Assessing Pronunciation Improvement in Students of English Using a Controlled Computer-Assisted Pronunciation Tool

Cristian Tejedor-García, David Escudero-Mancebo, Enrique Cámara-Arenas, César González-Ferreras, and Valentín Cardeñoso-Payo, *Member, IEEE*

***Abstract*—Over the last few years, we have witnessed a growing interest in computer-assisted pronunciation training (CAPT) tools and the commercial success of foreign language teaching applications that incorporate speech synthesis and automatic speech recognition technologies. However, empirical evidence supporting the pedagogical effectiveness of these systems remains scarce. In this study, a minimal-pair based CAPT tool that implements exposure–perception–production cycles and provides automatic feedback to learners is tested for effectiveness in training adult native Spanish users (English level B1–B2) in the production of a set of difficult English sounds. Working under controlled conditions, a group of users took a pronunciation test before and after using the tool. Test results were considered against those of an in-classroom group who followed similar training within the traditional classroom setting. Results show a significant pronunciation improvement among the learners who used the CAPT tool, as well as a correlation between human rater's assessment of post-tests and automatic CAPT assessment of users.**

***Index Terms*—Automatic assessment tools, computer-assisted pronunciation training (CAPT), learning environments, automatic speech recognition, speech synthesis.**

I. Introduction

Over the last decade, the assumption that information and communication technology is at a stage where it can contribute to the teaching of second language (L2) pronunciation and to the automatic diagnosis of L2 goodness of pronunciation (GOP) has fueled much debate and has led to a number of interesting practical proposals [1], [2], [3]. This assumption is partly based on the quality attained by current automatic speech recognition (ASR) and text-to-speech (TTS) systems. Google's machine-learning voice recognition, for example, has achieved a word accuracy rate of 95% for the English language, therefore reaching the threshold of human accuracy [4]. On the other hand, research on the quality of TTS has led Google to assert that deep neural network (DNN) technology already produces near-human speech for some speaking situations [5]. Furthermore, these technologies are offered out-of-the-box as part of the Android operating system, and the open application programming interface (API) allows easy development of multi-platform L2 teaching tools incorporating ASR and TTS.

Nevertheless, while computer-assisted learning of L2 grammar and vocabulary have been thoroughly measured and studied [6], [7], [8], [9], there have been relatively few attempts to empirically measure the extent to which these, or any other available speech technologies, may assist in the teaching and diagnosis of L2 pronunciation. There is virtually no reflection on how such systems could actually be integrated with pronunciation teaching protocols, from the methodological point of view.

In previous work, we reported on the design of a learning game for teaching L2 segmental pronunciation (i.e., the teaching of single speech sounds, like vowels, disregarding intonation and other suprasegmental aspects of connected speech [10]), articulated into freely selected minimal-pair tasks of exposure, discrimination, and production. The system was able to discriminate the pronunciation competence of different users: those with a certified higher level consistently obtained higher scores in the game [11]. We have also reflected on the degree of engagement generated by the tool [12], [13]. However, our findings in relation to the actual teaching efficiency of the system have been less conclusive: the introduction of corrective feedback [14], [15] allowed us to confirm that there was pronunciation improvement among users after the first few turns, while protracted use of the tool seemed to invariably lead to stagnation. An extra complication concerning the assessment of pronunciation improvement among users had to do with the freedom of movement granted to them and, therefore, with the already mentioned lack of control on the part of the system. While the more game-oriented users tended to boost their score by repeating those tasks they found easy and avoiding more challenging exercises, those who were more interested in actual learning insistently returned to the difficult sounds, even at the expense of their final scores. Though not necessarily faulty per se, these dynamics made it difficult to reach final conclusions concerning the tool's global effectiveness and efficiency. For that reason, in the present study we have decided to temporarily move away from the game paradigm, and work with a version of our computer-

Manuscript received MONTH XX, YYYY; revised MONTH XX, YYYY; accepted MONTH XX, YYYY. Date of publication MONTH XX, YYYY; date of current version MONTH XX, YYYY. (Corresponding author: C. Tejedor-García).

C. Tejedor-García, D. Escudero-Mancebo, C. González-Ferreras, and V. Cardeñoso-Payo are with the ECA-SIMM group, department of Computer Science, University of Valladolid, 47011, Valladolid, Spain. E-mail: {cristian, descuder, cesargf, valen}@infor.uva.es.

E. Cámara-Arenas is with the department of English Philology, University of Valladolid, 47011, Valladolid, Spain. E-mail: ecamara@fyl.uva.es.

For information on obtaining reprints of this article, please send e-mail to: reprints@ieee.org, and reference the Digital Object Identifier below.

Digital Object Identifier no. XX.XXXX/TLT.2020.XXXXXX

assisted pronunciation (CAPT) tool that should no longer be considered a learning game, but rather, a computer-assisted teaching tool.

Our present goal is to empirically determine whether the use of the tool allows students to improve their pronunciation at the segmental level, and to ponder the degree to which any performance improvement is comparable to that obtained through similarly designed and structured in-classroom training. Our goal has nothing to do with predicating the superiority of one particular teaching procedure, but with determining and assessing the range of improvement procured by the use of our CAPT tool.

In the present study, we set out to generate knowledge on a number of relevant issues according to three main research questions (RQ):

RQ1: Does our CAPT tool contribute to the teaching/learning of pronunciation at the segmental level?

- **Issue 1.1:** Is there a relative improvement in the students' pronunciation after using the tool?
- **Issue 1.2:** Is this improvement, if detected, sufficient/worth considering from a quantitative point of view?
- **Issue 1.3:** Are the students' difficulties (most difficult phonemes, perception vs. production) identified by the tool?

RQ2: Can off-the-shelf TTS and ASR systems be successfully and non-obstructively used by segment-focused CAPT tools?

RQ3: To what extent can methodologically sensitive design issues such as the use of minimal pair exercises within the exposure–discrimination–production cycle affect users' segmental pronunciation improvement?

II. Background and Related Work

Concern with pronunciation teaching methods has increased over the last decade. In the survey of 75 L2 pronunciation studies analyzed in Thomson and Derwing [1], sparse research on pronunciation instruction was reported until 2005 [16]. In particular, 26 empirical and replicable L2 pronunciation studies were published before 2005, and 49 more appeared between that year and the end of 2014 [1] (most of the analyzed material had been published as journal articles, Ph.D. theses and conference proceedings). Many of these studies (56) were concerned with L2-English pronunciation. However, only 29 studies featured the use of CAPT tools, and few resorted to ASR or TTS technologies. The meta-analysis of most of the studies surveyed in Thomson and Derwing [1], carried out by Lee et al. [17], evidenced pronunciation improvement following instruction. In our present study, we are particularly concerned with the methods that incorporate such systems, and the way they have been made subservient to the teaching goal: Section II-A shows the innovative value of the exposure–discrimination–production cycle in the state of the art (RQ3); and it shows why the quality of ASR and TTS constitutes a relevant concern for CAPT tools (RQ2). Also of particular interest to us, in our search for antecedents, is the way in which different CAPT systems have faced the challenge of assessing the pronunciation of users: Section II-B highlights the importance of measuring the users' improvements, legitimizing our approach (RQ1).

A. CAPT Methodologies

Over the last two decades, CAPT tools have been designed for a variety of environments, from desktop applications in multimedia laboratories [18], [19], [20], to Internet [21], [22] and mobile devices [23]. The list of designs includes such systems as Talk-to-me [24], [25], PLASER [26], or PARLING [27], among others. Some of these systems rely on available software like the Nuance Dragon Dictation software or NaturalReader TTS. The possibilities for integrating a large variety of promising software within the rich domain of pronunciation teaching methods turns CAPT into a complex and very creative venture [1].

Systems that attempt to emulate traditional classroom procedures [1] seem to be adopting a substitution agenda, or at least considering the possibility of a replacement of human-led pronunciation teaching. Other systems are, like ours, more concerned with the possibility of complementing traditional teaching with the incorporation of innovative strategies that rely on the use of speech technologies (e.g., ASR, TTS, etc.) [28], [29].

The use of computers in the teaching of pronunciation has been welcomed by those who perceive in this practice a potential for stress-free learning, learning-autonomy, adaptation to learners' needs and pace, practically unlimited input, intensive practice, immediate automated and individualized feedback, and dynamic forms of assessment [30], [31]. Risks of obstructive use are less actively noticed. Mainly due to the limitations of current automatic recognition and diagnosis, obstructive CAPT systems would accept wrong pronunciations, or reject acceptable pronunciations, or provide users with inadequate feedback [32]. Interestingly, even with their limitations, current ASR systems might, in the end, prove useful; this is so under the unchecked premise that whatever is difficult for the ASR must also be so for our students: homophones, minimal pairs, allophones occurring at word boundaries, and negative cases. If this is the case, ASR systems at their current level of development would help us anticipate pronunciation difficulties experienced by prospective L2 learners [33].

In applying computer technology to the teaching of pronunciation, designers must find their way around a vast field of methodological choices. There are systems committed to the segmental level [26], [28], [34], [35], and systems that focus on suprasegmental aspects such as stress or intonation [36], [37], [38], [39]. Designers who focus on the suprasegmental level usually take sides with comprehensibility in relation to another classic dilemma: that of defining the final goal either as native-like pronunciation or as the attainment of global intelligibility [40], [41].

Issues of intelligibility within the CAPT domain become particularly interesting when ASR technologies are used to diagnose pronunciation and provide feedback [20], [23], [24], [25]. Typically, a given pronunciation is marked as correct when it is recognized as the expected word by the integrated ASR system. Current ASR technologies may also allow for a

regulation of the system that translates into different levels of difficulty [24], [25]. It is also possible, for example, to show in real time, as a form of diagnostic feedback, the hidden Markov model (HMM) scores obtained by a user in the production of a given set of minimal pairs [27], [28]. In Akahane-Yamada et al. [28], users are also automatically provided with a numerical score from the ASR. However, there are grounds for questioning the usefulness of a feedback format where, as Hincks [24] points out, users are systematically headed towards iterative trial-and-error cycles, and nothing more, whenever a low score is attained.

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 [42]. CAPT systems are very adaptable in this sense: they may discard explicit articulatory instructions [27], they may mandatorily incorporate them [20], [28], or they may let the user decide whether they want them or not [24], [25]. When explicit descriptions are incorporated, recourse to fixed and moving images (describing the movement of articulators) are easily incorporated through available technologies [43]. Explicit information for preparation or feedback, on the other hand, need not be restricted to articulatory descriptions: tools for acoustic analysis may be integrated, providing the spectrographic description and formantic values of the model and produced sounds [28].

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 [35], [44]. Qualitative issues concerning the nature and quality of the model voice must be considered. Some have been working with recordings by native speakers [18], [19], [45], [46], [47]; some have used manipulated natural speech [34], [35]. While natural voice predominates, recent designs are introducing synthetic voice through TTS systems [48], [49]. 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 [50], [51], [52].

B. Assessment of Pronunciation Improvement

As regards the evaluation of the improvement in pronunciation when using CAPT systems, there are no common guidelines. In addition, there are barely any objective studies involving the use of ASR and TTS systems. The first measurement approach is to evaluate different activities using scores assigned by human evaluators. Different rating scales are used. The most usual is to compare the pre-test score with the post-test score. This approach varies according to whether it is segmental, suprasegmental, both, or conversational. As for the segmental level, the use of minimal pairs to evaluate the perception and production of isolated phonemes is tested by Wang [35] and Bradlow et al. [44]. There are studies that, instead of using minimal pairs, use lists of words to evaluate perception and production [18], [34], only perception [53], or only production [36], [39], [45], [46], [54], even at a suprasegmental level. Some systems analyze the words read by learners in sentences and the beginning and end of sentences: in Tanner and Landon [41], only improvement in the perception was considered; whereas in other works only production [19], [21], [37], [38], [39], [47]. Less controlled methods are also used, such as oral presentations [40] and the analysis of conversations [18], [19], [22]. For the latter method, spoken dialogue systems are also used [55].

Few studies measure the improvement after training with an ASR-based CAPT tool. In Akahane-Yamada et al. [28], the effectiveness of the ASR tool used to evaluate the minimal pair production is assessed. Improvement in perception and production of words and sentences is evaluated by Burlerson [56] and Liakin et al. [23]. The production of words in sentences is analyzed in Liakin et al. [48], Neri et al. [20], [27], and Tomokiyo et al. [43]. Finally, a mix of minimal pair activities and reading sentences is presented in Mak [26].

Studies on the application of TTS to L2 pronunciation teaching are still scarce, some of them proposing a combination of natural and synthetic speech [35], [48], and some exploring the effectiveness of using TTS in self-study [57].

An alternative approach to measuring pronunciation improvement uses objective measures and optionally correlates them with the scores of human evaluators. These objective measures can be those that indicate if a pronunciation is acceptable, measures of the quality of the pronunciation, such as GOP or ASR confidence scores. In Moustroufas and Digalakis [58], it is claimed that pronunciation scores (Gaussian mixture models log-likelihood and HMM confidence score) based on both L1 and L2 language characteristics of learners have a better correlation with human scores than those based only on characteristics of the L2 language. In Neri et al. [20] and Witt and Young [59], a phone-level comparison (likelihood-based GOP) is done to assess the pronunciation mistakes by comparing non-native speech to native speech. The objective scores of PhonePass, an HMM-based ASR software, were used by Hincks [24]. Subsequently, they were correlated to human rater scores [25]. Different ASR system outputs were used for the assessment of young children's basic English vocabulary [60], [61]. They based their work on phoneme-level language modeling and proved that this can be used to obtain good classification results, even with a relatively small amount of acoustic training data.

III. Description of the CAPT System

A. Pedagogical Basis

Our approach is strictly segmental and based on the minimal-pair technique. In our tool, users' progression through the different tasks is automatically decided by the system depending on their performance. A core of mandatory tasks is combined with an at-will detour through a series of reinforcement exercises. We use word-based exercises and feedback in the form of articulatory instructions [43]. Our instructions are not just presented in written format, but as audio-visual

events, a feature already used by Neri et al. [20]. 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 (see [19], [20], [28], [43]), few systems give such relevance to, or benefit so much from, the minimal-pair technique [26].

Like other systems, our CAPT tool adapts to the user. Our adaptation mechanism does not incorporate recommendations based on specific learner models (see ProTutor, [62]), but responds to the user's will to move in one direction or another within the training program (see PARLING, [27]). It also adapts to a user's performance, hence it incorporates an element of progress monitoring and control. This performance-dependent next-move strategy constitutes a novel automation feature that is absent from other tools, where users also receive diagnosis, but can only advance manually. Our tool imposes constraints on (1) the phonemes to be practiced, (2) the order in which they are confronted, (3) the number of mandatory sessions to be undertaken, and (4) task progression within sessions. It also gathers relevant information about user improvement, a feature not contemplated by otherwise similar tools [23], [48]. Although we share in Hincks' [24] suspicions about the reliability of automatic scores of ASR in assessing users' production, our tool resorts to ASR to obtain binary (right–wrong) ratings of word production, implementing error-based feedback that goes beyond the mere iteration of trial-and-error cycles.

Our training protocol, the design of the pre- and post-tests used in our experiment, and the criteria used in their assessment, are partially based on the native cardinality method (NCM) [63], [64], and other similar programs [65], [66], [67]. A core feature in NCM is the use of the mixed minimal pair, containing a Spanish monosyllabic word, like *san* (saint) and a similarly sounding monosyllabic English word, like *sun*.

Our tool follows NCM dynamics in its implementation of exposure–discrimination–production cycles, in this particular order. The rationale behind the cycle could be described as follows: first, we raise the user's awareness of a particular English phoneme, and then we prompt them to produce it (for a theoretical discussion of the method see [63], [64]).

As for the raising of phonological awareness, our tool presents users with model pronunciations of mixed pairs, recorded by a proficient bilingual speaker. This allows them to experience and reflect upon perceptual contrasts between confusable Spanish and English phonemes. The strategic decrease of speech rates, both in natural speech and TTS renderings, further favors reflexive listening by the user. Once users have been made aware of fundamental inter-linguistic contrasts, discrimination and production tasks revert to the traditional L2–L2 minimal pair drill. Carefully designed exposure, then, mediates the inductive discovery of the L2 phonemes from first-hand perceptual experience. When this experience is integrated and memorized, success at recognition and identification through discrimination exercises favors confirming and deepening acquired knowledge of L2 phonemes.

The last step consists in producing the new phonemes that are already mentally acquired (as attested by the fact that they can be identified). Recent research has emphasized the importance of getting the learner to notice her or his own errors [68], [69], [70]. Placing production tasks at the final stages of the training process ensures that learners are no longer imitating an externally presented model, but trying to build the sound by accommodating to a mental representation of it, already acquired in the previous stages. In this way, students are expected to detect mismatches between mental and physical forms; they should be able to self-diagnose accuracy and know when self-correction is in order.

The notion of different learners learning differently, according to individual styles and abilities, has been gaining relevance among researchers in the field over the last few years [71]. It is not surprising that many students manage to jump from perceptive memory to accurate production by dint of sheer intuition. For others, articulatory instructions might be necessary. The topic of whether explicit instruction in phonetics assists improvement remains rather controversial [46]. In NCM, rigorous and detailed articulatory descriptions are offered, although they do not constitute a necessary, defining or exclusive characteristic of the approach. Each of our tool sessions, on the other hand, is prefaced by a brief theory video. It is not so much the articulatory descriptions provided, but the exposure to mixed minimal pairs spoken by a single speaker, proficient in both languages. As well as providing the perceptive induction-oriented experience mentioned above, our videos incorporate instructions in the NCM style, that is, they indicate the kind of transformations to be practiced upon an L1 sound in order to turn it into an L2 sound.

Both articulation and perception cues are used. In this sense, we try to address different learning styles. In the videos and during the training sessions, our tool follows NCM in its use of phonetic transcriptions that follow the International Phonetic Alphabet's conventions [72], under the assumption that any particular aural memory will benefit in terms of recollection from attachment to a specific non-ambiguous visual form. There are other less obvious but essential NCM features that inform our training tool. For example, English vowels are ascribed to one of five elemental regions, coinciding with the five Spanish vowels. The version of the tool that we tested for this research project teaches English monophthongs of the {a}, {e}, and {i} regions; that is, English vowels whose articulation and acoustic properties are relatively close to Spanish {a}–/ɑː, æ, ʌ/, Spanish {e}–/e/, and Spanish {i}–/iː, ɪ/.

B. User Interaction Stages

User interaction with the CAPT tool can be described in terms of eight fundamental steps or stages. After logging in (stage 1), the user selects the next available lesson (stage 2). Each lesson is associated to a particularly difficult vowel contrast and twelve minimal pairs are available for the tasks of that lesson [73]. Lessons must be undertaken in a consecutive order.

Each lesson includes a range of task types organized around five overarching training modes, which are presented to the user through the Modes menu (stage 3) after selection of the next available lesson in stage 2. A user's navigation through those five modes is described in Fig. 1. For each lesson,

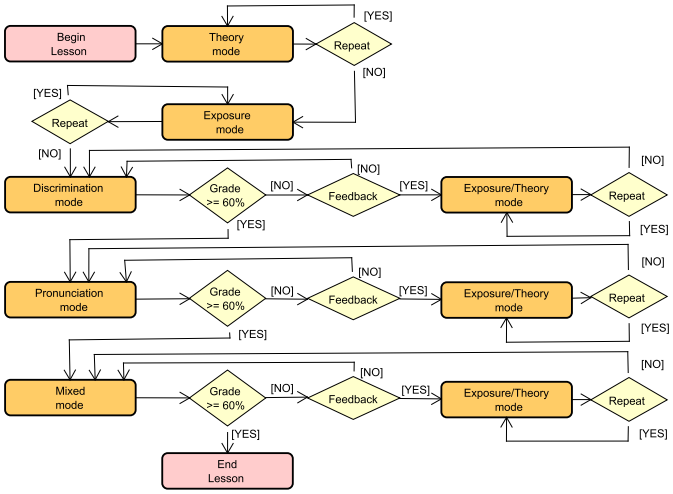

Fig. 1. Flowchart of the training modes within a lesson.

TABLE I
NUMBER OF TASK TOKENS PER TRAINING MODE

Mode	Theory	Exposure	Discrimination	Pronunciation	Mixed
Abbreviation	THE	EXP	DIS	PRO	MIX
# Task tokens	1	3	10	10	9

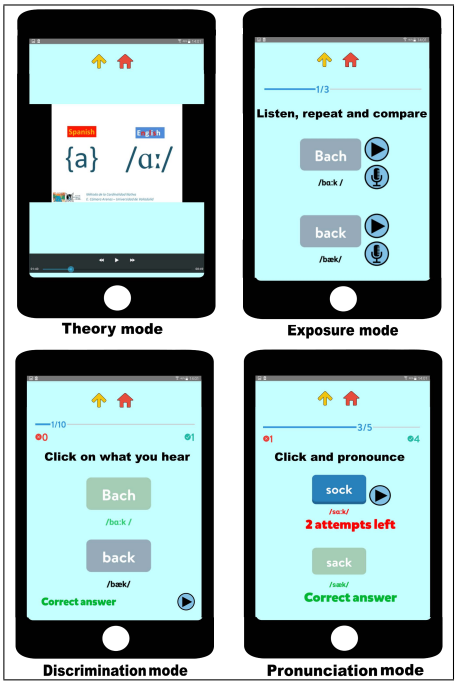

Fig. 2. User interface of THE, EXP, DIS and PRO modes.

the user goes through Theory, Exposure, Discrimination, and Pronunciation modes in that strict specific order. Then, a final Mixed mode follows where discrimination and production tasks alternate randomly. Each mode contains a fixed number of mandatory task tokens or instantiations of its characteristic task type (see Table I). Workflow with the tool is subjected to strict control by the system: neither lessons, nor modes within lessons, nor tasks within modes can be skipped or carried out in an order which is different from that established by the tool.

Fig. 2 shows screen captures corresponding to the four basic modes associated to each lesson: Theory, Exposure, Discrimination, and Production. The Mixed mode includes a random sequence of Discrimination and Production tasks which proceed through screens similar to those shown in Fig. 2 for the corresponding modes.

Stage 4 is accessed through the **Theory** (THE) link on the Modes menu. This is the first training mode, where a video describing the target vowels of the lesson in the NCM fashion is presented to the user. The option to advance to the next mode only becomes available at the end of the video. Within the 60 minutes afforded to each session and at their own discretion, users may choose to review this material as many times as they want. Work distribution and time control, in this group, are left in the hands of participants.

Although preliminary exposure to the contrasts constitutes an essential component of the theory videos, this aspect is further reinforced through the second training mode, **Exposure** (EXP), at stage 5. In this mode, three tokens of a minimal-pair task type are presented to the user, divided into listen–repeat–compare tasks for each minimal pair. The Exposure task type features the orthographic and phonemic forms of the two components of a minimal pair. Clicking on the respective play-buttons activates synthetically generated model pronunciations of each word to which the user must pay attention. Synthetic output is produced by Google's offline TTS tool for Android. Upon clicking, each word is automatically produced five times, each repetition being noticeably slower than the previous one. This mode offers users a first-hand unmediated aural experience of each contrast in order to assist their assimilation. After listening to each minimal pair five times (listen), users are invited to record their own versions of the two words (repeat) and to compare these to the synthesized outputs, at least once per word (compare). Synthesized and recorded versions of the word are consecutively reproduced when the student clicks on the 'play' button. Thus, users are implicitly prompted to carry out a subjective evaluation of their own productions, through comparison with a sanctioned model. After all previous required events per minimal pair (listen–repeat–compare) are completed, participants are allowed to remain in this mode for as long as they wish, listening, recording and comparing at will, before returning to the Modes menu.

Stage 6 is accessed from the Modes menu when the **Discrimination** (DIS) button becomes active and is clicked on. In this mode, participants are presented with written and transcribed minimal pairs, where only one of its constituents is synthetically generated. The challenge consists in identifying which of the words is being produced by the TTS. In this task type, the element to be synthesized is randomly selected by the tool. Users are allowed to listen to the synthesized words as many times as they want. The speed of delivery of the synthetic models varies alternately between normal and slow production rates. A total of ten discrimination tokens are presented within each discrimination mode. The tool reacts

to the right and wrong choices by issuing a characteristic right/wrong sound, and by highlighting the word in green (success) or red (failure).

Clicking on the **Pronunciation** (PRO) button in the Modes menu leads to stage 7. The task type in this mode challenges users to produce the words of a minimal pair, separately, with as much precision and distinctiveness as possible. Here we rely on Google's ASR for Android to automatically discern acceptable from non-acceptable inputs. The tool reacts to the user's pronunciation of each word with a right/wrong sound and by changing to green or red. Up to five attempts per word are allowed, rather than a single attempt, in order to avoid discouragement. After three consecutive failures, a 'play' button appears on the screen, inviting the user to listen to a synthesized version of the target word, as a means to provide corrective feedback.

After passing the Pronunciation mode (see Fig. 1), the user is directed to the Modes menu and allowed to enter stage 8, **Mixed** (MIX) mode. This works as a review mode, since it incorporates again both discrimination and pronunciation tasks. Notice, however, that the random succession of these two types of tasks in the Mixed mode brings forth an extra demand: it measures not only the skills acquired in Discrimination and Pronunciation modes, but also the ability to shift between them. The Mixed mode takes participants one step closer to real communication environments, where speakers must be ready both to discriminate and produce effective language. In this mode, production and discrimination tasks alternate randomly, summing up a total of nine task tokens (four discrimination tasks and five pronunciation tasks).

The score achieved in each of the five training modes accumulates, in percentage terms, and is shown in the Modes menu for each lesson. The accumulated score achieved for each lesson (expressed as a percentage) is regularly updated in the Lessons menu. Access to a new lesson is only activated when a minimum score of 60% is attained in each of the modes of the previous lesson. When reaching a score below 60% in either Discrimination, Pronunciation, or Mixed modes, users must do it again. A threshold over 50% reduces the incidence of success by chance, particularly in two-choice tasks, while keeping the threshold at 60% still offers the possibility of maximally discriminating up to five levels of success (6, 7, 8, 9, 10). When this threshold is not achieved, the tool prompts the user to go back to the Theory or Exposure modes before attempting the mode again, in order to review the theory of problematic vowels (THE), and to perceive again (EXP) the contrasting sounds practiced in the failed mode. When the review is over, users are brought back to the pending mode (see Fig. 1).

C. User's Activity Logs

Our CAPT tool monitors all user activities and gathers data associated to all low-level interaction events. These data are saved into local log files and automatically uploaded to a web server. We have identified and defined a set of experimental variables, from which quantitative measures can later be derived in order to provide consistent answers to our RQs. The experimental variables of our study are the following:

1) **Training intensity** (directly related to research questions RQ3, RQ1.2, RQ1.3). This computes the number of events tracked in each session of the experiment. It is derived from the number of exposure, discrimination, and recording/production tasks; the number of times a particular phoneme is practiced; the number of attempts in each training mode; the number of lessons and sessions in which the user has participated; and the times a word is listened to (including both listening events imposed by the system and those requested by the user).
2) **Training performance** (RQ3, RQ1.2, RQ1.3). This measures the success attained by the participant during a specific time of each event tracked. The variable encompasses right and wrong discrimination tasks; right and wrong pronunciation tasks; success rates in discrimination and production tasks per phoneme; number of training modes and lessons passed and failed; and the time spent on watching videos and performing training events, modes, and lessons.
3) **Pronunciation improvement** (RQ1, RQ2, RQ3). This considers the scores achieved in each training task, mode and lesson. Our tool also provides a final software score, that is, a total score granted by the application to each user at the end of the last session (see Section IV-C). Our experimental design includes a groupwise comparison of these scores with those assigned in the human-rated post-test.

IV. Method

We have implemented an approach similar to that used in Akahane-Yamada et al. [28] to evaluate pronunciation improvement; in our case, we compared the results obtained by users in a pre-test and a post-test, both based on minimal-pair production. In accordance with the NCM and other experts [58], our tool is specifically designed for L1 Spanish speakers learning L2 English pronunciation. In previous work, we determined that the N-best list of candidates and the scores provided by Google ASR could be satisfactorily used to correctly classify different levels of pronunciation [11] and reported on the evolution of pronunciation improvement over time [12]. In the present study, we focus on the correlation between the scores obtained during the training and those obtained in the post-test, on the one hand, and pronunciation improvement between pre- and post-test scores, on the other. The reliability of results rests on independent scoring by three raters, the verification of inter-rater agreement, the comparisons of scores in the pre/post tests and on the correlation between subjective and automatic assessment.

A. Participants

The participants in our study were recruited from a group of 20 students who had qualified and registered for an English as foreign language (EFL) B1–B2 student course at the Language Center of the University of Valladolid. This institution distributes its prospective students to its different courses by

means of a rigorous level test. By recruiting participants in this way we ensured (1) that our experiment realistically reproduced the diversity of students that attend the same course of the Language Center of the University of Valladolid; and (2) that all participants had initially the same level of English; by choosing to work with a course for strictly selected B1–B2 students, we also ensured (3) that any recruited students would have very little or, more likely, no previous training in English phonetics (eventually, pre-test results proved this to be the case).

An offer was made to all the students who had registered for the course, through the mediation of the instructor, and with the acquiescence of the institution's authorities, to cover a small part of the course program (more specifically, the teaching of a few English phonemes) by using a CAPT tool. The first 10 students who applied for it were welcomed as our experimental group. Two of them (20%) were female, and the other 8 (80%) were male. Their average age was 26 years (M = 26.4, SD = 5.5). These students worked with the CAPT tool for three one-hour-maximum sessions.

Additionally, an agreement was reached with the instructor of the course to cover the same material, during the same weeks, and through the same amount of one-hour sessions, with the rest of the students. This in-classroom group consisted of 8 members (two students dropped out for personal reasons unrelated to our study) of whom 5 (63%) were female and 3 (37%) were male; their average age was 24 years (M = 23.9, SD = 4.31). The instructor also agreed not to use any computer-assisted interactive tools in her sessions.

All 18 course students (in both experimental and in-classroom groups) were explicitly requested not to do any extra work in English (extra lessons, conversation exchanges with natives, etc.) during the four weeks that our experiment lasted. All of them agreed to comply with this condition. And all 18 students graciously agreed to and actually did take a pronunciation pre-test shortly before the beginning of the experiment, and a pronunciation post-test shortly after the end of the last session covered in both teaching conditions.

B. Protocol Description

The four-week experiment included a pre-test, a post-test, an experimental group, and an in-classroom group, as shown in Fig. 3. Pre- and post-test recordings were carried out using professional recording equipment in a quiet room.

Subjects took a 25-contrast **pre-test**, under the sole supervision of a member of the research team, and were asked to read aloud the minimal pairs/triplets presented in the contrasts (see Section IV-C for scoring details). They were free to repeat each contrast as many times as they wanted whenever they thought they might have mispronounced them. We imposed no time limitations. As described in Section III-A, the test included contrasts of the English pure vowel /ɑː, ʌ, æ/ that are usually reduced to Spanish {a}, vowel /e/, that is often realized as a closer Spanish {e}, and vowels /iː, ɪ/ usually reduced by Spanish natives {i}. Each participant took an average of 79.72 seconds to complete the pre-test, with a duration ranging from 59 to 107 seconds.

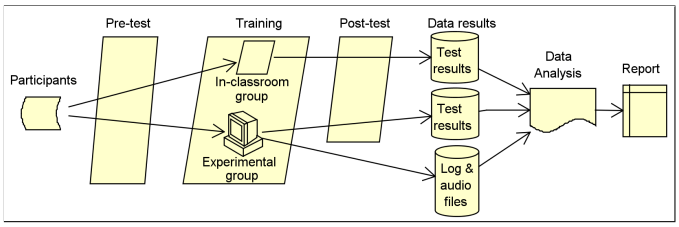

Fig. 3. Steps of the experiment protocol.

The words used in this version of the tool constitute a closed list of minimal pairs selected and supervised by an expert, ensuring that they are all recognized by the ASR system, and that homophones are adequately processed.

A week after running the pre-test, participants in the experimental group took part in three 60-minute training sessions. These sessions were separated by a 72-hour gap in order to avoid fatigue and to allow for learning consolidation [74]. In-classroom and CAPT sessions took place at the same time in different locations (classroom and laboratory). In both conditions students were given up to 60 minutes to complete the proposed training session. The tool's approach to the practice schedule is based on lessons. Each session contains two lessons and a minimal pair is practiced in each lesson (block distribution). Nevertheless, most phonemes were retaken in later sessions (spaced distribution) as follows: in the first session (lessons 1 and 2), phonemes /ɑː/–/æ/ and /æ/–/ʌ/ were contrasted. In the second session (lessons 2 and 3), /ɑː/–/ʌ/ were dealt with, as well as /e/–/æ/. The last session (lessons 3 and 4) involved the contrasts /ɪ/–/iː/ and /ɪ/–/e/. Only /iː/, a vowel that is pretty much interchangeable with the Spanish /i/, was left out of a repeated practice scheme.

Students in the **CAPT condition** exclusively used the CAPT system during the training sessions. The software application was installed in the computing laboratory at the Languages Center and run on each desktop using an Android emulator [75]. Before starting the first session, CAPT users were briefly instructed on how to use the software. Participants worked on their own and without human interaction in all sessions. Each student in the CAPT condition worked inside a cubicle, separated by glass dividers, and used a headset with microphone (*Logitech H390*). Along the three experimental sessions, a total of 72 minimal pairs were presented to participants (12 in each lesson). We took some precautions not to repeat too many of the pre-test and post-test words in the working sessions: just 10% of the total 144 words (15 words) used by the application were also present in the pre-test and post-test.

In-classroom group training sessions were guided by an L2-English teacher with a vast experience in English phonetics. The program included the same phonemes covered by the tool: /ɑː, ʌ, æ, e, iː, ɪ/. Each 60-minute session began with around 10 minutes of explicit articulatory instructions and auditory descriptions of the sounds, with ample exposure to contrasting examples. These examples were both produced by the instructor and extracted from the audio materials of an EFL handbook. All pronunciations were consistently within the range of standard American accent (General American). After

exposure, students practiced auditory discrimination activities using materials from the same handbook; they listened to word recordings produced by the instructor's computer and were asked to match them with printed pictures representing the meaning of the words, and/or with their printed orthographic/phonetic representations. Then they were asked to read aloud pairs and trios of contrasting words, containing the targeted vowels. They carried out this task publicly, and under the supervision of the instructor, who provided real-time feedback and prompted repetition when necessary. Throughout the lesson, the instructor allotted participation turns in a uniform way, ensuring the equal participation of all 8 students; she answered questions and produced model pronunciations upon request; she also diagnosed pronunciation errors publicly (that is, for the benefit of all attending students) and provided the necessary corrective feedback. Each session was closed with a 5–10 minute review.

Finally, one week after the last training session a **post-test** was given to participants in the two groups. The post-test contents and conditions were identical to those of the pre-test. Each student took an average of 95.61 seconds to complete the post-test, with a duration ranging from 62 to 140 seconds.

C. Scoring Procedures

Our conclusions concerning pronunciation improvement derive from the computation and analysis of the differences between the pre-test and post-test results, and consequently rest upon the reliability of the human rating involved in the assessment of both tests.

The testing and assessing used for pre-test and post-test rating were grounded in the NCM perspective (see Section III-A). Three specialized English-L2 pronunciation teachers from the University of Valladolid were recruited for rating the tests. The three of them were trained to perceive and assess slight deviations from American standard pronunciation and subtle processes of L1–L2 feature transfer, and the three applied the same criteria for scoring the tests. They revised a jumbled array of pre-test and post-test audio files without any indication as to which files were pre and which were post; the files were strictly anonymous. During the process, raters neither interacted with each other, nor with the participants, and we did not impose any extraneous assessment procedure on raters.

The three raters had experience in the academic assessment of pronunciation at the segmental level, and all of them were familiar with the scoring system we are about to describe, which is actually their usual way of scoring. Raters evaluated the pronunciation of isolated words, presented in random order and independently. They applied a 3-point scale, in 0.5 increments. According to their experience, the five-level scale [1–1.5–2–2.5–3] they usually apply in their courses provides a reasonably good frame for the discerning and assessment of segmental pronunciation. Since the Spanish educational system traditionally works with a [0, 10] scale, raters usually apply a linear scaling of the marks in order to map the scores onto the traditional [0, 10] scale.

For students in the CAPT condition, quantitative scores were computed from the success results gathered from the tool in completing DIS, PRO and MIX modes. For each student, the CAPT tool provides a **quantitative software score** $\overline{L}$, based on the score achieved in each one of the six lessons:

$$\overline{L}_s = \frac{1}{6}\sum_{i=1}^{6} L_{s,i} \in [0, 10]. \quad (1)$$

where s is the speaker, i is the lesson and each lesson score, $L_{s,i}$ is defined as:

$$L_{s,i} = \frac{\sum_{j=1}^{10}(\mathrm{DTT}_j + \mathrm{PTT}_j) + \frac{10}{9}\sum_{k=1}^{9}\mathrm{MTT}_k}{3} \in [0, 10]. \quad (2)$$

where s is the speaker and i is the lesson. The score in the Discrimination, Pronunciation, and Mixed modes are based on the number of successfully performed discrimination (DTT), pronunciation (PTT) and mixed (MTT) task tokens, respectively (see Table I).

TABLE II
DESCRIPTIVE STATISTICS FOR LEARNER TIME (MINUTES) PER TASK

	THE		EXP		DIS		PRO		MIX		Sum	
Task	$\overline{n}$	SD	$\overline{n}$	SD	$\overline{n}$	SD	$\overline{n}$	SD	$\overline{n}$	SD	$\overline{n}$	SD
Video	31.2	6.6	-	-	-	-	-	-	-	-	31.2	6.6
Pair listening	-	-	3.6	1.3	-	-	-	-	-	-	3.6	1.3
Word listening	-	-	2.4	1.3	0.6	0.5	2.5	2.0	1.1	0.8	6.6	2.6
Discrimination	-	-	-	-	2.6	1.0	-	-	1.0	0.7	3.6	1.2
Production	-	-	3.9	1.6	-	-	37.2	15.8	14.9	6.7	56.0	17.2
Times-out	0.1	0.1	7.0	1.1	2.3	1.1	1.8	0.9	2.0	0.8	13.2	2.0
Sum	31.3	6.6	16.9	2.7	5.5	1.6	41.5	16.0	19.0	6.8	114.2	6.6

V. RESULTS

A. CAPT Tool Users' Interaction

Table II shows high rates of active user-time invested in interactive tasks (101 minutes out of the total 180 minutes in three sessions). The remaining time (Times-out row, 13.2 minutes) is spent on transitions between tasks. Production across tasks (EXP, PRO, MIX) registers more time investment, on average, than other forms of involvement (a total of 56.0 minutes in three sessions); at the other end, listening to pairs in exposure and pair discrimination registered only 3.6 minutes each. A third of the total time was spent on the theory phase (viewing videos, 31.2 minutes). Even more time was spent on PRO tasks (39.7 minutes).

Table III reports the use of the CAPT tool by the users (see Section III-B for events description). Mand. and Req. mean mandatory and requested listening events; the former are imposed on the user as part of training, the latter are freely requested by users. In both cases, TTS synthesized versions of model words are used. On average, users listened to the TTS system 831.2 times (calculated as the sum of the EXP, DIS, PRO, and MIX $\#Mand.listenings$ and $\#Req.listenings$ values of column $\overline{n}$) and used the ASR system 615.6 times (calculated as the sum of the PRO and MIX $\#Productions$ values of column $\overline{n}$), giving a rate of 8.04 uses of the TTS/ASR per minute.

It also shows important differences in the use of the tool depending on the user. For instance, the user who performed

TABLE III
NUMBER OF EVENTS PER USER OF THE CAPT TOOL ALONG THE WHOLE EXPERIMENT

	THE			EXP			DIS			PRO			MIX		
	$\overline{n}$	MIN	MAX	$\overline{n}$	MIN	MAX	$\overline{n}$	MIN	MAX	$\overline{n}$	MIN	MAX	$\overline{n}$	MIN	MAX
Time (min)	31.3	20.1	39.2	16.9	11.1	29.6	5.5	3.7	7	41.5	19.2	65.1	19.0	3.7	34.1
#Tries	6.4	6	8	11.9	7	17	7.2	6	9	12.6	6	21	9	6	18
#Mand.listenings	-	-	-	347	210	510	69.5	60	82	-	-	-	26.8	15	54
#Req.listenings	-	-	-	146.9	64	292	29.9	0	75	147.9	25	426	63.2	20	178
#Discriminations	-	-	-	-	-	-	69.5	60	82	-	-	-	26.8	15	54
#Productions	-	-	-	-	-	-	-	-	-	441.5	166	806	174.1	87	382
#Recordings	-	-	-	90.2	56	134	-	-	-	-	-	-	-	-	-

TABLE IV
SUCCESSFUL AND FAILING TASK TYPE EVENTS, AND NUMBER OF LISTENING EVENTS AS A FUNCTION OF TARGET PHONEME

	Successful (S) and Failing (F) Events													
Task	ɑː		æ		ʌ		e		ɪ		iː		Total	
	S (%)	F	S (%)	F	S (%)	F	S (%)	F	S (%)	F	S (%)	F	S (%)	F
Discrimination	143 (75.7%)	46	198 (81.1%)	46	114 (77.0%)	34	144 (86.2%)	23	105 (78.9%)	28	78 (95.1%)	4	782 (81.2%)	181
Production	151 (36.2%)	266	261 (53.6%)	226	127 (42.1%)	175	195 (76.5%)	60	115 (58.4%)	82	103 (85.1%)	18	952 (53.5%)	827
All Productions	151 (8.9%)	1543	261 (15.5%)	1424	127 (10.6%)	1066	195 (31.1%)	433	115 (17.0%)	563	103 (37.1%)	175	952 (15.5%)	5204
	Mandatory (M) and User-Requested (U) Listening Events													
	ɑː		æ		ʌ		e		ɪ		iː		Total	
	M	U	M	U	M	U	M	U	M	U	M	U	M	U
Discrimination	189	86	244	89	148	62	167	75	133	74	82	24	963	410
Production	-	562	-	552	-	374	-	218	-	241	-	53	-	2000

the activities of the PRO mode fastest took 19.2 minutes and the one who spent the most time took 65.1 minutes. This contrast can be observed in the rest of the modes and in the number of times they interacted with the tool. The inter-user differences affect both the number of times users make use of the ASR (253 minimum vs. 1188 maximum, calculated as the sum of the PRO and MIX $\#Productions$ values of columns MIN and MAX, respectively), and the number of times they use the TTS (109 vs. 971 times, calculated as the sum of the EXP, DIS, PRO, and MIX $\#Mand.listenings$ and $\#Req.listenings$ values of columns MIN and MAX, respectively).

The variety in the use of the tool according to differences between users is motivated by the number of repetitions that users are required to do: the more they fail activities, the more they have to repeat. Table IV details the number of successful and failing interactions per tested phoneme. The final column indicates that discrimination activities were easier than pronunciation ones: 81.2% vs. 53.5% success rate, respectively. This difference is higher when the success rate of all the production attempts is compared: 15.5% (it should be remembered that users have a maximum of five attempts to produce the proposed word correctly).

From Table IV, we can also see that the differences between success and failure for production and discrimination activities are highly dependent on the target phoneme. The most difficult phoneme seems to be **ɑː** with only a 75.7% success rate for discrimination tasks and 36.2% for production tasks, with an 8.9% success rate when all productions carried out by the user are taken into account. The easiest one seems to be **iː** with a 95.1% success rate for discrimination tasks and a 37.1% for production tasks, with 37.1% success rate for all productions. These differences affect the number of times the users requested the use of the TTS for listenings: 648 for ɑː vs. 77 for iː.

Table V shows the confusion matrices between the phonemes of the minimal pair contrasts in discrimination and productions. We have included TPR (true positive rate or recall) and PPV (positive predictive value or precision) as quality indicators. In discrimination tasks, /ɑː/ has the lowest recall and /æ/ the lowest precision, which means /ɑː/ was the hardest to predict (TPR = 75.7%) and /æ/ the most commonly confused (PPV = 77.0%), while /iː/ got both the highest precision and recall (PPV = 87.6% and TPR = 95.1%). Concerning production tasks, the lowest precision and recall values were obtained for /ɑː/ (PPV = 43.1% and TPR = 36.2%), whereas the highest recall was obtained for /iː/ (TPR = 85.1%), and the highest precision for /ɪ/ (PPV = 73.2%).

B. Results of the Pre-test and Post-test

In order to evaluate the degree of inter-rater variability in the scoring done by human experts in pre-test and post-test, we carried out a Kendall's coefficient analysis. A relevant inter-rater agreement has been obtained (Kendall's coefficient $W = 0.493, items = 900, raters = 3, p - value = 3.1e - 19$) with a high Pearson correlation between the scores assigned to speakers by the different pairs of raters. For pre-test scores, $r = 0.87$ (Ev1|Ev2), 0.73 (Ev1|Ev3), and 0.79 (Ev2|Ev3),

TABLE V
CONFUSION MATRIX OF DISCRIMINATION (UPPER) AND PRODUCTION (LOWER) TASK TOKENS. ROWS: EXPECTED PHONEMES. COLUMNS: SELECTED/PRODUCED PHONEMES, RESPECTIVELY

Discrimination task tokens							
	ɑː	æ	ʌ	e	ɪ	iː	TPR (%)
ɑː	***143***	34	12	-	-	-	75.7
æ	19	***198***	11	16	-	-	81.1
ʌ	20	14	***114***	-	-	-	77.0
e	-	11	-	***144***	12	-	86.2
ɪ	-	-	-	17	***105***	11	78.9
iː	-	-	-	-	4	***78***	95.1
PPV (%)	78.6	77.0	83.2	81.4	86.8	87.6	

Pronunciation task tokens							
	ɑː	æ	ʌ	e	ɪ	iː	TPR (%)
ɑː	***151***	143	123	-	-	-	36.2
æ	78	***261***	35	113	-	-	53.6
ʌ	121	54	***127***	-	-	-	42.1
e	-	36	-	***195***	24	-	76.5
ɪ	-	-	-	33	***115***	49	58.4
iː	-	-	-	-	18	***103***	85.1
PPV (%)	43.1	52.8	44.6	57.2	73.2	67.8	

TABLE VI
PRE/POST TEST SCORES

Group	Ev.	Pre-test		Post-test		Difference (Wilcoxon signed rank test)				
		mean	N	mean	N	mean	N	z-score	r	p-value
Exp.	1	0.82	250	2.53	250	1.71	250	-7.864	0.50	<0.001
Exp.	2	0.99	250	2.45	250	1.46	250	-8.148	0.52	<0.001
Exp.	3	0.55	250	2.38	250	1.83	250	-7.422	0.47	<0.001
Exp.	1,2,3	0.85	750	2.59	750	1.74	750	-13.551	0.50	<0.001
In-Class.	1	0.41	200	0.68	200	0.27	200	-2.281	0.16	0.023
In-Class.	2	0.63	200	0.86	200	0.23	200	-3.056	0.22	0.002
In-Class.	3	0.27	200	0.61	200	0.34	200	-2.597	0.19	0.009
In-Class.	1,2,3	0.41	600	0.75	600	0.34	600	-4.566	0.20	<0.001

always with $p < 0.001$; for the post-test, $r = 0.97$ (Ev1|Ev2), 0.94 (Ev1|Ev3), and 0.95 (Ev2|Ev3), also with $p < 0.001$ in all cases.

Table VI shows the mean values of the raters' assessment of pre-test and post-test, where the z-score is the z statistic, Ev. means human evaluator, and the p-value is 2-tailed. There are a total of 1500 scores for the experimental group (10 participants x 25 minimal pairs x 2 tests x 3 raters) and 1200 scores for the in-classroom group (8 participants x 25 minimal pairs x 2 tests x 3 raters). Since our data did not pass the Kolmogorov–Smirnoff nor Levene's standard tests, we carried out several non-parametric tests for non-normally distributed data, in order to detect statistically significant differences. A comparison of pre-test and post-test scores, granted by the three human raters (Ev: 1,2,3), shows that there is improvement in both groups: from 0.85 to 2.59 in the experimental group and from 0.41 to 0.75 in the in-classroom group. Since pre-test and post-test contain the same words, a word-by-word comparison could be carried out between pre- and post-test realizations of the same items by each student. Here, a Wilcoxon signed-rank test found statistically significant differences between pronunciation improvement in both groups. The CAPT-group obtained

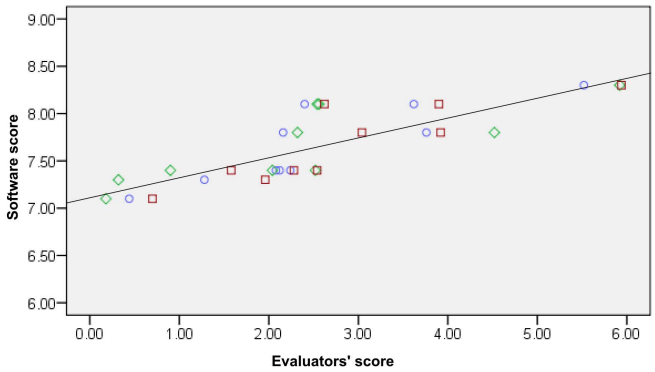

Fig. 4. Correlation between the post-test score assigned by evaluators (1:circle, 2:square, 3:rhombus) and the score of the application.

a 1.74 improvement mean ($Z = -13.551; p < 0.001$), with a large effect size ($r = 0.50$); the in-classroom group obtained a 0.34 improvement mean ($Z = -4.566; p < 0.001$) with a small effect size ($r = 0.20$).

Comparisons between the two independent groups were carried out using the Mann–Whitney U test. There are no statistically significant differences in the scores of the pre-test between the experimental group and the in-classroom group ($U = 18.0, p = 0.055$) with a moderate effect size ($r = 0.46$). Thus, although the pre-test scores of the experimental group are on average higher than the scores of the in-classroom group, it cannot be stated that these two groups were heterogeneous before the experiment. However, the Mann–Whitney U test between post-test results in both groups showed significant differences ($U = 9.0, p = 0.001$), with a large effect size ($r = 0.65$). Students that used the tool, therefore, outperformed students in the in-classroom group, both in absolute terms (1.74 vs. 0.34 of improvement) and in relative terms with respect to the initial level (205% vs. 82% of improvement).

Also, correlation has been found between the objective software scores and the post-test average scores of evaluators 1, 2 and 3 of the experimental group, as shown in Fig. 4. The Pearson correlation is $r = 0.84$ for evaluator 1 ($p = 0.002$), $r = 0.86$ for evaluator 2 ($p = 0.001$), and $r = 0.79$ for evaluator 3 ($p = 0.007$). Moreover, the correlation of evaluators 1, 2, and 3 together is $r = 0.84$ ($p = 0.002$). Finally, a potential evaluator score, Ev', can be obtained from the CAPT tool score (SCORE), with an average error of 5.5% using a linear regression model (see Fig. 4):

$$Ev' = -21.724 + 3.171 * SCORE \quad (3)$$

VI. DISCUSSION

Our first research question (**RQ1**) was concerned with finding out whether our CAPT tool contributes to the teaching/learning of users' pronunciation of English at the segmental level. Results clearly show that post-test scores among the tool users are significantly higher than those obtained in the pre-test; **Issue 1.1** is then positively settled. As for **Issue 1.2**, significant differences are also observed between the performance of the students that use the tool and those who do

not. In this respect, a disclaimer is in order: neither the efficacy of the in-classroom method, nor the competence of a certified and qualified instructor have been under investigation. On the contrary, in-classroom improvement data constitutes the yardstick against which we can confirm, by way of a reinforcement of our statistical processing, the pedagogical usefulness of our CAPT. Concerning **Issue 1.3**, the most difficult phonemes for users were identified and it has been shown that users encounter more difficulties in activities related to production. Results suggest that discrimination and production skills are asymmetrically interrelated. Participants are usually better at discrimination (7.2 tries per user) than production (12.6 tries per user, see Table III, #Tries). While a good production level seems to be preceded by a good performance in discrimination, a good discrimination does not guarantee an equally good production. The system is, therefore, sensitive to the expected difficulty of each type of activity.

Our second research question (**RQ2**) addressed the use of off-the-shelf TTS and ASR as pedagogical tools at the segmental level. The quality of synthetic voice in the rendering of minimal pairs seems to have been satisfactory. Participants in the CAPT condition consistently resorted to TTS models when faced with difficulties both in perception and production modes (#Req.List. rows of Table IV). Alternatively, traditional approaches to pronunciation teaching rely on real human voice (live or recorded). The process of recording clean and properly equalized model pronunciations is cumbersome and, therefore, the possibility of using synthetic voice grants unprecedented feasibility to future CAPT projects, provided that it proves to be non-obstructive to the process of learning. While we cannot assert that the quality of TTS was, by itself, responsible for pronunciation improvement, it does not seem to have affected the learning outcomes negatively [48], [50]. Despite the use of synthetic instead of natural voices in TTS, the results in terms of improvement after using our CAPT tool are comparable to those obtained by in-classroom training.

The role of ASR is even more important inasmuch as it offers diagnosis to users and triggers automatic feedback. Although current DNN techniques reach high rates of recognition [4], [76], [77], we had to carefully explore the capabilities of the ASR, for its use to be effective, by feeding it many properly and wrongly pronounced words in order to discard those that the system consistently failed to recognize or recognized even when wrongly pronounced—pronunciation experts assisted us in this prior word selection [33]. We found that the system had problems in recognizing (1) infrequent words and words rarely pronounced in one-word sentences; on the other hand, there were (2) words that would be properly identified by the system even when pronounced with L1 transferred pronunciation, probably due to a lack of likely alternatives among possible and frequent one-word sentences. Tool users performed a high number of successful interactions with an ASR system that had been previously explored and tested. We are in a position to assert that such shortcomings as environmental noise do not seem to adversely affect learning when practice words are properly selected and when the system is integrated within a well-designed CAPT. This is, we believe, a relevant issue which is logically dependent on the conclusion that our CAPT succeeds in mediating improvement. While the quality of the ASR system used by our CAPT is not causally related to improvement as an independent variable, it is enough not to constitute an impediment to improvement.

We must not forget, however, that the primary function of TTS and ASR systems is not usually pedagogical. By themselves, these instruments will not mediate learning in any particularly effective way. In order to do so, both systems must be integrated within carefully designed, research-based and well-informed teaching programs. Others have reported similar results regarding the usefulness of TTS and ASR technologies in CAPT applications. Participants in the study run by Eksi and Yesilcinar [57] registered an improvement in pronunciation using TTS in self-study websites. Liakin et al. [48] reported pronunciation improvement of liaison in L2 French after using a mobile TTS system. In the case of ASR, Neri et al. [27] proved the effectiveness of a CAPT tool that used ASR technology for training users in the pronunciation of decontextualized isolated words. In their study, the experimental group that used the tool attained a degree of improvement that was comparable to that of a control group that followed traditional, i.e., non-computerized, teacher-led training. Finally, Liakin et al. [23] reported improvement in the production of the L2 French vowel /y/ after training with a mobile ASR system. The results described in the present study and those reported in the literature confirm the view that TTS and ASR systems are ready to be used in language learning activities. In our case, the key to its effectiveness seems to lie in the sequencing of activities and the mandatory nature of corrective feedback. The system guides students to undertake a series of exercises in sequence according to theory-informed recommendations [63]. Moving forward requires success, while failure imposes revision and further practice. There is no stopping until challenges are overcome. This is, we believe, the circular dynamics ultimately responsible for improving in pronunciation. Following the method leads to improvement, as post-tests reveal.

Results show that the tool led users to carry out a significantly large number of listening, discrimination and production exercises (**RQ3**). With an effective and objectively registered 57% of the total time, per student, devoted to training (101 minutes out of 180), high training intensity is confirmed in the CAPT condition. Each of the users in the CAPT-group listened to a mean of 831.2 synthesized utterances and produced an average of 615.6 word-utterances (Table III), which were immediately diagnosed, triggering, when needed, automatic corrective feedback. This intensity of training (hardly attainable within a conventional classroom) implies a significant level of time investment on tasks which might constitute a relevant factor in explaining the larger gain mediated by the CAPT tool.

Participants generated valuable objective information concerning the quality of their interactions with the system. The correlation between the scores assigned by the system and those assigned by human raters shows consistency. Although automatic evaluation is not perfect, human assessment also has its weaknesses, and the evaluation of some of the speakers' pronunciations can be a difficult task even for human evalua-

tors, which explains rating discrepancies.

A. Limitations and Further Research

Results show that the adequate integration of a pedagogically informed design, with a pre-selection of pairs, and the use of TTS and ASR technologies in a CAPT tool, may successfully complement traditional in-classroom L2 English segmental pronunciation training. Such a tool further promotes a high level of training intensity and a corresponding increase in learning. Nevertheless, an explanation of the specific role of any of those elements in isolation would require further experimentation to obtain quantitative measurements of the effect of each one. This obviously requires the comparison of the results of different controlled versions of the tool, which is left for future work.

Although the recruitment procedure described in Section IV-A provides a realistic, representative and research-operational sample, a more conventional sampling procedure (e.g., random or criteria-driven selection of a control-group) and a closer systematic monitoring of the activities of the control-group, would have allowed a deeper exploration of relevant details about the impact of using our CAPT tool. Future experimentation on a larger scale should also consider these aspects.

One of the key goals of this work was to analyze the extent to which the use of our CAPT tool provides segmental pronunciation improvement comparable to that obtained after a similarly designed and structured in-classroom training, led by an experienced instructor. Although the results give a satisfactory answer to this question, a closer comparative study between the operation and results of the tool and those of human-led instruction might help us gain a more detailed knowledge of the possibilities of CAPT.

In this study, a heuristic value of 60% success was required for progressing to new tasks (see Fig. 1) and performance below that threshold led to corrective feedback and repetition. Although such a procedure has not introduced any noticeable bias in our results, as discussed above, it would be interesting to analyze the effect of choosing different thresholds or their adaptation to different activity types and user profiles and results. This would add a promising dimension to CAPT, to be explored in future work.

In order to cope with a certain degree of latency of pronunciation improvement, the post-test was performed one week after the training sessions and contained both words that had been practiced during the sessions and fresh new words. However, the analysis of the potential generalization of this learning and its long term persistence would require further future testing, probably incorporating only new words which include the target phonemes.

VII. Conclusions

Our CAPT system and the methodological choices it implements have allowed us to reliably measure the relative pronunciation improvement of the users who trained with it. The guidelines for an ideal quantitative experiment of pronunciation training described by Thomson and Derwing [1] have been followed: we have provided enough detail about participants and the protocol followed to allow replication, we have gathered a large activity log to be statistically analyzed, and we have compared the results with a reference group.

Speech technologies have proved to be particularly useful for increasing the amount of involved participation, immediate diagnostic and guiding feedback, and model pronunciations available to the students. The tool mediates short-term improvements in pronunciation that are comparable to those achieved through in-classroom instructor-led pronunciation training. These results support the notion that novel CAPT technology might have a place in future pronunciation teaching. However, we believe that, for any technological complement to be truly effective, it must be subordinated to well-thoughtout and carefully designed methodological frameworks that also encompass human interaction. The in-classroom learning experience, with its combination of involved and distant participation, its social nature, and the intervention of adaptive human agents, is rich in ways that cannot be easily emulated by technology. Still, efficient CAPT systems, like the one we have designed, may complement in-classroom teaching in a variety of ways: they can be used in the classroom and outside, they can be part of the student's homework, etc. The benefits of CAPT systems are particularly worth considering when dealing with large groups of students and, consequently, low rates of involved participation.

Current TTS systems and synthetic voice have reached the point where they can be seriously considered by developers of CAPT applications who need to generate pronunciation models for isolated words. The use of ASR systems as part of CAPT tools requires a careful pre-selection of language elements (words, sentences, ...) to be included in the exercises, so as to avoid bias that could cause the system to skip pronunciation errors and/or reject acceptable pronunciations of infrequent words, and words not found in one-word sentences.

Acknowledgment

This study is supported in part by the Ministry of Economy, Industry and Competitiveness of Spain – project key TIN2014-59852-R, and by Consejería de Educación of Junta de Castilla y León (VA050G18). We would like to thank the Language Center of the University of Valladolid, and the university for the PhD Research Grant funding of Cristian Tejedor (2015).

References

[1] 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, Jun. 2015, doi: 10.1093/applin/amu076.

[2] A. Kukulska-Hulme, "Mobile-assisted language learning," in *The Encyclopedia of Appl. Linguistics*, C. Chapelle, Ed. New York, NY, USA: Wiley Online Library, 2012, pp. 3701–3709, doi: 10.1002/9781405198431.wbeal0768.

[3] D. Escudero-Mancebo and M. Carranza, "Nuevas propuestas tecnológicas para la práctica y evaluación de la pronunciación del español como lengua extranjera," in *Proc. 50th Int. Congr. Asociación Europea Profesores Español*, Burgos, Spain, Jul. 20–24, 2015, pp. 218–227.

[4] M. Meeker, "Internet trends 2017," Los Angeles, CA, USA, Rep. May 2017.

[5] J. Shen *et al.*, "Natural TTS synthesis by conditioning WaveNet on MEL spectrogram predictions," in *IEEE Int. Conf. Acoustic Speech, Signal, and Processing*, Apr. 15–20, 2018, pp. 4779–4783, doi: 10.1109/ICASSP.2018.8461368.

[6] K. Al-Seghayer, "The effect of multimedia annotation modes on L2 vocabulary acquisition: a comparative study," *Lang. Learn. Technol.*, vol. 5, pp. 202–232, Jan. 2001, doi: 10125/25117.
[7] H. Yoon, "More than a linguistic reference: The influence of corpus technology on L2 academic writing," *Lang. Learn. Technol.*, vol. 12, no. 2, pp. 31–48, Jun. 2008.
[8] M. Levy, "Technologies in use for second language learning," *The Modern Lang. J.*, vol. 93, no. s1, pp. 769–782, Jan. 2009, doi: 10.1111/j.1540-4781.2009.00972.x.
[9] S. Sauro, "Computer-mediated corrective feedback and the development of L2 grammar," *Lang. Learn. Technol.*, vol. 13, no. 1, pp. 96–120, Feb. 2009, doi: 10125/44170.
[10] R. L. Trask, *A dictionary of phonetics and phonology*. London, U.K.: Routledge, 2004.
[11] D. Escudero-Mancebo, E. Cámara-Arenas, C. Tejedor-García, C. González-Ferreras, and V. Cardeñoso-Payo, "Implementation and test of a serious game based on minimal pairs for pronunciation training," in *Proc. SLaTE*, Leipzig, Germany, Sep. 4–5, 2015, pp. 125–130.
[12] 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. 8–12, 2016, pp. 1178–1179.
[13] 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. 23–25, 2016, pp. 177–186.
[14] C. Tejedor-García, V. Cardeñoso-Payo, E. Cámara-Arenas, C. González-Ferreras, and D. Escudero-Mancebo, "Playing around minimal pairs to improve pronunciation training," in *Proc. IFCASL*, ser. Feedback in Pronunciation Training Workshop, Saarland, Germany, Nov. 5–6, 2015.
[15] A. Rauber *et al.*, "TipTopTalk!: A game to improve the perception and production of L2 sounds," in *Abstracts New Sounds Aarhus, 8th Int. Conf. Second Language Speech*, Aarhus University, Aarhus, Denmark, Jun. 10–12, 2016, p. 160.
[16] T. M. Derwing and M. J. Munro, "Second language accent and pronunciation teaching: A research-based approach," *TESOL Quart.*, vol. 39, no. 3, pp. 379–397, Sep. 2005, doi: 10.2307/3588486.
[17] J. Lee, J. Jang, and L. Plonsky, "The effectiveness of second language pronunciation instruction: A meta-analysis," *Appl. Linguistics*, vol. 36, no. 3, pp. 345–366, Jul. 2015, doi: 10.1093/applin/amu040.
[18] 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 Second Language Learning Teaching Conf.*, Iowa State Univ. Press, Ames, IA, USA, Oct. 7–8, 2011, pp. 169–80.
[19] A. Weinberg and H. Knoerr, "Learning French pronunciation: Audio-cassettes or multimedia?" *CALICO J.*, vol. 20, no. 2, pp. 315–336, Jun. 2003, doi: 10.1558/cj.v20i2.215-336.
[20] 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. 02, pp. 225–243, May 2008, doi: 10.1017/S0958344008000724.
[21] G. Lord, "Podcasting communities and second language pronunciation," *Foreign Lang. Ann.*, vol. 41, no. 2, pp. 364–379, Mar. 2008, doi: 10.1111/j.1944-9720.2008.tb03297.x.
[22] M. C. B. Alastuey, "Synchronous-voice computer-mediated communication: Effects on pronunciation," *CALICO J.*, vol. 28, no. 1, pp. 1–20, Jan. 2010, doi: 10.11139/cj.28.1.1-20.
[23] 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, Jan. 2015, doi: 10.1558/cj.v32i1.25962.
[24] R. Hincks, "Speech technologies for pronunciation feedback and evaluation," *ReCALL*, vol. 15, no. 1, pp. 3–20, Jun. 2003, doi: 10.1017/S0958344003000211.
[25] R. Hincks, "Computer support for learners of spoken English," Ph.D. dissertation, KTH Royal Institute Technol., Stockholm, Sweden, 2005.
[26] B. Mak *et al.*, "PLASER: pronunciation learning via automatic speech recognition," in *Proc. HLT-NAACL Conf.*, Edmonton, Canada, May 27–Jun 1, 2003, pp. 23–29, doi: 10.3115/1118894.1118898.
[27] A. Neri, O. Mich, M. Gerosa, and D. Giuliani, "The effectiveness of computer-assisted pronunciation training for foreign language learning by children," *Comput. Assisted Lang. Learn.*, vol. 21, no. 5, pp. 393–408, Nov. 2008, doi: 10.1080/09588220802447651.
[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, Australia, Jun. 1998, pp. 1–4.
[29] C. D. Epp, D. Watanabe, and S. Swann, "Integrating a mobile vocabulary learning tool into a senior high school English class: challenges and opportunities," in *IAmLearning: Mobilizing Supporting Educator Pract.*, R. Power, M. Ally, D. Cristol, and A. Palalas, Eds., 2017, E-book.
[30] M. El Tatawy, "Corrective feedback in second language acquisition," *Working papers TESOL and Appl. Linguistics*, vol. 2, no. 2, pp. 1–19, Oct. 2002.
[31] A. Neri, C. Cucchiarini, H. Strik, and L. Boves, "The pedagogy-technology interface in computer-assisted pronunciation training," *Comput. Assisted Lang. Learn.*, vol. 15, no. 5, pp. 441–467, Aug. 2010, doi: 10.1076/call.15.5.441.13473.
[32] A. Neri, C. Cucchiarini, and H. Strik, "Selecting segmental errors in non-native Dutch for optimal pronunciation training," *IRAL-Int. Rev. Appl. Linguistics Lang. Teaching*, vol. 44, no. 4, pp. 357–404, Dec. 2006, doi: 10.1515/IRAL.2006.016.
[33] M. S. Mirzaei, K. Meshgi, and T. Kawahara, "Automatic speech recognition errors as a predictor of L2 listening difficulties," *CL4LC Workshop*, p. 192, Dec. 11, 2016.
[34] 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. of Kansas, Lawrence, KS, USA, 2009.
[35] 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, Canada, 2002.
[36] D. M. Chun, Y. Jiang, and N. Ávila, "Visualization of tone for learning Mandarin Chinese," in *Proc. 4th PSLLT Conf.*, Vancouver, British Columbia, Canada, Aug. 24-25, 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, Jan. 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, Aug. 2005, 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. Assisted Lang. Learn.*, vol. 17, no. 3–4, pp. 357–376, Aug. 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.
[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.
[42] M. Celce-Murcia and J. M. Goodwin, *Teaching Pronunciation*. London, U.K.: Thomson Learn., 2014.
[43] 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. 16-20, 2000, pp. 677–680.
[44] 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," *The J. Acoustical Soc. America*, vol. 101, no. 4, pp. 2299–2310, Apr. 1997, doi: 10.1121/1.418276.
[45] 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. of Texas Press, Austin, TX, USA, 1997.
[46] E. M. Kissling, "Teaching pronunciation: Is explicit phonetics instruction beneficial for FL learners?" *The Modern Lang. J.*, vol. 97, no. 3, pp. 720–744, Aug. 2013, doi: 10.1111/j.1540-4781.2013.12029.x.
[47] G. Lord, "(How) can we teach foreign language pronunciation? On the effects of a Spanish phonetics course," *Hispania*, vol. 88, pp. 557–567, Sep. 2005, doi: 10.2307/20063159.
[48] D. Liakin, W. Cardoso, and N. Liakina, "The pedagogical use of mobile speech synthesis (TTS): focus on French liaison," *Comput. Assisted Lang. Learn.*, vol. 30, no. 3–4, pp. 325–342, Apr. 2017, doi: 10.1080/09588221.2017.1312463.
[49] C. Munteanu *et al.*, "Hidden in plain sight: low-literacy adults in a developed country overcoming social and educational challenges through mobile learning support tools," *Pers. Ubiquitous Comput.*, pp. 1–15, Aug. 2014, doi: 10.1007/s00779-013-0748-x.
[50] T. Bione, J. Grimshaw, and W. Cardoso, "An evaluation of text-to-speech synthesizers in the foreign language classroom: learners' perceptions,"

in *CALL communities culture – short papers EUROCALL 2016*, St. Raphael Resort/Limassol, Cyprus, 24–27 Aug. 2016, pp. 50–54, doi: 10.14705/rpnet.2016.eurocall2016.537.

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

[52] G. Smith, W. Cardoso, and C. G. Fuentes, "Text-to-speech synthesizers: Are they ready for the second language classroom?" in *Proc. Meeting English Language Teaching*, Concordia University, Montréal, Canada, Oct. 20–24, 2015, doi: 10.14705/rpnet.2015.000318.

[53] R. I. Thomson, "Improving L2 listeners' perception of English vowels: a computer-mediated approach," *Lang. Learn.*, vol. 62, no. 4, pp. 1231–1258, Aug. 2012, doi: 10.1111/j.1467-9922.2012.00724.x.

[54] R. I. Thomson, "Computer-assisted pronunciation training: Targeting second language vowel perception improves pronunciation," *CALICO J.*, vol. 28, no. 3, pp. 744–765, May 2011, doi: 10.11139/cj.28.3.744-765.

[55] D. Litman *et al.*, "Towards using conversations with spoken dialogue systems in the automated assessment of non-native speakers of English," in *Proc. 17th Annu. Meeting Special Interest Group Discourse Dialogue*, Los Angeles, LA, USA, Sep. 13–15, 2016, pp. 270–275, doi: 10.18653/v1/W16-3635.

[56] D. Burleson, "Training segmental productions for second language intelligibility," Ph.D. dissertation, Indiana Univ. Press, Bloomington, IN, USA, 2007.

[57] G. Y. Eksi and S. Yesilcinar, "An investigation of the effectiveness of online text-to-speech tools in improving EFL teacher trainees' pronunciation," *English Lang. Teaching*, vol. 9, no. 2, pp. 205–214, Jan. 2016, doi: 10.5539/elt.v9n2p205.

[58] N. Moustroufas and V. Digalakis, "Automatic pronunciation evaluation of foreign speakers using unknown text," *Comput. Speech Lang.*, vol. 21, no. 1, pp. 219–230, Jan. 2007, doi: 10.1016/j.csl.2006.04.001.

[59] S. M. Witt and S. J. Young, "Phone-level pronunciation scoring and assessment for interactive language learning," *Speech Commun.*, vol. 30, no. 2, pp. 95–108, Feb. 2000, doi: 10.1016/S0167-6393(99)00044-8.

[60] M. Eskenazi and G. Pelton, "Pinpointing pronunciation errors in children's speech: examining the role of the speech recognizer," in *ISCA Tutorial Resources Workshop Pronunciation Model. Lexicon Adaptation Spoken Language Technologies*, Colorado, CO, USA, Sep. 14–15, 2002, p. 48–52.

[61] S. Pakhomov, J. Richardson, M. Finholt-Daniel, and G. Sales, "Forced-alignment and edit-distance scoring for vocabulary tutoring applications," in *Int. Conf. Text, Speech and Dialogue*, Brno, Czech Republic, Sep. 8–12, 2008, pp. 443–450, doi: 10.1007/978-3-540-87391-4_57.

[62] C. D. Epp, "ProTutor: A pronunciation tutor that uses historic open learner models," Master's thesis, Univ. Saskatchewan, Saskatoon, Saskatchewan, Canada, 2010.

[63] E. Cámara-Arenas, *Native Cardinality: on teaching American English vowels to Spanish students*, ser. Historia y sociedad. Ediciones Univ. Valladolid, 2013.

[64] E. Cámara-Arenas, "The NCM and the reprogramming of latent phonological systems: A bilingual approach to the teaching of English sounds to Spanish students," *Procedia - Social Behav. Sci.*, vol. 116, pp. 3044–3048, Feb. 2014, doi: 10.1016/j.sbspro.2014.01.704.

[65] A. Baker, S. Goldstein, and P. Dolgin, *Pronunciation Pairs: An Introductory Course for Students of English. Student's Book*. Cambridge, U.K.: Cambridge Univ. Press, 1990.

[66] A. Baker and L. Marshall, *Ship or sheep?* Cambridge, U.K.: Cambridge Univ. Press, 1981.

[67] J. D. O'Connor and C. Fletcher, *Sounds English-A pronunciation practice book*. Harlow, U.K.: Longman Group U.K., 1999.

[68] Y. Sheen and R. Ellis, "Corrective feedback in language teaching," in *Handbook Res. Second Lang. Teaching Learn.* New York, NY: Routledge, 2011, vol. 2, pp. 593–610.

[69] P. Kerswill and A. Williams, "New towns and koineization: linguistic and social correlates," *Linguistics*, vol. 43, no. 5, pp. 1023–1048, Sep. 2005, doi: 10.1515/ling.2005.43.5.1023.

[70] S. Loewen, "Focus on form," *Handbook Res. Second Lang. Teaching Learn.*, vol. 2, pp. 576–592, 2011.

[71] R. L. Oxford, "Language learning styles and strategies," in *Teaching English as a Second or Foreign Lang.*, M. Celce-Murcia, Ed. Heinle & Heinle, 2001, pp. 359–366.

[72] I. P. Association, I. P. A. Staff *et al.*, *Handbook of the International Phonetic Association: A guide to the use of the International Phonetic Alphabet*. Cambridge, U.K.: Cambridge Univ. Press, 1999.

[73] List of minimal pairs included in the CAPT tool English Vowels for Spanish Speakers. GitHub (eca-simm_minimal-pairs-enus-eses), 2019. [Online] Available: ttps://github.com/eca-simm/minimal-pairs-enus-eses - (Accessed: Nov. 3, 2019).

[74] K. Nader, "Reconsolidation and the dynamic nature of memory," *Cold Spring Harbor Perspectives Biol.*, vol. 7, no. 10, p. 1–20, Sep. 2015, doi: 10.1101/cshperspect.a021782.

[75] NoxPlayer - Free Android Emulator on PC and Mac, 2019. [Online] Available: https://www.bignox.com/ (Accessed: Nov. 2, 2019).

[76] R. Prabhavalkar, K. Rao, T. N. Sainath, B. Li, L. Johnson, and N. Jaitly, "A comparison of sequence-to-sequence models for speech recognition," in *Proc. Interspeech*, Stockholm, Sweden, Aug. 20–24, 2017, pp. 939–943, doi: 10.21437/Interspeech.2017-233.

[77] Y. Zhang, W. Chan, and N. Jaitly, "Very deep convolutional networks for end-to-end speech recognition," in *IEEE Int. Conf. Acoustic Speech, Signal, and Processing*, Mar. 5–9, 2017, pp. 4845–4849, doi: 10.1109/ICASSP.2017.7953077.

Cristian Tejedor-García is a Ph.D. student in Computer Science. He received the B.Sc., and the M.Sc. degrees with honours in computer science in 2014 and 2016, respectively, from the University of Valladolid. He currently works in the ECA-SIMM research group of the same university with a pre-doctoral grant. His research interests include speech technology, learning games and human-computer-interaction (HCI). He has published and achieved international awards.

David Escudero-Mancebo received the B.A. and the M.Sc. degrees in computer science in 1993 and 1996; and the Ph.D. degree in information technologies in 2002 from the University of Valladolid, Spain. He is an Associate Professor of computer science in the University of Valladolid. He is co-author of several publications in the field of computational prosody (modeling of prosody for TTS systems and labeling of corpora).

Enrique Cámara-Arenas has been teaching English phonetics and pronunciation for the last twenty years at the University of Valladolid. He has designed the NCM, for teaching English sounds to L1 Spanish students of English. He is also interested in the exploration and curricular integration of English grapho-phonemics within EFL environments.

César González-Ferreras received the B.Sc., M.Sc., and Ph.D. degrees in computer science from the University of Valladolid, Spain, in 1998, 2000, and 2009, respectively. He is an Assistant Professor at the Department of Computer Science at the University of Valladolid. His research interests include human–computer interaction, spoken language processing and prosody recognition.

Valentín Cardeñoso-Payo (M'02) received the M.Sc. and the Ph.D. in physics from the University of Valladolid, Spain. He has been the ECA-SIMM group director since 1998. His current research interests include machine learning techniques applied to human language technologies, HCI 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.