Addressing Confusions in Spoken Language in ESL Pronunciation Tutors

Oscar Saz, Maxine Eskenazi
Language Technologies Institute, Carnegie Mellon University, Pittsburgh PA, USA
osaz@cs.cmu.edu, max@cs.cmu.edu

Abstract
This paper presents a new approach for developing pronunciation tutors in Second Language (L2) learning. Applying the Basic Identification of Confusable Contexts (BICC) procedure we automatically generate curriculum that is rich in possible confusion contexts which can be practiced by L2 students in read-aloud tasks. This is the basis of a new pronunciation tutor where the student interacts orally in a scripted conversation with a virtual avatar. The student’s utterances are automatically evaluated to detect mispronunciations and the tutor provides feedback on the errors which produce a confusion in the message. An assessment was carried out with a group of English as Second Language (ESL) speakers to evaluate the tutor interface as well as to determine their motivation and engagement in practicing English with it.

Index Terms: pronunciation, intelligent tutors, error detection

1. Introduction
There is currently much literature on Computer-Aided Language Learning (CALL) tools for the pronunciation training of a Second Language (L2) [1]. These tools make use of various pinpointing techniques, Goodness of Pronunciation (GOP) [2] being the most popular one. They have shown a solid ability to detect phonetic errors and inaccuracies in speech in a variety of languages and tasks [3, 4]. When the pinpointing algorithm detects an error in the pronunciation of a student, feedback is provided showing the error as well as strategies on how to improve pronunciation.

The focus so far has been solely on the production of speech, checking whether the student’s speech is correct from the articulatory and acoustic points of view. What has been missing is the context, the overall effect that a mispronunciation has on the comprehension of the speech. Thus, all the errors made by an L2 student have been treated equally important when correcting the student. However, in many cases the error can be recovered by the listener due to our ability to predict meaning from context. On the other hand, there are cases where a single pronunciation error can be critical to the efficient transmission of information. This happens, for instance, with words that have minimal pairs (words differing in one phoneme). When this happens, the listener will ask for a clarification or will interpret a different meaning.

Minimal pairs have been used in L2 pronunciation tutors before [5]; but, again, they were only used to evaluate the production of the phonemic contrast.

A tool which could assess and correct the cases that present comprehension issues in a realistic environment can be an extension of the curriculum in regular L2 classes. In [6] we showed how Basic Identification of Confusable Contexts (BICC) could automatically identify cases where confusion is likely in a given text, selecting the most interesting ones for student practice. Now a new software tutor which incorporates this curriculum has been developed and tested. This tutor had to have an appealing interface, provide accurate error detection, and give feedback in a way which is helpful for the student.

The organization of this paper is as follows: Section 2 describes the dialog-based tutor for correcting spoken language confusions in English as a Second Language (ESL) speakers. A study carried out with a group of ESL speakers is presented in Section 3, and its results provided in Section 4. Finally, Section 5 gives the conclusions.

2. Dialog-based Tutor for ESL Pronunciation Training
This Section presents the dialog-based tutor for ESL pronunciation training where the curriculum is based on dialogs analyzed with BICC [6]. The dialogs are not open dialogs as in a spoken dialog system, but scripts prepared previously that the user has to pronounce correctly.

The tutor was designed with an easy interface where the user can choose its basic elements from the main window and directly access the practice session. Prior to the practice session, the user selects the native language and the dialog they want to practice. From that point on, the user’s speech is the only mode of interaction. The dialog is elicited as shown in Figure 1, where an avatar in the upper part of the screen plays the role of the conversation partner and the user’s avatar is at the bottom of the screen. In each turn, the virtual partner utters its turns using the Lernout & Hauspie TTS3000 speech synthesizer, then the sentence to be uttered by the user is shown on screen and the system starts the PocketSphinx toolkit [7] to process the user’s speech. To promote engagement, the user can personalize their avatar, providing it with a face and a name, while the computer avatar is randomized.
2.1. Evaluation and feedback

Each student utterance is analyzed as follows: First, if the quality of the input audio is low (high Signal-to-Noise Ratio (SNR) or clipping) or the user is not on-task, the user is reprompted with the same sentence. If the utterance passes this filter, the embedded speech verification method evaluates each target word in the sentence and decides whether it was correctly pronounced or not. If a mispronunciation which can create confusion is made, it is stored for later feedback and the next word is evaluated. When all the words have been evaluated, the next turn is prompted. The decision of whether a mispronunciation has occurred is made with a likelihood ratio between the posterior probability of the segment of speech \(X\) where the \(n\)-th word \(W_n\) occurs given its acoustic model \(\lambda_{W_n}\) and the posterior probability of the same segment given the model of the confusable word \(W'_n\), as in Equation 1. Techniques based on likelihood ratio have been used in utterance verification tasks and pronunciation evaluation [2].

\[
LR(X, \lambda_{W_n}, \lambda_{W'_n}) = \frac{P(X|\lambda_{W_n})}{P(X|\lambda_{W'_n})} \tag{1}
\]

After all the turns of the dialog are completed, feedback is provided. As long as there are errors in pronunciation which have not yet been reviewed, the virtual avatar presents the next error to correct. The user can see the word mispronounced in the context where it was uttered and the word that was actually said as in Figure 2. The user has to repeat the correct context and then the system advances to the next word. When there are no more errors to review, the list of words correctly pronounced by the student in the session is shown and the sessions finishes.

2.2. Generation of the curriculum

The dialogs practiced in the tutor are automatically generated with the BICC procedure [6]. This procedure takes an input sentence, \(S\), and for each one of its \(N\) words \((W_n)\) it finds its \(M\) minimal pairs \(\{W_{n1},...,W_{nm},...,W_{nM}\}\) in a pronunciation dictionary. A set of \(S'_n\) sentences are created where a word, \(W_n\), is replaced by a minimal pair \(W'_n\). The sentences which are grammatically correct after a Part-of-Speech (PoS) analysis are kept as confusion sentences of the original sentence \(S\), while the other sentences are discarded. Each of the user turns in a practice dialog is analyzed in this way and all the possible confusions are stored in the tutor to provide the error detection and feedback. The procedure was successfully evaluated in [6] and currently also affords personalization the curriculum to include only the minimal pairs that are difficult given the native language of the student.

3. Assessment with ESL Speakers

37 ESL speakers participated in a study to evaluate the new pronunciation tutor. They were volunteers recruited via posters and fliers. The native languages of the subjects can be seen in Figure 3, with a predominance of Chinese and Spanish speakers. The other languages were Arabic, Farsi, French, German, Hebrew, Japanese, Korean, Malay, Marathi and Telugu. The average age of the subjects was 26 (minimum 18, maximum 37). The distribution in gender was 23 male and 14 female. Regarding their self-assessment of their competence of spoken English, there were 12 intermediate speakers, 14 advanced speakers and 10 fluent speakers. Only 1 participant considered himself to be a beginner.

Each subject participated in two sessions with the dialog-based pronunciation tutor. On the first day, they took a pre-test which consisted of recording one minute of free speech summarizing a short text that they read before the recording. Then they practiced 3 of the dialogs and filled a short survey about their experience. The second day, they had another 3 dialogs, a post-test, which consisted of one minute of free speech after reading another short text, and the final survey. Participants had to
The main theme of the film was the connection between farming and rural areas.

<table>
<thead>
<tr>
<th>Confusable contexts</th>
<th>The main team of the film was the connection between farming and rural areas</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>The main theme of the film was the collection between farming and rural areas</td>
</tr>
<tr>
<td></td>
<td>The main theme of the film was the connection between harming and rural areas</td>
</tr>
</tbody>
</table>

The main theme of the film was the connection between farming and rural areas.

<table>
<thead>
<tr>
<th>Confusable contexts</th>
<th>I developed a way to allow a quick and simple access to the company resources</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>I developed a way to allow a quick and simple access to the company resources</td>
</tr>
<tr>
<td></td>
<td>I developed a way to allow a quick and simple access to the company resources</td>
</tr>
</tbody>
</table>

Table 1: Examples of sentences with confusable contexts used in the study (AWL words in boldface, GSL words in italic).

wait at least one week to come the second time. All the actions of the user and the tutor were logged in.

Each dialog had a total of 10 turns, with one sentence per turn to be uttered by the subjects. The set of words selected to be practiced by the subjects included 26 words from the Academic Word List (AWL), which were practiced by all the subjects, and several words from the General Service List (GSL) differing from one subject to another depending on their native language. The sentences contained these words in contexts which were confusable in the case of a mispronunciation, as in Table 1.

4. Results

The questions that the participants had to answer after the first session of the study concerned their experience and their perception of the interaction with the software. The answers were given according to a Likert scale of 1 to 7 points or 1 to 5 points. The results are provided in terms of the mean and standard deviation of the responses.

While the synthetic voice was not considered natural (3.9±1.5/7) nor friendly (4.2±1.7/7), participants answered that they did not have problems understanding it (5.9±1.3/7), although non-natives usually find it more difficult to understand synthesized speech than their native peers. Participants considered that the tutor could recognize their speech properly (5.0±1.4/7) and considered the feedback on their speech as accurate (5.4±1.3/7). The tutor was considered to be intuitive to use (6.1±1.1/7) and they always understood what to do next during the session (6.4±0.8/7).

Participants rated highly the ability to change their avatar’s face and name (5.5±1.2/7). They thought that the dialog had a natural flow (3.5±1.0/5) that kept them engaged (5.0±1.4/7) and that it was more interesting than reading sentences aloud (5.6±1.3/7). They considered receiving feedback to be very instructive (4.1±0.9/5) and when asked if they would have preferred to skip the feedback, they answered negatively (2.2±1.3/7).

The survey at the end of the second day was aimed at learning the subjects’ perception of how useful the tutor would be for them and of what they had learned with it.

They felt that the activities of the tutor were useful (5.6±0.9/7) and they would be interested in continuing to use it (5.4±1.1/7), although in general they did not seem to consider working with tutors as useful as speaking with real people (3.8±1.7/7). They self-reported a good level of motivation (5.4±1.1/7) when working with the tutor. Participants considered that they understood better how certain words get confused (6.1±0.9/7) and discovered words which sound like other words (5.9±0.8/7). Participants were not explicitly told that the tutor only corrected pronunciation on confusable contexts, so it helped them to find out these difficult contexts.

30 of the 37 participants were able to remember words that they had been practicing during the sessions with the tutor. 3 participants remembered one word, 11 participants remembered 2 words, 11 participants could remember 3 words and 5 remembered 4 words (they were asked to write down no more than 4-5 words). This indicates that they retained some words they need to further improve on and that they valued the feedback provided by the tutor sufficiently to keep them in memory. When they were asked to indicate which further improvements they would include in the tutor, 24 asked for a more natural synthesized voice, 22 asked for more extensive feedback on words and contexts, 33 asked to be able to listen to their own pronunciation and 28 asked for graphic tips on pronunciation (like a talking head).

4.1. Analysis of the log files

One of the elements that the tutor logged was when the user was made to repeat a sentence. As explained previously, this happened when the utterance had noise or clipping or the actual utterance did not correspond to the sentence prompted on screen. We studied the average percentage of utterances the tutor had the users repeat per day and per dialog in Figure 4. The first dialog of each day had an average of 8-9% reprompted turns, this number was reduced in the following dialogs below 5%. This decrease was significant with p < 0.05, indicating the entrapment of the users to the system.

Another element logged by the tutor was how many mispronounced words the user had uttered in each dialog. The actual utterances of the subjects have not been manually transcribed or labeled, so a full evaluation of the performance of the speech evaluation technique has not yet been possible. However, Figure 5 shows the average number of mispronunciations per dialog for the 3 groups of participants for their self-assessed level of spoken English. Intermediate speakers produced 3.5 mispro-
nunciations on average per dialog, while advanced speakers produced 3 mispronunciations and fluent speakers 2.5. The average number of words that were practiced in each dialog was 21. Fluent speakers outperformed the other groups significantly ($p < 0.05$) and advanced speakers also outperformed their intermediate peers ($p < 0.1$).

Finally, the tutor also logged the words that each subject had mispronounced during the dialogs. For the Chinese and Spanish speakers we looked for the AWL words in which each group underperformed the other. The results, showed how Chinese speakers significantly ($p < 0.05$) produced more errors than the Spanish speakers in the pairs ‘core’-‘call’, ‘expand’-‘expend’ and ‘phase’-‘face’; while Spanish speakers underperformed their Chinese peers in the pairs ‘varied’-‘buried’ and ‘clause’-‘close’. These results showed a consistency with the a priori knowledge of phonology of both languages. For instance, native Spanish speakers produce both graphemes ‘B’ and ‘v’ as the plosive sound /B/, which lead to errors pronouncing the word ‘varied’. This emphasizes the need for personalized curriculum in practice.

5. Conclusion

This paper has presented a new perspective for tutors for ESL learners with a focus on scripted dialog scenarios where student correction is focused on pronunciation errors that impede communication. A dialog-based pronunciation tutor has been developed and tested with a group of ESL speakers. The study has been a successful proof of concept of our novel ideas for pronunciation training, as shown by the surveys and the log files datamining. In the future, a classroom study could be carried out to assess the possibilities in a blended education environment.

The main contribution made here is the use of the BICC procedure [6] to automatically create curriculum for these dialog tasks. BICC can analyze sentences and texts and accurately find the words with minimal pairs where a confusion could appear, additionally allowing personalization to the needs of the learner. According to the surveys, the subjects learned and understood how confusions appear in specific contexts in spoken language and this is a novel result that we have not seen previously.

However, there is still work to do to enhance the capabilities of the tutor and solve some of the issues raised by the participants in the study. Further improvement can change how the feedback is presented, giving the L2 learners more tips on how to avoid their confusion errors; for instance, allowing the users to listen to themselves. The total integration of the BICC procedure in the dialog-based tutor (so that teachers and students can directly create their curriculum) is a necessary task as well, in order to encourage students to use it outside the lab.

Another outcome of this work is the acquisition of the speech samples uttered by the participants. This data creates a new database of more than 1 hour of free speech from the pre and post tests and approximately 6 hours of sentences uttered by the subjects during the practice. The data can be used to evaluate the capabilities of the system to provide accurate detection of mispronunciations.

6. Acknowledgments

Oscar Saz was supported by a Fulbright/MEC fellowship.

7. References