Iberspeech 2016
23-25 November 2016, Lisbon, Portugal

Improving L2 Production with a Gamified Computer-Assisted Pronunciation Training Tool, TipTopTalk!

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

[1]Department of Computer Science
[2]Department of English Philology
University of Valladolid
cristian@infor.uva.es

Abstract. We present a foreign language (L2) pronunciation training serious game, TipTopTalk!, based on the minimal-pairs technique. We carried out a three-week test experiment where participants had to overcome several challenges including exposure, discrimination and production, while using Text-To-Speech (TTS) and Automatic Speech Recognition (ASR) systems in a mobile application. The quality of users' production is measured in order to assess their improvement. The application implements gamification resources with the aim of promoting continued practice. Preliminary results show that users with poorer initial performance levels make relatively more progress than the rest. However, it is desirable to include specific and individualized feedback in future versions so as to avoid the performance drop detected after the protracted use of the tool.

Keywords: speech technology, computer assisted pronunciation training, gamification, leaning analytics, L2 pronunciation, minimal pairs

1 Introduction

In recent years, the use of Computer Assisted Pronunciation Training (CAPT) applications during the process of acquiring new languages is becoming widespread [5]. They have been proved to constitute effective resources for the improvement of L2 (foreign language) perception and production [7][4].

The popularization of smartphones and other smart devices has led to the extension of technological services to users [1]. Nowadays, the most popular mobile and desktop operating systems grant users a free access to several Text-To-Speech (TTS) and Automatic Speech Recognition (ASR) systems. If properly integrated within pedagogical routines, TTS systems will allow users to focus on the constituent sounds of target words, easily and immediately [8]. On the other hand, ASR systems may be used for the detection and assessment of pronunciations errors among non-native speakers [13]. There have been, however, very

few attempts to objectively measure the real improvement attained by users of pedagogically oriented speech technology [10][9].

Moreover, the combination of adequate teaching methods and gamification strategies will increase user engagement, provide an adequate feedback and, at the same time, keep users active and comfortable [12][11].

In this paper, we show the performance results of users of TipTopTalk! [6][14][15] - a second generation serious game application designed for L2 pronunciation training and testing. First, we will describe the software tool, starting with the main dynamics, continuing with the visual interface and the gamification elements included, and finishing with the technology applied. Then, we will present the results obtained after the test campaign. Finally, we will analyze the information thus gathered and suggest some recommendations for future development.

2 Overview of CAPT system

2.1 Application dynamics

The design of our serious game supports a learning methodology based on the sequencing of three different learning stages: exposure, discrimination and pronunciation [3] (see figure 1). These strategies are built into two separate modules: *Training* and *Challenge yourself*. Both include the same essential dynamics, although only the second one incorporates gamification features.

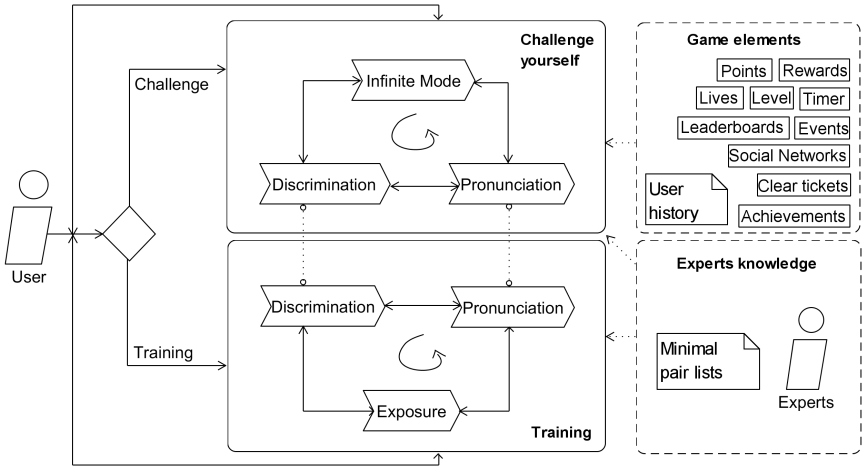

Fig. 1. Flow chart of the activities proposed to users.

From a pedagogical point of view, the use of minimal pairs [2] raises users' awareness of the potential risks of generating wrong meanings when phonemes

are not properly produced. The discrimination of the elements of a minimal pair often constitutes a challenging task for the ASR, since the phonetic distance between each couple of words, despite being clearly perceptible for native speakers, can be rather small in quantitative terms. In order to maximize efficiency, the lists of minimal pairs used by the tool are selected by expert linguists, for each language

Firstly, in the exposure mode, players become familiar with the distinctive phonemes within sequences of minimal pairs selected by a native linguist and presented at random. The aural correlate of each word is played a maximum of five times. Then, users decide whether to move on to next round of words, or to record their own realization of the words to compare it with the TTS versions. This mode is only available in the *Training* module.

Secondly, in the discrimination mode, users test their ability to discriminate between the elements of minimal pairs. They listen to the aural correlate of any of the words in each pair and must match it with the correct written form on the screen. As part of the gamification strategy, the game randomly asks users to pick the word that has not been uttered, rather than the uttered one. At higher levels of difficulty, the phonetic transcription of each word, otherwise visible, is removed. These strategies aim at the promotion of user adaptation and engagement. This discrimination mode is available both in *Training* and *Challenge yourself* modules.

Finally, in the pronunciation mode, participants are asked to separately read aloud (and record) both words of each minimal pair. A real-time feedback is provided instantly. Native model pronunciations of each word can be played as many times as the user needs. Speech is recorded and played using third party ASR and TTS applications.

The *Challenge yourself* module includes an extra mode, called *Infinite mode*, in which the aim is to complete the highest number of rounds possible. Discrimination and pronunciation challenges are presented randomly in each round. Users start with a finite number of lives that will decrease in one each time they fail. Also, the game's difficulty level increases with each round. For instance, from the tenth round on, the chance that the orthographic representation a word is substituted by asterisks is raised to 50%. From the twentieth round on, a 50% chance that the TTS button is absent is introduced. The amount of time allotted for round completion is also progressively reduced.

2.2 User interface

Each TipTopTalk! teaching strategy has its visual user interface containing different game elements. Figure 2 shows three visual user interface screenshots of the main game modes, that is, exposure, discrimination and pronunciation.

The first screenshot of Figure 2 shows a standard round within the exposure training mode. There is a menu-options bar at the top through which users can exit the current game, go forward to the next round or go back at will. There is a status bar beneath the menu-options bar indicating users the round they are in. The system allows us to register whether users play the model for both

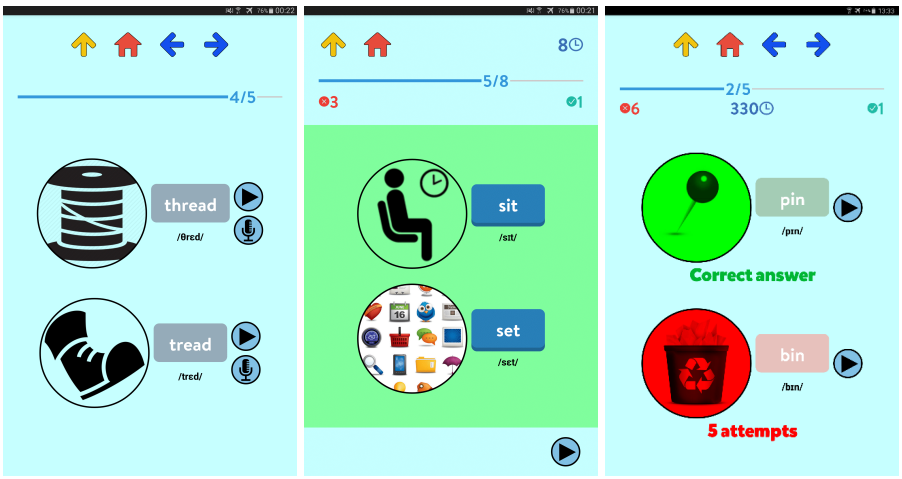

Fig. 2. Visual user interface of exposure, discrimination and pronunciation modes.

words at the beginning of each round. Pictures, orthographic forms and phonetic transcriptions are displayed at the center of the screen. Finally, we keep track of the number of times users synthesize a word or record themselves. We save the recorded voice in a file for subsequent analyses and corpus compilation.

The second screenshot of Figure 2 (discrimination mode) includes new elements such as a timer in the right top corner and both discrimination wrong and correct counters. There is a background colour as a gamification element. If the colour is green, users must choose the word they think is being played. However, if the colour is red, they must choose the wrong one. In the right bottom corner there is a button that plays another time the sound of the word.

The last screen capture in Figure 2 represents a snapshot of a pronunciation mode round. This part of the game, introduces more feedback elements than the previous. When the user utters the test word correctly, the corresponding icon and other related elements change their base color to green, and the word gets disabled as a positive feedback message appears. Otherwise, a message appears containing the words recognized by the ASR (different from the test word) together with a non-positive feedback. The mispronounced word changes its base color to red and remains active before it gets disabled only after five unrecognized realizations by the user. When a word is repeatedly pronounced in an unrecognizable way, we limit to five the number of attempts before moving on to the next word, so as not to discourage users.

We gather relevant quantitative data from all emerging events in the visual interface of the application with which we feed a personalized daily log for each user in order to determine whether her or his pronunciation skills are improving. In addition, we send depersonalized events to our Google Analytics account, from our application, in order to compute how often a given event has occurred.

2.3 Gamified sessions

The main advantages of using a gamification design strategy are: (i) an increase in learners' engagement, and (ii) the possibility compiling a comprehensive and individualized feedback while keeping users active and relaxed, while free to progress at their own pace in an anxiety-free context. As a function of correct and wrong answers, TipTopTalk! adapts to the player. New training modes are suggested based on the results of the current one. For instance, in the discrimination mode, if a user achieves the maximum score, advancement to a pronunciation mode will be suggested. Contrarily, going back to the exposure mode will be automatically recommended after a low score has been attained in discrimination.

As a strategy for enhancing encouragement and engagement, users add points to their *phonetic level* and gain several achievements (dependent on the mode and difficulty level). There are also different language-dependent leader boards, based on scores attained and the number of completed rounds, where all players are ranked to increase engagement through competition. On the one hand, sharing results via social networks plays an important role in the gamification strategy by virtue of the competitiveness that it promotes. On the other hand, social networks will allow a worldwide expansion of the application.

Other gamification elements include: a limited time to complete the current round or a game; the granting of more or less points depending on the difficulty level and the number of attempts required for completion; the allotting of a number of reserve lives to allow further playing; the dispensation of an amount of *clear tickets* which allow users to skip the current round and move on to next one; and the graphical display of the visual percentage of a game list result. Finally, we incorporate a system of push notifications that sends users motivational and challenging messages in order to trigger their engagement.

2.4 Technology

Several elements belong to our system. Figure 3 represents the architecture of TipTopTalk! From left to right, *UserAndroidDevice* represents an Android device in which TipTopTalk! is installed. It connects to an external TTS application which users can freely choose. We integrate some Google services such as *GoogleVoiceSearch* for ASR system, *GoogleAnalytics* for registering user interactions with the system, and *GooglePlayGames* for the introduction of gamification elements.

Besides, results are also saved as a JSON format log file that compiles all possible depersonalized data diachronically to be sent to a *WebServer*. The application runs with a list of 793 minimal pairs for American English and 168 for Simplified Chinese. Currently it is being enriched with words for German, European Spanish and European and Brazilian Portuguese. Each exercise in all modes displays approximately 8 pairs. All pairs are classified within categories of phonemic contrasts. Finally, the icons that illustrate the meanings of the words used by the system and all user's log files are stored in our own *WebServer*.

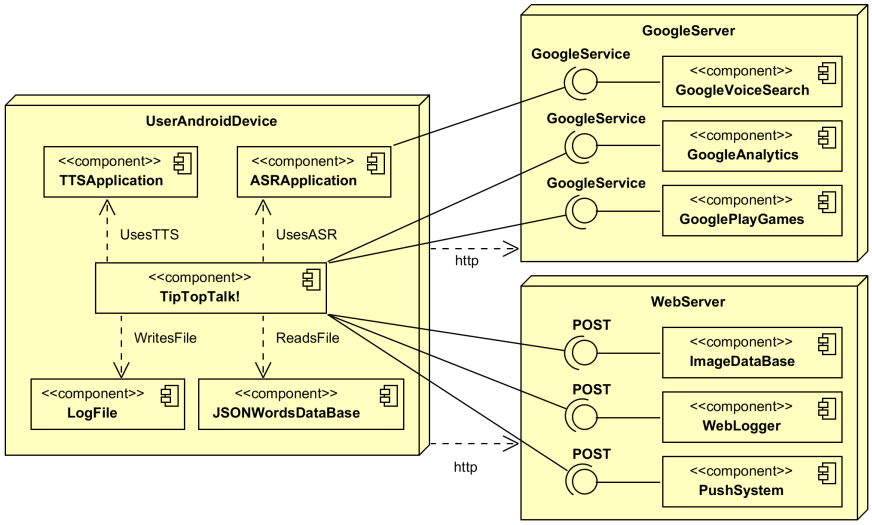

Fig. 3. System's architecture. There are three main different components: an Android device (left-hand component), Google's server to provide some online services (top-right component) and a private web server to collect data (bottom-right component).

3 Assessment

The experiment was structured as follows: up to 100 native Spanish students of computer engineering and English philology and native Chinese students of English participated in the project. All of them received the application via email, and installed it on their own devices. Then, they were given permission to interact with the application as they wished. During three-week test campaign, 58% of users made extensive use of TipTopTalk! with up to 6000 interactions during the first week. Some of the participants remained engaged in the game for several weeks, registering more than 11000 events.

This campaign has generated a database with approximately 88000 entries containing information about the use of the tool made by each user in relation to the different exercises. This set of records R, can be defined as

$$R = E \cup D \cup P \cup O \tag{1}$$

where E represents the amount of entries corresponding to exposure turns, D stands for those corresponding to discrimination exercises, P for user productions and O for control manipulations (such as activity transitions, logging in or out of the system, etc.). Discrimination exercises are characterized as

$$D = \cup_u \cup_k D_{u,k} \tag{2}$$

where $D_{u,k}$ expresses a sequence of chronologically ordered discrimination attempts by user u=1..U of the words of a kind of pair k=1..K, so that

$$D_{u,k} = (d_1..d_{N_{u,k}}) \tag{3}$$

where $N_{u,k}$ represents the number times a user u tries to discriminate words of a kind of pair k. A function of quality $f_D(D_{N_{u,k}}, w, s)$computes the average number of correct answers attained within a window of w attempts, beginning at the position s =1..$N_{u,k}$–w, in $D_{u,k}$. For user u, the production of words of a kind of pair k, is represented by the sequence

$$P_{u,k} = (p_1..p_{M_{u,k}}) \tag{4}$$

where p_i represents the attempts to pronounce words of a kind of pair k taking into account the fact that the game allows up to five attempts for each word and $M_{u,k}$ stands for the number times that user u tries to pronounce words of a kind of pair k. Quality of pronunciation is captured by the function $f_P(P_{u,k}, w, s)$ where s=1..$M_{u,k}$ measures the quality of a user u's pronunciation attempts in relation to the words of a kind of pair k within a window of w words (with up to five attempts) beginning at position s. Function f_P accounts for the position of the target word within a list of predictions made by the ASR, the reliability indicators generated by the ASR system, the number of attempts made by the user, and the possible existence of homophone words.

The contrast between the value of f at a given s, relative to the value of f for s=0 will tell us about the user's performance progression in both the discrimination and production phases of the different pairs and their contrasting phonemes.

4 Results and discussion

We have analyzed all data from the discrimination and production modes with the improvement functions f_D and f_P. These functions integrate significant data concerning the user, the kind of pair that is being discriminated or the word that is being produced, and the number of trials.

Figure 4 represents the evolution of functions f_D and f_P at s. They show their average values varying u and k with a window size of w=6. For the interpretation of the dependence of u, we class users into three groups depending on the values of f in the initial window s=6. We consider this value to be representative of the initial competence of each user before using TiptTopTalk! for the first time.

In general, user's performance shows improvement along time both in discrimination and pronunciation tasks. On the one hand, in the discrimination mode the three categories of users tend towards significant improvement from start to end. On the other, up until s=12 there is a significant improvement in the production mode. We can conclude that users with a poorer initial level make the most significant progress. The average user progresses initially towards an optimal point after which the values of f begin to fall. Users with a higher

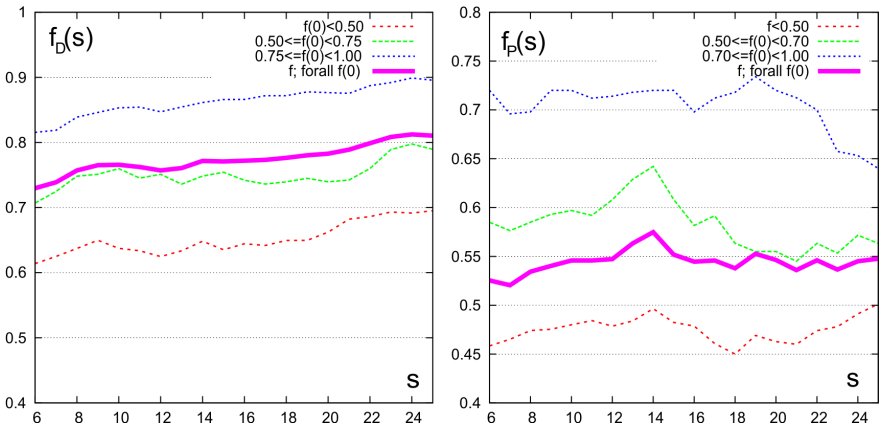

Fig. 4. Progression of the function of quality along time of use in discrimination (left-hand diagram) and production (right-hand diagram).

initial level (some of them are, in fact, native speakers) register an initial drop in performance which we think attributable to the lack of individualized feedback and the playability variables introduced in order to make the game more challenging (for instance, in discrimination exercises users must click on one or the other word within a pair depending not only on the word they hear, but also on the background color displayed). We also believe this decrease in performance has to do with habituation and gradual loss of interest in the game.

5 Conclusions and future work

Empirically obtained results reveal that TipTopTalk! helps users with a low initial level of competence to improve L2 both pronunciation and phoneme discrimination. The specification of gamification elements present in our CAPT system has proved to be useful in order to keep users active for some time. Nevertheless, acclimatization factors lead to a fall in interest and performance after protracted use. This suggests the convenience of introducing specific feedback mechanisms to assist and guide users, especially when a performance drop is detected.

A major obstacle for the future progress of TipTopTalk! might be its dependence on Google ASR and an external TTS for assessing speech production. As these are black-box systems within the application, future improvement may be somewhat compromised. A possible solution could lie in the use of open-source tools.

Our next step will take us to focusing on particular difficulties concerning specific contrasts and phonemes. A new version of the application will allow us to analyze concrete aspects of use in relation to exposure and perception when the same kind of production difficulties is repeatedly encountered.

Acknowledgments. We would like to thank the Ministerio de Economía y Competitividad y Fondos FEDER – project key: TIN2014-59852-R Videojuegos Sociales para la Asistencia y Mejora de la Pronunciación de la Lengua Española – and Junta de Castilla y León – project key: VA145U14 Evaluación Automática de la Pronunciación del Español Como Lengua Extranjera para Hablantes Japoneses.

References

1. Campbell, S.W., Park, Y.J.: Social implications of mobile telephony: The rise of personal communication society. Sociology Compass 2(2), 371–387 (2008)
2. Celce-Murcia, M., Brinton, D.M., Goodwin, J.M.: Teaching pronunciation: A reference for teachers of English to speakers of other languages. Cambridge University Press (1996)
3. Cámara-Arenas, E.: Native Cardinality: on teaching American English vowels to Spanish students. S. de Publicaciones de la Universidad de Valladolid (2012)
4. Ehsani, F., Knodt, E.: Speech technology in computer-aided language learning: Strengths and limitations of a new call paradigm. Language Learning & Technology 2(1), 45–60 (1998)
5. Escudero-Mancebo, D., Carranza, M.: Nuevas propuestas tecnológicas para la práctica y evaluación de la pronunciación del espanol como lengua extranjera. Actas del L Congreso de la Asociación Europea de Profesores de Espanol, Burgos (2015)
6. Escudero-Mancebo, D., Cámara-Arenas, E., Tejedor-García, C., González-Ferreras, C., Cardenoso-Payo, V.: Implementation and test of a serious game based on minimal pairs for pronunciation training. SLaTE-2015 pp. 125–130 (2015)
7. Eskenazi, M.: An overview of spoken language technology for education. Speech Communication 51(10), 832–844 (2009)
8. Handley, Z.: Is text-to-speech synthesis ready for use in computer-assisted language learning? Speech Communication 51(10), 906 – 919 (2009), spoken Language Technology for Education Spoken Language
9. Kartushina, N., Hervais-Adelman, A., Frauenfelder, U.H., Golestani, N.: The effect of phonetic production training with visual feedback on the perception and production of foreign speech sounds. The Journal of the Acoustical Society of America 138(2) (2015)
10. Linebaugh, G., Roche, T.: Evidence that L2 production training can enhance perception. Journal of Academic Language & Learning. 9(1), A1–A17 (2015)
11. McFarlane, A., Sparrowhawk, A., Heald, Y.: Report on the educational use of games. TEEM (Teachers evaluating educational multimedia), Cambridge (2002)
12. Muntean, C.I.: Raising engagement in e-learning through gamification. In: Proc. 6th International Conference on Virtual Learning ICVL. pp. 323–329 (2011)
13. Neri, A., Cucchiarini, C., Strik, H.: Automatic speech recognition for second language learning: How and why it actually works. In: Proc. ICPhS. pp. 1157–1160 (2003)
14. Tejedor-García, C., Cardenoso-Payo, V., Cámara-Arenas, E., González-Ferreras, C., Escudero-Mancebo, D.: Playing around minimal pairs to improve pronunciation training. IFCASL (2015)
15. Tejedor-García, C., Cardenoso-Payo, V., Cámara-Arenas, E., González-Ferreras, C., Escudero-Mancebo, D.: Measuring pronunciation improvement in users of CAPT tool TipTopTalk! Interspeech (2016)