Implementation and test of a serious game based on minimal pairs for pronunciation training

David Escudero-Mancebo[1] *, Enrique Cámara-Arenas*[2]*,*
Cristian Tejedor-García[1]*, César González-Ferreras*[1]*, Valentín Cardeñoso-Payo*[1]

[1]Department of Computer Science [2]Department of English Philology
Universidad de Valladolid
descuder@infor.uva.es

Abstract

This paper introduces the architecture and interface of a serious game intended for pronunciation training and assessment for Spanish students of English as second language. Users will confront a challenge consisting in the pronunciation of a minimal-pair word battery. Android ASR and TTS tools will prove useful in discerning three different pronunciation proficiency levels, ranging from basic to native. Results also provide evidence of the weaknesses and limitations of present-day technologies. These must be taken into account when defining game dynamics for pedagogical purposes.
Index Terms: Computer assisted pronunciation training

1. Introduction

Speech technologies have proved to constitute useful resources in the field of second language learning and pronunciation improvement [1, 2, 3]. Using text-to-speech conversion systems (TTS), students may be easily and instantly exposed to model pronunciations of the words of a language [4]. Also, automatic speech recognition systems (ASR) designed for the use of natives, may indirectly help to filter inadequate (non-recognizable) pronunciations produced by non-natives. Non-natives faced with such ASR devices will consciously strive to make themselves understood [5, 6]. Most of the systems referred to in the state of the art section use TTS and ASR applications that have been adapted to deal with the pronunciation of L2 students. In fact, some of them have been trained ad-hoc to confront non-native speech. However, operating systems nowadays provide free access to their general purpose TTS and ASR services so that these resources may be integrated in applications. In this paper, we present an entertainment application for pronunciation training/assessment that uses native Android ASR and TTS APIs.

By virtue of their transportability, the popularization of smartphone and tablet terminals has also contributed to the expansion of the range of technological services available for users [7]. Applications for language learning and pronunciation improvement have also proliferated, often linking their services to online courses [8, 9]. However, online courses register high drop-out rates, and it is now known that many people will abandon such services after a few uses [10]. Service gamification attempts have been made in order to lessen abandonment by designing attractive applications that generate pleasant and beneficial attachment [3]. There exist good examples of games that have been designed for learning language in the state of the art: [11] presents a game for vocabulary acquisition, [12] a game for practicing oral skills. We have designed an application that challenges the user by assigning a score to their pronunciation, so that an improvement of the score represents an objective betterment of their skills. As we are about to show, this challenge will also help us ascertain the efficacy of a particular ASR system as a tool for assessing the quality of users' pronunciation and the adequateness of TTS systems in providing users with pronunciation models.

In our application, pronunciation challenges are presented in the form of minimal pairs [13]. From a pedagogical point of view, the use of minimal pairs promotes the users' awareness of the the potential risks of producing the wrong meanings when the correct phonemes are not properly executed. Distinguishing between the words that compose the minimal pairs constitutes, a priori, a difficult task for the ASR system as the phonetic distance between each couple of words is small. Thus, they are easily confused if the pronunciation is not sufficiently clear. The presentation of minimal pairs allows us to focus on specific phonemic contrasts which in most cases require serious practice on the part of Spanish students of English as second language due to their difficulty. The result is a test battery that allows the user to listen to each minimal pair before trying to correctly pronounce each of its components, until success is attained.

The architecture of the resulting game is shown in section 2.1. Section 2.2 describes the challenge presented to the users and the set of minimal pairs that have been employed for our first testing of the system. In section 3, we present the results obtained after exposing three specifc populations to the game. In the discussion section, as well as in the conclusions, we will argue for the benefits of this kind of approximations to pronunciation teaching, and a number of relevant issues and considerations will be taken into account and noted for prospective research.

2. Experimental procedure

2.1. Serious game definition

2.1.1. Arquitecture of the system

Figure 1 represents the conceptual architecture of the system. The *Control* module includes the application's business logic. The *Minimal Pairs' Database* is accessed by the *Control* component in order to extract the minimal pairs. The *Game Interface* component will present each pair to the users in accordance with the game dynamics, to be explained in later sections of this paper. The interface manages the speaking turns of the user and responds to his/her demands of the TTS service. The *Control* component makes use of an *ASR component* that translates spoken words into text. When the patterns produced by the ASR

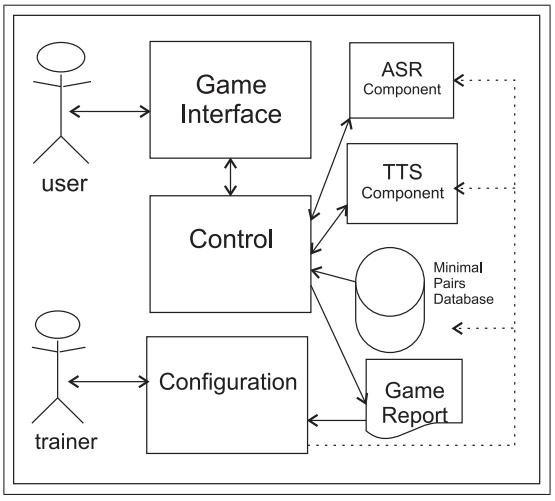

Figure 1: Arquitecture of the serious game: conceptual components of the system.

component match those of the target words, the pronunciation is correct. The *TTS component* is used to generate a spoken version of any required word. It allows users to listen to a model pronunciation of the words before they try to pronounce them themselves.

A *Configuration* component selects the language in which the ASR and TTS components operate. Furthermore, it allows selecting among different sets of minimal pairs according to the language to be tested. Results will show the capital importance of a proper selection of minimal pairs. The *minimal pairs' database* –which constitutes the knowledge database of the system– can be updated in order to improve the system or to include new challenges.

Finally, a *Game Report* is generated at the end of each game. This report registers user dynamics, including the timing of the oral turns (both for recognition and for synthesis) and the results obtained.

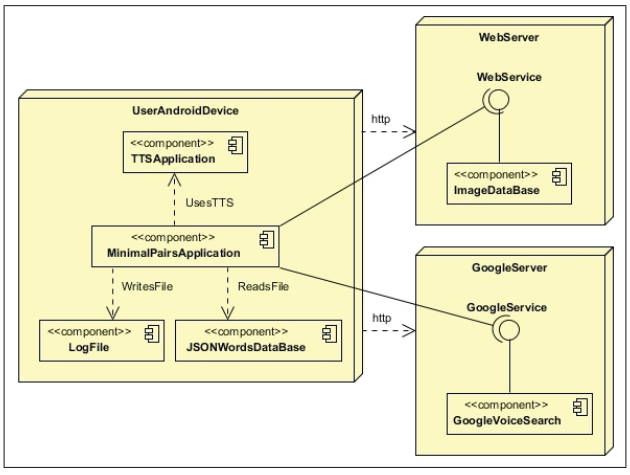

Figure 2: Android implementation components of the serious game.

2.1.2. *Implementation in Android*

Figure 2 shows the use we have made of the Android resources in implementing the game. *UserAndroidDevice* represents an Android operating system device which installs the game and the Android TTS application. *JSONWordsDataBase* constitutes the local database that contains the lists of minimal pairs. This database consists of a group of JSON files classified by languages and list types. The *LogFile* component is a local file intended to obtain useful statistics of the played games and to improve the application.

The *UserAndroidDevice* communicates with a web server via http in order to obtain icons for each target word, contributing to make the interface more attractive. These images are captured in cache memory by the Android device depending on its system memory capability. The *UserAndroidDevice* makes use of the Android Speech API, which connects with *GoogleServer* in order to perform the ASR process. TTS is locally generated.

Future versions might access the *Google Analytics* service in order to enrich the presentation of results. In fact, a web server might be used for socializing the application and allowing for the incorporation of several players to the same game.

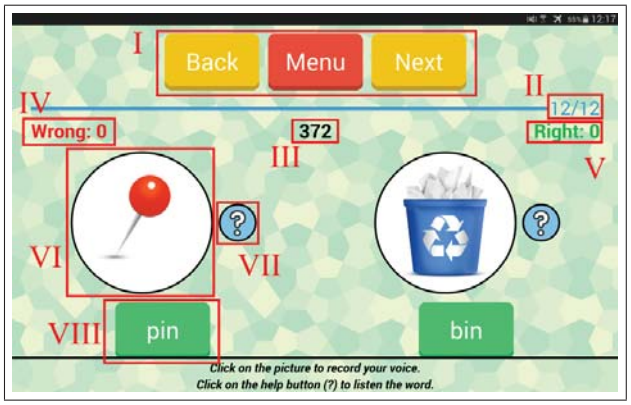

Figure 3: Example of main visual interface of a game. Buttons in I allow users to navigate freely during the game. II displays the current status of the game, that is, the current pair of words presented in relation to the total pairs to be presented. III displays the maximum remaining time to end the game (in seconds). IV and V represent the total number of wrong and right attempts respectively. VI displays an icon of the word to be pronounced. VII represents a help button that is used for listening to the linked word (with TTS module). VIII is a clickable button to start the recognition mode (with ASR module) of the word.

2.1.3. *Interface of the game*

Figure 3 shows the different parts of the game's interface. Subjects were asked to separately read aloud (and record) both words of 10 pairs randomly selected from the twenty pairs contained in table 1. They could freely choose to listen to each of the words separately –that is, they would not listen to the pair in a sequence unless they decided to click on them sequentially. On the other hand, they could freely choose to record the words without listening to the model. In the event of a realization detected as *wrong answer* by the application, subjects could repeat again up to five times if necessary. After that, the application would shut the recording mechanism and force users to continue with the rest of the words. Alternatively, subjects might decide to continue with the test, leaving behind those items they felt they were not going to be able to produce properly. The recording mechanism was also shut when any realization was detected as *right pronunciation* by the application.

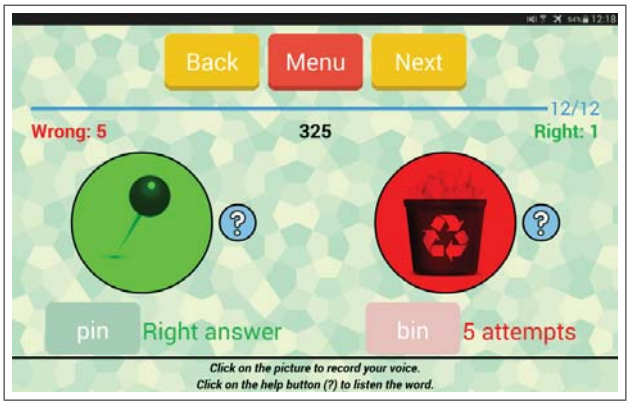

Figure 4: Example of a right spoken word and a word with five wrong attempts.

If a user pronounces the selected word correctly, the corresponding icon changes its base color to green, and gets disabled as a positive feedback message appears. Otherwise, a message with the recognized words appears on the graphical interface and a non-positive feedback message is presented. Also if the user has tried to pronounce five times the same word without success, the element VI changes its base color to red and it gets disabled as seen in Figure 4. Speakers had a maximum of 7 minutes to complete the test. The challenge for users is to obtain as many right pronunciations as possible, in as little time as possible.

When each of the words of a minimal pair is judged as correct by the ASR system, or when the player has reached the maximum limit of failed attempts in one of the words –the other being correct–, or when the maximum limit of failed attempts has been reached in both of them, a new minimal pair automatically appears on the graphical interface.

2.2. Minimal pairs selection for L1 Spanish L2 English speakers

The intersection between the phonological systems of American English and European Spanish roughly encompasses bilabial, alveolar and velar nasals; voiceless dental, alveolar and labiodental fricatives; and, to some extent, voiceless affricates. In other words, only the sounds /m, n, ɲ, θ, s, f, ʧ/ are pronounced in both languages with remarkable similarity, to the point of possible interchangeability [14, 15]. All other consonants, and certainly all vowels, contrast quite perceptively, at least from the perspective of trained human judges. Sounds also deploy significantly different behaviors within the speech chain in each language; to give just a few well-known examples: voiceless plosives in stressed onset positions are released with aspiration in English but not in Spanish [16, 17]; on the other hand voiced plosives turn into voiced fricatives or approximants when intervocalic in Spanish, but not in standard versions of American English. Particularly, while vowel length in Spanish is not phonemically significant, the real length of all English vowels is largely dependent on whether they are closed by voiced or voiceless consonants, to the point that such feature often plays an essential role in the identification of pairs like lose-loose or peck-peg, where the closing voiced consonant is often subjected to total devoicing. All in all, the transference of Spanish segments and their distribution to the articulation of English words brings about a strongly flavored accent that is somewhat represented in the ETP transcriptions of Table 1.

Minimal Pair	NT	ETP
sock - suck	sɑˈk - sʌk	sak/sok-sak/suk
dunce - dance	dʌnᵗs - dænᵗs	*d*anᵗs/*d*unᵗs - *d*anᵗs
mess - mass	mɛs - mæs	mɛs-mas
curse - course	kʰᵡɝːs - kʰᵡɔɔˠ s	kɛrs - kors
were - where	wɝː - wɛəˠ	gwɛr - gwɛr
will - wheel	wiːəɬ- wiːəɬ	gwil - gwil
soot - suit	sʷʊtˢ - sʷuːtˢ	sʷ *ut* - sʷ *ut*
don - dawn	dɑːn - dɒːn	*d*an/*d*on - *d*on/*d*aun
sit - set	sɪtˢ - sɛtˢ	si*t* - s*e*t
caper - caber	ˈkʰᵡeɪpəˠ- ˈkʰᵡeˈɪbəˠ	ˈkeiper - ˈkeiβer
mat - mad	mætˢ - mæˈdᶻ	ma*t* - maθ/mað
letch - ledge	letʃ - leˈdʒ	letʃ - letʃ
lose - loose	lʷuːz - lʷuˈs	lʷ*u*s - lʷ*u*s
luff - love	lʌf - lʌˈv	laf - laf
read - wreathe	ˈriːdᶻ - riːð	riθ/rið - βrið/wrið
waiter - wader	ˈweɪɾəˠ- ˈweɪdəˠ	gweiter - gweiðer
peck - peg	pʰekˣ - pʰeˈgᵞ	pɛk - pɛx/pɛk
sue - zoo	sʷʊu - zʷʊu	sʷ*u* - sʷ*u*
sun - shun	sʌˈn - ʃʌˈn	san - san
when - Gwen	weˈn - gweˈn	gwɛn - gwɛn

Table 1: List of minimal pairs to be used. NT: Narrow transcription of the words according to standard pronunciation ETP: Expected Transferred Pronunciation for Spanish ESL students.

The visual differences between the NT and the ETP columns in Table 1 –expressed mostly in traditional diacritic signs– represent every articulatory and, consequently, acoustic feature that distinguishes a proper Standard pronunciation from a transferred one. A properly trained human agent will be able to perceive the correctness of a particular realization regardless of the minimal pair where it is included. So, in the realization of, for example, wheel / wiːəɬ/ the particular timbre of all harmonic sounds, and aspects such the velarization of /l/ and the l-coloring transition, may be ascertained, or their absence detected, quite independently.

The many differences represented by the NT and ETP columns in Table 1 attest to its interest and relevance as a basis for testing and diagnosing the pronunciation skills of ESL students who speak Spanish as their first language. It is worth pointing out, nevertheless, that while human agents concerned with the goodness of pronunciation of a particular student would judge the presence or absence of each and all the features present in the NT column, most present-day ASR systems would be only concerned with the recognizability of each item, and the degree to which realizations may be confused with one another.

2.3. Testing population

Three different groups of users are distinguished according to their a-priori English pronunciation proficiency:

Group A North American native speakers. The speakers of this group are used as a baseline for checking the limitations of the ASR system. They are L2 Spanish students visiting our University.

Group B Spanish students of English philology. All of them had passed an specific course on English phonetics so that they were supposed to be high level English speakers.

Group	Speakers	# Tries	# Listens	Time (s)
A	12	372	35	2431
B	21	1033	400	6677
C	20	1094	606	7492
Total	53	2499	1041	16600

Table 2: Number of participants in the test. # Tries in the number of times that the participants attempted to pronounce a given word. # Listens is the number of times that the participants use the TTS system to listen to the word. Time is the total duration of the participation of the players.

Group C Spanish students of Computer Science. Despite the fact that some of these student may have an acceptable or a good English level, generally the pronunciation of Spanish university students is not as good as desirable.

It is expected that the informants of Group A will play the game without any mistakes. The informants of the Group B are expected to play better than the speakers of the Group C. Their results are expected to be comparable to those obtained by the speakers of the Group A. All participants are volunteers. Table 2 summarizes the number of speakers that collaborated and their implication. We kept record of the contact and declared level of the speakers.

Group	$\overline{Tries}$	$\overline{Success}$	$\overline{Fails}$	$Recall(\%)$
A	31±7	21±4	10±6	69±17
B	49±14	18±3	31±15	41±15
C	55±9	15±4	40±10	28±10

Table 3: Success rate of the participants. The format of the cells is mean value ± standard deviation. $\overline{Tries}$ refers to the number of times that the speakers attempt to pronounce the total set of word (24 words in total). $\overline{Success}$ refers to the number of successful attempts: the ASR identifies the expected word. $\overline{Fails}$ refers to the number of times the ASR system does not identify the expected word. $Recall$ is the relation among the number of times the ASR system identify the expected word and the number of attempts.

3. Results

Table 3 presents the mean scores obtained by the speakers of the three groups after playing one game. The speakers of Group A obtain excellent results when compared with those of the speakers of the other two groups with a mean success of 21± 4 out of a maximum of 24. Speakers in Group C obtain the worst results. They are worse than the ones obtained by the speakers of the Group B (28% vs. 41% of Recall), with statistically significant differences (95% confidence level) for all the variables of the table when the t-test with asymmetric hypothesis is applied (except for the variable $\overline{Tries}$ where p-value=0.06) .

The results obtained by Group B speakers are significantly different from those obtained by the speakers in Group A with a confidence level above 99%, except for the variable $\overline{Success}$ whose p-value is 0.057. Interestingly, Group B speakers increased $\overline{Success}$ –more so than Group C speakers (18 vs 15)– without a corresponding increase in the number of attemps (49 vs 55). On the other hand, a higher number of attempts (49 vs 31 of Group A speakers) justifies the significant differences between Group A and B speakers.

	Group A		Group B		Group C	
1	wreathe	100	luff	100	wreathe	100
2	luff	94	wreathe	100	luff	98
3	wader	73	letch	97	letch	98
4	soot	64	loose	90	wader	96
5	sock	58	wader	88	sock	96
6	caber	56	peck	84	soot	96
7	letch	50	sue	84	Gwen	89
8	mass	38	sock	83	shun	88
9	don	33	dunce	81	sue	86
10	mess	33	dawn	80	dawn	85
11	Gwen	31	soot	79	were	83
12	shun	30	Gwen	76	peg	83
13	were	20	were	72	peck	82
14	dunce	12	don	71	loose	81
15	mat	11	zoo	70	dunce	81

Table 4: Most frequent words that are not recognized by the ASR system in percentage

In order to understand why the speakers in Group A (the English native speakers) also fail, we present Table 4 with their most frequent mistakes. More than a half of the errors occur when the words luff, wader and wreathe are pronounced: 68 wrong answers out of a total of 122 unrecognized words. Being rather infrequent in everyday English, these words are penalized by the language model upon which the ASR system is based, and are not, therefore, identified by it. Indeed, the word wreathe is never identified by the system (100% of fails). Of course, speakers in Groups B and C failed massively in these words as well. The rest of errors in Group A seem to be anecdotal and mostly due to environmental noise or misuse of the interface. The same table allows us to identify words like peck, sue, or dawn, that are never confused by the speakers in Group A but very frequently so by the Spanish players.

Every prediction of the ASR system is supplied with an n-best list of 5 possible words, where each of them is followed by a numeric value named $gscore$, which is proportional to the reliability of the prediction. The realization is considered correct as long as it is within the list of 5 elements returned by the ASR. Thus, for example, a speaker in Group C tried to pronounce the word $mass$ and the system outputted the following n-best list of possible recognitions, with a gscore=0.25.

$$"math","nas","mass","nice","myass"$$

Although the target word $mass$ is, in fact, contained within the n-list, a low gscore value evidences the poor quality of the

Group	gscore Right	gscore Wrong	gscore Total	Time (s)
A	0.70±0.3	0.59±0.3	0.67±0.3	203±66
B	0.65±0.3	0.59±0.3	0.61±0.3	318±82
C	0.58±0.3	0.55±0.3	0.56±0.3	375±54

Table 5: Mean value ± standard deviation. $gscore$ is a value returned by the ASR system that indicates the quality of the prediction. $Time$ stands for the time that the user devotes to finish the test. $Right$ registers the values obtained when the output of the system predicts correctly the expected word. $Wrong$ represents the values obtained when the ASR system does not predict the expected word correctly.

Group	Position 1	2	3	4	5
A	63.6	18.8	8.4	6.8	2.4
B	51.8	21.8	13.7	10.6	2.1
C	47.3	23.8	13.8	9.7	5.4

Table 6: Distribution in percentage of the position of the correct prediction in the n-best list of predictions returned by the ASR.

pronunciation of this particular speaker in relation to this particular item. For the same word a native speaker obtained the following n-best list with a $gscore = 0.85$:

$$"mass","Mass","masse","masts","mass."$$

Table 5 shows that the values of the $gscore$ index is clearly representative of each of the groups. There are statistical significant differences (asymmetric t-test with 95% confidence level) across the different groups of speakers except when the Spanish players fail (column $Wrong$, rows C and B).

For the construction of Table 6, we have considered the position of the target word within the n-best list as a possible indicator of the quality each speakers' pronunciation. Thus, the number of times that the target word appears at the first position within the n-best list is higher for Group A speakers (63.6%) than for Group B (51.8%) and Group C (47.2%) speakers. These differences became statistically significant when an χ-Square test was applied (p-value = 0.005326 χ^2 = 21.7869, df = 8).

Going back to Table 5, significant differences can also be observed in relation to the game duration for the three types of speakers (p-value $<$ 0.0065 assimetrix t-test). Group A speakers finish their game long before the rest of speakers. The slowest players are those in Group C as a consequence of their higher number of attemps and wrong-answer feedbacks (see resutls of Table 3).

Finally, Table 7 shows the use that the players make of the TTS system (we have removed from the table the listening events concerning the words *wreathe* and *luff* which are, as we said, problematic for the ASR system) Players listen to the models when they have doubts about the way they are to be pronounced. The use of the TTS system by the speakers in Group A is negligible (2$\pm$2 on average). They use it when the system is not identifying their utterance. In these cases, listening does not turn up to be very useful (only 5.5% rate) as the problem is not with the speaker as much as with the ASR system. Speakers in Group C use the TTS more frequently than Group B speakers

Group	$\overline{Count}$	Rate	Pos.	Neg.
A	2$\pm$2	5.5	26.3	73.7
B	18$\pm$12	27.6	29.4	70.6
C	28$\pm$17	35.5	30.3	69.7

Table 7: Use of the text-to-speech system. $\overline{Count}$ stands for the mean number of times the speakers use the TTS system during her/his game. $Rate$ computes the percentage of listening actions with respect to the total number of actions (listen plus spoken events). Pos. Neg. is the percentage of times that after listening the result of the ASR system is positive or negative.

(35.5% vs 27.6%). Indeed, the TTS system is used a significantly larger number of times by the speakers in the Group C: more than once every three turns (35.5%). The percentage of times that the word is correctly pronounced after listening to its TTS version rises above 29% for speakers of both Groups B and C.

4. Discussion

As it has been pointed out, Group A speakers (the native speakers) obtain recognition rates that are significantly higher than the ones obtained by the non-native speakers. Nevertheless we must point out the fact that the failing rate of native speakers, although small, was not equal to zero. This reveals a weakness of the system: perfect pronunciations may not be recognized by the ASR system. There are several reasons for this. The first one has its origin in the environmental conditions in which the test is performed: background noise or disfluencies in the speakers' utterances cause the system to fail. On the other hand, and more importantly, we must take into account the fact that we are using a real open ASR system, which is configured to identify free speech; there are some words that are more difficult to be identified than others, not because of their phonetic structure but because of their low frequency of appearance in the language. By virtue of the language models used as reference by the ASR, those words that are not usually found in one-word sentences are penalized when pronounced in isolation, as the game requires. This fact is partially attenuated when the system outputs more than one prediction. As our experiment has shown, some words are very difficult (or impossible) to be identified by the ASR system.

This fact must be taken into account when the tests are configured. The words that compose the battery of minimal pairs must be tested by native speakers before entering the list. Otherwise, the game might lead to a situation where the system declares that a correct native pronunciation is a wrong one. Furthermore, the alternative predictions that the system outputs must be taken into account since the appearance of the target word within the n-best list does guarantee that the system actually identifies it so that its production is correct.

Despite these limitations, our paper proves that on the whole, the tests generate objective measurements that can be representative of the quality of the pronunciations. The $gscore$ values and the position of the expected word in the list of predictions can be an indicator of the quality of the pronunciation. Also the time devoted to complete the test depends on the proficiency of the speaker, being very low when proficiency is high. The availability of these objective metrics combined with others related to the prosodic production [18] can be used to suggest training activities to those speakers that present unsatisfactory results.

In a somewhat looser way, since synthetic models are more frequently used by those with a lower pronunciation level, the number of times that a given player uses the TTS service can also be taken as an indicator of his/her level. Nevertheless this resource does not seem to be as useful as desirable because in many cases the pronunciation after several repetitions and with the use of the TTS service does not improve. This result evinces the need for future extensions of the system that must include extra helps such as visual reinforcements and selective listening of the phonemes that are causing the confusions.

5. Conclusions

The definition of challenges based on ASR and TTS tools may help us assess a user's pronunciation level in an L2: We have

proved that the results provided by the application strongly correlate with the expected level of the user in the sense that, as predicted, native speakers got the best scores, Spanish speakers in Group C the worst, and Spanish speakers of Group B consistently remained between the two extremes.

An adequate definition of the activities involved is essential in the design of truly effective tools. There are, as of today, serious handicaps to overcome due to the limitations of ASR and TTS systems. Words correctly pronounced by natives may not be properly recognized by the system, while words badly pronounced may be recognized and accepted, and will consequently generate misleading positive feedbacks for non-native users.

Our Minimal-pairs setting incorporates a pedagogic gesture that could be further developed. In its present form, it conveniently warns users that pairs of words that may be wrongly though habitually reduced to the same pronunciation –for example were/where– are in fact to be pronounced contrastively, that is, with different consonants or vowels in each case. This is likely to prompt non-native users to listen to particular words before attempting pronunciation. On the other hand, an obvious conclusion derived from our research is that the previous selection of pairs must be responsive not only to the user's needs and pedagogic targets –consonants and vowels which are known to be difficult– but also to the limitations, scope and working procedures of present ASR systems. However, to the extent that this double restriction can be accounted for, our conclusion is that present systems may be successfully used in the teaching of pronunciation.

In its present state, our application falls short of the term *serious game* provided that its gaming elements are reduced to a scoreboard with punctuation and timing. Although we should not underestimate the motivational effects of this strategy in combination with an attractive interaction board, and the possibilities of diversive exploration offered by the TTS device, the fact remains that further gamification would be welcome. Actually, a new version is currently being developed that will allow to rank registered users in a hall of fame, and the possibility that users may issue and launch pre-designed challenges among themselves within a social network. We believe that such strategies, among others being considered, would have a significant impact in terms of motivation for a continuing use of the application.

In moving from the present prototype into a more final version of the game, it is also clear that a more adequate pedagogic phase must be implemented, by incorporating activities of guided exposition and discrimination exercises. A more natural progression must be designed: before trying out their own pronunciations, players should be allowed to become familiar with the sounds in previously calculated sequences (minimal pair or trios, etc.), and their ability to discriminate between them should be tested. While incorporating these two stages might present some difficulties, the greater challenge would be to be able to incorporate mechanisms to provide real particularized feedback based on automatically identified errors.

6. Acknowledgements

This work has been partially supported by Consejería de Educación de la Junta de Castilla y León (project SAMPLE VA145U14) and the Spanish Ministery of Economy and Competitiveness (project TIN2014-59852-R).

7. References

[1] M. Eskenazi, "An overview of spoken language technology for education," *Speech Communication*, vol. 51, no. 10, pp. 832–844, 2009.

[2] F. Ehsani and E. Knodt, "Speech technology in computer-aided language learning: Strengths and limitations of a new call paradigm," *Language Learning & Technology*, vol. 2, no. 1, pp. 45–60, 1998.

[3] A. McFarlane, A. Sparrowhawk, and Y. Heald, *Report on the educational use of games*. TEEM (Teachers evaluating educational multimedia), Cambridge, 2002.

[4] Z. Handley, "Is text-to-speech synthesis ready for use in computer-assisted language learning?" *Speech Communication*, vol. 51, no. 10, pp. 906 – 919, 2009, spoken Language Technology for Education Spoken Language.

[5] A. Neri, C. Cucchiarini, and H. Strik, "Automatic speech recognition for second language learning: How and why it actually works," in *Proc. ICPhS*, 2003, pp. 1157–1160.

[6] I. McGraw and S. Seneff, "Immersive second language acquisition in narrow domains: a prototype island dialogue system." in *SLaTE*, 2007, pp. 84–87.

[7] S. W. Campbell and Y. J. Park, "Social implications of mobile telephony: The rise of personal communication society," *Sociology Compass*, vol. 2, no. 2, pp. 371–387, 2008.

[8] F. Lys, "The development of advanced learner oral proficiency using ipads," *Language Learning and Technology*, vol. 17, no. 2, pp. 94–116, 2013.

[9] B. Pellom, "Rosetta stone ReFLEX: Toward improving english conversation fluency in asia," in *Proceedings of the Intenational Symposium on Automatic Detection of Errors in Pronunciation Training*, 2012, pp. 1–20.

[10] Y. Levy, "Comparing dropouts and persistence in e-learning courses," *Computers & education*, vol. 48, no. 2, pp. 185–204, 2007.

[11] I. McGraw, B. Yoshimoto, and S. Seneff, "Speech-enabled card games for incidental vocabulary acquisition in a foreign language," *Speech Communication*, vol. 51, no. 10, pp. 1006–1023, 2009.

[12] H. Strik, J. Colpaert, J. van Doremalen, and C. Cucchiarini, "The disco asr-based call system: practicing l2 oral skills and beyond." in *LREC*, 2012, pp. 2702–2707.

[13] M. Celce-Murcia, D. M. Brinton, and J. M. Goodwin, *Teaching pronunciation: A reference for teachers of English to speakers of other languages*. Cambridge University Press, 1996.

[14] E. Cámara-Arenas, *Native Cardinality: on teaching American English vowels to Spanish students*, S. de Publicaciones de la Universidad de Valladolid, Ed., 2012.

[15] 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 and Behavioral Sciences*, vol. 116, pp. 3044–3048, 2014.

[16] D. F. Finch and H. O. Lira, *A course in English phonetics for Spanish speakers*. Heinemann Educational Books, 1982.

[17] A. Cruttenden, *Gimson's pronunciation of English*. Routledge, 2014.

[18] D. Escudero-Mancebo, C. González-Ferreras, and V. Cardeñoso Payo, "Assessment of non-native spoken spanish using quantitative scores and perceptual evaluation," in *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*. European Language Resources Association (ELRA), may 2014, pp. 3967–3972.