A Game on Pronunciation Feedback in Facebook

Nikos Tsourakis
ISSCO/TIM/FTI
University of Geneva
Geneva, Switzerland
Nikolaos.Tsourakis@unige.ch

***Abstract*—The proliferation of social networking services has revolutionized the way people around the world interact and communicate. The need to overcome language barriers that impose hurdles to this human communication is more acute than ever. Computer Assisted Language Learning systems try to bridge the gap between the need and the availability of language services in education. The convergence of both worlds is an obvious choice. In this work we extend a successful CALL platform and integrate it to Facebook, the most popular social network. After creating a pronunciation feedback game we investigate the interaction patterns of four users and present some preliminary results from ongoing work.**

Keywords-component; Pronunciation Feedback; Computer Assisted Language Learning; Social Networks

I. INTRODUCTION

Social networks are a useful tool to study social relationships and social phenomena through the connections among individuals instead of examining the properties for each one of them. It has been shown that all people are connected to one another by an average of "six degrees of separation" (your friend is one degree away from you) [1]. A recent study [2] reports that the average separation distance in Facebook, the most popular social network, is even less. After performing a world-scale social-network graph-distance computation they have shown that this number has been reduced to 3.74 "degrees of separation". The flow of information in our social entourage is constant as what we do and think has an impact as far as our friends' friends' friends, following the "three degrees of influence rule" [3].

Little attention has however been paid to exploring the full potential of integrating CALL systems in social networks. Benefits like performance feedback by your friends and the ability to comment, to request help, or to share content, would offer a more pleasant and engaging experience to learners. In this work we have integrated our CALL-SLT system into Facebook, the most popular social network service with more than 800 million active users. CALLSLT [4] is a spoken conversational partner designed for beginner- to intermediate-level language students who wish to improve their spoken fluency in a limited domain. It offers about a dozen combinations of L1s and L2s and is freely available for use (cf. callslt.org for detailed instructions).

CALL systems that rely on the output of the ASR to assess language skills engender the risk of accepting a sentence when in reality the pronunciation was incorrect (false positive) or rejecting a correctly pronounced sentence (false negative). This can increase the confusion of users concerning their pronunciation competence. Different tools for pronunciation training for second language learning try to alleviate this problem by either examining the production of speech [5] or by identifying confusable contexts [6].

This paper presents some early results of a work in progress after designing a pronunciation feedback game in Facebook. The idea behind the game was simple. Users spoke sentences in L2 (French) and acquired a pronunciation score in a scale between 0 (non-native) to 100 (native). Their score was juxtaposed with the one of their Facebook friends that had also used the system. We tried to investigate possible implications of this interaction scheme and to extract interaction patterns.

The rest of the paper is organized as follows. Section 2 describes the CALL-SLT system, and Section 3 the design of the experiment. Section 4 presents some initial results. The final section concludes.

II. THE CALL-SLT SYSTEM

CALL-SLT [4] is an Open Source speech-based CALL application for beginning to intermediate-level language students who wish to improve their spoken fluency. The system is deployed in the Web using a server/client architecture as described in [7]. Most processing, in particular speech recognition and language understanding, occurs on a remote server. The core idea is to give the student a prompt, formulated in their own (L1) language, indicating what they are supposed to say; the student then speaks in the learning (L2) language, and is scored on the quality of their response.

A. Presentation of Prompts

One way that CALL-SLT differs from other work (e. g. [8]) is in its presentation of prompts to the students. Instead of giving students complete prompts in their own language (the L1), our system uses interlingua representations in L1. In this way we avoid the undesirable effect of tying the language being studied (the L2) too closely to the L1 in the student's mind. In this work, the interlingua is realized in a telegraphic textual form but it is possible to produce graphical and video realizations of it without changing the underlying architecture. We support forms for different L1s, including English, French, Japanese, Chinese and Arabic.

The system is loaded with a set of possible prompts that represent the target content for a given lesson. Each turn starts

with the student asking for the next prompt. The system responds by showing a surface representation of the underlying interlingua for a target L2 sentence. For example, a student whose L1 is French and whose L2 is English might be given the following textual prompt:

COMMANDER DE_MANIERE_POLIE SALADE

Valid responses to this prompt would "I would like a salad", "Could I have the salad?", or simply "A salad, please"; the grammar supports most of the normal ways to formulate this type of request.

B. Pronunciation Module

In order to test our ideas in practice, we implemented a module that could elicit, in a simplified way, the pronunciation competence of each user. Each input sentence by the learner was assessed with respect to some reference utterances from native speakers. Specifically, we utilized data from 6 native French female subjects and from 6 female intermediate-level language students. Each one of them provided 30 sentences; half of the subjects in each group were used for training the module and the other half for testing it. After normalizing the waveforms, we used the software program praat [9] to extract the mean pitch, the mean intensity and the mean of the first-to-fifth formant frequencies of each utterance and used different combinations of them as our feature space.

Support Vector Machines have proven effective in a wide range of classification tasks, and were also chosen as the candidate method for our pronunciation module. We experimented with different combination of features and results of the SVM method (polynomial kernel and trade-off between training error and margin 5000) using the WEKA Toolkit [10] are shown in Table 1. The dyad of mean intensity and mean fifth formant provided the lowest error rate (98.88%), comparable with the result obtained using as feature the mean of the fifth formant (98.33%). When the mean intensity was chosen as the single feature, the correct classification reached to 83.33%, This suggests that non-native subjects did speak less loud, an indication of feeling less confident. The role of second (F2) and third (F3) formants frequencies for foreign accent determination have been reported in previous studies [11]. In our case however the mean value of formants was calculated for the whole spoken sentence and not for isolated phonemes, something that might explain why we did not encounter improved performance using the F2 or the F3.

The binary output of the classifier (native/non-native) was a restrictive factor for our experiment as ideally we would like a specific score for the input sentence between 0-100. In order to alleviate this deficiency, the module randomly produced a score between 30%-70% when the subject was classified as non-native and between 70%-95% when she was considered as native (Figure 1). The strategy of using narrow ranges of scores was imposed in order to avoid confusion when users uttered a sentence in a similar way and would expect a similar if not identical score. The module, bundled with praat and WEKA, constituted yet another step in the processing chain of CALL-SLT online system, with an average overhead latency of around 300 msec.

TABLE I. CLASSIFICATION ERROR (PERCENTAGE) OF USERS' NATIVENESS / NON- NATIVENESS

Features	Mean of: Pitch (P), Intensity (I), Formant (F1-F5)			
	Correctly Classified	Precision	Recall	F-Measure
P	61.11%	63.20%	53.30%	57.80%
I	83.33%	76.80%	95.60%	85.10%
P-I	83.88%	79.60%	91.10%	85.00%
P-I-F1	84.44%	79.80%	92.20%	85.60%
P-I-F2	83.33%	78.80%	91.10%	84.50%
P-I-F3	84.44%	79.20%	93.30%	85.70%
P-I-F4	96.11%	97.70%	94.40%	96.00%
P-I-F5	**98.88%**	**100.0%**	**97.80%**	**98.90%**
P-I-F1-F2	83.33%	78.80%	91.10%	84.50%
P-I-F1-F2-F3	84.44%	81.00%	90.00%	85.30%
I-F5	98.33%	100.0%	96.70%	98.30%
F5	**98.33%**	**100.0%**	**96.70%**	**98.30%**

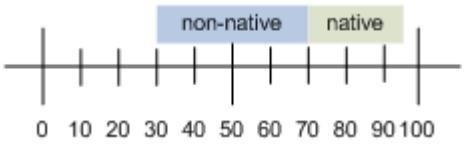

Figure 1. Score range for the native/non-native target groups

III. EXPERIMENTAL DESIGN

According to previous studies an opponent makes players more motivated and focused during a programming course [12] or a translation game [13], something that improves the effectiveness of the game per se. We therefore organized our experiment as a pronunciation game where social contacts constitute the potential opponents of users.

For each subject we constructed her contact's network as shown in Figure 2. The topology of the network was a concentric structure of two levels comprised of friends (one level) and friend's friends (two levels). There was also a third level with contacts outside user's network. We recruited subjects from our friends list in order to be able to utilize their social entourage. To avoid biases due to stereotypes all opponents were chosen from the same ethnic group (Greeks) and gender was approximately balanced. Moreover we picked image profiles with faces that didn't expose strong emotions. All participants had intermediate-level French language skills.

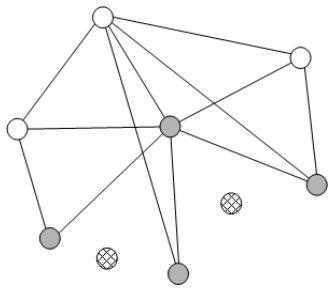

Figure 2. Topology of the contacts network. All nodes participated in the first round of the study (data collection) and only the grey modes in the second one (testing). Grid-filled nodes represent not connected subjects

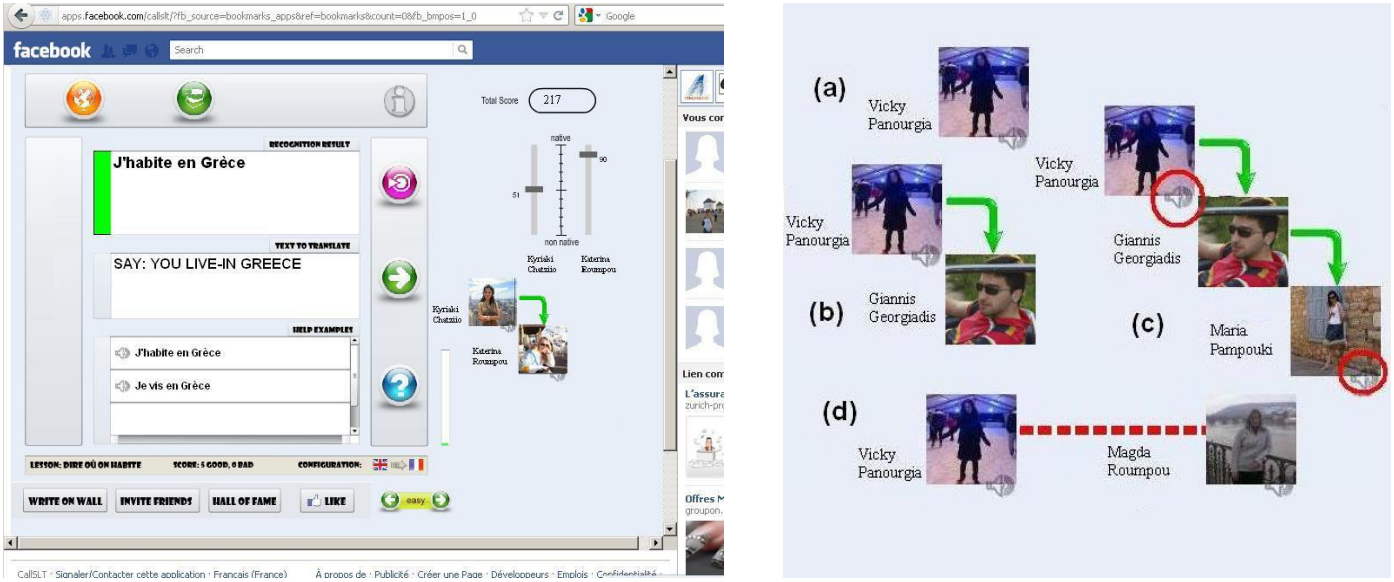

Figure 3. CALL-SLT as a Facebook application (left). On the left side the middle pane shows the prompt; the top pane, the recognition result; the bottom pane, text help examples. Pronunciation related widgets are presented on the right side. From bottom down, total score, pronunciation scale, contact connections. Presentation of the connections patterns (right). Alone (a), friend (b), friend of friend (c) not-connected to the subject (d). A speaker icon in the downright corner of the image signifies that the audio file is available for listening

The study was presented to the subjects as if we wanted to test a new feature, namely our new pronunciation module. The game was split into two rounds. During the first one all participants were asked to contribute to the database of sounds by using the application as they wished. Interacting with the standard system (no pronunciation assessment) was an essential exercise to familiarize themselves with the offered functionalities. Two weeks later the system was ready to be used for assessing users' pronunciation skills.

The interaction with the application is performed as follows. The prompt is presented in the middle textbox of Figure 3 (left). The user decides what she is going to say, presses the "recognize" (purple) button, and speaks. She may ask for acoustic help at any time by clicking the "help" (blue) button. After each turn the pronunciation score for the user and the one of the opponent are presented side by side in two vertical slide bars. We also animate the movement of the two sliders from the lower level (non-native) towards the upper one (native) to grab user's attention and focus on competition. The connection between the user and the opponent is presented as a cascaded sequence of profile images. The green arrow signifies connection, whereas the red dotted line the opposite (Figure 3(right)). For presenting scores and connections we considered the following:

- Results were offered only if the speech recognition was successful.
- Users were randomly exposed to four patterns in the same frequency, namely "Alone" (no comparison with a contact), "First" (comparison with a first level contact), "Second" (comparison with a second level contact) and "No Contact" (comparison with a not-connected subject).
- Users could repeat the same sentence multiple times while the opponent and his score remained the same.
- After a repetition of the same sentence the new user's score was in the range ± 15 of the score obtained in the previous turn, given that the classifier provided the same class (native or non-native).
- The total score was calculated as a sum of the pronunciation score at each turn and the game ended when it reached to 2500 points.
- The opponent's score using recognized waveforms was a random number between 15%-95%.

Our intention was not to impose direct competition among users by requesting the repeat of the same phrase if the subject got lower score compared to the one of the opponent but to let them decide when to do so. They had also the opportunity to listen to the opponent's waveform by clicking on his profile image or listen to their own last recorded utterance by clicking on their photo.

IV. Preliminary Results & Discussion

In this section we will provide some preliminary results obtained from 4 female users that interacted with the system. This first analysis focused on questions like: "Do users decide to repeat a prompt after they see the score of their contact?", "Does the difference in score matter?", "Is the distance of the contact important?", "Do users listen to the prompts of their opponents? ", etc. The results presented in Table 2 provide a first insight to the aforementioned questions.

As already mentioned the experiment was designed in a way so that participants would be equally exposed to the same connection patterns. In the "*Exposed to …*" rows we can observe that this was more or less accomplished. Each subject seems to have followed her own interaction style, which provides an initial distinction of user types:

TABLE II. RESULTS OF THE INTERACTION FROM 4 USERS

	User 1	User 2	User 3	User 4
Exposed to "Alone"	8	6	8	11
Exposed to "First"	6	6	9	11
Exposed to "Second"	6	8	9	12
Exposed to "No Contact"	8	6	9	12
Repeats after "Alone"	37.5%	66%	37.5%	9.09%
Repeats after "First"	50%	33.33%	11.12%	18.18%
Repeats after "Second"	37%	50%	44.45%	0%
Repeats after "No Contact"	25%	33.33%	22.23%	0%
Repeats after lower score	79.31%	65.38%	33.33%	5%
Repeats after higher score	0%	50%	28.57%	6.67%
Listen "Alone" prompt	0%	0%	12.5%	90.9%
Listen "First" prompt	0%	50%	66.67%	72.72%
Listen "Second" prompt	0%	0%	55.56%	58.34%
Listen "No Contact" prompt	0%	16.67%	22.23%	58.34%
Average prompt repetition	0	3	7.11	1.63

Competitive (User 1). The real incentive for her is the score balance. 79% of the times she got a lower score than the opponent's, she decided to retry. She feels so confident that she can do better by retrying that she doesn't need to listen to waveforms, even her own.

Experimenter (User 2). This subject uses all the functionalities offered by the system without explicitly focusing on one aspect of it.

Scholastic (User 3). On average this user listened to each contact's prompt 7.1 times, probably due to curiosity and in order to compare. She was not motivated too much to repeat the same sentence.

Social Spectator (User 4). The user doesn't show any signs of competition, which is why she rarely repeats a prompt. On the other hand, she exposes strong interest in listening to the prompts of her contacts (1.63 times).

In the previous categorization there is one to one association between user types and the four users. This distinction is our recommendation; it is indicative and by no means exhaustive. It might be useful to others intending to implement a similar protocol. More data is required to determine behavioral patterns and to form better distinctions.

A comment expressed by all participants was about the sensitivity of the pronunciation module. Their concerns were not related to its randomness, a fact they were not aware of, but mostly that they might not receive the same score when the sentence was spoken in the same way. Moreover there were times when it was not obvious what made an opponent's score better. Subjective self-assessment in language proficiency is always an issue even if a real pronunciation analysis takes place. Providing a cumulative score on pronunciation competence without specifying problematic regions in the spoken utterance would inevitably raise objections. In our case a more appropriate solution might be the substitution of the fine grained scale of the sliders with a scale of only 3-4 levels. The development of a more sophisticated pronunciation module would permit exploring the full potentials of our ideas.

V. CONCLUSIONS

In this work we tried to design a pronunciation game on Facebook and investigate the effects of integrating a CALL system with a social network. A core contribution is the addition of an element of motivation by involving another player/learner.

There were some inherent deficiencies in the game design however. Experimental subjects were recruited from the author's contact list in order for us to exploit their social entourage. An additional requirement was that they had an intermediate level in the French language. Both restrictions limited the study to few participants.

Finally, when deploying applications in social networks issues related to privacy must be addressed. During the recruitment phase a fifth subject politely rejected to contribute due to the fact that our Facebook application specifically requested access to her friends list.

REFERENCES

[1] P. S. Dodds, R. Muhamad and D. J. Watts, "An Experimental Study of Search in Global Social Networks," Science 301 (2003): 827-29.

[2] L. Backstrom, P. Boldi, M. Rosa, J. Ugander, S. Vigna, "Four Degrees of Separation," WebSci 2012.

[3] N. A. Christakis and J. H. Fowler, "Connected: The Suprising Power of Our Social Networks and How They Shape Our Lives," Little Brown: New York, 2009.

[4] P. Bouillon, S. Halimi, M. Rayner, N. Tsourakis, "Evaluating a web-based spoken translation game for learning domain language," Proceedings of INTED, Valencia, Spain, 2011.

[5] H. Strik, K. Truong, F. deWet, and C. Cucchiarini, "Comparing different approaches for automatic pronunciation error detection," Speech Communication, vol. 51, no. 10, pp. 845–852, 2009.

[6] O. Saz, M. Eskenazi, "Identifying Confusable Contexts for Automatic Generation of Activities in Second Language Pronunciation Training," Proceedings of the SLaTE Workshop, Venice, Italy, 2011.

[7] M. Fuchs, N. Tsourakis, M. Rayner, "A Lighweight Scalable Architecture For Web Deployment of Multilingual Spoken Dialogue Systems," Proceedings of LREC 2012, Istanbul, Turkey.

[8] C. Wang and S. Seneff, "Automatic assessment of student translations for foreign language tutoring," Proceedings of NAACL/HLT 2007.

[9] P. Boersma and D. Weenink. Praat: Doing phonetics by computer. http://www.praat.org/.

[10] M. Hall, E. Frank, E. G. Holmes, B. Pfahringer, P. Reutemann, I. Witten, "The WEKA Data Mining Software: An Update," SIGKDD Explorations, Volume 11, Issue 1, 2009.

[11] L. M. Arslan, J. H. L Hansen, "A Study of Temporal Features and Frequency Characteristics in American English Foreign Accent," The Journal of the Acoustical Society of America, July 1997.

[12] R. Lawrence, "Teaching data structures using competitive games," Education, IEEE Transactions on, vol. 47, no. 4, pp. 459 – 466, 2004.

[13] W. Ling, I. Trancoso, R. Prada, "An Agent Based Competitive Translation Game for Second Language Learning," SLaTE Workshop, Venice, Italy, 2011.