Automatic evaluation of the pronunciation with callslt, a conversation partner exclusively based on speech recognition

EICHENBERGER, Fanny, et al.

Abstract

At a time when markets are globalizing, second language acquisition is increasingly important in current society. More than a way to stand out, multilingualism has become a necessity, especially in multilingual countries such as Switzerland. Like many aspects of the day to day life, language learning is being revolutionized by technological progress. Today, CALL (Computer-assisted language learning) software allows users to acquire a new language at their own pace, without any constraints of time and place. Thanks to the great evolution of speech recognition technology, it is now possible to interact orally with computers. On this basis, the Faculty of Translation and Interpreting of Geneva University developed its own software called CALL-SLT, a conversation partner exclusively based on speech recognition. Since its usefulness as a learning coach has already been proven, we propose a new experiment aiming to discover if CALL-SLT can also play the role of an automatic evaluator of the L2 pronunciation. 17 French speakers were given access to the tool and took a test specially developed for this purpose. In this paper, we [...]
AUTOMATIC EVALUATION OF THE PRONUNCIATION WITH CALL-SLT, A CONVERSATION PARTNER EXCLUSIVELY BASED ON SPEECH RECOGNITION

F. Eichenberger¹, P. Bouillon¹, J. Gerlach¹, M. Déjos²

¹University of Geneva (SWITZERLAND)
²Intercountry (FRANCE)

Abstract

At a time when markets are globalizing, second language acquisition is increasingly important in current society. More than a way to stand out, multilingualism has become a necessity, especially in multilingual countries such as Switzerland. Like many aspects of the day to day life, language learning is being revolutionized by technological progress. Today, CALL (Computer-assisted language learning) software allows users to acquire a new language at their own pace, without any constraints of time and place. Thanks to the great evolution of speech recognition technology, it is now possible to interact orally with computers. On this basis, the Faculty of Translation and Interpreting of Geneva University developed its own software called CALL-SLT, a conversation partner exclusively based on speech recognition.

Since its usefulness as a learning coach has already been proven, we propose a new experiment aiming to discover if CALL-SLT can also play the role of an automatic evaluator of the L2 pronunciation. 17 French speakers were given access to the tool and took a test specially developed for this purpose.

In this paper, we will present the method we used and summarize the results. While these are not perfect, they are very encouraging and suggest that an evaluation of pronunciation with speech recognition is indeed possible.

Keywords: pronunciation, L2, CALL, computer-assisted pronunciation training, CAPT, CALL-SLT, speech recognition, automatic evaluation.

1 INTRODUCTION

At a time when markets are globalizing, second language acquisition is increasingly important in current society. More than a way to stand out, multilingualism has become a necessity, especially in multilingual countries such as Switzerland. Like many aspects of the day to day life, language learning is being revolutionized by technological progress. Today, CALL (Computer-assisted language learning) software allows users to acquire a new language at their own pace, without any constraints of time and place. Thanks to the great evolution of speech recognition technology, it is now possible to interact orally with computers. On this basis, the Faculty of Translation and Interpreting of Geneva University developed its own software called CALL-SLT, a conversation partner exclusively based on speech recognition. CALL-SLT is a speech enabled conversation partner, designed for the easy development of different types of language courses. The basic principle of the system consists in prompting the language learner, by a text or multimedia prompt, to produce a spoken utterance. The system then uses grammar based speech recognition to determine whether the spoken utterance is an adequate response to the prompt, and thus accepts or rejects it. This principle is used to create courses ranging from simple vocabulary training to elaborate dialogs.

Several experiments on CALL-SLT [1] have shown that speech recognition allowed users not only to learn grammar and vocabulary in second languages, but also to practice oral production and to improve pronunciation skills [2][3][4][5]. Since the particularity of this tool is that users can only interact orally with the system, we decided to focus on pronunciation. As the usefulness of CALL-SLT as a learning coach has already been shown, we chose to explore the possibility of using CALL-SLT for the evaluation of pronunciation skills. This experiment was carried out in collaboration with Intercountry, a French company specializing in teaching engineering. Our experiment has three major aims: 1. to discover whether grammar based speech recognition technology can assess pronunciation; 2. to determine whether automatic evaluation can be as reliable as human evaluation; 3. to see whether it is possible to automatically attribute a language level to each candidate.
The article is structured as follows: Section 2 gives the methodology; section 3 presents the experiment; section 4 gives the main results; section 5 concludes this article.

2 METHODOLOGY

This study was carried out in three main steps: design of the test, pre-test and experiment.

The first step was to develop a test with the aim of measuring the pronunciation. To be as accurate as possible, we decided to combine three kind of tasks: discrimination, where the candidate must identify which word differs among two very similar sentences he hears; repetition, where he must repeat the word he hears; and reading, where he must read aloud a sentence written. These three types of exercise allow us to get a better analysis of the potential mistakes made by the candidates: the discrimination task enables us to check the perception of sounds; the repetition task the production of sounds; and the reading task the grapheme-morpheme association. Table 1 (below) summarizes the general structure of the different tasks.

<table>
<thead>
<tr>
<th>Table 1. Summary of the different tasks.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Task</td>
</tr>
<tr>
<td>Type of stimulus</td>
</tr>
<tr>
<td>Form of stimulus</td>
</tr>
<tr>
<td>Form of answer</td>
</tr>
<tr>
<td>Number of prompts</td>
</tr>
</tbody>
</table>

In addition to the form of the evaluation, we also had to define the specific sounds we wanted to test. To meet the needs of Intercountry, we targeted French learners of English. To this end, we decided to test some typical mistakes made by French speakers in English [3]. We selected four English-specific types of difficulty: 1. the production of the dental fricatives /θ/ et /ð/; 2. the production of the glottal fricative /h/; 3. the distinction between long and short vowels such as /ɑː/-/æ/ and /iː/-/ɪ/; 4. the production of diphthongs such as /aɪ/ et /əʊ/. As CALL-SLT evaluation is binary (accepted vs. rejected answers), these four types of sounds were then integrated into minimal pairs. This way, we could test the production of a sound by telling the system which element of the pair should be considered as correct (the correct reference pronunciation, positive answer) and which one should be rejected (the “French pronunciation” of the word, negative answer). For example, to test the production of the dental fricative /θ/, we used the minimal pair mouth-mouse since the French speakers often tend to substitute /θ/ by the French phoneme /s/. Figures 1 and 2 present CALL-SLT interface for the three tests.

![Figure 1. On the left: interface of the discrimination task. On the right: interface of the repetition task.](image-url)
The second step corresponds to a phase of pre-test. Before starting the experiment, we had to check the accurateness of the speech recognition and, if needed, to adjust its degree of severity. To this end, we asked three native English speakers to complete the test. They were expected to pass the test perfectly, since their pronunciation is considered as the reference. The results of this pre-test were very encouraging: 94% of the answers (66 out of 70 answers) were correctly evaluated by the program. This suggests that the speech recognition is quite accurate. As the four rejected answers corresponded to four different prompts and because of the small number of participants, we decided to leave the test unchanged and to proceed to the experiment.

The third step, which we will detail in the next section, is the experiment.

3 EXPERIMENT

Our main objective is to discover if CALL-SLT, already well-known as a learning coach, can be turned into a pronunciation evaluator. In this experiment, the objective was to compare the judgements given by CALL-SLT with those given by human evaluators. 17 native French speaking learners of English took part in the experiment: 10 women and 7 men from 24 to 65 years old without any specific background in language technologies. All candidates completed the three tasks as described in Section 2. The experiment was carried out in two steps: first, candidates took the test and were assessed by CALL-SLT. Next, the data collected by CALL-SLT was assessed by an Intercountry English teacher. As CALL-SLT evaluation is binary, we asked the human evaluator to follow the same principle by indicating for each prompt if she would accept or refuse the pronunciation heard.

4 RESULTS

The results of both types of evaluation were combined in the same Excel document. We created one table per candidate and indicated for each prompt the evaluation provided by CALL-SLT and the evaluation provided by the English teacher. A total of 442 data – 26 prompts for each candidate – were collected and assessed under the three status “accepted”, “rejected” or “null”.

To answer our three research questions, we analysed the results under two different angles: CALL-SLT responses vs the human evaluation of individual answers, and CALL-SLT scores vs language level attribution.

4.1 CALL-SLT responses vs human evaluation

The quantitative analysis of the results, which relates directly to our two first research question, aims to discover if speech recognition accepts and rejects answers as a human does.

As already mentioned, CALL-SLT assesses pronunciation by accepting or rejecting the user’s answers based on the pronunciations defined in its grammar. A third case occurs when the recording of the answer is bad due for example to poor microphone quality or manipulation errors, or if the confidence score of the recognition result for an answer is too low, i.e. is below a predefined confidence threshold. In those cases, the answer is rejected and the system returns a “nothing recognized” response. We will refer to these cases as “null” responses. Similarly, the human evaluator
was asked to mark items as “null” when pronunciation assessment was difficult or impossible. Figure 3 shows combined results for the three categories.

![Figure 3. Accept, Reject and Null results for CALL-SLT and human evaluator.](image)

We observe that CALL-SLT accepted less answers than the human evaluator and rejected about the same number. The largest difference is found for the null answers, with 126 answers considered as “null” by CALL-SLT against only 3 by the English teacher. For the user, these are equivalent to rejections, making CALL-SLT globally more severe than the human evaluator. Since the system’s null answers include answers that the system recognised but discarded due to a too low confidence score, an adjustment of the confidence threshold would reduce the null responses, and distribute these items among the two other categories.

Because of the high amount of null answers in CALL-SLT, we decided to compare, for each prompt, the assessment provided by the speech recognition and by the human evaluator. This way, we will be able to answer our second research question and discover if both types of evaluation can coincide. Table 2 classifies the answers according to the status that were attributed by CALL-SLT and by the teacher.

<table>
<thead>
<tr>
<th></th>
<th>TEACHER</th>
<th>Accepted</th>
<th>Rejected</th>
<th>Null</th>
</tr>
</thead>
<tbody>
<tr>
<td>CALL-SLT</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Accepted</td>
<td>261</td>
<td>4</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>Rejected</td>
<td>29</td>
<td>21</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>Null</td>
<td>96</td>
<td>28</td>
<td>2</td>
<td></td>
</tr>
</tbody>
</table>

Table 2. Comparison of the status attributed by CALL-SLT and the teacher.

This analysis provides us a clearer view of CALL-SLT 126 null answers (numbers with a black circle in Table 2). Indeed, we can see that the teacher also invalidates 2 and rejects 28 of them. This means that 30 null answers from CALL-SLT correspond to items that must rightly be refused.

In parallel, we can notice that CALL-SLT evaluation is reliable, namely coincide with the human evaluation, in 312 out of 442 cases (71%). These correspond to the yellow boxes in Table 2. In 5 cases (the green boxes), automatic evaluation was too tolerant by accepting answers that were refused by the coach. On the contrary, automatic evaluation was too severe in 125 cases (the red boxes): it refused answers that were accepted by the teacher. This allows us to conclude that the
quality of the automatic evaluation is reliable in almost 3 out of 4 cases and that, when it is not, the assessment is mainly too severe.

4.2 Language level attribution

As part of our third research question, we asked the teacher to give an approximate language level to each candidate according to the Common European Framework of Reference for Languages. The objective of this is to discover if a correlation between the score obtained on CALL-SLT by the candidates, namely the number of accepted answers, and their language level is possible. Table 3 shows a summary of the results.

<table>
<thead>
<tr>
<th>Level given by teacher</th>
<th>Number of candidates</th>
<th>Average CALL-SLT score</th>
</tr>
</thead>
<tbody>
<tr>
<td>A2</td>
<td>3</td>
<td>11.3</td>
</tr>
<tr>
<td>B1</td>
<td>6</td>
<td>15.6</td>
</tr>
<tr>
<td>B2</td>
<td>6</td>
<td>15</td>
</tr>
<tr>
<td>C1</td>
<td>2</td>
<td>23.5</td>
</tr>
</tbody>
</table>

Because of the large amount of null answers with CALL-SLT, the correlation is difficult, particularly for intermediate levels B1 and B2. However, we noticed a certain trend in extreme levels: C1 and A2 levels oscillate around 23 and 11 good answers respectively. This is quite positive and suggests that a correlation between score and level could be possible.

5 CONCLUSION

Before concluding this paper, it is important to keep in mind that our experiment was conducted with a small panel of participants. In addition, we had to face the withdrawal of one English evaluator. Therefore, the human evaluation was provided by only one evaluator, which gives a subjective character to the results of this experiment.

Even if they are not perfect, we can nevertheless note that the results are very encouraging.

Although the evaluation made by CALL-SLT does not perfectly match the evaluation made by the native speaker, we found out that automatic evaluation of the pronunciation by CALL-SLT is actually possible. Moreover, the speech recognizer and the severity of the automatic evaluation can still be adjusted, notably through the confidence threshold of the system. This would allow the automatic evaluation to agree more closely with the human evaluation. Finally, we noticed a correlation between the score obtained on CALL-SLT by the candidates and their language level, which suggests that an automatic level attribution is also practicable.

REFERENCES

[1] [http://callslt.unige.ch/](http://callslt.unige.ch/)