IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received March 11, 2020, accepted April 2, 2020, date of publication April 16, 2020, date of current version May 1, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.2988406

Using Challenges to Enhance a Learning Game for Pronunciation Training of English as a Second Language

CRISTIAN TEJEDOR-GARCÍA, DAVID ESCUDERO-MANCEBO, VALENTÍN CARDEÑOSO-PAYO, (Member, IEEE), AND CÉSAR GONZÁLEZ-FERRERAS

Departamento de Informática, Universidad de Valladolid (UVa), 47011 Valladolid, Spain

Corresponding author: Cristian Tejedor-García (cristian@infor.uva.es)

This work was supported in part by the Ministerio de Economía y Empresa (MINECO), in part by the the European Regional Development Fund FEDER under Grant TIN2014-59852-R, in part by the Consejería de Educación de la Junta de Castilla y León under Grant VA050G18, and in part by the University of Valladolid (Ph.D. Research Grant 2015 and MOVILIDAD DOCTORANDOS UVa 2019).

ABSTRACT Learning games have a remarkable potential for education. They provide an emergent form of social participation that deserves the assessment of their usefulness and efficiency in learning processes. This study describes a novel learning game for foreign pronunciation training in which players can challenge each other. Native Spanish speakers performed several pronunciation activities during a one-month competition using a mobile application, designed under a minimal pairs approach, to improve their pronunciation of English as a foreign language. This game took place in a competitive scenario in which students had to challenge other participants in order to get high scores and climb up a leaderboard. Results show intense practice supported by a significant number of activities and playing regularity, so the most active and motivated players in the competition achieved significant pronunciation improvement results. The integration of automatic speech recognition (ASR) and text-to-speech (TTS) technology allowed users to improve their pronunciation while being immersed in a highly motivational game.

INDEX TERMS Computer-assisted pronunciation training, mobile learning game, mobile application, English L2 pronunciation, challenges, motivation.

I. INTRODUCTION

Globalization leads to a fast increase in second language (L2) acquisition needs. Although traditional in-classroom lessons and one-to-one tutoring are still preferred as high quality formal learning approaches, computer-assisted language learning (CALL) is becoming a very useful resource as language and speech processing technologies advance [1].

Correct pronunciation is recognized as one of the most important issues in language learning, first because it permits speech comprehensibility and intelligibility to be enhanced and second, it is a means to acquiring a native-like pronunciation [2], [3]. Due to its importance and the fact that an appropriate training could significantly improve it [4], computer-assisted (aided) pronunciation training (CAPT) becomes a very important sub-area of CALL. CAPT is a computer-based language learning approach which enables users to access self-training in a timely and ubiquitous way [5]. A correct design of foreign pronunciation training and learning activities with a CAPT tool is crucial to ensuring effective learning development (performance) and user motivation [6].

One of the main arguments in favor of using games for learning is that they motivate and engage users [7], [8]. Well-designed games deploy techniques that encourage players to achieve a state of intense concentration and full involvement, when challenges are closely paired to ability [9]. Also, the potential of games to create effective social practices can provide means for users to participate in communities sharing learning interests [10], [11]. There is some empirical research about the motivational impact of social learning games [12]–[15], and the effect of competition on user's

The associate editor coordinating the review of this manuscript and approving it for publication was Muhammad Imran Tariq.

performance and learning [16]–[19]. Nevertheless, in the domain of pronunciation training in language learning, to the best of our knowledge, it is non-existent.

In this work, we present the design and evaluation of a mobile learning game which offers a CAPT tool for learners of English as L2, named Clash of Pronunciations (COP). The game focuses on pronunciation training at the segmental level (i.e., the teaching of single speech sounds, such as vowels, disregarding intonation and other suprasegmental aspects of connected speech [3], [20]). It implements the well-established perception / production / perception–production sequence of activities cycle [21] over an adequately chosen set of English words associated to minimal pairs for practicing the pronunciation of English vowel and consonant sounds [22]. In order to ensure a higher level of motivation, we propose that the system should allow users to launch and answer challenges between each other. As compared to our previous challenge-free version of the game [23], [24], the alternative presented here ensures a higher and more stable level of motivation, while also providing a measurable increase in correct pronunciation of the phonemes addressed in the game.

Our previous results have shown that the use of speech technologies, providing immediate audio and visual feedback, contribute to improving the pronunciation competence of the user when they are integrated into appropriate methodologies [23], [25], [26].

Here, we analyze a competitive variant of the CAPT learning game and address the question of how the intensive use of the game allows highly motivated players to achieve significant second language pronunciation improvement.

The rest of the paper is structured as follows. We begin by reviewing the literature about speech technologies and social learning games for CAPT. Then, we describe our game design in detail by presenting the pedagogical fundamentals, the architecture, the game dynamics, and the scoring system. Section IV explains the case study set-up, including details about the recruitment campaign, the participants, and the analysis variables. Finally, we discuss our results across all the relevant issues mentioned above, reaching a set of conclusions that are summarized in the last section.

II. RELATED WORK

From a methodological point of view, our approach is strictly segmental and based primarily on the minimal-pair technique. The game consists of word-based exercises and feedback in the form of articulatory instructions [22], [27]. However, we go one step further in relation to both elements. Our instructions are not just presented in the written format, but as audio-visual texts—a feature already used by [5]. More importantly, our exercises are pair-based rather than word-based and, consequently, their reliance on phonemic contrasts promotes the increase of phonological awareness. Although sporadically used [5], [27]–[29], few systems give such protagonism to, or benefit so much from, the minimal-pair technique [30], [31].

A. SPEECH TECHNOLOGIES FOR PRONUNCIATION TRAINING

Designers must find their way around a vast field of methodological choices for applying computer technology to the teaching of pronunciation. There are systems that commit to the segmental level and the improvement of phoneme perception and production [28], [30], [32]–[35]; whereas other systems focus on such suprasegmental aspects as stress or intonation [36]–[39]. Designers who focus on the suprasegmental level usually take sides with comprehensibility in relation to another classic dilemma: defining the final goal as either standard native-like pronunciation or the attainment of global intelligibility [40], [41].

Information and communication technology can be seen to be at a stage where it can contribute to the teaching of L2 pronunciation by simply looking at the quality attained by current automatic speech recognition (ASR) systems [42]. For instance, Google declares that its machine-learning voice recognition system has achieved a word accuracy rate of 95% for the English language, therefore reaching the threshold of human accuracy in some applications [43]. Issues of intelligibility within the CAPT domain become particularly interesting when ASR technologies are used to diagnose pronunciation and provide feedback [5], [44]–[47]. Typically, a given pronunciation is marked as correct when it is recognized as the expected word by the integrated ASR system. Although ASR systems can also show the hidden Markov model (HMM) scores obtained for each utterance as a means of corrective feedback [28], [48], it has also been pointed out that it heads users toward simplistic iterative trial and error cycles when a low score is attained [44].

Whenever users are purposely exposed to models of good pronunciation, there are also important decisions to be made concerning quantitative and qualitative aspects of the voice to be used. Different designs tend to include a single voice, a reduced number of them, or, as in the case of high variability phonetic training (HVPT), a large gallery of different voices [33], [49], [50]. Qualitative issues concerning the nature and quality of the model voice must be considered. Some have been working with recordings by native speakers [29], [51]–[54]; while others have used manipulated natural speech [32], [33]. While natural voice has generally predominated, recent designs are introducing synthetic voice through the use of text-to-speech (TTS) systems [26], [55]. Research on the quality of TTS has led Google to assert that deep neural network (DNN) technology already produces near-human speech [56]. As mentioned earlier in the introduction, despite a certain amount of controversy concerning the pedagogical use of TTS systems in L2 teaching, there is empirical evidence that supports its applicability [57]–[59].

Humans are able to intuitively learn sounds through simple exposure and imitation, without any theoretical explanations. Nevertheless, many defend the convenience of explicitly describing and teaching the articulation of sounds [60]. CAPT systems are very adaptable in this sense: they may discard explicit articulatory instructions [48], they may mandatorily

incorporate them [5], [28], or they may let the user decide whether they want them or not [44], [45]. When explicit descriptions are incorporated, recourse to fixed and moving images (describing the movement of articulators) are easily incorporated through available technologies [27]. Explicit information for preparation or feedback, on the other hand, need not be restricted to articulatory descriptions: tools for acoustic analysis may be integrated that provide the spectrographic description and formantic values of the model and the produced sounds [28].

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

Gamification refers to the use of game design elements within non-game contexts [61]. In particular, gamification uses game-based mechanics, aesthetics and game thinking to engage individuals, promote learning, and solve problems [62]. Furthermore, educational gamification helps individuals to be immersed in learning, enhances their motivation and brings them playfulness [63].

Dropouts and early abandonment in e-learning are important concerns for the current education industry [64]. Despite the great effort of companies and education centers to reduce these negative aspects by designing engaging and motivational distance and digital courses and applications, these problems have not yet been solved [65].

Duolingo has marked a turning point in the industry of language learning [66]. Before Duolingo, commercial language learning tools such as Sanako[1] or Rosetta Stone[2] developed strategies based on course content selection and the design of interfaces that targeted either academic institutions or self-training. However, Duolingo changed the rules by introducing gamification elements with the clear intention of motivating their users [66].

Even though the number of software applications with gamification elements intended for L2 language learning is increasing, there are few empirical studies that validate their effectiveness and even fewer in the case of pronunciation training. For instance, in McGraw *et al.* [67], a card-based game is presented for L2 vocabulary acquisition with speech technology. In Neri *et al.* [48], a word game based on stories for children with an ASR system, PARLING, is reported. In Strik *et al.* [68], a CAPT game for practicing Dutch oral and grammar skills based on speaking practice and feedback is reported. In Danowska-Florczyk and Motowski [69], the Polish language is taught as a user role-based game, in which its complex grammar system and lexical problems are presented to learners as tasks and activities. There are other innovative ways of pronunciation teaching with gamification, such as a multi-language karaoke application, SLIONS [70], or a recursive dialogue game for personalized pronunciation training [71].

Gamification elements included in educational language learning tools are typically points badges, leaderboards, performance graphs and avatars. A characterization of these elements can be found in the general review presented in Sailer *et al.* [72]. Points are the most extended game elements in tools for language learning. They serve as a resource to evaluate the goodness of user interventions and permit user proficiency to be measured. Most of the time, however, the result is a simple binary output, reporting whether the user correctly performed or failed the activity [69], [73], [74]. In Murad *et al.* [70], an overall score (0 to 100) from a user's karaoke performance is shown to the learner. In the ''Say It Again, Kid!'' digital game [75], a 0–5 star score is used as feedback to the learners [76]. In Duolingo and Babbel,[3] points can involve the accomplishment of user levels and badges and can lead to earning rewards.

Badges are commonly used to present messages or symbols that represent user achievements. These pop-up messages are not only used to give information about users' performance, but also to provide feedback or encourage them to keep on training [67]. For instance, in Duolingo, a user earns a badge if she/he obtains 50 points in a day. In Busuu,[4] a user can obtain different types of badges by performing such learning activities as completing a course or taking a test. In Danowska-Florczyk and Motowski [69], ''The Lord of Memory'' badge is given to the student who remembered most of the new words from previous classes.

Avatars are included to enrich the user experience during games. They are widely used in virtual environments of language learning games [77], [78]. An avatar represents the user in the game, and allows interaction with other users through their respective avatars. For instance, in Wang *et al.* [78], human-based avatars allow more than two people to be involved in conversations. In Danowska-Florczyk and Motowski [69], a warrior-based avatar represents each group of students in the game. In Duolingo, Duo is an owl which can serve as a coach to motivate users to achieve higher learning goals and it can also instruct users.

Leaderboards display a ranking that permits users to compare their performance with that of other players. This is specially interesting in the present work as it permits users to compare their own level to that of their counterparts, contributing to building self-awareness of level. It also allows competitions to be established as a means to promote social interaction between users [15]. This element is commonly used in social learning applications and courses, but is almost non-existent in the state-of-the-art about pronunciation training studies with CAPT. In Van Hentenryck and Coffrin [79], some results suggesting the positive effect of the leaderboard in motivating students to push their solutions to a problem are presented. In Duolingo and Babbel, progress is measured by gaining experience points (XP) and going up levels, which affects their social leaderboards.

[1] http://www.sanako.com/
[2] http://www.rosettastone.com
[3] https://babbel.com
[4] https://www.busuu.com

Performance graphs offer information about the student's progress over time. The difference with leaderboards is that, in this case, performance graphs do not compare the user's performance to other players. For instance, Sanako offers a complete dashboard for teachers to follow up student progress in the language laboratory. In Strik *et al.* [68], a final report about all the mistakes made by the learner is generated after each conversation. In Duolingo and Babbel, graph statistics and historical records are available to users.

III. THE COP APPLICATION

A. PEDAGOGICAL BASIS

The game relies on the use of a set of minimal pairs, that is, two words frequently monosyllabic, which are identical except for one sound, while their meanings are completely different. In English, «bet»-«bed» or «pen»-«pan» are minimal pairs. Although minimal pairs were originally conceived as a linguistic strategy, in the structuralism paradigm, for collecting the phonemic repertory of different languages, they naturally became integrated in teaching pronunciation methods [80], [81]. Our proposed tool implements the traditional program that consists of (1) **exposure** activities with minimal pairs, synthesizing both words in the pair at varying paces, and allowing the student to directly experience the perceptive differences between two particular phonemes; followed by (2) **discrimination** (perception) activities in which the player must decide to which word of the pair a given synthesized audio corresponds, and finally, (3) **production** activities, in which users must pronounce the words of the minimal pair correctly so that the ASR system will accept them. The main goal of this three-step process is to increase students' self awareness, making them realize that there are relevant acoustic characteristics that were not correctly perceived before; and to ensure that such acoustic characteristics are assimilated by the speakers so they are able to accurately articulate the target sounds.

B. SYSTEM'S COMPONENTS

The client-server software architecture of COP includes several elements (see Fig. 1). From left to right, *AndroidClient* represents an Android device (version 4.4 or higher) in which the COP application is installed. Interaction results with the application are saved as a JSON format (*LogFile*) that compiles all possible depersonalized data diachronically and sends it to our *WebLogger* in a *WebServer*. The lists of minimal pairs are defined by English phonetic experts in a simple text file (*JSONWordsDababase*) which includes in each line the following information on the pair: orthographic transcription, phonetic transcription, and the possible homophone words. They consist of 329 minimal pair words (English words and their phonetic transcription). These lists can easily be extended to new words and languages for future experiments. They are organized in lists corresponding to a pair of phonemes to contrast: twelve consonant and eight vowel contrasts with at least ten paired words each.

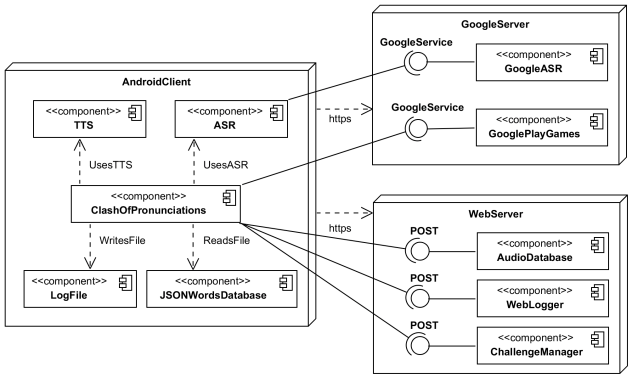

FIGURE 1. COP system's environment. There are three main different components: an Android device where COP is installed (left-hand component), the Google server to provide some online services (top-right component), and a private web server to collect data (bottom-right component).

Regarding speech technology, synthetic speech output is produced by Google's offline TTS application for Android. It offers an instant synthesized version of the minimal pair words after receiving their orthographic transcription from our CAPT tool. Google's ASR system provides real-time feedback about user utterances. In particular, the audio file of the utterance is sent to Google's ASR through the Internet. The prediction of the ASR system provides an n-best list of n possible string hypotheses (in the case of COP, this parameter is $n = 3$), ordered from highest to lowest confidence rates. These values represent numeric values called g-scores, which are proportional to the reliability of the prediction of each text hypothesis (from 0% to 100% in a scale [0, 1]). Thus, the utterance is considered correct as long as it is within the list of n elements returned by the ASR.

GooglePlayGames is an online Google platform that provides gaming service and software development kits of ready to use game features in digital applications. In particular, we have integrated some functionalities, such as the user's game profile, achievements, leaderboards, and turn-based multiplayer support. This platform is complemented by the *ChallengeManager* component, specifically developed for COP. It controls the user's sign-up and log-in, the scoring system, the list of possible players to challenge, and keeps track of the results of each challenge (player, score, and date).

C. USER INTERFACE

Once the COP application is opened, a welcome screen is shown which consists of a menu of six options (see Fig. 2 and the upper-left screenshot of Fig. 3). The first one is "Playing". In this menu option, players can launch new challenges (this process is described in Section III-D). The second is "Pending challenges". Players can review the challenges they are involved in (both finished and pending). They can also answer pending challenges by playing their respective match. The third option is "Training". Users can practice isolated phonemes by selecting the specific training activity and

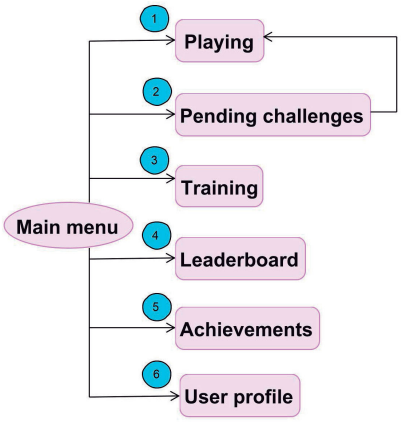

FIGURE 2. Interaction options available in the main menu of the COP application. Each number corresponds to the same number in the upper-left graphical user interface screenshot of Fig. 3.

the list of words (this process is described in Section III-D). The leaderboard of points is also available from this menu. A grid of players with their position, points, nickname, and avatar is displayed (see upper-right screenshot of Fig. 3).

The fifth option displays the achievements obtained. There are two types of achievement. First, those related to the points obtained during the competition: ''The phonetic beginner, magician, expert, master, and hero'', available on reaching 50, 100, 200, 300, and 500 points, respectively. Second, those related to the participation in challenges: ''Sociable'' (play 5 challenges), ''Known among all'' (play 50 challenges), ''World Idol'' (play 500 challenges), ''Full house!'' (launch a challenge with at least 4 other participants), and ''Never-ending game'' (win a challenge). Finally, the ''user profile'' data such as nickname, avatar, remaining daily challenges, points, and number of finished challenges is presented.

D. DYNAMICS OF THE GAME

The competition proposed in our experimentation consists of an asynchronous turn-based social game in which users challenge each other and their results are reflected on a leaderboard. The main goal of a user playing with the developed CAPT game is to achieve points by successfully performing some pronunciation **activities** based on the exposure–perception–production cycle (see Section III-A) in **challenge matches**, trying to reach the best position possible on a **leaderboard** (see the upper-right screenshot of Fig. 3). A match can be played in two modes: **Playing** and **Training**. In the Playing mode, users get points by participating in challenges against other participants (see the interaction flow in Fig. 4). A minimum of two and a maximum of five participants perform the same activities in their respective matches.

The idea behind the game is that students can practice the target minimal pair words through pronunciation challenges with other players by performing perception and production activities. The minimal pairs list presented in the

FIGURE 3. Screenshots of the main menu (upper-left), leaderboard (upper-right), discrimination (lower-left) and production (lower-right) activities in a challenge match. The numbers of the main menu correspond to the interaction functionalities described in Fig. 2. "Juega" means *play*, "Retos pendientes" means *pending challenges*, "Entrena" means *training*, "Salir" means *exit*, "Desconectar" means *log out*, "partidas" means *challenges*, "puntos" means *points*, "Ronda" means *round*, "Pulsa en la que oigas" means *click on the word you have heard*, "Pulsa y pronuncia" means *click and pronounce the word*, "Correcto" means *correct*, "2 intentos restantes" means *two attempts left*, "Has pronunciado" means *You have pronounced*.

challenges is randomly selected by the system to try to keep the same variety of the difficulty level along the competition. Users obtain points by performing discrimination and production activities in which a minimal pair appears. In the **discrimination activities** of the Playing mode, the system synthesizes one of the words in the minimal pair. The player must select the word that she/he thinks has been uttered, using the interface displayed in the lower-left screenshot of Fig. 3. Users can listen to the sound as many times as they want, but the final score will be penalized (see Section III-E). In the **production activities** of the Playing mode (see the lower-right screenshot of Fig. 3), users must try to pronounce correctly the two words of the minimal pair. When there is no match, the output of the ASR system is displayed as an emerging badge. Additionally, the system invites users to a

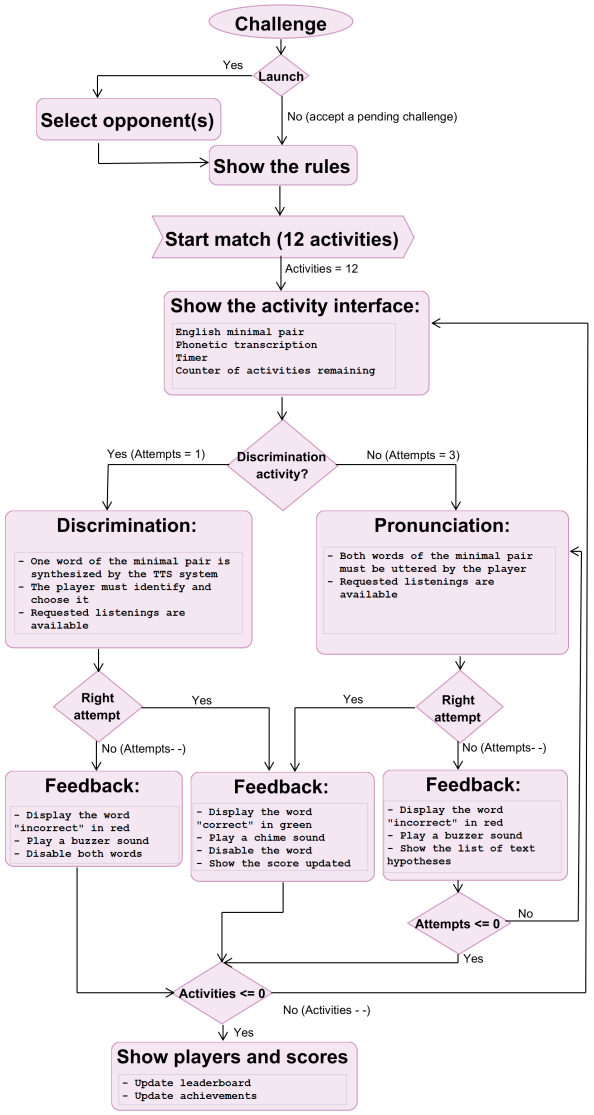

FIGURE 4. Flow diagram of a challenge in the Playing mode.

requested listening of the correct pronunciation via the TTS system. There is a maximum number of three tries per word of the pair. There is a time limit of 100 and 10 seconds per production and discrimination activity, respectively.

When playing a match in the Training mode, users can freely select the minimal pairs contrast to practice and the activity type (exposure, discrimination or production). In exposure Training activities, a minimal-pair is presented to the user and she/he has to (1) listen to the words up to five times, each repetition being noticeably slower than the previous one; (2) repeat them at least once; and (3) compare both, the synthesized sound and her/his own utterance. Exposure activities offer users a first hand unmediated aural experience of each contrast in order to assist their assimilation. On the other hand, in discrimination and production Training activities, the dynamics are the same as in the Playing mode. However, each type of activity is practiced separately (not mixed up as in the Playing mode). In these activities, the orthographic and phonemic forms of the two components of a minimal pair are displayed. However, there is no time limit and users do not obtain extra points while training. To summarize, a COP user has the following options:

- *Play matches*. Users must perform different discrimination and production activities with minimal pairs to obtain a score. When a player finishes the match, a message with the points and right and wrong attempts is displayed. She/he must wait for the other participants' results (asynchronously) to declare winners and losers and add the points achieved in the match to the leaderboard (see the complete flow of a challenge in Fig. 4).
- *Training matches*. Besides playing matches, users can choose training activities as an unlimited option. They can choose exposure, discrimination or production activities of the minimal pair contrasts they want. They do not add points to the leaderboard in this mode.
- *Launch challenges*. Each user can challenge up to five other users (from those who are 10 positions below or above her/him in the ranking). Then the procedure is the same as playing a match. Each player can launch challenges until she/he reaches the 30 daily maximum challenges (launched or accepted).
- *Answer the received challenges*. The user receives a message about a player who is challenging him/her. The incoming challenge can be accepted to perform the match or ignored without computing in the maximum of 30 per day (but one is deducted from the challenge creator).

Finally, in order to establish the competitive game configuration, we have followed the design guidelines stated by Johnson and Johnson [82].

- *Interaction with other parties*. When a user wants to participate in the Playing mode, he or she must **launch** a new challenge or **accept** the incoming challenge sent by another game player (it could be rejected as well). All subjects of a given challenge perform the same discrimination and production activities included in the match in their respective turn (twelve activities, six for discrimination and six for production, interspersed).
- *Winning rules*. The winner of a challenge is the player who achieves the highest *MatchScore* (see Section III-E). Ties (more than one winner) can occur in challenges involving more than two players. Users get points only from finished challenges, when the last user performs her/his match. The winner of a competition is the player who achieves the most points during all the competition days, reaching the first position of the leaderboard.
- *Negative goal interdependence*. If a user wins the challenge, the others lose (except in challenge ties of more than two players).
- *Comparability among participants*. The common leaderboard is updated with the final scores achieved in each challenge (see the upper-right screenshot of Fig. 3).

In order to guarantee a similar game level, the possible available opponents for a challenge belong to a range of ten positions above and below the creator's current leaderboard position. There is a limit of 30 matches per user per day in order to avoid counterproductive extra working load.

- *Perceived scarcity*. Only the best players can reach rewards and achievements since their quantity is limited. Only the players who finish at least 60 challenges obtain an academic certification. Players who reach one of the fifteen first positions of the leaderboard obtain a reward.
- *Quantity of winners*. The number of winners of a challenge are the participants with the highest *MatchScore*. The number of winners of the competition is one (the first position on the leaderboard).

E. SCORING SYSTEM

The points achieved by a player in a challenge depend on the performance in her/his corresponding match and the rest of the players. It can be defined as:

$$MatchScore = BaseScore + ExtraScore \tag{1}$$

The *BaseScore* is established with the performance result in each one of the discrimination and pronunciation activities in each player's match:

$$BaseScore = \sum_{D=1}^{6} u_D + \sum_{P=1}^{6} v_P; \quad u_D, v_P \in \{\alpha, \beta, \gamma\} \tag{2}$$

where u_D and v_P are the weight values assigned to the activity performance value according to the result: α is the value assigned to a wrong attempt (0), β is the value referred to a right attempt with some help, such as a request for a word listening or performing more than one production attempt (1), and γ is the value assigned to a right attempt without help (2).

TABLE 1. Extra points scoring system of COP (*ExtraScore* value). *IsCreator* value is true when the player launched the challenge, *MaxBaseScore* value is true when the *BaseScore* achieved by the player is the highest one of the challenge, and *BetterRank* indicates if the leaderboard position of the player is higher than the position of the opponent(s). *rank*1 and *rank*2 are the leaderboard position of the player with the higher and lower position of the leaderboard, respectively. *n* is the number of players in the challenge.

Condition			Extra points	
IsCreator	MaxBaseScore	BetterRank	$n = 2$	$n \in \{3, 4, 5\}$
T	T	T	0	0
T	T	F	$rank1 - rank2$	$3 * (n - 1)$
T	F	T	$-(rank1 - rank2)$	$-(n - 1)$
T	F	F	0	0
F	T	T	0	0
F	T	F	$rank1 - rank2$	$(n - 1)$
F	F	T	0	0
F	F	F	0	0

The *ExtraScore* is the value added to the *BaseScore* after all players in the challenge finish their matches. As shown in Table 1, this value depends on the number of players, the player who launches the challenge (*IsCreator*) and the leaderboard position difference between the player and her/his opponents (*BetterRank*) and the *BaseScore*.

In particular, the extra points scoring system has the premise of rewarding courageous players. Those who challenge players above in the leaderboard and achieve the victory obtain more extra points (no penalties in case of losing the challenge). However, when top players challenge worse ones in terms of leaderboard position and lose the challenge, they are penalized.

IV. CASE STUDY

In previous work, we reported on the design and testing of a serious game for teaching L2 pronunciation, TipTopTalk! (TTT), articulated into minimal pair activities of exposure, discrimination, and production that counted toward a final score in an individual self-competition framework [23], [24]. Initially, 100 native Spanish (L1) undergraduate students from the University of Valladolid (Spain) signed up online to collaborate in the TTT competition. However, the final number of users from whom we could gather representative data (those who performed every stage of the competition) was 45; 22 (48.9%) were female and 23 (51.1%) were male. The average age was 23 years (M = 23.13, SD = 3.58).

Although TTT did not include challenges, COP and TTT share the basic gamification features, such as points, badges, leaderboards, profile avatar, and performance graphs. TTT also included such specific features as number of remaining lives, clear tickets and difficulty level selection. Despite the inclusion of these elements in TTT, a decreasing trend in the quantity of pronunciation activities was generally observed after the first days of the 24 day competition time, most probably due to habituation and gradual loss of interest in the game [23]. This led us to analyze the differential impact on the user's second language pronunciation improvement of a second version of the game, COP, which evolves from individual to social by introducing a challenge-driven system. The case study set-up and game instrumentation allowed relevant data to be gathered to analyze user performance depending on their degree of involvement in the competition, expressed in terms of the user's motivation and pronunciation improvement.

First, we describe the user's enrollment in the COP competition process. We then explain the sources for gathering data. Finally, we declare and define the metrics related to our study.

A. PARTICIPANT ENROLLMENT

Initially, 347 native Spanish (L1) undergraduate students from the University of Valladolid (Spain) signed up online to collaborate in the COP competition. 214 (61.7%) were female and 133 (38.3%) were male. The average age was 21 years (M = 21.3, SD = 1.96). However, the final number of users for whom we could gather representative data (those who performed every stage of the competition) was 165.

The participant recruitment process in the COP competition (as in TTT) was made by sending invitation emails to their corporate University email addresses and by means of invitation talks in selected classrooms. To participate in the competition, students had to fill in a registration form with some personal information and sign an informed consent. Additionally, participants had to complete the pre/post-competition questionnaires at the beginning and end of the competition (see Section IV-C). A prize was given to the first 15 players classified on the leaderboard. In addition, an academic certification was also offered to players who participated in at least 60 challenges and filled in both questionnaires.

After registering, students received the instructions to download the game installation file from Google Play. A time window with a start and end date for playing was established for a total of 24 days of competition after the first week of enrollment. Users could play anytime and anywhere, using their own smart devices. Finally, during the whole competition, the research team answered mails from users asking for technical help about the installation and execution of the game.

B. GATHERING DATA

We gathered data from four different sources:

- Registration form: users' demographic information such as name, age, sex and studies.
- Pre-test questionnaire: declared level of English proficiency and a likert-type questionnaire to evaluate the degree of competitiveness of the users.
- Post-test questionnaire: a final questionnaire for users who finished the corresponding competition about (1) the usability of the tool (adapted from Brooke *et al.* [83]), (2) declared reasons for playing, and (3) attitude toward competition. Furthermore, we designed another questionnaire to obtain information from users who abandoned the game before completing all the stages (4). Questionnaires (2), (3) and (4) were developed by members of the research team who are experts in psychology and education.
- User interaction log file: the game monitors users' interaction by recording all relevant events related to the metrics of our study (see Section IV-C).

C. METRICS

In this section, we focus on measuring and analyzing game intensity, motivation, performance, and learning improvement. They are described in the following subsections.

1) GAME INTENSITY AND MOTIVATION

We characterize the user's game intensity in terms of declared reasons for participating in the competition (motivation), and quantity and regularity in match participation. Data related to the user's motivation is gathered into log files:

- Number of active days: quantity of days in which a user participates in Training or Playing matches.
- Number of activities: amount of discrimination and production attempts performed by a user in Training or Playing matches.
- Degree of motivation: subjective answers to the questionnaire at the end of the competition (motivation for participating, feelings during the competition and reasons for abandoning).

2) PERFORMANCE

Related to the number of events tracked from each participant. Different indicators of the CAPT tool characterize a user's performance:

- Production attempt: every attempt to produce correctly the proposed word of a pair. Binary value (true, false) indicating whether the orthographic transcription of the word matches the user's utterance result of the n-best list of hypotheses of the ASR.
- Production success rate: percentage of right production attempts according to the total number of attempts of the user in the competition.
- Discrimination attempt: every attempt to select correctly the word of a pair synthesized by the system. Binary value (true, false) indicating whether the user chooses the word of the minimal pair that the system synthesizes in the activity.
- Discrimination success rate: percentage of right discrimination attempts according to the total number of attempts of the user in the competition.
- Number of matches: quantity of matches (Playing or Training mode) in which the user participates (launched or answered).
- Match duration: time a user spends on performing the activities of a match.
- Challenge win rate: number of challenges won by a player divided by the total number of challenges participated in.
- Leaderboard position: place on the competition's leaderboard that a player occupies during a challenge.
- Number of points: quantity of points obtained from a finished challenge in the Playing mode (see Section III-E).

3) PROFICIENCY IMPROVEMENT

The learning improvement analyzed in this case study is related to the perception and production skills involved in the activities of the competition. In our study, player skills are the success rates of production and discrimination activities at a specific moment of time. In particular, we compare inter and intra-group success rates of the same quantity of these activities at the beginning and at the end of the competition — equivalent to the first and last two days of the competition.

4) USER GROUPS

We use the metric value of the number of matches (Playing mode) to divide COP users into three statistical tertiles in terms of quantity level of activity performance:

TABLE 2. Mean number of activities (production and discrimination attempts) performed by the users in each competition, classified by mode (Playing and Training). Column $\overline{\#A}$ contains the mean and percentage number of activities performed by an average user in the mode. Column #PA/#PM includes two values: #PA is the number of participants in the corresponding activity in the mode. #PM refers to the total number of participants in the mode. The percentage number refers to the rate of users in the activity compared to the total in the mode. The number in parenthesis following the case study acronym refers to the number of participants with representative data.

		COP (165)		TTT (45)	
	Activity	$\overline{\#A}$	#PA/#PM	$\overline{\#A}$	#PA/#PM
Training	Production	81.3 (53.0%)	102/128 (80%)	24.3 (39.6%)	18/29 (62%)
	Discrimination	72.1 (47.0%)	108/128 (84%)	37.0 (60.4%)	25/29 (86%)
Playing	Production	2168.0 (60.5%)	165/165 (100%)	349.9 (46.3%)	23/36 (64%)
	Discrimination	1413.2 (39.5%)	165/165 (100%)	405.2 (53.7%)	36/36 (100%)

TABLE 3. Declared reasons for playing. The question in the final questionnaire was: select the statements that fit your motivation for playing; you can choose as many as you consider appropriate. The symbol * indicates statistically significant inter-group differences with a Chi-Square test at 95% confidence. Bold values are explained in the results section.

Question	Casual	Habitual	Constant	Total	Answer
Q1.1	1.9%	0.0%	0.0%	0.6%	- (No answer)
Q1.2	**11.1%**	**16.4%**	**37.5%**	**21.8%***	I was hooked on the game, and played as much as I could.
Q1.3	74.1%	80.0%	71.4%	**75.2%**	The prospect of improving my pronunciation was an important incentive in using the app.
Q1.4	20.4%	18.2%	**5.4%**	14.5%*	I felt obliged to play, basically because I needed the academic certificate.
Q1.5	18.5%	21.8%	26.8%	22.4%	I have played a lot because I felt that my pronunciation was getting better.
Q1.6	**50.0%**	**69.1%**	**91.1%**	**70.3%***	Climbing up the leaderboard was my main incentive in using the app.
Q1.7	50.0%	41.8%	35.7%	**42.4%**	The possibility of challenging my mates was an important incentive for using the app.
Q1.8	**20.4%**	**47.3%**	**76.8%**	**48.5%***	Winning the prize was an important incentive for using the app.

T1 (Constant), T2 (Habitual), and T3 (Casual). They are intended to classify users by a high, medium, and low participation in the competition, respectively.

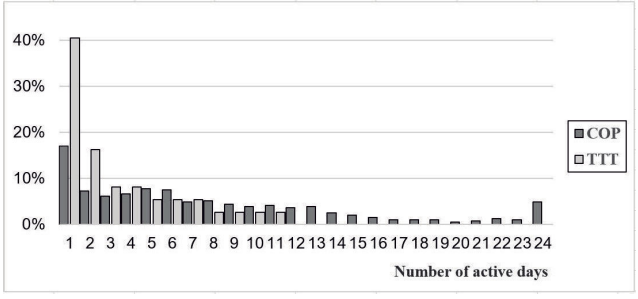

FIGURE 5. User distribution by number of active days with participation in each competition.

V. RESULTS

A. GAME INTENSITY AND MOTIVATION

First, we analyze the activity traced per day in the competition in order to discern the implication of the challenging competition as a motivational factor. Fig. 5 shows the number of days in which some players' activity was traced, concerning the distribution of player activity throughout the 24 competition days. In both competitions, a high number of users only played the game one day, this quantity being greater in TTT than in COP (41% vs. 17% of the users only participated one day, respectively). Regarding the rest of the users, the plot describes clear differences: no TTT user played more than 11 days, while 25% of COP users played more than 11 days and 5% were active during the whole 24 day competition period.

Second, Table 2 shows the total workload activity during the competitions by means of quantity of perception and production attempts in both Playing and Training activities. The data and the statistical tests in this manuscript were computed with the SPSS software [84]. In the Playing mode, an average user of COP performs more than six and three times production and discrimination activities than an average TTT user (2168.0 vs. 349.9 and 1413.2 vs. 405.2, respectively). These differences are statistically significant in both cases ($U = 186.0$, $p < 0.001$ for productions and $U = 907.0$, $p < 0.001$ for discriminations, Mann–Whitney U tests). In the Training mode, an average user of COP performs almost four times more production activities (81.3 vs. 24.3; $U = 606.5$, $p = 0.022$, Mann–Whitney U test) and two times more discrimination activities (72.1 vs. 37.0, no significant differences) than an average TTT user.

Third, Table 3 displays the reasons provided by the users for taking part in the COP competition. The majority of players declared that improving English pronunciation (Q1.3 75.2%) and climbing up the leaderboard (Q1.6 70.3%) were the principal reasons for playing. Winning a prize (Q1.8 48.5%) and beating a known player (Q1.7 42.4%) were the third and fourth reported reasons.

Concerning the inter-group differences, winning a prize (Q1.8) was the predominant motivational reason for playing for Constant players: 76.8%, being clearly higher than in the other two groups, Habitual and Casual, 47.3% and 20.4%, respectively. There are statistically significant differences in all cases: between Casual and Habitual ($\chi^2(1) = 8.795$,

TABLE 4. Attitude toward competition. The question in the final questionnaire was: check the statements expressing your feelings toward competition during the game; you can choose as many as you consider appropriate. The symbol * indicates statistically significant inter-group differences with a Chi-Square test at 95% confidence. Bold values are explained in the results section.

Question	Casual	Habitual	Constant	Total	Answer
Q2.1	31.5%	40.0%	44.6%	**38.8%**	- (No answer)
Q2.2	3.7%	5.5%	0.0%	**3.0%**	I enjoyed the training matches more than the pressure of the game.
Q2.3	16.7%	12.7%	10.7%	13.3%	I would rather play against the machine and against myself privately.
Q2.4	7.4%	5.5%	1.8%	4.8%	Challenging others and the leaderboard make me feel bad and puts me off participating.
Q2.5	**13.0%**	**10.9%**	**28.6%**	**17.6%***	I feel uncomfortable with so many challenges.
Q2.6	1.9%	3.6%	3.6%	**3.0%**	The challenges have caused me anxiety and discomfort.
Q2.7	**51.9%**	**43.6%**	**23.2%**	**39.4%***	I like to advance at my own pace without comparing myself with others.
Q2.8	7.4%	5.5%	5.4%	6.1%	It makes me uncomfortable to find myself in a situation where I must prove that my pronunciation is good.

TABLE 5. Early dropout questionnaire results (anonymous). The question was: check the statements that fit your reasons for early abandonment of the competition.

Question	Total	Answers
Q3.1	42.42%	Technical reasons (my mobile device is not Android, I cannot install the game...).
Q3.2	41.67%	Lack of time.
Q3.3	27.27%	It bothered me that other users did not accept the challenges or the delayed time until accepting them.
Q3.4	18.94%	The game bothered me or I did not like it.
Q3.5	12.88%	I did not find it useful to learn English.
Q3.6	4.55%	I was frustrated not being able to win.
Q3.7	3.79%	The competition made me feel bad.

$p = 0.003$); between Casual and Constant ($\chi^2(1) = 35.010$, $p < 0.001$); and between Habitual and Constant ($\chi^2(1) = 10.275$, $p = 0.001$).

The motivation for climbing up the leaderboard (Q1.6) made the difference: 91.1% of Constant players vs. 69.1% (Habitual) and 50.0% (Casual). There is a statistically significant difference between Constant and Casual ($\chi^2(1) = 22.481$, $p < 0.01$) and between Constant and Habitual ($\chi^2(1) = 8.436$, $p = 0.04$). The percentage of players that declared being engaged (Q1.2) reached 37.5% in the Constant group, this percentage being lower for the other groups: 16.4% and 11.1%. There is a statistically significant difference between Constant and Casual players ($\chi^2(1) = 10.337$, $p < 0.002$) and between Constant and Habitual players ($\chi^2(1) = 6.285$, $p = 0.018$).

Few players (less than 21% in any case) declared the academic certificate was the principal reason (Q1.4). This fact had less impact for Constant players: 5.4%, with a statistically significant difference with respect to Casual players ($\chi^2(1) = 5.579$, $p = 0.023$).

Table 4 shows the possible reasons that reflect why participants did not enjoy the competition. None of the possible answers depicted in this table were over 50% (except the Casual answer for Q2.7). In fact, close to 40% of the users did not declare any negative opinion toward competition (Q2.1). Only 3% of the players declared more enjoyment during training than in competition (Q2.2) and again only 3% of players affirmed they suffered anxiety during the game (Q2.6).

Concerning inter-group differences, Constant users declared feeling uncomfortable with such a number of matches (Q2.5), 28.6%, with statistically significant differences with Casual players (13.0%, $\chi^2(1) = 4.050$, $p = 0.044$) and with Habitual ones (10.9%, $\chi^2(1) = 5.447$, $p = 0.020$). Of particular interest is the answer related to practice at their own pace (Q2.7), not only because it was the most selected answer overall, 39.4%, but also because it made the difference between the Constant group (23.2%) and the others: Habitual (51.9%, $\chi^2(1) = 5.208$, $p = 0.022$) and Casual (43.6%, $\chi^2(1) = 9.643$, $p < 0.002$).

Concerning players who quit the COP competition before completing all mandatory protocol steps (N = 182), Table 5 shows that most early dropout reasons were (N = 129): technical reasons (Q3.1 = 42.42%) and lack of time (Q3.2 = 41.67%). Few players reported being uncomfortable during the competition (Q3.7 = 3.79%) or frustrated by losing (Q3.6 = 4.55%).

B. PERFORMANCE

Table 6 displays a summary of the user's performance in the COP and TTT competitions. It is analyzed by the different indicators and groups defined in Section IV-C2, since not all users interact with the tool in the same way. A non-parametric Kruskal–Wallis test reports statistically significant inter-group differences ($p < 0.001$, see Table 7) in ten of the eleven performance indicators, except for the training match mean duration time ($H(3) = 5.851$, $p = 0.119$). The group named Casual is the one with the least implicated users during the COP competition with an average of 91.6 matches (Table 6). On the other hand, players of the group named Constant are very active, reaching 468.7 matches on average. The 20 most active users of this last group performed an average of 29 matches per day (30 is the maximum allowed). In this group, we also find the users that competed until the end to get to the top of the leaderboard. The Habitual group

TABLE 6. Indicators of activity per type of user. COP participants are divided into three groups using tertiles: Constant, Habitual and Casual.

		COP								TTT	
		Constant (56)		Habitual (55)		Casual (54)		Total (165)		Total (45)	
		Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD
Training	Production success rate	50.7%	33.4	47.5%	33.4	21.8%	30.4	40.2%	34.8	24.8%	35.1
	Discrimination success rate	58.1%	33.6	52.6%	32.4	30.3%	36.3	47.2%	36.0	62.6%	34.8
	Number of matches	30.2	35.3	13.6	18.8	6.0	13.5	16.7	26.4	6.3	8.1
	Match mean duration (s)	36.8	17.8	50.9	41.4	39.3	33.9	42.3	32.9	113.0	346.5
Playing	Production success rate	67.4%	9.4	55.5%	11.6	53.2%	9.9	58.8%	12.0	35.1%	35.5
	Discrimination success rate	83.3%	4.8	76.0%	5.9	72.2%	6.1	77.2%	7.3	60.0%	34.0
	Number of matches	468.7	188.8	144.3	29.7	91.6	11.3	237.1	201.1	53.3	88.8
	Match mean duration (s)	68.3	14.1	85.6	27.6	97.0	65.6	83.5	43.0	43.6	48.8
Leaderboard	Challenge win rate	49.4%	11.5	43.4%	17.4	41.0%	16.4	44.6%	15.6	-	-
	Mean position	29.4	17.8	98.3	37	149.2	33.0	91.6	57.8	23.0	13.1
	Mean number of points	17.5	3.0	12.7	3.2	12.2	3.1	14.1	3.9	13.1	8.57

TABLE 7. Kruskal–Wallis test statistics of indicators of activity for the groups of both competitions (Constant, Habitual, Casual, and TTT).

		H	df	p
Training	Production success rate	26.915	3	< 0.001
	Discrimination success rate	24.712	3	< 0.001
	Number of matches	33.172	3	< 0.001
	Match mean duration (s)	5.851	3	0.119
Playing	Production success rate	49.041	3	< 0.001
	Discrimination success rate	65.489	3	< 0.001
	Number of matches	164.739	3	< 0.001
	Match mean duration (s)	83.014	3	< 0.001
Leaderboard	Challenge win rate*	9.590	2	0.008
	Mean position	166.682	3	< 0.001
	Mean number of points	54.993	3	< 0.001

* Only in the COP competition

is made up of users who kept on playing after reaching the minimum to obtain the certification, in spite of having little chance of achieving the rewards.

The comparison between groups reveals statistically significant differences in the number of matches played in all cases (see Table 8). The Constant players are the most efficient ones, since they achieve the highest rates of wins and production success attempts, having statistically significant differences with respect to the other two groups. Regarding the success rate in discrimination activities, the Mann–Whitney U tests indicate significant differences between the three groups. The differences in the performance of the player groups have an impact on the final positions of the leaderboard. Constant players occupy higher leaderboard positions than Habitual, and Habitual positions are, at the same time, higher than Casual ones: 29.4, 98.3, and 149.2, mean positions, respectively, being statistically significant differences in all cases. These positions are in line with the mean number of points obtained per match: 17.5 vs. 12.7 vs. 12.2 for Constant, Habitual, and Casual groups, respectively. In this case, there are statistically significant differences between the Constant group and the others. The time spent in matches (Playing mode) is another indicator which also evidences a higher skill level of Constant players. It is shorter than the rest of the groups (68.3s vs. 85.6s, and 97.0s, respectively) and this difference is statistically significant.

Regarding training activities, Constant players train more than the rest of the users: 30.2 vs. 13.6 vs. 6 (number of Training matches). There are statistically significant differences between the three groups in all cases. There are also statistically significant differences in the success rate of Training production and discrimination activities when a Mann–Whitney U test is applied in two cases (Constant vs. Casual and Habitual vs. Casual).

Table 8 also shows the statistical comparison between the COP groups and the TTT one concerning the performance indicators presented in Table 6. The TTT group members played fewer matches (53.3 as an average) than the other three COP groups, this difference being statistically significant in all cases. This low participation in Playing mode is also in line with the success rate achieved in production activities of matches. Constant and Habitual groups outperformed TTT players with statistically significant differences. According to the success rate in discrimination activities, there is only one case with statistically significant differences when applying a Mann–Whitney U test, Constant vs TTT.

In the Training mode, the least active group's players of COP (Casual) trained a similar amount to the TTT ones, 6 and 6.3 Training matches, respectively. However, the Mann–Whitney U test results indicate there are important statistically significant differences when comparing the number of Training matches of the TTT group with the two best COP ones, Constant and Habitual. TTT players reached a 24.8% success rate of training production activities, slightly better than the Casual COP players (21.8%). However, the Constant and Habitual groups outperformed the TTT one in production success rate values. On the other hand, the TTT group reached the best discrimination activities success rate in the Training mode (62.6%), with statistically significant differences in all cases except for Constant vs. TTT.

The last indicator that also evidences a higher implication of COP players is the time spent in matches of the Playing

TABLE 8. Pairwise comparisons using Mann–Whitney U test (U, p-value) for the groups Constant, Habitual, Casual, and TTT of both competitions.

Dimension	Indicator	Constant–Habitual	Constant–Casual	Habitual–Casual	Constant–TTT	Habitual–TTT	Casual–TTT
Training	Production success rate	–	(834.0, < 0.001)	(885.5, < 0.001)	(787.5, < 0.001)	(830.0, 0.003)	–
	Discrimination success rate	–	(875.5, < 0.001)	(976.0, < 0.001)	–	(941.0, 0.039)	(638.0, < 0.001)
	Number of matches	(1196.5, 0.042)	(702.0, < 0.001)	(839.5, < 0.001)	(710.5, < 0.001)	(896.5, 0.018)	(535.0, < 0.001)
	Match mean duration (s)	–	–	–	–	–	–
Playing	Production success rate	(659.5, < 0.001)	(449.5, 0.004)	–	(565.5, < 0.001)	(889.5, 0.015)	–
	Discrimination success rate	(512.5, < 0.001)	(168.0, < 0.001)	(987.5, 0.003)	(642.5, < 0.001)	–	–
	Number of matches	(687.0, < 0.001)	(435.0, < 0.001)	(876.0, < 0.001)	(43.0, < 0.001)	(360.0, < 0.001)	(535.0, < 0.001)
	Match mean duration (s)	(746.0, < 0.001)	(514.0, < 0.001)	–	(481.5, < 0.001)	(322.0, < 0.001)	(252.0, < 0.001)
Leaderboard	Challenge win rate*	(1136.5, 0.017)	(1028.0, 0.004)	–	–	–	–
	Mean position	(37.0, < 0.001)	(0.0, < 0.001)	(429.0, < 0.001)	(0.0, < 0.001)	(0.0, < 0.001)	(0.0, < 0.001)
	Mean number of points	(413.5, < 0.001)	(345.5, < 0.001)	–	–	(755.0, < 0.001)	–

* Only in the COP competition

TABLE 9. Production and discrimination success rates at the beginning and end of the competitions. *First* and *Last* represent the right success rates in the first and last 15% of the total number of activities. *Z* and *p* values are derived from a Wilcoxon signed-rank test at the 95% confidence level (2-tailed).

		COP								TTT	
		Constant (56)		Habitual (55)		Casual (54)		Total (165)		Total (45)	
		Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD
Production	First (%)	66.7	7.8	53.7	8.8	52.9	7.9	57.8	10.3	34.1	9.1
	Last (%)	73.4	5.7	56.1	10.4	55.2	8.0	61.7	11.8	36.7	11.9
	Diff (%)	6.7	6.5	2.4	8.4	2.3	5.5	3.8	7.2	2.6	9.6
	(Z, p)	(-5.7, <0.001)		(-3.2, 0.001)		(-2.9, 0.004)		(-7.1, <0.001)		(-1.2, 0.233)	
Discrimination	First (%)	76.9	5.7	73.7	7.3	71.8	6.2	74.1	6.7	58.2	14.0
	Last (%)	87.0	5.7	78.1	6.7	73.6	7.9	79.7	8.8	61.3	11.7
	Diff (%)	10.1	4.7	4.4	6.4	1.8	6.6	5.4	6.9	3.1	15.6
	(Z, p)	(-6.5, <0.001)		(-4.3, <0.001)		(-2.9, 0.003)		(-8.6, <0.001)		(-0.8, 0.403)	

mode (68.3s vs. 85.6s vs. 97.0s, Constant, Habitual, and Casual groups, respectively) in comparison to the TTT group (43.6s). There are statistically significant differences in all cases.

C. PROFICIENCY IMPROVEMENT

We examine players' perception and production improvement rates at the end of both competitions. In particular, we measure the activities performed during the first and last two days of the competitions. Table 9 displays a comparison of the mean values of production and discrimination success rates at the beginning and end of the competitions.

To test the interaction effect between session (first vs. last) and group (Constant, Habitual, Casual, TTT), a non-parametric mixed model was used. We employed the R-package nparLD that has been shown to be a robust method [85]. The model shows that there is an interaction effect between session and group in pronunciation ($p < 0.001$) and discrimination ($p < 0.001$). There is an improvement in all groups (Diff row of Table 9), in both production and discrimination, but the improvement depends on the group. A Wilcoxon signed-rank test shows that these differences are statistically significant only for the three COP groups in production ($p < 0.001$, $p = 0.001$, and $p = 0.004$ for Constant, Habitual, and Casual groups, respectively), and in discrimination ($p < 0.001$, $p < 0.001$, and $p = 0.003$ for Constant, Habitual, and Casual groups, respectively).

Inter-group differences increase from the most active users to the least active ones. In particular, the most active players in the COP competition (Constant group) improved the most in both activity types: 6.7% vs. 2.4% (Habitual) and 2.3% (Casual) for production, and 10.1% vs. 4.4% (Habitual) and 1.8% (Casual) for discrimination. These differences are statistically significant (production: $H(2) = 15.149$, $p = 0.001$; discrimination: $H(2) = 47.303$, $p < 0.001$; Kruskal–Wallis test). In particular, pairwise comparisons with a Mann–Whitney U test show that the Constant group production success difference is higher than the Habitual group ($U = 1014$, $p = 0.002$) and Casual group ($U = 918.0$, $p = 0.001$), being statistically significant. In discrimination, there are also differences in the same cases: between Constant and Habitual ($U = 740.0$, $p < 0.001$); and between Constant and Casual ($U = 380.0$, $p < 0.001$). Finally, the improvement value reached by the Constant players, which is higher than that of the TTT group, is also statistically significant in both pronunciation ($U = 325.0$, $p = 0.045$) and discrimination activities ($U = 403.0$, $p = 0.005$).

VI. DISCUSSION

High-performance rates of success were achieved by the participants of the competition by carrying out a significantly large number of perception and production activities with feedback (Table 6). These values are really high compared to the time of individual attention that a teacher can provide

TABLE 10. Pearson correlation between variables of different domains.

Domain	Player skills	Proficiency improvement	Performance	Game intensity and motivation	Variable used
Player skills	1	-0.134	-0.484**	0.506**	Production First (%) as in Table 9
Proficiency improvement	-0.134	1	-0.287*	0.160*	Production Diff (%) as in Table 9
Performance	-0.484**	-0.287**	1	-0.639**	Leaderboard position as in Table 6
Game intensity and motivation	0.506**	0.160*	-0.639**	1	Number of active days as in Figure 5

**. Correlation is significant at the 0.01 level (2-tailed). *. Correlation is significant at the 0.05 level (2-tailed).

to each student in a traditional class. In particular, the most active and motivated participants in terms of game intensity (the Constant group) achieved the highest success rates in both production and discrimination activities in the Playing mode, with statistically significant differences with the Habitual and Casual groups (Table 7 and Table 8). Regarding performance in the Training mode, our results showed that, in both types of activity (production and discrimination), the Constant and Habitual groups were better than the Casual one, but we did not find any difference between the Constant and Habitual groups.

In comparison with the previous non-social experience (TTT), COP players had statistically significant higher rates of success than the TTT ones in production activities, in both the Playing and Training modes (Table 8). Besides, COP players carried out significantly more production and discrimination activities than in TTT, both in the Playing and Training modes (Table 2). These results support the statement reported in Sepehr and Head [8], which declares the more competitive the game, the better the player's performance due to the positive effect that winning the challenge has on self-concept and in the sense of competence. In this regard, it seems that pursuing a goal, such as beating other players in a competition, improves performance and proficiency in the case of highly motivated users (Constant users), since players are more focused and put more effort into the activity [15].

Although the ambiguous results on user learning of introducing a competitive element in learning games such as in Chen *et al.* [17], in which students in their non-competition condition performed significantly better on the learning achievement test than those in the competition condition, and in Vrugte *et al.* [18], in which the learning outcomes comparison of a collaborative condition and a competitive one did not show any difference between these two independent conditions; our results showed that only COP players (competition condition) had a statistically significant proficiency improvement in both production and discrimination activities (Table 9). When we compared the amount of improvement between the two games, our results showed that COP participants outperformed TTT ones, since statistically significant differences were found in production and discrimination activities. In the case of the learning improvement among the three COP groups (Table 9), our results showed that, in both types of activities, the Constant group had a significantly higher improvement than the others. This supports the idea of the most active players achieving better learning outcomes.

Table 10 has been included to analyze the impact of player skills on learning and enjoyment. Although game intensity and motivation seem to correlate with proficiency improvement and performance more than players' skills (last row values in the table are higher than first row values), players' skills seem to have an impact on users' behavior, resulting in a high correlation value between players' skills and motivation variables ($r = 0.506$ between the number of active days and pronunciation success rate at the beginning of the competition). This could be justified by the fact that the stimulus of the prize is more powerful for the most skilled players; as they are the ones who, a priori, have more options to win the competition. The correlation between players' skills and improvement is negative because the more skilled the user is, the less the margin for improvement she/he has. Correlations with performance, are negative because the better the performance the lower the Leaderboard position.

The social competition designed turned out to be a key positive motivational factor. Results seem to indicate that the competition had a more powerful motivational effect on the most active students for training difficult activities and sounds in order to achieve a great performance in the Playing mode. On the other hand, the competitive configuration of the game could have decreased motivation in users without options to obtain the final prize (Casual and Habitual players). Furthermore, We found a more constant student's activity in the challenging competition condition (COP) since, in TTT, the students' activity was quite irregular, with sudden high peaks some days and low activity in the rest of the days (Fig. 5). Therefore, our first results comparing levels and time (days) of game activity agree with other studies, such as Sepehr and Head [8] and Cagiltay *et al.* [15], which found a positive effect of competition on player motivation and learning. The answers to the final questionnaire support this idea since, for most COP players, the two main declared reasons for playing were improving their English pronunciation and climbing up the leaderboard (Table 3). Intergroup differences marked in question Q2.7 (Table 4) reveal that having a negative attitude toward competition (being uncomfortable with comparing with others) could have been a key factor in decreasing motivation because it was marked by almost 53.1% of Casual users and 43.6% of Habitual users. Regarding question Q2.5 (Table 4), 28.6% of Constant users declared being uncomfortable with challenges, probably due to the high workload required to be in the highest position of the leaderboard. Besides, less than 7%

declared anxiety, discomfort and bad feelings about proving that their pronunciation is good in challenges (Table 4). In that sense, we believe that the introduction of controlled elements of competitiveness, such as winning rules, comparability among participants, and interaction with other participants, among others [82], is a motivational trigger for some students (Constant players), since it requires students to improve their skills to win the challenges [16]. As Sepehr and Head [8] reported, competitiveness can contribute to experience flow and increasing one's sense of competence when the challenges are overcome. Unlike those previous studies which stated that competition and gamification elements, such as points and leaderboards, could diminish the intrinsic motivation and engagement [12], [86], our results showed a positive effect on student motivation, in agreement with other previous studies that considered competition as a motivational trigger stimulating motivation, engagement, and persistence [13], [14].

The high early abandonment rate (182 of 347 players did not finish data collection) is in line with the current tendency in online learning courses (some universities report drop-out rates as high as 80% [87]); video games (Robinson [88] reports that only 40% of players return after the first game session) and learning games (Virvou and Katsionis [89] points out that learning games are at a disadvantage with respect to pure entertainment games). In our case, 22 players abandoned after a one day session (12.1% of the total of 182 dropouts). We tried to clarify the abandonment reasons by asking users about it in the questionnaire reported in Table 5: apart from technical reasons (Q3.1), lack of time (Q3.2) and likeability (Q3.3 and Q3.4) are the main reasons behind the abandonment. We understand lack of time because of the high game intensity required from Constant users. Likeability is also understandable as not all users feel equally comfortable when competing. User skills could have been another reason for abandonment, as not all the players are equally competent for competing. Nevertheless, we observed that a high percentage of users that dropped out were as skilled as the best ones: the 30 most active players in the Constant group (56 players) had an initial production success rate higher than 51%, with a maximum of 87% and mean of 70%; while in the dropout group (182 players), there were 45 users in this interval (> 51%), 35 over the Constant mean value (70%) and 10 players over 87%. Further research should be done to offer sufficient incentives to the different types of user so as to stimulate all of them in training in order to improve their competences.

A. LIMITATIONS AND FUTURE WORK

More research is necessary to explore, in greater depth, the role of different personal variables, such as dispositional competitive personality traits, skill levels in the activities and previous levels of motivation for learning, among others. It could be, as stated in some studies [18], that competition has positive effects only in highly skilled students and, on the contrary, it could have a negative effect on the less skilled or insecure ones. We also need more research to explore the different levels and conditions of the competition with respect to the type of learning activities.

Other limitations related to the two competitions of this study concern the different sample sizes and time conditions. In particular, the great quantity of participants in the COP competition allowed us to analyze their results by dividing them into different groups. Also, even though both competitions were not carried out at the same time, they followed the same day-protocol schedule, and there were differences in the initial success rates of both competitions. However, more research is needed to analyze the impact on learning outcomes of other possible factors, such as students' practice behavior or their educational level [90].

Although the competitions reported in this article have been tested only for an L1–Spanish/L2–English environment and for a limited number of phonemes, we are confident that the methodology based on minimal pairs is extensible to most of the other English sounds, and it might be helpful for teaching some other languages as a useful resource for foreign pronunciation training. We are currently designing similar CAPT versions for minority languages. While, in this work, we have integrated the speech technology of a particular operating system, the TTS and ASR systems of other commercial off-the-shelf services offer similar resources which also predict similar effective results.

VII. CONCLUSION

In this study, we have described and analyzed a novel learning game for pronunciation training in which players can challenge each other. The mobile game application turned out to be a useful resource for English pronunciation training. It relied on speech technologies (ASR and TTS) which have proved to be particularly useful for increasing the amount of game intensity, immediate feedback, and model pronunciations available to the students. It was also based on a specific cycle of pronunciation activities following the minimal pairs paradigm.

Native Spanish speakers played the game in a competition for English as foreign language pronunciation training, where a performance and motivation analysis was done to examine the effects of challenging.

We have studied an important issue, with the current drive by educators to discover new ways to motivate students to encourage effective uptake. Despite the fact that collaborative and competitive strategies in second language learning continue to invite discussion and disagreement on which one should be included as an effective element of motivation and performance for students, we have observed that the explicitly competitive structure of the game resulted in more positive effects on student performance and motivation than a previous version of the game, in terms of a higher number of activities (game intensity) and greater playing regularity. The same pattern was detected for pronunciation improvement, in which the most active players achieved better pronunciation results. The most active players also trained

more than the rest, performing the most difficult activities, despite being more focused on the competition than the others.

ACKNOWLEDGMENT

The authors would like to thank the University of Valladolid and its Language Learning Center for providing access to invite students to participate in this study. We also would like to thank Enrique Cámara-Arenas, Valle Flores-Lucas and Mario Corrales-Astorgano for their contributions to the work presented in this manuscript.

REFERENCES

[1] A. Dina and S.-I. Ciornei, "The advantages and disadvantages of computer assisted language learning and teaching for foreign languages," *Procedia-Social Behav. Sci.*, vol. 76, pp. 248–252, Apr. 2013, doi: 10.1016/j.sbspro.2013.04.107.

[2] M. G. O'Brien, T. M. Derwing, C. Cucchiarini, D. M. Hardison, H. Mixdorff, R. I. Thomson, H. Strik, J. M. Levis, M. J. Munro, J. A. Foote, and G. M. Levis, "Directions for the future of technology in pronunciation research and teaching," *J. Second Lang. Pronunciation*, vol. 4, no. 2, pp. 182–207, 2018, doi: 10.1075/jslp.17001.obr.

[3] R. I. Thomson and T. M. Derwing, "The effectiveness of L2 pronunciation instruction: A narrative review," *Appl. Linguistics*, vol. 36, no. 3, pp. 326–344, Jul. 2015, doi: 10.1093/applin/amu076.

[4] T. Bongaerts, "Ultimate attainment in L2 pronunciation: The case of very advanced late L2 learners," in *Second Language Acquisition and the Critical Period Hypothesis*, D. Birdsong, Ed. Mahwah, NJ, USA: Lawrence Erlbaum, 1999, pp. 153–159.

[5] A. Neri, C. Cucchiarini, and H. Strik, "The effectiveness of computer-based speech corrective feedback for improving segmental quality in L2 Dutch," *ReCALL*, vol. 20, no. 2, pp. 225–243, May 2008, doi: 10.1017/S0958344008000724.

[6] C. Nagle, "Motivation, comprehensibility, and accentedness in L2 Spanish: Investigating motivation as a time-varying predictor of pronunciation development," *Mod. Lang. J.*, vol. 102, no. 1, pp. 199–217, Mar. 2018, doi: 10.1111/modl.12461.

[7] A. Mitchell and C. Savill-Smith, *The Use of Computer and Video Games for Learning: A Review of the Literature*. London, U.K.: Learning and Skills Development Agency, 2004.

[8] S. Sepehr and M. Head, "Understanding the role of competition in video gameplay satisfaction," *Inf. Manage.*, vol. 55, no. 4, pp. 407–421, Jun. 2018, doi: 10.1016/j.im.2017.09.007.

[9] M. Prensky, *Digital Game-Based Learning*. New York, NY, USA: McGraw-Hill, 2004, doi: 10.1145/950566.950596.

[10] K. D. Squire, "Video games and education: Designing learning systems for an interactive age," *Educ. Technol.*, vol. 48, no. 2, p. 17, 2008.

[11] D. W. Shaffer, K. R. Squire, R. Halverson, and J. P. Gee, "Video games and the future of learning," *Phi Delta Kappan*, vol. 87, no. 2, pp. 105–111, Oct. 2005, doi: 10.1177/003172170508700205.

[12] J. Koivisto and J. Hamari, "Demographic differences in perceived benefits from gamification," *Comput. Hum. Behav.*, vol. 35, pp. 179–188, Jun. 2014, doi: 10.1016/j.chb.2014.03.007.

[13] H. N. H. Cheng, W. M. C. Wu, C. C. Y. Liao, and T.-W. Chan, "Equal opportunity tactic: Redesigning and applying competition games in classrooms," *Comput. Educ.*, vol. 53, no. 3, pp. 866–876, Nov. 2009, doi: 10.1016/j.compedu.2009.05.006.

[14] S. Deterding, "Eudaimonic design, or: Six invitations to rethink gamification," in *Rethinking Gamification*, M. Fuchs, S. Fizek, P. Ruffino, and N. Schrape, Eds. Lüneburg, Germany: Meson Press, Jun. 2014, pp. 305–331, doi: 10.25969/mediarep/727.

[15] N. E. Cagiltay, E. Ozcelik, and N. S. Ozcelik, "The effect of competition on learning in games," *Comput. Educ.*, vol. 87, pp. 35–41, Sep. 2015, doi: 10.1016/j.compedu.2015.04.001.

[16] S. Sampayo-Vargas, C. J. Cope, Z. He, and G. J. Byrne, "The effectiveness of adaptive difficulty adjustments on students' motivation and learning in an educational computer game," *Comput. Educ.*, vol. 69, pp. 452–462, Nov. 2013, doi: 10.1016/j.compedu.2013.07.004.

[17] C.-H. Chen, J.-H. Liu, and W.-C. Shou, "How competition in a game-based science learning environment influences students' learning achievement, flow experience, and learning behavioral patterns," *J. Educ. Technol. Soc.*, vol. 21, no. 2, pp. 164–176, 2018.

[18] J. ter Vrugte, T. de Jong, S. Vandercruysse, P. Wouters, H. van Oostendorp, and J. Elen, "How competition and heterogeneous collaboration interact in prevocational game-based mathematics education," *Comput. Educ.*, vol. 89, pp. 42–52, Nov. 2015, doi: 10.1016/j.compedu.2015.08.010.

[19] S. Y. Chen and Y.-M. Chang, "The impacts of real competition and virtual competition in digital game-based learning," *Comput. Hum. Behav.*, vol. 104, Mar. 2020, Art. no. 106171, doi: 10.1016/j.chb.2019.106171.

[20] R. L. Trask, *A Dictionary of Phonetics and Phonology*. Evanston, IL, USA: Routledge, 2004.

[21] A. N. Meltzoff, P. K. Kuhl, J. Movellan, and T. J. Sejnowski, "Foundations for a new science of learning," *Science*, vol. 325, no. 5938, pp. 284–288, Jul. 2009.

[22] E. Cámara-Arenas, *Native Cardinality: On Teaching American English Vowels to Spanish Students* (Historia y Sociedad). Valladolid, Spain: Ediciones Universidad de Valladolid, 2013.

[23] C. Tejedor-García, V. Cardeñoso-Payo, E. Cámara-Arenas, C. González-Ferreras, and D. Escudero-Mancebo, "Measuring pronunciation improvement in users of CAPT tool TipTopTalk!" in *Proc. Interspeech*, San Francisco, CA, USA, Sep. 2016, pp. 1178–1179.

[24] C. Tejedor-García, D. Escudero-Mancebo, C. González-Ferreras, E. Cámara-Arenas, and V. Cardeñoso-Payo, "Improving L2 production with a gamified computer-assisted pronunciation training tool, TipTopTalk!" in *Proc. IberSPEECH*, Lisbon, Portugal, Nov. 2016, pp. 177–186.

[25] C. Tejedor-García, D. Escudero-Mancebo, C. González-Ferreras, E. Cámara-Arenas, and V. Cardeñoso-Payo, "Evaluating the efficiency of synthetic voice for providing corrective feedback in a pronunciation training tool based on minimal pairs," in *Proc. SLaTE*, Stockholm, Sweden, Aug. 2017, pp. 26–30, doi: 10.21437/SLaTE.2017-5.

[26] C. Tejedor-García, D. Escudero-Mancebo, E. Camára-Arenas, C. González-Ferreras, and V. Cardeñoso-Payo, "Assessing pronunciation improvement in students of English using a controlled computer-assisted pronunciation tool," *IEEE Trans. Learn. Technol.*, early access, Mar. 12, 2020, doi: 10.1109/TLT.2020.2980261.

[27] L. M. Tomokiyo, L. Wang, and M. Eskenazi, "An empirical study of the effectiveness of speech-recognition-based pronunciation training," in *Proc. 6th ICSLP*, Beijing, China, Oct. 2000, pp. 677–680.

[28] R. Akahane-Yamada, E. McDermott, T. Adachi, H. Kawahara, and J. S. Pruitt, "Computer-based second language production training by using spectrographic representation and HMM-based speech recognition scores," in *Proc. 5th ICSLP*, Sidney, BC, Australia, Jun. 1998, pp. 1–4.

[29] A. Weinberg and H. Knoerr, "Learning French pronunciation: Audiocassettes or multimedia?" *Calico J.*, vol. 20, no. 2, pp. 215–336, Jan. 2013, doi: 10.1558/cj.v20i2.215-336.

[30] B. Mak, M. Siu, M. Ng, Y.-C. Tam, Y.-C. Chan, K.-W. Chan, K.-Y. Leung, S. Ho, F.-H. Chong, J. Wong, and J. Lo, "PLASER: Pronunciation learning via automatic speech recognition," in *Proc. HLT-NAACL Conf.*, Edmonton, AB, Canada, May/Jun. 2003, pp. 23–29, doi: 10.3115/1118894.1118898.

[31] L. Chen, Q. Gao, Q. Liang, J. Yuan, and Y. Liu, "Automatic scoring minimal-pair pronunciation drills by using recognition likelihood scores and phonological features," in *Proc. SLaTE*, Graz, Austria, Sep. 2019, pp. 25–29, doi: 10.21437/SLaTE.2019-6.

[32] J.-Y. Lee, "The effects of pronunciation instruction using duration manipulation on the acquisition of English vowel sounds by pre-service Korean EFL teachers," Ph.D. dissertation, Univ. Kansas, Lawrence, KS, USA, 2009.

[33] X. Wang, "Training Mandarin and Cantonese speakers to identify English vowel contrasts: Long-term retention and effects on production," Ph.D. dissertation, Simon Fraser Univ., Burnaby, BC, Canada, 2002.

[34] E. Pyshkin, J. Blake, A. Lamtev, I. Lezhenin, A. Zhuikov, and N. Bogach, "Prosody training mobile application: Early design assessment and lessons learned," in *Proc. 10th IEEE Int. Conf. Intell. Data Acquisition Adv. Comput. Systems: Technol. Appl. (IDAACS)*, Metz, France, vol. 2, Sep. 2019, pp. 735–740.

[35] E. Boitsova, E. Pyshkin, T. Yasuta, N. Bogach, I. Lezhenin, A. Lamtev, and V. Diachkov, "StudyIntonation courseware kit for EFL prosody teaching," in *Proc. 9th Int. Conf. Speech Prosody*, Poznań, Poland, Jun. 2018, pp. 413–417.

[36] D. M. Chun, Y. Jiang, and N. Ávila, "Visualization of tone for learning Mandarin Chinese," in *Proc. 4th PSLLT Conf.*, Vancouver, BC, Canada, Aug. 2012, pp. 77–89.

[37] D. M. Hardison, "Generalization of computer assisted prosody training: Quantitative and qualitative findings," *Lang. Learn. Technol.*, vol. 8, no. 1, pp. 34–52, 2004.

[38] D. M. Hardison, "Contextualized computer-based L2 prosody training: Evaluating the effects of discourse context and video input," *Calico J.*, vol. 22, no. 2, pp. 175–190, Jan. 2013, doi: 10.1558/cj.v22i2.175-190.

[39] Y. Hirata, "Computer assisted pronunciation training for native english speakers learning Japanese pitch and durational contrasts," *Comput. Assist. Lang. Learn.*, vol. 17, nos. 3–4, pp. 357–376, Jul. 2004, doi: 10.1080/0958822042000319629.

[40] R. Hincks and J. Edlund, "Promoting increased pitch variation in oral presentations with transient visual feedback," *Lang. Learn. Technol.*, vol. 13, no. 3, pp. 32–50, Oct. 2009, doi: 10125/44190.

[41] M. W. Tanner and M. M. Landon, "The effects of computer-assisted pronunciation readings on ESL learners' use of pausing, stress, intonation, and overall comprehensibility," *Lang. Learn. Technol.*, vol. 13, no. 3, pp. 51–65, Oct. 2009, doi: 10125/44191.

[42] Y. Zhang, W. Chan, and N. Jaitly, "Very deep convolutional networks for end-to-end speech recognition," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Mar. 2017, pp. 4845–4849, doi: 10.1109/ICASSP.2017.7953077.

[43] M. Meeker, "Internet trends 2017," Kleiner Perkins, Los Angeles, CA, USA, Tech. Rep., May 2017. [Online]. Available: https://www.kleinerperkins.com/files/INTERNET_TRENDS_REPORT_2017.pdf

[44] R. Hincks, "Speech technologies for pronunciation feedback and evaluation," *ReCALL*, vol. 15, no. 1, pp. 3–20, May 2003, doi: 10.1017/S0958344003000211.

[45] R. Hincks, "Computer support for learners of spoken English," Ph.D. dissertation, KTH Royal Inst. Technol., Stockholm, Sweden, 2005.

[46] A. Neri, C. Cucchiarini, and H. Strik, "Pronunciation training in Dutch as a second language on the basis of automatic speech recognition," *Stem-, Spraak-Taalpathologie*, vol. 15, no. 2, pp. 1–10, Oct. 2007.

[47] D. Liakin, W. Cardoso, and N. Liakina, "Learning L2 pronunciation with a mobile speech recognizer: French /y/," *Calico J.*, vol. 32, no. 1, pp. 1–25, Jun. 2014, doi: 10.1558/cj.v32i1.25962.

[48] A. Neri, O. Mich, M. Gerosa, and D. Giuliani, "The effectiveness of computer assisted pronunciation training for foreign language learning by children," *Comput. Assist. Lang. Learn.*, vol. 21, no. 5, pp. 393–408, Dec. 2008, doi: 10.1080/09588220802447651.

[49] A. R. Bradlow, D. B. Pisoni, R. Akahane-Yamada, and Y. Tohkura, "Training japanese listeners to identify English /r/ and /l/: IV. Some effects of perceptual learning on speech production," *J. Acoust. Soc. Amer.*, vol. 101, no. 4, pp. 2299–2310, Apr. 1997.

[50] Y. Shinohara and P. Iverson, "High variability identification and discrimination training for japanese speakers learning english /r/–/l/," *J. Phonetics*, vol. 66, pp. 242–251, Jan. 2018, doi: 10.1016/j.wocn.2017.11.002.

[51] N. C. Guilloteau, "Modification of phonetic categories in French as a second language: Experimental studies with conventional and computer-based intervention methods," Ph.D. dissertation, Univ. Texas Press, Austin, TX, USA, 1997.

[52] E. M. Kissling, "Teaching pronunciation: Is explicit phonetics instruction beneficial for FL learners?" *Mod. Lang. J.*, vol. 97, no. 3, pp. 720–744, Sep. 2013, doi: 10.1111/j.1540-4781.2013.12029.x.

[53] G. Lord, "(How) can we teach foreign language pronunciation? On the effects of a Spanish phonetics course," *Hispania*, vol. 88, no. 3, p. 557, Sep. 2005, doi: 10.2307/20063159.

[54] P. Pearson, L. Pickering, and R. Da Silva, "The impact of computer assisted pronunciation training on the improvement of Vietnamese learner production of English syllable margins," in *Proc. 2nd Pronunciation 2nd Lang. Learn. Teaching Conf.* Ames, IA, USA: Iowa State Univ. Press, Oct. 2011, pp. 80–169.

[55] D. Liakin, W. Cardoso, and N. Liakina, "The pedagogical use of mobile speech synthesis (TTS): Focus on French liaison," *Comput. Assist. Lang. Learn.*, vol. 30, nos. 3–4, pp. 325–342, May 2017, doi: 10.1080/09588221.2017.1312463.

[56] J. Shen, R. Pang, R. J. Weiss, M. Schuster, N. Jaitly, Z. Yang, Z. Chen, Y. Zhang, Y. Wang, R. Skerrv-Ryan, R. A. Saurous, Y. Agiomvrgiannakis, and Y. Wu, "Natural TTS synthesis by conditioning WaveNet on MEL spectrogram predictions," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Apr. 2018, pp. 4779–4783.

[57] T. Bione, J. Grimshaw, and W. Cardoso, "An evaluation of text-to-speech synthesizers in the foreign language classroom: Learners' perceptions," in *Call Communities and Culture—Short Papers From Eurocall*. Kato Pyrgos, Cyprus: St. Raphael Resort/Limassol, Aug. 2016, pp. 50–54, doi: 10.14705/rpnet.2016.eurocall2016.537.

[58] Z. Handley, "Is text-to-speech synthesis ready for use in computer-assisted language learning?" *Speech Commun.*, vol. 51, no. 10, pp. 906–919, Oct. 2009, doi: 10.1016/j.specom.2008.12.004.

[59] G. Smith, W. Cardoso, and C. G. Fuentes, "Evaluating text-to-speech synthesizers," in *Proc. Meeting English Lang. Teach.* Montréal, QC, Canada: Concordia Univ., Oct. 2015, pp. 108–113. [Online]. Available: https://research-publishing.net/manuscript?10.14705/rpnet.2015.000318, doi: 10.14705/rpnet.2015.000318.

[60] M. Celce-Murcia and J. M. Goodwin, *Teaching Pronunciation*. London, U.K.: Thomson Learning, 2014.

[61] 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)*, New York, NY, USA, 2011, pp. 2425–2428, doi: 10.1145/1979742.1979575.

[62] K. M. Kapp, *The Gamification of Learning and Instruction: Game-Based Methods and Strategies for Training and Education*. Hoboken, NJ, USA: Wiley, 2012.

[63] D. Codish and G. Ravid, "Academic course gamification: The art of perceived playfulness," *Interdisciplinary J. E-Learn. Learn. Objects*, vol. 10, pp. 131–152, Jan. 2014, doi: 10.28945/2066.

[64] E. Colferai and S. Gregory, "Minimizing attrition in online degree courses," *J. Educators Online*, vol. 12, no. 1, p. n1, 2015.

[65] M. G. Gomez-Zermeno and L. A. De la Garza, "Research analysis on MOOC course dropout and retention rates," *Turkish Online J. Distance Educ.*, vol. 17, no. 2, pp. 3–14, Apr. 2016, doi: 10.17718/tojde.23429.

[66] D. Huynh and H. Iida, "An analysis of winning Streak's effects in language course of 'Duolingo,'" *Asia–Pacific J. Inf. Technol. Multimedia*, vol. 6, no. 2, pp. 23–29, 2017.

[67] I. McGraw, B. Yoshimoto, and S. Seneff, "Speech-enabled card games for incidental vocabulary acquisition in a foreign language," *Speech Commun.*, vol. 51, no. 10, pp. 1006–1023, Oct. 2009, doi: 10.1016/j.specom.2009.04.011.

[68] H. Strik, J. Colpaert, J. van Doremalen, and C. Cucchiarini, "The DISCO ASR-based CALL system: Practicing L2 oral skills and beyond," in *Proc. 8th Int. Conf. Lang. Resour. Eval. (LREC)*, Istanbul, Turkey, May 2012, pp. 2702–2707.

[69] E. Danowska-Florczyk and P. Mostowski, "Gamification as a new direction in teaching Polish as a foreign language," in *Proc. 5th Int. Conf. ICT Lang. Learn.*, Florence, Italy, Nov. 2012, pp. 1–4.

[70] D. Murad, R. Wang, D. Turnbull, and Y. Wang, "SLIONS: A karaoke application to enhance foreign language learning," in *Proc. ACM Multimedia Conf. Multimedia Conf. (MM)*, Seoul, South Korea, Oct. 2018, pp. 1679–1687, doi: 10.1145/3240508.3240691.

[71] P.-H. Su, C.-H. Wu, and L.-S. Lee, "A recursive dialogue game for personalized computer-aided pronunciation training," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 23, no. 1, pp. 127–141, Jan. 2015, doi: 10.1109/TASLP.2014.2375572.

[72] M. Sailer, J. U. Hense, S. K. Mayr, and H. Mandl, "How gamification motivates: An experimental study of the effects of specific game design elements on psychological need satisfaction," *Comput. Hum. Behav.*, vol. 69, pp. 371–380, Apr. 2017, doi: 10.1016/j.chb.2016.12.033.

[73] S. Schwab and J.-P. Goldman, "MIAPARLE: Online training for discrimination and production of stress contrasts," in *Proc. 9th Int. Conf. Speech Prosody*, Poznan, Poland, Jun. 2018, pp. 572–576, doi: 10.21437/SpeechProsody.2018-116.

[74] Y. Zhang, H. Ding, P. Zelchenko, X. Cui, Y. Lin, Y. Zhan, and H. Zhang, "Prosodic disambiguation by Chinese EFL learners in a cooperative game task," in *Proc. 9th Int. Conf. Speech Prosody*, Poznań, Poland, Jun. 2018, pp. 979–983, doi: 10.21437/SpeechProsody.2018-198.

[75] R. Karhila, S. Ylinen, S. Enarvi, K. J. Palomäki, A. Nikulin, and O. Rantula, "SIAK—A game for foreign language pronunciation learning," in *Proc. Interspeech*, Stockholm, Sweden, Aug. 2017, pp. 3429–3430.

[76] R. Karhila, A.-R. Smolander, S. Ylinen, and M. Kurimo, "Transparent pronunciation scoring using articulatorily weighted phoneme edit distance," in *Proc. Interspeech*, Graz, Austria, Sep. 2019, pp. 1866–1870, doi: 10.21437/Interspeech.2019-1785.

[77] A. Berns, A. Gonzalez-Pardo, and D. Camacho, "Game-like language learning in 3-D virtual environments," *Comput. Educ.*, vol. 60, no. 1, pp. 210–220, Jan. 2013, doi: 10.1016/j.compedu.2012.07.001.
[78] C. X. Wang, B. Calandra, S. T. Hibbard, and M. L. M. Lefaiver, "Learning effects of an experimental EFL program in second life," *Educ. Technol. Res. Develop.*, vol. 60, no. 5, pp. 943–961, Oct. 2012, doi: 10.1007/s11423-012-9259-0.
[79] P. Van Hentenryck and C. Coffrin, "Teaching creative problem solving in a MOOC," in *Proc. 45th ACM Tech. Symp. Comput. Sci. Educ. (SIGCSE)*, New York, NY, USA, Mar. 2014, pp. 677–682, doi: 10.1145/2538862.2538913.
[80] A. Baker and S. Goldstein, *Pronunciation Pairs Teacher's Book*. Cambridge, U.K.: Cambridge Univ. Press, 2008.
[81] A. Baker, *Ship Or Sheep? Student's Book: An Intermediate Pronunciation Course*, vol. 1. Stuttgart, Germany: Ernst Klett Sprachen, 2006.
[82] D. W. Johnson and R. T. Johnson, "The impact of cooperative, competitive, and individualistic learning environments on academic achievement," in *International Guide to Student Achievement*, J. Hattie and E. M. Anderman, Eds. Evanston, IL, USA: Routledge, 2013, pp. 372–374.
[83] J. Brooke, "SUS—A quick and dirty usability scale," *Usability Eval. Ind.*, vol. 189, no. 194, pp. 4–7, Jun. 1996.
[84] *IBM SPSS Statistics for Windows*, Version 25.0, IBM Corp., New York, NY, USA, 2017.
[85] K. Noguchi, Y. R. Gel, E. Brunner, and F. Konietschke, "nparLD: An R software package for the nonparametric analysis of longitudinal data in factorial experiments," *J. Stat. Soft.*, vol. 50, no. 12, pp. 1–23, Sep. 2012, doi: 10.18637/jss.v050.i12.
[86] R. Van Eck and J. Dempsey, "The effect of competition and contextualized advisement on the transfer of mathematics skills a computer-based instructional simulation game," *Educ. Technol. Res. Develop.*, vol. 50, no. 3, pp. 23–41, Sep. 2002, doi: 10.1007/BF02505023.
[87] T. W. Atchley, G. Wingenbach, and C. Akers, "Comparison of course completion and student performance through online and traditional courses," *Int. Rev. Res. Open Distrib. Learn.*, vol. 14, no. 4, pp. 1–14, Sep. 2013, doi: 10.19173/irrodl.v14i4.1461.
[88] M. Robinson. (2014). Analytics Drive Design. GamesAnalytics, Edinburgh, U.K. Accessed: Mar. 2, 2020. [Online]. Available: https://www.gdcvault.com/play/1019527/Analytics-Driven
[89] M. Virvou and G. Katsionis, "On the usability and likeability of virtual reality games for education: The case of VR-ENGAGE," *Comput. Educ.*, vol. 50, no. 1, pp. 154–178, Jan. 2008, doi: 10.1016/j.compedu.2006.04.004.
[90] B. W. F. P. de Vries, C. Cucchiarini, H. Strik, and R. van Hout, "Spoken grammar practice in CALL: The effect of corrective feedback and education level in adult L2 learning," *Lang. Teaching Res.*, vol. 23, pp. 1–22, Jan. 2019, doi: 10.1177/1362168818819027.

CRISTIAN TEJEDOR-GARCÍA received the B.Sc. and M.Sc. degrees (Hons.) in computer science from the University of Valladolid, Valladolid, Spain, in 2014 and 2016, respectively, where he is currently pursuing the Ph.D. degree in computer science. He is currently an Associated Member of the ECA-SIMM Research Group, University of Valladolid, with a Predoctoral Grant. He has published and achieved international awards. His research interests include speech technology, learning games, and human–computer interaction.

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 computational prosody (modeling of prosody for TTS systems and labeling of corpora).

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

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.