Improving Speech Recognition for a Communicative CALL Task

Michael Emonts

Deryle Lonsdale

C. Ray Graham

Michael Rushforth

Improving Speech Recognition for a Communicative CALL Task

Michael Emonts, Deryle Lonsdale, C. Ray Graham, Michael Rushforth,
PSST! Research Group, Brigham Young University

When speech application developers produce communicative activities for computer-assisted language learning (CALL), they often experience difficulties in getting the off-the-shelf programs to recognize learner responses with a high degree of accuracy. This paper describes the measures we have taken to improve recognition rate with the OGI toolkit and chronicles the improvement based on each change made. Changes include (1) limiting the universe of possible (probable) responses by making the task more explicit; (2) applying linguistic knowledge to the altering of word pronunciations; (3) converting .wav files to other file types; (4) altering the recognizer used for speech recognition.

Need for Interaction in Language Learning

Language acquisition research has strongly suggested that there are two major learner activities that contribute to second language acquisition: receiving comprehended input and producing comprehensible output (for summaries of this research see Ellis 1994, 273–84; Lightbown and Spada 1999, 39; and Gass and Selinker 1994, 200–201, 276–78). For many years now, computers with interactive multimedia capabilities have provided an excellent source of comprehensible input for language learners. And interactive exercises provided through the computer interface have provided a mechanism through which comprehension could be verified, thus assuring that the material was not only comprehensible but also comprehended by the particular learner. The computer interface has also provided a number of activities through which learners can practice speaking skills in a rather mechanical way.

CALL, however, has lacked the capacity to provide opportunities for the learner to practice what language acquisition specialists have called comprehensible or forced output—that is, oral output that is comprehended and responded to appropriately. This inability of computers to respond appropriately to oral output from learners has been a major drawback to CALL (Egan 1999, 280–81; Harless, Zier, and Duncan 1999, 313–14).

Why Use the OGI Toolkit

The CSLU/OGI toolkit[1] best suited our needs in designing interactive language learning tasks. The OGI toolkit provides a user-friendly graphical interface with high-level functionality that enables users to easily design dialog scenarios. Along with the toolkit comes Baldi, an agent featuring modern facial animation technologies that are articulatorily correct (see Figure 1). The use of Baldi provides learners an opportunity to interact with a visible agent, thus making the overall interaction more realistic and meaningful. The fact that it is freely distributable and has a well-documented website further adds to the benefits of using the OGI toolkit.

Figure 1. Baldi

Improving Speech Recognition Accuracy

The "directions" activity was chosen as the task in which to improve the system. In this activity, a map is presented showing the location of the student's house and the current location of the student's friend who is trying to visit (see Figure 2). The learner's goal is to provide step-by-step directions that will lead the friend to the student's house. When a correct direction is given, the friend then follows that direction, thus coming closer to the student's house (see Figure 3).

In order to improve the recognition accuracy for this activity, a program needed to be created that would enable us to quantify the accuracy of each speaker's utterances and to eliminate the need for extensive testing. To do so, we collected samples of correct directions from various adult speakers and stored them in .wav format. These files were then used as examples of input to the system that should be recognized, thus speeding up the testing process. Recognition accuracy was then quantified by simply determining what percentage of utterances were recognized correctly, as specified by a phrase-structure grammar. Although the corpus collected was relatively small, it provided a benchmark upon which we could progress. The program that was developed for testing accuracy allowed us to isolate factors that lead to a poor recognition rate.

Figure 2. Directions need to be given to lead the student's friend to the correct house

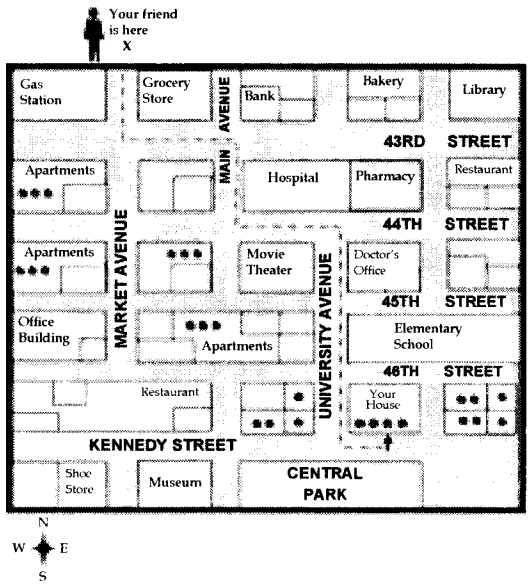

Figure 3. As correct directions are given, progress is shown

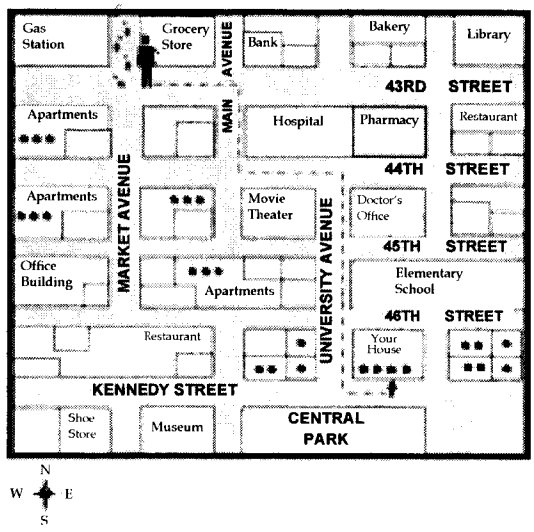

Table 1. Effect on accuracy by various recognition system input parameters

Test	Number Correct	Total Number of Files	Percent Correct
Grammar			
Controlled grammar	24	24	100%
One vs. two	13	24	54%
Go vs. walk	6	24	25%
North, south, east, west	19	24	79%
North, south, east, west	33	100	33%
Optional [kh] in "walk"	38	100	38%
Optional [kh] in "block"	34	100	34%
Recognizers(N,S,E,W)			
Adult_8kHz_0	33	100	33%
Adult_8kHz_0	39	100	39%
Child_16kHz_0	77	100	77%
Child_16kHz_1	92	100	92%
Adult_16kHz_0	97	100	97%
Grammar			
One vs. two	94	100	94%
One through five	86	100	86%
File types			
LINEAR	86	100	86%
RIFF	86	100	86%
NIST	92	100	92%
Grammar			
Walk vs. go	93	100	93%
Flexible syntax	91	100	91%
Optional [kh] in "walk"	93	100	93%

The first step was to simplify the grammar to the extent that it consisted only of the correct direction (i.e., "go south one block"). It is no surprise that 100% accuracy was obtained with a deterministic grammar. From this point, however, any change made to the grammar would lower accuracy percentage and allow us to observe the sources of recognition error. Table 1 shows the changes made to the controlled grammar and their influence on recognition rate. Introducing the number "two" to the grammar, for instance, made the system have to determine whether the input utterance was "go south one block" or "go south two blocks." Unfortunately, this small change in grammar reduced the response accuracy (i.e., the system's ability to correctly understand the speaker's utterance) to only 54%. Introducing "walk" to the controlled grammar reduced the accuracy to only 25%. Allowing the grammar to determine which direction was spoken (i.e., north, south, east, or west) was not as troublesome, yielding 79% accuracy.

At this point, it became clear that a larger number of test files was required to ensure reliability of testing results. The number of files tested was increased

to 100, and surprisingly, the above grammar distinguishing the four directions only yielded 33% accuracy.

Because each word in the grammar was recognized using the default pronunciations that came with the toolkit, changes made to the pronunciations appeared to marginally improve results. For example, the representation of "walk" in Worldbet, the transcription scheme used in the toolkit, is {w > kc kh}. When a speaker says, "walk two blocks south," however, the "k" is rarely aspirated as indicated by the "kh" in the toolkit's pronunciation (cf. Ladefoged 1993, 49–55). Making the aspiration on the "k" optional (i.e., {w > kc [kh]}) improved recognition accuracy from 33% to 38%. The inclusion of optional aspiration in the "k" of "block," however, damaged recognition rate, lowering accuracy from 38% to 34%. This is clearly due to the fact that stops at the ends of utterances are typically released.

Using the basic grammar that included only the four directions (33% accuracy), changes were made by replacing the default recognizer (i.e., Adult_8kHz_0, an 8kHz sample adult model) with other recognizers provided in the toolkit. As Table 1 illustrates, huge differences in recognition rate were exhibited, depending on the recognizer being used. Recognizers that utilized a 16kHz sampling rate exhibited the best recognition accuracy, with the adult version leading the way with a 97% recognition rate.

Although the accuracy was reasonably high, it must be remembered that the grammar was still very simple. The recognizer at this point only has to choose among four possibilities (i.e., "go north one block," "go east one block," "go south one block," and "go west one block"). Further complicating the grammar by adding the flexibility of numbers one through five lowers the accuracy to 86% but greatly increases the grammar's complexity.

The choice of file formats for the input .wav file also affects recognition. Converting the file type from the default LINEAR format to NIST, for example, raises the overall recognition rate from 86% to 92%.

Adding a flexible syntax rule to the grammar greatly increased the complexity of the grammar yet only marginally affected recognition accuracy. The flexible syntax rule allows the user to speak in either of two correct syntaxes (i.e., "go south one block" and "go one block south"). The inclusion of this syntactic rule only weakened the recognition rate to 91%.

Figure 4. Final grammar that distinguishes eighty possible directions

```
set G1 {
        {"instruction"
        "$dir = south | east | west | north;
        $dist = one | two | three | four | five;
        $displace = go | walk;
        $x = [*sil%% | *any%%] $displace
[*sil%%] $dist [*sil%%] $block [*sil%%] $dir [*sil%%
| *any%%];
        $y = [*sil%% | *any%%] $displace
[*sil%%] $dir [*sil%%] $dist [*sil%%] $block [*sil%%
| *any%%];
        $path = ($x | $y);"}
}
```

After the optional aspirated "k" was added to the pronunciation model for "walk," we attained a final accuracy rate of 93% for a reasonably complex grammar. Following all of these alterations, the grammar (see Figure 4) can distinguish among the four directions, five possible numbers of blocks, two ways of traveling (walk and go), and two syntactic methods (described above). The recognizer is now able to use this grammar to determine whether the utterance is one of the few correct directions among the eighty different possibilities.

Possible Areas of Future Research

Although 93% accuracy is a vast improvement, an error rate of 7% is very serious in language pedagogy. Having correct responses rejected 7% of the time may be a serious detriment to learning and provide much discouragement (Mostow and Aist 1999, 415–16). In order to further increase the recognition rate of this activity, we make the following suggestions.

First, it was found that many errors were caused by a poor recording environment. Problematic .wav files were often saturated with loud background noise or even truncated so that they didn't contain the entire utterance. A better recording environment or better recording equipment (or both) would probably increase accuracy. Second, the recognition rate can be enhanced further by finding the optimal settings of various internal parameters, such as penalties and rejection medians. Third, developing a new recognizer specifically created for the "directions" activity would likely reduce the error rate immensely. To do so, the neural network would need to be retrained on a large corpus of applicable utterances. Furthermore, since the activity is geared toward learners of English, developing a recognizer specifically trained on foreign-accented English would also prove beneficial.

Conclusion

To summarize, we have successfully improved the recognition rate of our "directions" activity to an accuracy level of 93%. We have found that by altering the task design, simplifying the grammar, changing the speech recognizer, converting the format of the .wav files, and altering the word pronunciations, recognition rate may be increased substantially.

Notes

1. See the website for the Center for Spoken Language Understanding at the Oregon Graduate Institute (www.cslu.ogi.edu/toolkit).

References

Egan, Kathleen B. 1999. Speaking: A critical skill and a challenge. *CALICO Journal* 16(3): 277–93.

Ellis, Rod. 1994. *The study of second language acquisition.* Oxford: Oxford University Press.

Gass, Susan M., and Larry Selinker. 1994. *Second language acquisition: An introductory course.* Hillsdale, NJ: Erlbaum.

Harless, William G., Marcia A. Zier, and Robert C. Duncan. 1999. Virtual dialogues with native speakers: The evaluation of an interactive multimedia method. *CALICO Journal* 16(3): 313–37.

Ladefoged, Peter. 1993. *A course in phonetics,* 3rd ed. Fort Worth: Harcourt Brace Jovanovich College Publishers.

Lightbown, Patsy M., and Nina Spada. 1999. *How languages are learned.* Oxford: Oxford University Press.

Mostow, Jack, and Gregory Aist. 1999. Giving help and praise in a reading tutor with imperfect listening—Because automated speech recognition means never being able to say you're certain. *CALICO Journal* 16(3): 407–24.