IberSPEECH 2018
21-23 November 2018, Barcelona, Spain

Improving Pronunciation of Spanish as a Foreign Language for L1 Japanese Speakers with Japañol CAPT Tool

Cristian Tejedor-García[1], *Valentín Cardeñoso-Payo*[1], *María J. Machuca*[2], *David Escudero-Mancebo*[1], *Antonio Ríos*[2], *Takuya Kimura*[3]

[1]Department of Computer Science, University of Valladolid, Spain
[2]Department of Spanish Philology, Autonomous University of Barcelona, Spain
[3]Department of Spanish Language and Literature, Seisen University, Japan

cristian@infor.uva.es

Abstract

Availability and usability of mobile smart devices and speech technologies ease the development of language learning applications, although many of them do not include pronunciation practice and improvement. A key to success is to choose the correct methodology and provide a sound experimental validation assessment of their pedagogical effectiveness. In this work we present an empirical evaluation of Japañol, an application designed to improve pronunciation of Spanish as a foreign language targeted to Japanese people. A structured sequence of lessons and a quality assessment of pronunciations before and after completion of the activities provide experimental data about learning dynamics and level of improvement. Explanations have been included as corrective feedback, comprising textual and audiovisual material to explain and illustrate the correct articulation of the sounds. Pre-test and post-test utterances were evaluated and scored by native experts and automatic speech recognition, showing a correlation over 0.86 between both predictions. Sounds [s], [fl], [ɾ] and [s], [fɾ], [θ] explain the most frequent perception and production failures, respectively, which can be exploited to plan future versions of the tool, including gamified ones. Final automatic scores provided by the application highly correlate (r>0.91) to expert evaluation and a significant pronunciation improvement can be measured.
Index Terms: Computer-Assisted Pronunciation Training, corrective feedback, pronunciation training, second language learning, speech recognition, speech synthesis, minimal pairs

1. Introduction

Steady advances in automatic speech recognition (ASR) and speech synthesis (TTS) lead to an increase of their use in Computer-Assisted Pronunciation Training (CAPT) tools [1]. Availability and usability of mobile smart devices motivates the development of foreign language learning applications [2, 3], some of them include pronunciation practice and improvement as the main focus. Although existing literature on CAPT indicate that these systems tell us how good learners perform and how to improve their pronunciation [4, 5], choosing the correct methodology and elaborating a sound experimental validation of the assessment of pedagogical effectiveness of these tools are essential to further develop the field [6, 7].

Japañol[1] is a CAPT tool we have specifically designed for native Japanese learners of Spanish as a foreign language. Recent studies suggest that these learners mispronounce vowels and consonants [8, 9], one of the most frequent errors being the substitution of Spanish round back high vowel [u] by Japanese unrounded central-back vowel [ɯ]. Consonant mispronunciations usually stem from different phonological processes in both languages: phoneme substitutions between allophonic variants or changes in the place of articulation are the most common errors. For instance, [l] and [ɾ] are allophones of the same rhotic phoneme /r/ in Japanese. By contrast, both sounds are different phonemes in Spanish. Therefore, the trill Spanish phoneme /r/ is often pronounced by Japanese speakers as a flap ([ɾ]) or even as a lateral ([l]) [10].

Most frequent mistakes are found in fricatives: [f] is often confused with [x], especially when these sounds are followed by [u]. In this way, two different Spanish words "*fuego*" (fire) and "*juego*" (play) can be pronounced in the same way by Japanese speakers (/f/ is realised as labiodental and /x/ as velar) and none of these sounds exist in Japanese. The only similar fricative phoneme is /h/, and its realization depends on the vocalic context, for example, [ç] appears before palatal vowel or semivowel (/i/ or /y/) and [ɸ], before velar one [ɯ]. [θ] is another fricative sound which causes pronunciation problems. The substitution of this phoneme by [s] is a very common mistake, a phenomenon known as Asian "seseo" (lisp) [11], although it should not be considered a characteristic of non-nativeness, since this pronunciation is accepted in Spanish American variants [12]. Finally, according to phonotactic rules, Spanish allows consonant followed by consonant in onset clusters (i.e., '**pr**ado´) and coda ones (i.e., '**ins**cripción´). The combination of fricative /f/ and /l/ or /r/ in onset also triggers mispronunciation. This problem is not only due to the fact that these consonants exhibit errors in simple onset position, but also because most of the Japanese syllables have a consonant followed by vowel structure.

In previous works, we presented the development of serious games for training foreign pronunciation based on minimal pairs tasks [13, 14]. We were able to assess user's pronunciation level in a L2, that is, those with a certified higher level consistently reached better scores in the game [15]. Besides, we also have discussed that the introduction of corrective feedback [16] allowed us to confirm that there was pronunciation improvement among users after the first stages of use. However, continued use of the tool seemed to invariably lead to stagnation. Finally, while the freedom of movement on game-oriented tools lead users to boost their score by repeating those tasks they found easy, learning-oriented tools insist on users' difficulties offering guided and corrective feedback and achieving better effectiveness and efficiency pedagogical results [17].

[1]This study has been partially supported by the Ministerio de Economía y Empresa (MINECO) and the European Regional Development Fund FEDER (TIN2014-59852-R) and by Consejería de Educación of Junta de Castilla y León (VA050G18).

10.21437/IberSPEECH.2018-21

In this paper we present an empirical evaluation of effectiveness in pronunciation training in Japanese students of Spanish as a foreign language. Corrective feedback in pronunciation training of second language acquisition could be implicit or explicit, in terms of whether or not the learner is informed of the corrected form of the error [18]. Since it is recognized as an essential component of these kind of systems, an explanation mode has been included, comprising textual and audiovisual material which provides explanations and illustrations on the correct articulation of the sounds included in each lesson. We also provide a guided path to users to follow in order to complete all activities, based on their results. The preferable characteristics of corrective feedback are: unambiguous, understandable, detectable, short and should preferably take account of learner characteristics, both proficiency and literacy level [19]. We provide different types of feedback depending on the tasks and results (see subsection 2.2).

In section 2, the experimental procedure is described, which includes the participants and protocol stages, the speech material processing and Japañol description. Results section shows users' interaction degree with the CAPT system, the most difficult Spanish sounds found in perception and production tasks and user's human rater and ASR scores consistency, correlation and improvement. We end the paper with a discussion about the relevance of the results and the conclusions and future work.

2. Experimental procedure

2.1. Protocol description

A total of eight native Japanese students between 20 and 22 years old were selected as participants for our experiment. All participants qualified for the same Spanish as foreign language for beginners (A2-B1) course at the Language Center of the University of Valladolid. In this way we guaranteed (1) that all students had the same initial level of Spanish and (2) that our experiment realistically reproduced the variety of persons that attend to these type of courses.

Systematization of Spanish mispronunciations produced by Japanese speakers was the first step to choose minimal pairs in Spanish [20]. A set of 56 words, chosen according to the pairs to be worked out along the game and selected according to their phonetic difficulty, where spoken by each of the 8 participants, both before (pre-test) and after (post-test) training sessions. Participants exclusively used the CAPT system in the three 45-minutes-maximum training sessions; with a delay of at least 48 hours. A total of 84 minimal pairs corresponding to mispronunciations were presented to participants, 12 in each of the seven lessons. Students were asked not to complete more than three lessons per session. The software application was installed under an Android emulator (NOX App player[2]) in the computers of the Language Center multimedia laboratory. At the beginning of the first session, students were instructed on how to use the software. Then, they had to work individually and did not have any communication with either instructors or classmates. Each participant worked inside an individual cubicle equipped with a headset with microphone.

Once finished the post-test, a manual revision and isolation of the recorded words of pre and post-test was carried out in order to elaborate a perceptual listening test. Five expert phoneticians and native speakers assigned a correct/incorrect value to each word, plus some extra annotations about the pronunciation which are not used in this work. All data was presented to human raters randomly and without user association. The audio files were also processed with the same ASR used in the application (*Google* ASR). Human raters were asked to concentrate in the specific sound of each word which should be generated correctly, ignoring the bad pronunciations of the rest of sounds.

[2]Official website (last visited June, 24th 2018) https://www.bignox.com/

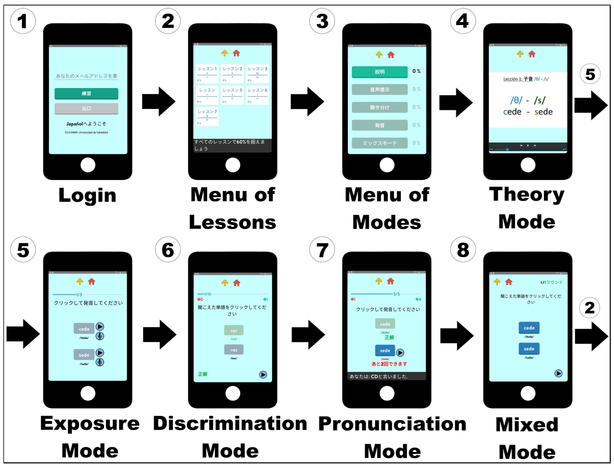

Figure 1: *Standard flow to complete a lesson in Japañol.*

2.2. Software tool of the CAPT system

Figure 1 shows the regular sequence of steps in order to complete a lesson in Japañol. After user's authentication in (step 1), seven lessons are presented at the main menu of the application (step 2). Each lesson includes a pair of Spanish sound contrasts and users achieve a particular score, expressed in percentage. Lessons are divided in five main training modes: Theory, Exposure, Discrimination, Pronunciation and Mixed modes (step 3) in which each one proposes several task-types with a fixed number of mandatory task-tokens. The final lesson score is the mean score of the last three modes. Users are guided by the system in order to complete all training modes of a lesson. When reaching a score below 60% in Discrimination, Pronunciation or Mixed modes, users are recommended to go back to exposure mode as a feedback resource and then return to the failed mode. Besides, next lesson is enabled when users reach a minimum score of 60%.

The first training mode is **Theory** (step 4). A brief and simple video describing the target contrast of the lesson is presented to the user as a first contact with feedback. At the end of the video, next mode becomes available; but users may choose to review the material as many times as they want. **Exposure** (step 5) is the second mode. Users strengthen the lesson contrast experience previously introduced in Theory mode, in order to support their assimilation. Three minimal pairs are displayed to the user. In each one of them, both words are synthetically produced by *Google* TTS for five times (highlighting the current word), alternately and slowly. After that, users must record themselves at least one time per word and listen to their own and system's sound. Words are represented with their orthographic and phonemic forms. A replay button allows to listen to the specified word again. Step 6 refers to **Discrimination** mode, in which ten minimal pairs are presented to the user consecutively. In each one of them, one of the words is synthetically produced, randomly. The challenge of this mode consists on identifying which word is produced. As feedback elements,

Table 1: *Events by user with the CAPT tool in the whole experiment. THE, EXP, DIS, PRO and MIX correspond to Theory, Exposure, Discrimination, Pronunciation and Mixed modes, respectively. $\overline{n}$, m and M are the mean, minimum and maximum values. Time (min) row represents the time spent in minutes per person in each mode in the whole experiment. #Tries is the number times a mode is practiced by each user. Mand. and Req. mean mandatory and requested TTS listenings. Productions use ASR.*

	THE			EXP			DIS			PRO			MIX		
	$\overline{n}$	m	M	$\overline{n}$	m	M	$\overline{n}$	m	M	$\overline{n}$	m	M	$\overline{n}$	m	M
Time (min)	16.96	11.13	20.75	18.73	12.77	21.90	7.26	6.03	8.93	39.82	22.43	72.85	17.35	9.78	21.43
#Tries	8.25	7	10	11.38	10	14	10.16	8	15	10.38	8	14	7.38	4	10
#Mand.listenings	-	-	-	308.75	220	390	94.75	80	134	-	-	-	22.13	12	30
#Req.listenings	-	-	-	97.5	70	126	35.0	0	82	43.13	0	99	27.25	9	52
#Recordings	-	-	-	64.6	47	81	-	-	-	-	-	-	-	-	-
#Discriminations	-	-	-	-	-	-	94.75	80	134	-	-	-	22.13	12	30
#Productions	-	-	-	-	-	-	-	-	-	166.5	124	242	80.63	38	113

words have their orthographic and phonetic transcription representations. Users can also request listen to the target word again with a replay button. Speed varies alternately between slow and normal speed rates. Finally, the system changes word color to green (success) or red (failure) with a chime sound. **Pronunciation** is the fourth mode (step 7) which aim is to produce as well as possible, both words, separately, of the five minimal pairs presented with their phonetic transcription. *Google*'s ASR determines automatically and in real time acceptable or non-acceptable inputs. In each production attempt the tool displays a text message with the recognized speech, plays a right/wrong sound and changes word's color to green or red. The maximum number of attempts per word is five in order not to discourage users. However, after three consecutive failures, the system offers to the user the possibility of request a word synthesis as an explicit feedback as many times as they want with a replay button. **Mixed mode** is the last mode of each lesson (step 8). Nine production and perception tasks alternate at random in order to further consolidate obtained skills and knowledge.

3. Results

Table 1 shows user's interaction degree with Japañol. After all training sessions, Japanese learners spent an average of 100.12 minutes performing the proposed tasks. Users consumed a 83.06% of the mean effective time for carry out interactive exercises in EXP, DIS, PRO and MIX modes. As a mean term, users listened to the TTS system 628.51 times and used the ASR system 247.13 times, giving a rate of 8.74 uses of the TTS/ASR per minute. Table 1 also shows important differences in the use of the tool depending on the user. For instance, the user who performed tasks of the PRO mode in the fastest way spent 22.43 minutes and the one who spent more time 72.85 minutes. This contrast can also 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 the users make use of the ASR (162 minimum vs. 355 maximum) and the number of times they requested the use of TTS (79 vs. 359 times).

Tables 2 and 3 display the confusion matrices between the sounds of the minimal pairs in perception and production events. The most confused pair in discrimination tasks is [l]-[ɾ] (54 times) and the least confused one is [θ]-[f] (5 times). The sounds with the lowest pronunciation success rate are [s] and [fl] (both $<$ 71%), and those with the highest pronunciation success rate are [f] and [θ] (both $>$ 92%), corresponding to the lowest number of requested listenings (5 and 20).

Table 3 shows pronunciation events results at first and last attempt per word utterance. A produced word is correct if ASR

Table 2: *Confusion matrix of discrimination task-tokens (diagonal: right discrimination task-tokens). The rows are the sounds expected by the tool and the columns are the sounds selected by the user. $\overline{success}$ refers to the success rate of the corresponding sound row. #Lis is the number of requested listenings to the sound row.*

		Discrimination task-tokens										
#Lis		[fl]	[fɾ]	[l]	[ɾ]	[rr]	[s]	[θ]	[f]	[fu]	[xu]	$\overline{success}$
28	[fl]	***55***	23	-	-	-	-	-	-	-	-	70.5%
33	[fɾ]	20	***57***	-	-	-	-	-	-	-	-	74.0%
62	[l]	-	-	***113***	25	9	-	-	-	-	-	76.9%
71	[ɾ]	-	-	29	***112***	14	-	-	-	-	-	72.3%
29	[rr]	-	-	6	18	***92***	-	-	-	-	-	79.3%
27	[s]	-	-	-	-	-	***45***	19	-	-	-	70.3%
20	[θ]	-	-	-	-	-	6	***109***	3	-	-	92.4%
5	[f]	-	-	-	-	-	-	2	***49***	-	-	96.1%
40	[fu]	-	-	-	-	-	-	-	-	***50***	18	73.5%
30	[xu]	-	-	-	-	-	-	-	-	15	***46***	75.4%

results provide in first place the desired word. Users can try a maximum of five wrong productions per word. We can observe a success improvement from first to last attempt (last column), being the highest ones [fl] (48.4%) and [fɾ] (27.7%) sounds. At first attempt, the most confused pair in production tasks is [fl]-[fɾ] (88 times) and the least confused one is [l]-[rr] (20 times). The sounds with the lowest discrimination success rate are [fl] and [s] (both $<$ 35%), and those with the highest discrimination success rates are [l] and [ɾ] (both $>$ 75%). At last attempt, the most confused pair in production tasks is [s]-[θ] (102 times) and the least confused one is [l]-[ɾ] (3 times). The sounds with the lowest production success rate are [s] and [fɾ] (both $<$ 65%). A higher number of requested listenings (first column) appears for last attempt pronunciation at lower success rates. The highest production rate success sounds are [l] and [ɾ] (both $>$ 96%).

Table 4 presents the scores for each user at any of the given stages of the experiment (pre-test, CAPT tool, post-test, and a delta score of pre and post-test). EXP and ASR scores refer to both tests learners' qualifications by human raters and *Google* ASR, respectively. These scores are computed by summing up the number of correct words per speaker and normalizing the result to the range [0,10]. JAP score is computed by the number of correct and incorrect task-tokens while doing the required task-types of the training modes (Discrimination, Pronunciation and Mixed modes) as a qualification to rank the participants.

Concerning to EXP values, a consistency test among human raters based on the Fleiss' Kappa statistical indicator was carried out both for pre-test and post-test evaluations. For lo-

Table 3: *Confusion matrix of pronunciation task-tokens at first and last attempt per word sequence (diagonal: right pronunciation task-tokens at first and last attempt per word sequence). The rows are the sounds expected by the tool and the columns are the sounds produced by the user.* $\overline{success}$ *refers to the success rate of the corresponding sound row. #Lis is the number of requested listenings to the sound row.*

<table>
<tr><th colspan="13">Pronunciation task-tokens (first attempt | last attempt)</th></tr>
<tr><th>#Lis</th><th></th><th>[fl]</th><th>[fr]</th><th>[l]</th><th>[ɾ]</th><th>[rr]</th><th>[s]</th><th>[θ]</th><th>[f]</th><th>[fu]</th><th>[xu]</th><th>$\overline{success}$</th></tr>
<tr><td>11 | 70</td><td>[fl]</td><td>***17 | 85***</td><td>48 | 29</td><td>-</td><td>-</td><td>-</td><td>-</td><td>-</td><td>-</td><td>-</td><td>-</td><td>26.2% | 74.6%</td></tr>
<tr><td>1 | 59</td><td>[fr]</td><td>40 | 38</td><td>***23 | 68***</td><td>-</td><td>-</td><td>-</td><td>-</td><td>-</td><td>-</td><td>-</td><td>-</td><td>36.5% | 64.2%</td></tr>
<tr><td>3 | 15</td><td>[l]</td><td>-</td><td>-</td><td>***94 | 128***</td><td>11 | 2</td><td>12 | 3</td><td>-</td><td>-</td><td>-</td><td>-</td><td>-</td><td>80.3% | 96.2%</td></tr>
<tr><td>1 | 14</td><td>[ɾ]</td><td>-</td><td>-</td><td>10 | 1</td><td>***89 | 126***</td><td>19 | 4</td><td>-</td><td>-</td><td>-</td><td>-</td><td>-</td><td>75.4% | 96.2%</td></tr>
<tr><td>3 | 36</td><td>[rr]</td><td>-</td><td>-</td><td>8 | 4</td><td>25 | 8</td><td>***84 | 124***</td><td>-</td><td>-</td><td>-</td><td>-</td><td>-</td><td>71.8% | 91.2%</td></tr>
<tr><td>6 | 92</td><td>[s]</td><td>-</td><td>-</td><td>-</td><td>-</td><td>-</td><td>***20 | 64***</td><td>38 | 48</td><td>-</td><td>-</td><td>-</td><td>34.5% | 57.1%</td></tr>
<tr><td>2 | 120</td><td>[θ]</td><td>-</td><td>-</td><td>-</td><td>-</td><td>-</td><td>41 | 54</td><td>***63 | 114***</td><td>24 | 7</td><td>-</td><td>-</td><td>49.2% | 65.1%</td></tr>
<tr><td>0 | 9</td><td>[f]</td><td>-</td><td>-</td><td>-</td><td>-</td><td>-</td><td>-</td><td>24 | 8</td><td>***40 | 56***</td><td>-</td><td>-</td><td>62.5% | 87.5%</td></tr>
<tr><td>3 | 62</td><td>[fu]</td><td>-</td><td>-</td><td>-</td><td>-</td><td>-</td><td>-</td><td>-</td><td>-</td><td>***35 | 58***</td><td>22 | 25</td><td>61.4% | 69.9%</td></tr>
<tr><td>2 | 29</td><td>[xu]</td><td>-</td><td>-</td><td>-</td><td>-</td><td>-</td><td>-</td><td>-</td><td>-</td><td>19 | 14</td><td>***34 | 64***</td><td>64.2% | 82.1%</td></tr>
</table>

Table 4: *Different user scores in pre-test, post-test and game by experts, ASR and game in a scale of [0, 10]. ID, EXP, ASR and JAP refer to user identifier, experts, Google ASR and Japañol, respectively.*

	Pre-test		Game	Post-test		Δ (Pre/Post)	
ID	EXP	ASR	JAP	EXP	ASR	EXP	ASR
07	9.8	5.1	9.4	9.4	7.5	-0.4	2.4
05	8.6	4.5	9.1	8.8	4.6	0.2	0.2
01	8.1	2.9	8.0	7.9	4.6	-0.3	1.7
02	7.3	3.7	8.4	7.8	4.8	0.4	1.2
03	6.9	1.9	6.3	7.6	2.7	0.8	0.8
08	6.8	2.8	6.5	7.1	3.4	0.4	0.6
06	5.8	0.8	5.9	6.5	3.2	0.8	2.4
04	5.4	2.1	6.9	7.3	2.9	1.9	0.8

gistic reasons, post-test evaluation was carried out first and a moderate agreement [21] was found (Kappa value=0.50). There is a considerable increment for pre-test, reaching a substantial agreement (Kappa value=0.63). Comparing subjective and objective scores, delta score values (last two columns) are positive in almost all cases both in EXP and ASR (only two of them decrease in EXP, both in top three). They also show a fair correlation with pre-test expert scoring (r=-0.856) and post-test expert scores with ASR ones (r=-0.735). Pre-test scores assigned by experts (EXP) show reasonable linear regression correlation with those obtained by applying the ASR in the same test (r=0.890). A similar correlation is found for post-test EXP and ASR (r=0.834). Game scores (JAP) show good correlation with EXP post-test results (r=0.912), clearly over the correlation found between JAP and pre-test human rater results (r=0.867).

4. Discussion and Conclusion

We have described the experiment results with a CAPT system of native Japanese learners of Spanish as a foreign language, that include a software tool, Japañol, that we believe is worth taking into account when thinking about possible teaching complement. The results have shown that Japanese learners of Spanish have difficulty with [f] in onset clusters position in both perception (Table 2) and production (Table 3). [s]-[θ] present similar results, they tend to substitute [θ] by [s], but this pronunciation is accepted in Latin American Spanish. About liquid consonants, Japanese speakers are more successful at phonetically producing [l] and [ɾ] than discriminating these phonemes. Japanese speakers have already acquired these sounds since they are allophones of a same liquid phoneme in Japanese. For this reason, it does not seem to be necessary to distinguish them in Japanese, whereas it is in Spanish.

Applied methodology has proved coherent and efficient because easier tasks (exposure, discrimination) were presented before tougher ones (pronunciation) in time and repetition terms (Table 1). It has also led to better final scores (Table 4). Participants consistently resorted to TTS models when faced with difficulties both in perception and production modes (#Lis column of Tables 2 and 3 and #Req.List.row of Table 1). The fact that our CAPT tool makes use of ASR and TTS systems is the principal pedagogical and operational concern. The quality of synthetic voice in the rendering of minimal pairs appears to have been adequate. Judging from gathered data, the TTS system employed seems to be beneficial for students [22, 23]: the rate of success significantly increases after undertaking the exposure activities imposed by feedback. The role of ASR is even more crucial as it offers diagnosis to users as real-time automatic feedback. Nowadays, ASR systems have some limitations such as isolated or infrequent words, adaptation to technology and L1 transferred pronunciation utterances as correct L2 words (false positives). However, [24] demonstrated the effectiveness of ASR-based CAPT tool for training users in the production of decontextualized isolated words and [25] reported L2 French vowel /y/ production improvement after training with a mobile ASR system.

Human raters' post-test scores fairly correlate with ASR and game ones, being useful in the future to be able to evaluate a large amount of users reducing human costs (Table 4). The lowest pre-test scores' users improved more than the best ones. However, they did not reach better results than the top ones. This is due to the fact that the tool does not give extra activities when the limit is reached. As a future work, we are collaborating with Seisen University with a bigger group of students divided into experimental, control and in-classroom groups. We are also considering the possibility of applying the methodology to other more exotic and minority languages. Finally, a comparison between a game-oriented version versus a learning-oriented one could be our next step.

5. References

[1] R. I. Thomson and T. M. Derwing, "The effectiveness of L2 pronunciation instruction: A narrative review," *Applied Linguistics*, vol. 36, no. 3, p. 326, 2014.

[2] A. Kukulska-Hulme, "Language learning defined by time and place: A framework for next generation designs," in *Left to My Own Devices: Learner Autonomy and Mobile Assisted Language Learning*, ser. Innovation and Leadership in English Language Teaching, J. E. Díaz-Vera, Ed. Bingley, UK: Emerald Group Publishing Limited, 2012, vol. 6, pp. 1–13. [Online]. Available: http://oro.open.ac.uk/30756/

[3] A. Pareja-Lora, C. Calle-Martínez, and P. Rodríguez-Arancón, *New perspectives on teaching and working with languages in the digital era*. Research-publishing. net, 2016.

[4] W. Li and D. Mollá-Aliod, "Computer processing of oriental languages. language technology for the knowledge-based economy," *Lecture Notes in Computer Science*, vol. 5459, 2009.

[5] M. Carranza, "Diseño de aplicaciones para la práctica de la pronunciación mediante dispositivos móviles y su incorporación en el aula de ele," *El español entre dos mundos: Estudios de ELE en Lengua y Literatura*, pp. 279–297, 2014.

[6] 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, 2008.

[7] J. Lee, J. Jang, and L. Plonsky, "The Effectiveness of Second Language Pronunciation Instruction: A Meta-Analysis," *Applied Linguistics*, vol. 36, no. 3, p. 345, 2015.

[8] M. Carranza, "Errores y dificultades específicas en la adquisición de la pronunciación del español le por hablantes de japonés y propuestas de corrección," *Nuevos enfoques en la enseñanza del español en Japón*, pp. 51–78, 2012.

[9] A. Blanco-Canales and M. Nogueroles-López, "Descripción y categorización de errores fónicos en estudiantes de español/l2. validación de la taxonomía de errores aacfele," *Logos: Revista de Lingüística, Filosofía y Literatura*, vol. 23, no. 2, pp. 196–225, 2013.

[10] 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 Journal of the Acoustical Society of America*, vol. 101, no. 4, pp. 2299–2310, 1997.

[11] J. M. Cabezas Morillo, "Las creencias de los estudiantes japoneses sobre la pronunciación española: un análisis exploratorio," *Trabajo de Máster, Universidad Pablo de Olavide*, 2009. [Online]. Available: http://www.mecd.gob.es/dam/jcr:5ebaa0b9-7b77-4a28-a732-b743be08dd06/2010-bv-11-05cabezas-pdf.pdf

[12] M. Carranza, C. Cucchiarini, J. Llisterri, M. J. Machuca, and A. Ríos, "A corpus-based study of spanish L2 mispronunciations by Japanese speakers," in *Edulearn14 Proceedings. 6th International Conference on Education and New Learning Technologies*. IATED Academy, 2014, pp. 3696–3705. [Online]. Available: http://liceu.uab.cat/~joaquim/publicacions/Carranza_et_al_14_Corpus_Spanish_L2.pdf

[13] 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!" *Interspeech*, pp. 1178–1179, September 2016.

[14] C. Tejedor-García, D. Escudero-Mancebo, E. Cámara-Arenas, C. González-Ferreras, and V. Cardeñoso-Payo, "Improving L2 production with a gamified computer-assisted pronunciation training tool, TipTopTalk!" *IberSPEECH 2016: IX Jornadas en Tecnologías del Habla and the V Iberian SLTech Workshop events*, pp. 177–186, 2016.

[15] 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," *SLaTE*, pp. 125–130, 2015.

[16] C. Tejedor-García, V. Cardeñoso-Payo, E. Cámara-Arenas, C. González-Ferreras, and D. Escudero-Mancebo, "Playing around minimal pairs to improve pronunciation training," *IFCASL 2015*, 2015.

[17] 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," *SLaTE*, 2017.

[18] Y. Sheen and R. Ellis, "Corrective feedback in language teaching," *Handbook of research in second language teaching and learning*, vol. 2, pp. 593–610, 2011.

[19] B. Penning de Vries, C. Cucchiarini, H. Strik, and R. van Hout, "The Role of Corrective Feedback in Second Language Learning: New Research Possibilities by Combining CALL and Speech Technology," *Proceedings of SlaTE 2010, Tokyo, Japan*, nov 2010.

[20] C. Tejedor-García and D. Escudero-Mancebo, "Uso de pares mínimos en herramientas para la práctica de la pronunciación del español como lengua extranjera," *Revista de la Asociación Europea de Profesores de Español. El español por el mundo*, no. 1, pp. 355–363, 2018.

[21] J. R. Landis and G. G. Koch, "The measurement of observer agreement for categorical data," *biometrics*, pp. 159–174, 1977.

[22] G. Smith, W. Cardoso, and C. G. Fuentes, "Text-to-speech synthesizers: Are they ready for the second language classroom?" *Concordia University*, 2015.

[23] D. Liakin, W. Cardoso, and N. Liakina, "The pedagogical use of mobile speech synthesis (TTS): focus on French liaison," *Computer Assisted Language Learning*, vol. 30, no. 3-4, pp. 325–342, 2017.

[24] A. Neri, O. Mich, M. Gerosa, and D. Giuliani, "The effectiveness of computer assisted pronunciation training for foreign language learning by children," *Computer Assisted Language Learning*, vol. 21, no. 5, pp. 393–408, 2008.

[25] D. Liakin, W. Cardoso, and N. Liakina, "Learning L2 pronunciation with a mobile speech recognizer: French /y/," *CALICO Journal*, vol. 32, no. 1, p. 1, 2015.