATIC ASSESSMENT OF ORAL PRODUCTION QUALITY

Fig. 1 (stage 1) shows how recordings were collected with the supervised version of the video game used to train an automatic classifier (named intelligent component in the figure, stage 3) that is responsible for assessing the quality of the players' oral productions in the autonomous version. Subsection IV-A describes the details on how the corpus is processed in order to be used for training the intelligent component. The automatic assessment module will use the recorded spoken answer of the player to decide whether the user must repeat the spoken answer related to the current activity or continue playing the next step of the graphic adventure. When training people with intellectual disabilities, an accurate prediction of answer quality has to be harmonized with the avoidance of the players' frustration, since too strict judgements could give way to early abandonment. Subsection IV-B describes how we trained a system as accurate as possible, while subsection IV-C details how a bias was introduced in the system in order to provide a more permissive decision after the first repetition.

A. TRAINING CORPUS

The training corpus was gathered in previous game sessions as described in [10]. All the informants played the game with a therapist supervising their activities. The therapists offered real time feedback to the users on how to try to improve their speech, and they decided whether the user could continue the game adventure or had to repeat the last spoken answer for the oral activity. Therapists used a specifically designed software button box to provide their evaluation and judgement in an easy and transparent way, following a specific evaluation rubric. The interaction of DS users was

automatically logged to a file where all input events were stored, including their timing, for further off-line tracking and analysis. Spoken answers for all oral turns were also recorded by the video game so that a collection of utterances were compiled at the end of game sessions, including the type of speaker, the particular targeted activity and the decision of the therapist for each case.

Real-time judgments of the therapists are greatly influenced by the type of speaker and training session conditions. Since the therapist knows the limitations of each speaker in order to generate a correct pronunciation, a personalized evaluation is carried out. Thus, he/she could ask a given user to repeat an exercise even when the quality of his/her utterance is comparatively better than the one accepted for another speaker. Furthermore, if the therapist observes that the speaker is tired or frustrated he could allow the game to continue while he/she gives extended instructions. Due to this user adapted evaluation, we preferred to assess the quality of the utterances in a second offline step, while using the real time judgments to calculate the coefficients of the bias depending on the turn, as described in subsection IV-C.

An expert in prosody was responsible of the off-line annotation of the quality of the utterances. With the aid of a web interface, she listened to each audio file and decided whether the speaker should have repeated the sentence, since the required quality was not achieved, or should have not done so, since a satisfactory utterance was produced. The expert delivered her judgments on a purely auditory perceptual basis, without any acoustic analysis of the sentences, and the focus was on the intonational and prosodic structure. As a consequence, aspects related to intelligibility, quality of the pronunciation or adjustment to the expected sentence were not taken into account. Even in the case of speakers with a low cognitive level and serious problems of intelligibility, the main criterion was whether they had modeled prosody with a certain success, even if the speech was not fully understood by the expert. Just like the therapist's evaluations, the sentences were judged as right or wrong according to the categories of intonational phonology and the learning objectives of the video game PRADIA. These objectives were well known by the evaluator, since she participated in the game design. In total, she labeled 996 sentences uttered by 23 different speakers following the procedure detailed above.

The openSmile toolkit [60] was used to extract acoustic features from each recording. The GeMAPS feature set [61] was selected due to the variety of acoustic and prosodic features contained in this set, which contains frequency related features, energy related features, spectral features and temporal features. The arithmetic mean and the coefficient of variation were calculated on these features. Furthermore, 4 additional temporal features were added: the silence and sounding percentages, silences per second and the mean length silences. A total of 34 prosodic features were used, and these are fully described in a previous work [9].

B. AUTOMATIC ASSESSMENT ALGORITHMS

We chose a support vector machine (SVM) classifier to automate the decision on whether the user must repeat an oral production activity or continue playing the game. For this, we used the offline expert annotations as the training dataset. As we showed in [62], SVM outperforms neural networks and decision trees in this task, reaching an accuracy close to 80% (78.8% recall, 0.291 false positive rate and 0.212 false negative rate).

Linear SVM machines are specially adequate for binary classification. The main idea is to use the training data to find the parameters of a hyperplane that divides the space in two parts, so that the samples in one side of the hyperplane belong to class -1 and those in the other belong to the class $+1$. In our case, a sample in the space $\bar{x} = (\bar{f}, A)$ is a p-dimensional vector containing the acoustic parameters $\bar{f}$ and the activity A, and the output values ($y = 1$ or $y = -1$) are associated, respectively, with the labels *right* (the user can continue the video game adventure) and *wrong* (the user must repeat the oral production). The training corpus is made of a dataset of n points of the form $(\bar{x}_1, y_1), \ldots, (\bar{x}_n, y_n)$ where y_i are either $+1$ or -1. The SVM classifier is trained to decide whether the output is $y = 1$ or $y = -1$, given a new sample x.

The hyperplane is defined by its normal vector $\bar{w}$ and a constant value b [63] that must satisfy the set of inequalities $y_i(\bar{x}_i \cdot \bar{w} + b) - 1 \geq 0 \quad \forall i$. The problem is solved by Lagrangian minimization to obtain a set of n Lagrangian coefficients $\alpha_i \geq 0, \forall i$ so that

$$\bar{w} = \sum_{i=1}^{n} y_i \alpha_i \bar{x}_i, \quad b = \bar{w} \cdot \bar{x}_k \quad \text{for some} \quad \alpha_k > 0. \tag{1}$$

and the formula for the output of a linear SVM is

$$u = \bar{w} \cdot \bar{x} - b, \tag{2}$$

so that the sign of u permits the class y to be assigned to the input $\bar{x}$. Note that once the hyperplane is computed in the training stage, the output can be computed very fast with the linear combination of the input parameters.

We use the implementation of the sequential minimal optimization (SMO) algorithm [64] provided by Weka tools [65] for computing the Lagrangian coefficients α. Feature selection is applied to reduce the dimensionality of $\bar{x}$ from 34 to 21 by discarding the attributes whose information gain $H(y) - H(y|x)$ is lower than 10^{-4}. The values of $\bar{x}$ are normalized. Even though the implementation permits the use of other kernel functions, we obtained the best classification results with the linear version: polynomial kernel $k(\bar{x}_i, \bar{x}_j) = (\bar{x}_i \cdot \bar{x}_j)^d$ with $d = 1$.

C. BIASING THE MODULE

The SVM output defined in equation 2 is converted into a probability by the Weka tools via a logistic function. As has been pointed out, the special characteristics of the DS users mean that false negatives usually lead to frustration and an

TABLE 1. Empirical data used to compute the bias $B(T)$, where r counts the number of times the therapist considers the student can continue playing and w counts the number of times the student must repeat.

	T=1	T=2	T=3	Total
r	325	114	9	448
w	133	19	1	153

P(T|r)	0.73	0.25	0.02
P(T|w)	0.87	0.12	0.01
B(T)	1.2	0.5	0.3

increase in the abandonment rates. Thus, entering a bias to modulate the SVM output probability in terms of the number of attempts is important. So, we introduced the following modified decision rule for continuing the game:

$$P(y = r|\bar{x}) > P(y = w|\bar{x}) \cdot B(T). \tag{3}$$

where y is the output of the classifier and r means right oral production, and w means wrong production; $B(T)$ is the (multiplicative) bias, so that as $B(T) < 1$, we reduce the probability of the classifier prescribing a repetition of the activity for the game player.

We make this bias a function of the turn T (number of times the user has already tried to perform correctly the oral activity in the same game session) and of the specific activity A, which affects the prosodic parameters $\bar{x} = (\bar{f}, A)$, as we already know. So, the probability of the output y (r or w) is given by:

$$P(y|T, \bar{f}, A) = \frac{P(T|y, \bar{f}, A)}{P(T|\bar{f}, A)} \cdot P(y|\bar{f}, A). \tag{4}$$

Assuming that the turn T is both statistically independent from the acoustic parameters $\bar{f}$ and of the activity A, the equation can be rewritten as:

$$P(y|T, \bar{f}, A) = \frac{P(T|y)}{P(T)} \cdot P(y|\bar{f}, A), \tag{5}$$

so that the decision rule to allow the user to continue becomes

$$P(y = r|\bar{f}, A) \cdot \frac{P(T|r)}{P(T)} > P(y = w|\bar{f}, A) \cdot \frac{P(T|w)}{P(T)}. \tag{6}$$

which can be rewritten as

$$P(y = r|\bar{f}, A) > P(y = w|\bar{f}, A) \cdot \frac{P(T|w)}{P(T|r)}. \tag{7}$$

If we compare this equation with 3, we get the expression $B(T) = P(T|w)/P(T|r)$ for the bias, and $P(y = r|\bar{f}, A)$ and $P(y = w|\bar{f}, A)$ are obtained from the SVM classifier trained using $\bar{x} = (\bar{f}, A)$ as input vectors. $P(T|w)$ and $P(T|r)$ are empirically computed from the real-time decision results of the therapists during the training sessions (see Table 1). The data in Table 1 have been computed with the subset of samples of campaign C3 of PRAUTOCAL (see Table 4 of [10]) which were available before the automatic module was released. In our case, $B(1) = 1.2$ for the first turn ($T = 1$), $B(2) = 0.5$ and we empirically make $B(3) = 0.3$. A maximum of three repetitions are allowed, as a design criteria of the video game to avoid discouragement.

TABLE 2. Characteristics of informants: CA is chronological age expressed in years, VA is verbal mental age expressed in years, NVCL is non-verbal cognitive level, MPercT is the mean percentage of success in perception and MProdT in production of the PEPS-C tasks.

Speaker	Gender	CA	VA	NVCL	MPercT	MProdT
MD045	M	20	6.75	27	68.8	60.4
FD046	F	36	9.33	20	64.6	46.3
MD047	M	37	8.25	21	64.6	48.4
FD048	F	25	8.5	23	76	68.2
FD049	F	25	7.08	24	66.7	61.3
FD050	F	13	8.08	18	75	76.3
MD051	M	15	6.83	15	60.4	55.6
MD052	M	20	4.17	16	52.1	37.5
MD053	M	17	5.75	20	56.3	46.9
FD055	F	33	6.42	15	54.2	31.3
FD057	F	21	7.25	19	74	70.3
FD058	F	21	8.33	13	69.8	79.7
MD059	M	42	6.17	16	66.7	45.3
FD060	F	18	5.58	12	64.6	57.8
FD061	F	20	6.83	17	84.4	66.1
MD062	M	19	7.08	13	66.7	82.8
FD063	F	18	4.17	14	50	42.2
MD064	M	18	5.75	13	58.3	51.6
FD065	F	18	3.75	14	55.2	18.8
mean		23	6.64	17	64.7	55.1

V. EVALUATION PROCEDURE

A. INFORMANTS

Informants were recruited from the special education center "Pino de Obregón" and from the Down syndrome Association ASDOVA. A professional carried out a previous characterization of informants. We used the Spanish adaptation of the Peabody Picture Vocabulary Test-Revised (PPVT-R; [66]), which is a measure of receptive vocabulary; the test gives us a measure of their verbal mental age equivalent. This test also gives an estimation of the rank of the participants' Intelligence Quotient. In order to get an assessment of their cognitive level, we used the Spanish version of the Raven's Colored Progressive Matrices (CPM; [67]), which is a non-verbal test of fluid intelligence. In order to have specific measurements of prosody level, the full PEPS-C battery in its Spanish version [68] was also administered to participants. The mean percentage of success in perception and production PEPS-C tasks is presented in Table 2.

Table 2 describes the characteristics of the informants. The table shows the high variability of the informants in our corpus, with $CA \in [13, 42]$ and $VA \in [3.75, 9.33]$ years old. NVCL goes from 12 to 27, while the results of the PEPS-C test report values go from 50 to 84.4 for MPercT and from 18.8 to 82.8 for MProdT.

B. TESTING SESSIONS

Informants performed testing sessions in the following modes:

- **Supervised mode**: the trainer is seated close to the student and helps him/her during the game activities. The trainer decides if the student must repeat the oral activities by using a specific hardware device that makes the process transparent for the users. The trainer can

FIGURE 3. Pictures of the different training sessions. In the upper picture, the trainer is supervising the interaction and decides with a supplementary device whether the student must repeat the activity. The picture in the middle shows a semi-supervised training session and the picture in the bottom shows an autonomous session that is monitored with a surveillance system. In these two last cases, the tool has an intelligent component that decides on the quality of the production activities.

also support the student during the perception and visual activities. The trainer takes advantage of the game session to give corrective feedback that permits the students to improve. The interventions of the trainer are recorded.

- **Semi-supervised mode** (or semi-autonomous mode): in contrast to the supervised mode, a group of students work together in a computer lab. All the students play the same version of the game in a different computer. The trainer in this case supervises the activities of all the students and only assists those students that require it. The decisions about the repetition of the oral turns are taken by the automatic module described in the previous section. The trainer was instructed to interfere as little as possible. One of the members of the research team took the role of technician, assisting in the case of a problem with the software or with the hardware associated (headset, mouse, etc…). Another member of the research group observed the training session that was programmed for one hour. The session was recorded.
- **Autonomous mode**: in this mode, students played the video game isolated in a room with a computer. The session was followed by the members of the research group and by the trainer with a video surveillance system. The students received the instructions to play the game without any help, and they did not know that they were observed. This time only the first chapter of the video game was performed. The individual sessions were also recorded.

The order followed for performing the training sessions was first supervised, next semi-supervised and finally autonomous, with about one weak distance between the different training sessions. Figure 3 presents pictures taken during the three types of training sessions.

C. EVALUATION DIMENSIONS, INSTRUMENTS AND METRICS

The following dimensions are taken into account: effectiveness, efficiency, engagement, easy to learn, and error tolerance. These dimensions are named five Es and are broadly used to interpret results of usability tests [13]. Quesenbery expanded the definition of usability found in the ISO standard (ISO 9241:1998) ''The extent to which a product can be used by specified users to achieve specified goals with effectiveness, efficiency and satisfaction in a specified context of use''. Effectiveness is defined as ''the completeness and accuracy with which users achieve their goals''. The goal for the trainer is that students make oral communication exercises. Tool is effective as long as players train with intensity and complete the tasks without abandonment. Efficiency is defined as ''the speed (with accuracy) with which users can complete their task''. The tool is efficient if players train with reduced resources; in this case, the main resource is the trainer's time. A priori, the more autonomous the interaction is, the more efficient the tool is. Nevertheless, the accuracy with which activities are performed must be taken into account. Easiness to learn is defined as ''how well the product supports both initial orientation and deepening understanding of its capabilities''. In previous works we focused on designing an interface easy to learn for players [8] that is crucial as users with Down syndrome are difficult users due to their short term memory and language deficits. A priori, it is expected the tool will be more difficult to be learnt by the players in absence of the therapist. In the different training sessions with real users we are computing how much this easiness decreases with autonomy. Engaging is ''the degree to which the tone and style of the interface makes the product pleasant or satisfying to use''. We consider the tool engages players as long as

they manifest they want to play again after training and the trainers consider it a useful tool to be used. This dimension is related to effectiveness because if the feedback provided by the tool causes frustration to users with Down syndrome the risk of abandonment is high as it is extensively discussed in [8]. Error tolerance is ''how well the design prevents errors, or helps with recovery from those that do occur''. Once again, the supervised version is a priori more error tolerant as the trainer is there to help when player mistakes or system helps occur. Experimentation is permitting to test how the software and player reacts when error occurs.

The final goal is to contrast the three versions of the video game, taking into account the five dimensions. Next we describe the instruments we use, taking into account the dimension affected and whether it affects the perspective of the trainer or the player.

- **Log files** of the video game that permit the evolution of the users to be monitored during the game sessions. The log files permit to monitor that players perform the training activities (effectiveness). They also permit to compute the time used to finish the training sessions, which we relate with the efficiency of the training, and the number of errors and interruptions as indicators of the easy to learn and error tolerance dimensions. From the log files, we obtain numerical values that permit the different modes to be compared (section VI-A).
- **Observations** of the game sessions, both in situ or using the video recordings, are used for evaluating how easy to learn the different versions of the system are and the reactions of users and system when errors occur (error tolerance). We use the recordings for annotating the interventions of the trainers in order to compare the three modes in terms of efficiency.
- **Questionnaires** based on the SUS standard usability test [69] for evaluating how easy to learn and engaging the different versions of the system are for the therapists.
- **Interviews** of the therapist with the players are used to retrieve information about players' user experience as far as the dimensions of easy to learn and engagement are concerned. The possibility of players filling in the questionnaires directly was discarded due to their intellectual deficits ([70] also discards the use of questionnaires and in [2] the questionnaires are filled up by the fathers). Nevertheless, the therapist uses the questionnaire as a script during the interviews. The direct interaction between players and the research team in interviews was discarded due to the special characteristics of the users.
- **Perceptual AB tests** to compare modes as far as the quality of the audios recorded during the oral productions of the players are concerned. These tests permit us to check that the players' effort while training oral productions (effectiveness) is compared throughout the different game modes (section VI-B).

TABLE 3. Instruments and dimensions for evaluating usability in the different video game modes. P means that the evaluation takes into account the perspective of the player and T the one of the trainer.

Instrument	Effectiveness	Efficiency	Easy to learn	Error tolerance	Engaging
Log files	P	P	P	P	
Observations	PT	PT	PT	PT	PT
Questionnaires			T	T	T
Interviews			PT	PT	PT
AB tests	T				

Table 3 summarizes the relationship of the different instruments with respect to the usability dimensions taken into account for the present study.

VI. RESULTS

A. ACTIVITY LOG FILES ANALYSIS

The workload of the users during game activity is presented in Table 4. The table computes the activities of the three chapters of the video game in the three different modes (only the first chapter is recorded in the autonomous mode). Informants performed a different number of activities because not all the informants completed all the testing sessions. Overall, we tracked 991 production activities, 331 comprehension activities and 643 visual activities. Errors are frequent; the number of errors made by typical users and children have been observed close to zero in previous studies [8]. Production activities are the most difficult ones, with 410 errors versus 113 in comprehension activities and 133 in visual activities. This result is in line with what is expected, as Down syndrome speakers have higher competencies in perception than in production: Table 2 shows that the mean value of MPercT is higher than the mean of MProdT (64.7>55.1); and studies found many individuals with Down syndrome who have a profile of stronger visuo-spatial than verbal processing skills [71]–[73]. The workload for some users is intensive in time: for example user FD046 devoted 2974 seconds to production activities, 727 seconds to discrimination activities and 1085 seconds to visual activities. On the other hand, other users such as FD058 practice very little.

Table 5 compares user performance per mode for activities in the first episode of the video game (8 production activities, 4 comprehension activities and 4 visual activities). 10 speakers completed this chapter in the three modes (only speaker FD065 had problems to complete all the activities in the autonomous mode, she performed 5 of the production activities and none of the visual ones). Mann-Whitney statistical tests have been done to compare some data among groups. Production training activities require more time in supervised mode than in the other modes. The clearest difference appears when supervised productions are compared with autonomous productions (388.2 vs 176.5 s, p-value < 0.05). Differences between supervised and semi-supervised modes

TABLE 4. Activities per user and type of activity. # is the number of activities, *Time* is the total time in seconds devoted, per user, to each of the activities and *#Err* is the number of errors committed by the user during the activity.

	Production				Comprehension			Visual		
User	#	Time	#Err	#audios	#	Time	#Err	#	Time	#Err
MD045	72	1964	26	92	24	457	9	47	911	2
FD046	73	2974	46	97	26	727	11	44	1085	2
MD047	34	1557	11	45	12	335	2	24	707	5
FD048	33	1585	13	45	10	176	1	24	445	2
FD049	29	1166	17	43	7	175	2	19	439	7
FD050	24	614	4	27	7	195	1	19	431	0
MD051	32	1280	11	41	11	189	0	26	432	1
MD052	31	2541	11	42	10	442	1	18	739	5
MD053	26	1174	7	31	6	280	1	19	548	2
FD055	76	2776	22	88	27	864	14	44	1946	18
FD057	72	1992	21	90	24	525	8	48	724	0
FD058	8	287	2	10	4	154	1	6	199	3
MD059	76	2722	40	104	24	670	11	47	1144	7
FD060	82	2640	38	105	26	605	12	51	1196	13
FD061	32	607	4	36	11	161	1	23	357	2
MD062	75	3343	28	92	26	743	5	49	1166	7
FD063	72	3006	45	85	25	784	16	48	1888	25
MD064	75	2455	34	64	27	802	10	49	1444	19
FD065	69	2908	30	0	24	671	7	38	1091	13
Total	991	37591	410	1137	331	8955	113	643	16892	133

TABLE 5. User performance in the different types of activities and modes. # is the number of activities, *Time* is the time in seconds devoted to each of the activities (expressed in mean and 95% confidence interval of the mean), *Errors* is the number of errors committed by the user during the activity, *Help Clicks* is the number of times that the players chose to play the reference sentence and *Audio Helps* is the number of audio helps played by the video game automatically after a period of time of inactivity (expressed as mean [min, max]).

	Supervised	Semi-supervised	Autonomous
Production			
#	80	80	77
Time	388.2 (293 , 483)	311.8 (217 , 406)	176.5 (147 , 205)
Errors	2.8 [0, 7]	3.2 [1, 8]	2.8 [1, 7]
Help Clicks	1.7 [0, 5]	1.9 [0, 6]	0.4 [0, 3]
Audio Helps	0.5 [0, 2]	2.1 [0, 9]	0.1 [0, 1]
Comprehension			
#	40	40	40
Time	133.4 (110 , 156)	143.9 (105 , 181)	106.5 (083 , 129)
Errors	1.2 [0, 2]	2.3 [0, 5]	2.2 [1, 6]
Help Clicks	1.6 [0, 6]	0.7 [0, 5]	0.5 [0, 2]
Audio Helps	0.3 [0, 1]	0.7 [0, 3]	0.0 [0, 0]
Visual			
#	40	40	36
Time	134.8 (110 , 158)	142.4 (106 , 178)	128.6 (045 , 211)
Errors	0.8 [0, 2]	1.7 [0, 6]	1.1 [0, 3]

are less marked, comparable in the case of comprehension activities (133.4 vs 143.9 s, $p > 0.05$). This is because, in general, the therapist guides the student through the execution of the activities, thus reducing the probability of failure (see the next subsection). For the same reason, visual activities take less time in supervised mode, since the teacher helps the student to get a successful result as soon as possible.

The column *Errors* in Table 5 refers to the number of times the player selects the wrong option in comprehension and/or visual activities or that she/he must repeat the oral production because its quality is not good enough. There is a high dependency on the speaker with minimums of 0 or 1 mistake along the whole chapter and a maximum of 7 mistakes. The number of errors grows with the degree of autonomy in the comprehension activities (from 1.2 to 2.3 and 2.2, only statistical differences between supervised and autonomous modes) and in visual activities (from 0.8 to 1.7 and 1.1, no statistical differences between any group pairs). More easily comparable is the number of errors in production mode, where it is similar in automatic, semi-supervised and supervised modes (no statistical differences between any group pairs).

More autonomy does not imply more use of help resources as its use decreases in all the type of activities. The use of help clicks goes down from 1.7 to 0.4 in production activities (only statistical differences between supervised and autonomous) and from 1.6 to 0.5 in comprehension activities (statistical differences between supervised and semi-supervised). Audio helps (the ones that appear when the user waits too much time before performing the activity) are very rare in the autonomous mode (once is the maximum), which is consistent with the shorter training periods indicated by the column *Time*.

B. PERCEIVED QUALITY OF THE ORAL PRODUCTIONS

The perception test was performed with a web interface in which 21 audio pairs are presented. The evaluators can listen to audios as many times as they want and they are asked to select one of the following five options: I prefer audio A, I prefer audio B, no preference as the quality is poor in both cases, no preference as the quality is good in both cases, no answer.

The audio pairs AB are randomly selected from the whole set recorded during the training sessions. In each pair, both A and B correspond to the same production activity and both are uttered by the same player. No specific qualification or age profile was required for the evaluators.

The AB test was performed by 28 evaluators reporting the information summarized in Table 6 and Fig. 4. In 36% of the total 588 answers, evaluators mantain that they do not have a preference between both utterances. When there is no

TABLE 6. Distribution of opinions in the AB test.

A vs B			Preference		Both are ...		No Answer
			A	B	Wrong	Right	
Supervised	vs	Semi-supervised	96(49%)	41(21%)	19(10%)	37(19%)	3(1%)
Supervised	vs	Autonomous	62(32%)	45(23%)	36(18%)	50(26%)	3(1%)
Supervised	vs	Others	158(40%)	86(22%)	55(14%)	87(22%)	6(2%)
Semi-supervised	vs	Supervised	41(21%)	96(49%)	19(10%)	37(19%)	3(2%)
Semi-supervised	vs	Autonomous	26(13%)	89(45%)	17(9%)	54(28%)	10(5%)
Semi-Supervised	vs	Others	67(17%)	185(47%)	36(9%)	91(23%)	13(3%)
Autonomous	vs	Supervised	45(23%)	62(32%)	36(18%)	50(26%)	3(1%)
Autonomous	vs	Semi-supervised	89(45%)	26(13%)	17(9%)	54(28%)	10(5%)
Autonomous	vs	Others	134(34%)	88(22%)	53(14%)	104(27%)	13(3%)

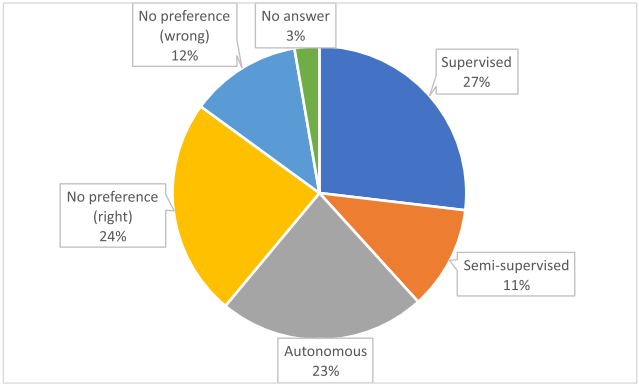

FIGURE 4. Pie chart with the distribution of opinions in the AB test for audio answer quality.

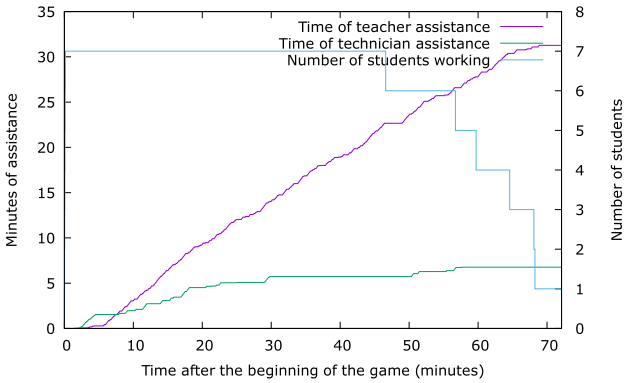

FIGURE 5. Percentage of the training session in which the teacher and the technician are assisting the students in semi-supervised mode.

preference, 66% of cases consider that both utterances are correctly pronounced.

When the evaluators manifest a preference, the supervised mode outperforms those collected in the semi-supervised and autonomous modes (27% vs 11% and 23% respectively). The autonomous version obtains remarkable results when compared with the supervised and the semi-supervised modes: very close to supervised mode (45 vs 62 cases) and better than semi-supervised mode (89 vs 26 cases). The worst results are obtained in the semi-supervised mode, which are far behind those obtained in the supervised (41 vs 96) and autonomous (26 vs 89) modes. This is probably due to the fact that students have to read aloud in a classroom, while other mates and the teacher are around, which could intimidate them.

C. TIMING AND RESOURCES ANALYSIS

As justified in section V-C, we analyze the active assistance time of the therapist to compare the efficiency among the different playing modes. The trainer was present in the different testing sessions and her/his activity was recorded. As described in section V-B, during the supervised sessions, the trainer was seated next to the player and usually advises about the best way to solve the proposed activities, specially when the student fails; however, in the semi-supervised sessions, the teacher is present in the lab while students work in parallel and assists students on demand or when she/he considers it is necessary; in the autonomous testing sessions, the students are remote monitored with a surveillance systems so that she/he only assists when a problem occurs. A technician was also present in the semi-supervised and autonomous sessions. As training sessions are recorded, we can compute the active time in which the therapist assists students in each mode.

In the autonomous version, the therapist had to assist the students twice: two students had difficulties in one of the graphic puzzles. The therapist solved the situation immediately and the student could resume the game without difficulties. The students were FD063 and MD064, both with poor NVCL and VA coefficients (see Table 2).

With respect to the supervised modality, in order to compute the therapist's time of assistance, 3 training sessions have been transcribed and aligned with the audio. These training sessions were randomly selected, one per chapter. 19% percent of the time contains audio recordings of students (both speaking with the machine or with the therapist); 11% corresponds to oral interventions of the therapist (both to assist the oral production or to give advice on using the interface); 70% is silence, corresponding to user interaction with the video game.

Fig. 5 shows the temporal evolution of a semi-supervised work session and Table 7 shows information about the assistance received by the students in that session. The therapist was busy during approximately 45% of the work session attending students, as shown in the red curve in Fig. 5 (31.2 of the overall 72.1 minutes). Even though the therapist had 169 interventions, most lasted less than 5 seconds (36% of the overall assistance time) and they concerned to monitoring and encouraging the users; only 5% of the trainer turns lasted

TABLE 7. Assistance received by the students in semi-supervised mode. # is the number of times the student received assistance, μ is the mean duration of the assistance in seconds, and *min max* are the minimum and maximum duration of the assistance in seconds.

	Therapist				Technician			
User	#	μ	min	max	#	μ	min	max
FO074	7	6.4	1.2	18.9	2	12.0	9.4	14.7
FD060	10	6.2	1.4	28.0	4	3.1	1.7	5.5
FO070	9	6.5	1.4	31.4	8	15.8	3.2	38.2
MD064	15	6.9	1.1	37.6	3	14.1	1.6	21.1
FD065	76	11.4	1.3	48.9	5	11.7	3.9	33.9
FD063	29	13.4	1.1	57.0	4	8.9	2.1	16.1
MD062	23	17.9	1.5	43.3	9	11.9	2.1	37.3
Total	169	11.1	1.1	57.0	35	11.6	1.6	38.2

more than 20 seconds. Her assistance is more intense in the initial period of the game with peaks between minutes 20 and 40. After minute 50 her activity gradually decreases, mainly because some of the students ended the training session and abandoned the room.

The technician assistance events occur mainly in the first 20 minutes and lasted, overall, only 6.8 minutes. They are mainly short (28% lasted less than 5 seconds and only 8% lasted more than 20 seconds) and they mainly concerned with problems with the placement of headphones.

Table 7 presents the intensity results for the assistance received by users, in ascending order and sorted in terms of the time each user spent to finish the game. There is an important diversity, with almost 26.5 minutes difference between the fastest (user FO074) and the slowest student (user MD062). These differences are not only related with the number of interventions, but also with the proficiency of the students. It was annotated that student MD062 was not very collaborative and the teacher had doubts about the capabilities of users FD063 and FD065 to finish the working session.

Thus, in terms of therapist occupation, there is a high contrast between the different modes (blue bar of Fig. 6). The semi-autonomous mode is the most intense mode for the trainer, as she/he is active during 44% of the time, in contrast to 19% in the supervised session or close to 1% in the autonomous one. Nevertheless, taking into account the fact that the semi-supervised session is performed with several students in parallel, the semi-supervised sessions increase throughput with respect to the supervised ones.

D. USER PREFERENCES (QUESTIONNAIRES, INTERVIEWS AND OBSERVATIONS)

In this section we analyze the dimensions related to engagement, easy to learn and error tolerance, as described in section V-C. Sources of information are the SUS questionnaire, interviews and observations. Easy to learn and error tolerance opinions are triangulated with the information collected in the log files. First, we focus on the trainer perspective and then we continue with the perspective of the player. We also report the opinions comparing the three different modes of the video game.

The therapists provided relevant information both by filling in the SUS questionnaire and in interviews. Both of them

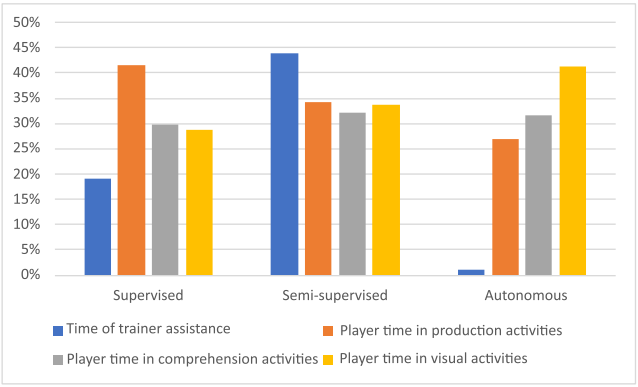

FIGURE 6. Distribution of students activity per mode (in percentages) and percentage of occupation of the trainer in the different modes.

reported comments per question and one of them assigned scores. The final score obtained by the supervised tool is 80 and 77.5 in the semi-supervised version (Grade B in the SUS scale). Concerning the dimension related to how easy to use the software is, they declared they were confident with the use and considered that there was no need to learn a lot of things before they could get going with this system. One of the therapists wrote in the form: *after using it a few times the game dynamics is understood. Others can use it depending on their interest, not on the complexity of the use*. She also declared that the activities were well integrated with a common thread as far as the question about inconsistencies in the interface were concerned.

As for the capabilities of the video game to tolerate errors, no relevant problems were reported. One of the therapists declared: *If the software is blocked I restart and resume where it was interrupted*. The other assigned a 5 to this topic.

Concerning the question about how frequently she would use the software, she declared that she was happy with the tool but it is not thought for a frequent use because it is not a resource covering all the oral competencies but a subset of them: the related with prosody.

Concerning to the preference for the different modes of the game, one of the therapists stated that it could be complicated to respond to a group of users. And she reported: *More difficult to use in group situations. Teachers of speech therapy are allowed to have individual cases.* Nevertheless, she was surprised with the performance of the students during the semi-supervised and autonomous modalities: *I can not believe how well they have behaved during the working sessions*; or *Definitively, we help them too much, they are able to do more than we think by themselves*. The other assistant stated that *it is possible for them to use it autonomously*.

With respect to the students' opinions, Tables 8 and 9 present the collected answers of the usability test. All the users considered it easy to use and were comfortable while using the game (Q9 of Table 9). Negative comments concerning the difficulty of the game (Q2 and Q3 of Table 8) seem to be related to the difficulty of the activity, not with usability concerns. In that sense, all the users said they faced difficult

TABLE 8. Answers to the SUS usability questionnaire given by users (questions Q1-Q5). SUS questions in rows.

#Q	**Question** and { User:Answer }
Q1	**I think that I would like to use this system frequently**
	MD064: I would like to play FD060: I would like frequently FD065: Yes, many times FO074: Yes FD063: I felt good, yes MD062: Ok, I would like to FO070: Yes, it was interesting
Q2	**I found the system unnecessarily complex**
	MD064: Sometimes a little bit complicated FD060: It is difficult, better with help FD065: Yes, there are difficult tasks FO074: It is easier with the teacher and sometimes alone FD063: Yes it is difficult, it is more difficult alone MD062: I do not understand everything and sometimes it is complicated FO070: Recording with the microphone overwhelms us
Q3	**I found the system was easy to use**
	MD064: There are fragments that are difficult. FD060: It is not, when I repeat it is easier. The microphone is the worst. FD065: I have to train. FO074: Sometimes need for help. FD063: It is not simple, I need help. MD062: Generally it is difficult, if I do not understand something and it helps me to talk. FO070: At the beginning it was more difficult, now it is easier.
Q4	**I think that I would need technical support to be able to use this system.**
	MD064: I need help FD060: I do not need help. I can give more, but sometimes I need help. FD065: Yes, I need the teacher. FO074: Yes, I understood the game. FD063: Yes I need support. MD062: Sometimes I need help. FO070: Yes sometimes I need the teacher
Q5	**I found the various functions in this system were well integrated.**
	MD064: It is ok. FD060: Activities are fine. FD065: Everything was ok. FO074: –No Answer– FD063: I do not understand everything. MD062: The game is easy to understand and it helps me to speak better. FO070: Everything is ok.

situations during the game and that the help of the therapist was important in those situations (Q4 of Table 8).

The feedback was positive as far as how frequently users would like to use the tool was concerned. They explicitly declared this fact in the questionnaire (Q1 of Table 8) and the teacher reported that students insistently asked her to play the video game frequently along the course. We observed a clear attitude of engagement during the game sessions.

All the users declared that the game was easier when the therapist assisted them. At the same time, as far as the preferred mode is concerned, there was a division of opinions: 5 out of 7 preferred using the game alone. Problems with the use of the system had to do with the wrong connection of the headphones and with difficulties in solving the visual activities. Only two students needed assistance during the autonomous sessions.

VII. DISCUSSION

With respect to the first research question that was proposed in the introduction of the paper asking whether the automatic evaluation component of the oral production could serve to drive the activities in the particular context of a game to train the communicative competencies of speakers with Down syndrome, the results show that the module permits students to perform the training exercises while playing. In the semi-supervised mode, all the players could finish the graphic adventure in a reasonable time (62 minutes mean with minimum of 47 minutes and maximum of 72 minutes), while the therapist had a reduced number of monitoring and assistance interventions (Table 7 reports 169 total, 7 users and 72 minutes session length which corresponds to about 1.7 interventions per user every 5 minutes). Workload on direct personal interventions seems to be not so demanding, so the trainer is able to monitor the progress of all the participants. In the autonomous mode, assistance was only required twice and it was due to a difficulty found in a visual activity (not in a production activity). As reported in section VI-D, engagement is not reduced when the automatic module is introduced, collecting opinions from players and trainers who say that the users want to play the graphic adventure again. At the same time, the perception test reported in section VI-B allows us to say that the quality of the audios collected in the non-supervised sessions is comparable to the quality of the audios collected in the supervised sessions. Thus, the module

TABLE 9. Answers to the SUS usability questionnaire given by users (questions Q6-Q10). SUS questions in rows.

Q6	**I thought there was too much inconsistency in this system.**
	MD064: Sometimes I get confused in the games. FD060: –No Answer– FD065: No. FO074: No. FD063: –No Answer– MD062: –No Answer– FO070: No.
Q7	**I think that most people would learn to use this system very quickly.**
	MD064: Other people can learn, I would help. FD060: Yes, if they are Down like me. FD065: I'd like other colleagues to play. FO074: I'd love to show it to other colleagues. FD063: Other colleagues could learn. MD062: Yes, I am a good person and I would help them. FO070: I would like others to play and I would help them.
Q8	**I found the system very cumbersome to use.**
	MD064: Sometimes it is confusing. FD060: There are difficult things. With help it is easier. FD065: Cumbersome. FO074: Not cumbersome. FD063: Some parts are difficult. MD062: I am learning step by step. FO070: Yes, sometimes
Q9	**I felt very confident using the system.**
	MD064: Comfortable. FD060: Comfortable. FD065: Comfortable. FO074: Comfortable. FD063: Yes, and I loved to do it. MD062: I felt comfortable playing. FO070: Yes, I would like to continue.
Q10	**I needed to learn a lot of things before I could get going with this system.**
	MD064: Sometimes. FD060: I need train my voice a little bit more. FD065: Yes. FO074: I did it with the group and it was easy. FD063: Many aspects are worked in the classroom. MD062: I need to listen and improve. FO070: Yes.
	Which modality do you prefer: assisted or autonomous?
	MD064: Alone. FD060: With teacher. FD065: Alone. FO074: Alone. FD063: With teacher. MD062: Alone. FO070: Alone.

for autonomous gaming has allowed students to practice the proposed exercises for training oral production while also performing the dynamics of the game with a high degree of engagement.

With respect to the second research question, in which we ask about the effects on the use of the learning tools caused by the introduction of the automatic component, changes have been observed both in the duration (rows labeled as Time in Table 5) and the intensity of the training sessions as described in section VI-A. Training sessions are shorter because the feedback of the teacher is eliminated (Figure 6). The use of assistance resources is reduced and the number of failures in visual and comprehension activities is higher (rows labeled as Errors in Table 5). We mainly focused on the automatic module for evaluating oral production activities and it is a positive result that the number of errors in production activities and the quality of audios is comparable in the three modes (Table5 and Figure 4). Nevertheless, the feedback reduction is an important concern as it has a prominent role on learning. Searching for ways to improve how the feedback is provided to players during autonomous game play is a concern for future works. Our goal is to use the information captured from the student's interaction like oral production mistakes, use of help resources… to provide personalized tips and messages that lead to a more intense training of participants in the autonomous mode.

From the training perspective, efficiency is the dimension in which the advantage of the inclusion of the automatic module is clearer. The automatic module means that there is no longer any need for a specialist to be seated with the students when they practice. The semi-supervised version seems to be the best option to maintain a compromise between training intensity and economizing time and resources. Establishing the optimal number of students practicing at the same time

with the presence of the teacher is beyond the scope of this paper. This variable is very dependent on the level of the students and should be set by the therapist or teacher with respect to the configuration of his/her laboratory and the specific profile of his/her students.

Usability and playability concerns (like dimensions, facets, attributes and properties...) are used for evaluating learning games not because there exists universal consensus about their definitions but because they are useful to systematize evaluation processes and to organize discussion about the results. In our case, the Quesenbery framework has permitted us to organize the instruments used for evaluating the impact of the introduction of the automatic module. Other frameworks exist that are specific for video games that could have been more suitable because they focus on player experience. One of the most used and referenced is [74] that has been used for proposing guidelines for educational video games [75]. Considering the attributes proposed in this framework including satisfaction, learnability, effectiveness, inmersion, motivation, emotion, and socialization, and its respective properties, would have enriched the discussion and presentation of results of our work. Nevertheless, the use of a more general framework (not specific for video games) has permitted us to compare the different versions of the video game taking into account both the perspective of the player and the perspective of the trainer. Observations, questionnaires and interviews have evidenced that the different aspects related with the player satisfaction, learnability, inmersion, motivation and emotion are not significatively affected, at the time that the number of resources (time and trainer assistance) is reduced. The counterpart is that feedback is reduced with a possible negative impact on learning.

LIMITATIONS

The order in which the users perform the 3 sessions is the same for all users: supervised, semi-supervised and autonomous. We opted for this order instead of a randomized order because it is the real training path the students follow in a typical course, going from easier to more difficult tasks, starting with a supervised presentation of the game and ending with autonomous playing. Repeating the experiment with a different group of players who perform training sessions in randomized order, would make results independent of the previous experience the users have of the game avoiding a possible bias of current results due to the users having experience of the game with supervision prior to the autonomous use.

The experimentation was planned to inquire about how the introduction of the automatic module affects usability and enables autonomous training. Other important concerns related with learning and proficiency improvement of players after training have not been evaluated. Here we have shown that players with Down syndrome can train oral productions autonomously with the tool, but evaluating how much the students learn with the video game requires longer term experimentation and is out of the scope of the present paper.

VIII. CONCLUSION

In this article, we have shown that the inclusion of a module for the automatic evaluation of oral production quality in a learning digital game permits players with Down syndrome to train autonomously at levels of efficiency comparable to those observed in the supervised version. The user interaction hints introduced into the tool permit users to play without the help of the trainers in most cases. Although some players had problems to finish the training session due to the difficulty of the visual activities, the degree of engagement did not decrease during relatively long training sessions.

Although the degree of attention during training could decrease when there is no supervision (an increase in the number of errors in comprehension activities and in visual games), the quality of the audios produced in the autonomous mode is comparable to those of the recorded audios during the training sessions supervised by the therapist.

The automatic version does not guarantee permanent control of users, but the high degree of engagement has been shown to be enough to motivate players to keep on training. The attractiveness of the game does not decrease with time, which is an opportunity to introduce specific feedback for students to improve while playing. Work for the future entails both a redesign of explicit feedback mechanisms in case of error and enriching the analysis performed by the automatic component in order to provide practical advice that allows speakers to improve with respect to their particular limitations.

ACKNOWLEDGMENT

The authors would like to thank the students and therapists of "El Pino de Obregón" and "ASDOVA" special education centers for their valuable participation and collaboration in the experimental activities carried their for this work.

REFERENCES

[1] E. S. Tanis, S. Palmer, M. Wehmeyer, D. K. Davies, S. E. Stock, K. Lobb, and B. Bishop, "Self-report computer-based survey of technology use by people with intellectual and developmental disabilities," *Intellectual Develop. Disabilities*, vol. 50, no. 1, pp. 53–68, Feb. 2012.

[2] J. Feng, J. Lazar, L. Kumin, and A. Ozok, "Computer usage by children with Down syndrome: Challenges and future research," *ACM Trans. Accessible Comput.*, vol. 2, no. 3, pp. 1–44, Mar. 2010.

[3] S. Caton and M. Chapman, "The use of social media and people with intellectual disability: A systematic review and thematic analysis," *J. Intellectual Develop. Disab.*, vol. 41, no. 2, pp. 125–139, Apr. 2016.

[4] A. R. Cano, Á. J. García-Tejedor, and B. Fernández-Manjón, "A literature review of serious games for intellectual disabilities," in *Design for Teaching and Learning in a Networked World*. Cham, Switzerland: Springer, 2015, pp. 560–563.

[5] S. Tsikinas and S. Xinogalos, "Studying the effects of computer serious games on people with intellectual disabilities or autism spectrum disorder: A systematic literature review," *J. Comput. Assist. Learn.*, vol. 35, no. 1, pp. 61–73, Feb. 2019.

[6] T. Martins, V. Carvalho, F. Soares, and M. F. Moreira, "Serious game as a tool to intellectual disabilities therapy: Total challenge," in *Proc. IEEE 1st Int. Conf. Serious Games Appl. Health (SeGAH)*, Nov. 2011, pp. 1–7.

[7] R. D. Kent and H. K. Vorperian, "Speech impairment in Down syndrome: A review," *J. Speech, Lang. Hearing Res.*, vol. 56, no. 1, p. 178, 2013.

[8] C. González-Ferreras, D. Escudero-Mancebo, M. Corrales-Astorgano, L. Aguilar-Cuevas, and V. Flores-Lucas, "Engaging adolescents with

Down syndrome in an educational video game," *Int. J. Hum.–Comput. Interact.*, vol. 33, no. 9, pp. 693–712, Sep. 2017.

[9] M. Corrales-Astorgano, D. Escudero-Mancebo, and C. González-Ferreras, "Acoustic characterization and perceptual analysis of the relative importance of prosody in speech of people with Down syndrome," *Speech Commun.*, vol. 99, pp. 90–100, May 2018.

[10] D. Escudero-Mancebo *et al.*, "PRAUTOCAL corpus: A corpus for the study of Down syndrome prosodic aspects," *Lang Resour. Eval.*, 2021, doi: 10.1007/s10579-021-09542-8.

[11] R. Yáñez-Gómez, D. Cascado-Caballero, and J.-L. Sevillano, "Academic methods for usability evaluation of serious games: A systematic review," *Multimedia Tools Appl.*, vol. 76, no. 4, pp. 5755–5784, Feb. 2017.

[12] J. Paavilainen, "Defining playability of games: Functionality, usability, and gameplay," in *Proc. 23rd Int. Conf. Acad. Mindtrek*, Jan. 2020, pp. 55–64.

[13] W. Quesenbery, "Dimensions of usability," in *Content and Complexity: Information Design in Technical Communication*, M. J. Albers and M. B. Mazur, Eds. Evanston, IL, USA: Routledge, 2003.

[14] S. McLeod, G. Daniel, and J. Barr, "When he's around his brothers. . . he's not so quiet: The private and public worlds of school-aged children with speech sound disorder," *J. Commun. Disorders*, vol. 46, no. 1, pp. 70–83, Jan. 2013, doi: 10.1016/j.jcomdis.2012.08.006.

[15] J. McCormack, S. McLeod, L. McAllister, and L. J. Harrison, "A systematic review of the association between childhood speech impairment and participation across the lifespan," *Int. J. Speech-Lang. Pathol.*, vol. 11, no. 2, pp. 155–170, Jan. 2009, doi: 10.1080/17549500802676859.

[16] D. R. Boone, S. C. McFarlane, S. L. V. Berg, and R. I. Zraick, *The Voice and Voice Therapy*, 10th ed. London, U.K.: Pearson, 2019.

[17] D. Le, K. Licata, C. Persad, and E. M. Provost, "Automatic assessment of speech intelligibility for individuals with aphasia," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 24, no. 11, pp. 2187–2199, Nov. 2016.

[18] Y. Qin, T. Lee, S. Feng, and A. P. H. Kong, "Automatic speech assessment for people with aphasia using TDNN-BLSTM with multi-task learning," in *Proc. INTERSPEECH*, 2018, pp. 3418–3422.

[19] Y. Qin, T. Lee, and A. P. H. Kong, "Automatic assessment of speech impairment in cantonese-speaking people with aphasia," *IEEE J. Sel. Topics Signal Process.*, vol. 14, no. 2, pp. 331–345, Feb. 2020.

[20] A. Maier, T. Haderlein, U. Eysholdt, F. Rosanowski, A. Batliner, M. Schuster, and E. Nöth, "PEAKS—A system for the automatic evaluation of voice and speech disorders," *Speech Commun.*, vol. 51, no. 5, pp. 425–437, May 2009.

[21] J. Kim, N. Kumar, A. Tsiartas, M. Li, and S. S. Narayanan, "Automatic intelligibility classification of sentence-level pathological speech," *Comput. Speech Lang.*, vol. 29, no. 1, pp. 132–144, Jan. 2015.

[22] G. Dimauro, V. Di Nicola, V. Bevilacqua, D. Caivano, and F. Girardi, "Assessment of speech intelligibility in Parkinson's disease using a speech-to-text system," *IEEE Access*, vol. 5, pp. 22199–22208, 2017.

[23] I. Laaridh, W. B. Kheder, C. Fredouille, and C. Meunier, "Automatic prediction of speech evaluation metrics for dysarthric speech," in *Proc. INTERSPEECH*, Aug. 2017, pp. 1834–1838.

[24] D. Martínez, E. Lleida, P. Green, H. Christensen, A. Ortega, and A. Miguel, "Intelligibility assessment and speech recognizer word accuracy rate prediction for dysarthric speakers in a factor analysis subspace," *ACM Trans. Accessible Comput.*, vol. 6, no. 3, p. 10, 2015.

[25] P. Janbakhshi, I. Kodrasi, and H. Bourlard, "Automatic pathological speech intelligibility assessment exploiting subspace-based analyses," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 28, pp. 1717–1728, 2020.

[26] N. P. Narendra and P. Alku, "Automatic assessment of intelligibility in speakers with dysarthria from coded telephone speech using glottal features," *Comput. Speech Lang.*, vol. 65, Jan. 2021, Art. no. 101117.

[27] H. M. Chandrashekar, V. Karjigi, and N. Sreedevi, "Spectro-temporal representation of speech for intelligibility assessment of dysarthria," *IEEE J. Sel. Topics Signal Process.*, vol. 14, no. 2, pp. 390–399, Feb. 2020.

[28] M. Tu, V. Berisha, and J. Liss, "Interpretable objective assessment of dysarthric speech based on deep neural networks," in *Proc. INTERSPEECH*, 2017, pp. 1849–1853.

[29] B. A. Al-Qatab and M. B. Mustafa, "Classification of dysarthric speech according to the severity of impairment: An analysis of acoustic features," *IEEE Access*, vol. 9, pp. 18183–18194, 2021.

[30] M. Li, D. Tang, J. Zeng, T. Zhou, H. Zhu, B. Chen, and X. Zou, "An automated assessment framework for atypical prosody and stereotyped idiosyncratic phrases related to autism spectrum disorder," *Comput. Speech Lang.*, vol. 56, pp. 80–94, Jul. 2019.

[31] O. Saz, S.-C. Yin, E. Lleida, R. Rose, C. Vaquero, and W. R. Rodríguez, "Tools and technologies for computer-aided speech and language therapy," *Speech Commun.*, vol. 51, no. 10, pp. 948–967, Oct. 2009.

[32] W. R. Rodríguez, O. Saz, and E. Lleida, "A prelingual tool for the education of altered voices," *Speech Commun.*, vol. 54, no. 5, pp. 583–600, Jun. 2012.

[33] A. Abad, A. Pompili, A. Costa, I. Trancoso, J. Fonseca, G. Leal, L. Farrajota, and I. P. Martins, "Automatic word naming recognition for an on-line aphasia treatment system," *Comput. Speech Lang.*, vol. 27, no. 6, pp. 1235–1248, Sep. 2013.

[34] T.-S. Tan, Helbin-Liboh, A. K. Ariff, C.-M. Ting, and S.-H. Salleh, "Application of malay speech technology in malay speech therapy assistance tools," in *Proc. Int. Conf. Intell. Adv. Syst.*, Nov. 2007, pp. 330–334.

[35] M. Shahin, B. Ahmed, A. Parnandi, V. Karappa, J. McKechnie, K. J. Ballard, and R. Gutierrez-Osuna, "Tabby talks: An automated tool for the assessment of childhood apraxia of speech," *Speech Commun.*, vol. 70, pp. 49–64, Jun. 2015.

[36] R. Patel, H. Kember, and S. Natale, "Feasibility of augmenting text with visual prosodic cues to enhance oral reading," *Speech Commun.*, vol. 65, pp. 109–118, Nov. 2014.

[37] D. Sztahó, G. Kiss, and K. Vicsi, "Computer based speech prosody teaching system," *Comput. Speech Lang.*, vol. 50, pp. 126–140, Jul. 2018.

[38] J. S. Brumberg, J. C. Thorson, and R. Patel, "The prosodic marionette: A method to visualize speech prosody and assess perceptual and expressive prosodic abilities," *Speech Commun.*, vol. 104, pp. 95–105, Nov. 2018.

[39] S. Fagel and K. Madany, "A 3-D virtual head as a tool for speech therapy for children," in *Proc. 9th Annu. Conf. Int. Speech Commun. Assoc.*, 2008, pp. 1–4.

[40] L. Czap, "Automated speech production assessment of hard of hearing children," *IEEE J. Sel. Topics Signal Process.*, vol. 14, no. 2, pp. 380–389, Feb. 2020.

[41] S. Van Vuuren and L. R. Cherney, "A virtual therapist for speech and language therapy," in *Proc. Int. Conf. Intell. Virtual Agents*. Cham, Switzerland: Springer, 2014, pp. 438–448.

[42] A. Parnandi, V. Karappa, T. Lan, M. Shahin, J. McKechnie, K. Ballard, B. Ahmed, and R. Gutierrez-Osuna, "Development of a remote therapy tool for childhood apraxia of speech," *ACM Trans. Accessible Comput.*, vol. 7, no. 3, pp. 1–23, Nov. 2015.

[43] K. M. Kapp, *The Gamification of Learning and Instruction: Game-Based Methods and Strategies for Training and Education*. Aßlar, Germany: Pfeiffer, 2012.

[44] S. Deterding, M. Sicart, L. Nacke, K. O'Hara, and D. Dixon, "Gamification. Using game-design elements in non-gaming contexts," in *Proc. Annu. Conf. Extended Abstr. Hum. Factors Comput. Syst. (CHI EA)*, 2011, pp. 2425–2428.

[45] M. Prensky, *Digital Game-Based Learning*. Wiltshire, U.K.: Paragon House Publishers, 2007.

[46] E. Boyle, T. M. Connolly, and T. Hainey, "The role of psychology in understanding the impact of computer games," *Entertainment Comput.*, vol. 2, no. 2, pp. 69–74, Jan. 2011.

[47] M. Ganzeboom, M. Bakker, L. Beijer, T. Rietveld, and H. Strik, "Speech training for neurological patients using a serious game," *Brit. J. Educ. Technol.*, vol. 49, no. 4, pp. 761–774, Jul. 2018.

[48] B. Ahmed, P. Monroe, A. Hair, C. T. Tan, R. Gutierrez-Osuna, and K. J. Ballard, "Speech-driven mobile games for speech therapy: User experiences and feasibility," *Int. J. Speech-Lang. Pathol.*, vol. 20, no. 6, pp. 644–658, Oct. 2018.

[49] A. Hair, P. Monroe, B. Ahmed, K. J. Ballard, and R. Gutierrez-Osuna, "Apraxia world: A speech therapy game for children with speech sound disorders," in *Proc. 17th ACM Conf. Interact. Design Children*, Jun. 2018, pp. 119–131.

[50] A. A. Navarro-Newball, D. Loaiza, C. Oviedo, A. Castillo, A. Portilla, D. Linares, and G. Álvarez, "Talking to teo: Video game supported speech therapy," *Entertainment Comput.*, vol. 5, no. 4, pp. 401–412, Dec. 2014.

[51] M. Cagatay, P. Ege, G. Tokdemir, and N. E. Cagiltay, "A serious game for speech disorder children therapy," in *Proc. 7th Int. Symp. Health Informat. Bioinf.*, Apr. 2012, pp. 18–23.

[52] M. Corrales-Astorgano, D. Escudero-Mancebo, C. González-Ferreras, Y. Gutiérrez-González, V. Flores-Lucas, V. Cardeñoso-Payo, and L. Aguilar-Cuevas, "The magic stone: A video game to improve communication skills of people with intellectual disabilities," in *Proc. INTERSPEECH*, 2016, pp. 1565–1566.

[53] L. Aguilar, "Learning prosody in a video game-based learning approach," *Multimodal Technol. Interact.*, vol. 3, no. 3, p. 51, Jul. 2019.

[54] V. Stojanovik, "Prosodic deficits in children with Down syndrome," *J. Neurolinguistics*, vol. 24, no. 2, pp. 145–155, Mar. 2011.
[55] S. J. Loveall, K. Hawthorne, and M. Gaines, "A meta-analysis of prosody in autism, Williams syndrome, and Down syndrome," *J. Commun. Disorders*, vol. 89, Jan. 2021, Art. no. 106055.
[56] L. Abbeduto, S. F. Warren, and F. A. Conners, "Language development in Down syndrome: From the prelinguistic period to the acquisition of literacy," *Mental Retardation Develop. Disabilities Res. Rev.*, vol. 13, no. 3, pp. 247–261, 2007.
[57] L. Abbeduto, "Pragmatic development," *Down Syndrome Res. Pract.*, vol. 13, pp. 57–59, 2008.
[58] G. E. Martin, J. Klusek, B. Estigarribia, and J. E. Roberts, "Language characteristics of individuals with Down syndrome," *Topics Lang. Disorders*, vol. 29, no. 2, p. 112, 2009.
[59] S. Tsikinas and S. Xinogalos, "Designing effective serious games for people with intellectual disabilities," in *Proc. IEEE Global Eng. Educ. Conf. (EDUCON)*, Apr. 2018, pp. 1896–1903.
[60] F. Eyben, F. Weninger, F. Gross, and B. Schuller, "Recent developments in openSMILE, the munich open-source multimedia feature extractor," in *Proc. 21st ACM Int. Conf. Multimedia (MM)*, 2013, pp. 835–838.
[61] F. Eyben, K. R. Scherer, B. W. Schuller, J. Sundberg, E. Andre, C. Busso, L. Y. Devillers, J. Epps, P. Laukka, S. S. Narayanan, and K. P. Truong, "The geneva minimalistic acoustic parameter set (GeMAPS) for voice research and affective computing," *IEEE Trans. Affect. Comput.*, vol. 7, no. 2, pp. 190–202, Apr. 2016.
[62] M. Corrales-Astorgano, P. Martínez-Castilla, D. Escudero-Mancebo, L. Aguilar, C. González-Ferreras, and V. Cardeñoso-Payo, "Automatic assessment of prosodic quality in Down syndrome: Analysis of the impact of speaker heterogeneity," *Appl. Sci.*, vol. 9, no. 7, p. 1440, Apr. 2019.
[63] C. J. C. Burges, "A tutorial on support vector machines for pattern recognition," *Data Mining Knowl. Discovery*, vol. 2, no. 2, pp. 121–167, 1998.
[64] J. Platt, "Sequential minimal optimization: A fast algorithm for training support vector machines," Microsoft, Redmond, WA, USA, Tech. Rep. MSR-TR-98-14, 1998.
[65] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten, "The WEKA data mining software: An update," *ACM SIGKDD Explor. Newslett.*, vol. 11, no. 1, pp. 10–18, 2009.
[66] L. Dunn, L. Dunn, and D. Arribas, *Test de Vocabulario en Imágenes Peabody*. Madrid, Spain: TEA, 2006.
[67] J. Raven, J. C. Raven, and J. Court, *Test de Matrices Progresivas: Manual/Manual for Raven's Progessive Matrices and Vocabulary Scalestest de Matrices Progresivas*. Barcelona, Spain: Paidós, 1993, no. 159.9. 072.
[68] P. Martínez-Castilla and S. Peppé, "Developing a test of prosodic ability for speakers of iberian spanish," *Speech Commun.*, vol. 50, nos. 11–12, pp. 900–915, Nov. 2008.
[69] J. Brooke, "SUS: A retrospective," *J. Usability Stud.*, vol. 8, no. 2, pp. 29–40, 2013.
[70] A. R. Cano, B. Fernández-Manjón, and Á. J. García-Tejedor, "Using game learning analytics for validating the design of a learning game for adults with intellectual disabilities," *Brit. J. Educ. Technol.*, vol. 49, no. 4, pp. 659–672, Jul. 2018.
[71] D. Fidler, S. Hepburn, and S. Rogers, "Early learning and adaptive behaviour in toddlers with Down syndrome: Evidence for an emerging behavioural phenotype?" *Down Syndrome Res. Pract.*, vol. 9, no. 3, pp. 37–44, 2006.
[72] C. Jarrold, A. Baddeley, and A. Hewes, "Genetically dissociated components of working memory: Evidence from Downs and Williams syndrome," *Neuropsychologia*, vol. 37, no. 6, pp. 637–651, Jun. 1999.
[73] P. P. Wang and U. Bellugi, "Evidence from two genetic syndromes for a dissociation between verbal and visual-spatial short-term memory," *J. Clin. Exp. Neuropsychol.*, vol. 16, no. 2, pp. 317–322, Apr. 1994.
[74] J. L. G. Sánchez, F. L. G. Vela, F. M. Simarro, and N. Padilla-Zea, "Playability: Analysing user experience in video games," *Behav. Inf. Technol.*, vol. 31, no. 10, pp. 1033–1054, Oct. 2012.
[75] A. Ibrahim, F. L. G. Vela, P. P. Rodríguez, J. L. G. Sánchez, and N. P. Zea, "Playability guidelines for educational video games: A comprehensive and integrated literature review," *Int. J. Game-Based Learn.*, vol. 2, no. 4, pp. 18–40, Oct. 2012.

DAVID ESCUDERO-MANCEBO received the B.A. and M.Sc. degrees in computer science, and the Ph.D. degree in information technologies from the University of Valladolid, Valladolid, Spain, in 1993, 1996, and 2002, respectively. He is currently a Regular Member of the ECA-SIMM Research Group and an Associate Professor with the Department of Computer Science, University of Valladolid. He is the coauthor of several publications in the field of speech processing, with special focus on prosody and on developing tools and learning games for training pronunciation.

MARIO CORRALES-ASTORGANO received the B.A. and M.Sc. degrees in computer science, and the Ph.D. degree in information technologies from the University of Valladolid, Valladolid, Spain, in 2013, 2015, and 2019, respectively. He is currently a Regular Member of the ECA-SIMM Research Group. His research interests include human–computer interaction, learning games for people with intellectual disabilities, and speech processing.

VALENTÍN CARDEÑOSO-PAYO (Member, IEEE) received the M.Sc. and Ph.D. degrees in physics from the University of Valladolid, Valladolid, Spain, in 1984 and 1988, respectively. Since 1998, he has been the Director of the ECA-SIMM Group, University of Valladolid. He has been an advisor of ten Ph.D. works in speech synthesis and recognition, online signature verification, and structured parallelism for high-performance computing. His research interests include machine learning techniques applied to human language technologies, human–computer interaction, and biometric person recognition.

CÉSAR GONZÁLEZ-FERRERAS received the B.Sc., M.Sc., and Ph.D. degrees in computer science from the University of Valladolid, Valladolid, Spain, in 1998, 2000, and 2009, respectively. He is currently a Regular Member of the ECA-SIMM Research Group and an Associate Professor with the Department of Computer Science, University of Valladolid. His research interests include human–computer interaction, spoken language processing, and prosody recognition.