



Software That Listens: It's Not a Question of Whether, It's a Question of How

Krystyna A. Wachowicz

University of Arizona, Sierra Vista

Brian Scott

Universal Interface, Denton, TX

ABSTRACT

Rapid advances in speech recognition technology have opened up new possibilities in computer-assisted language learning (CALL). From our perspectives as language teacher and applications developer (respectively), we review three levels of speech-interactive learning activities in selected commercial products: activities for vocabulary development, conversational practice, and pronunciation. Our review suggests that the effectiveness of speech-interactive CALL is determined less by the capabilities of the speech recognizer than by (a) the design of the language learning activity and feedback, and (b) the inclusion of repair strategies to safeguard against recognizer error.

KEYWORDS

Speech Recognition, Pronunciation, Feedback, Communicative Approach, Games, Dialogue, Vocabulary, Commercial Software

INTRODUCTION: GOALS AND PERSPECTIVE

The concept of applying Automatic Speech Recognition (ASR) to the teaching of languages is not new. As early as 1982, Professor Harry Wohlert of Oklahoma State University used ASR in his pioneering "German-by-Satellite" program (Wohlert, 1984, 1991). Between satellite lectures, his students practiced their German on shared Apple II Pluses equipped with ShadowVet recognition systems running courseware that he had written using the "Voice-Based Learning Systems" authoring language. ASR technology at that time was limited to recognizing small sets of single words.

© 1999 *CALICO Journal*



Nevertheless, Wohlerl reported that the more reticent students, who found it difficult to practice speaking in public, improved in speaking German. The ASR system never made fun of their faltering attempts to articulate. Wohlerl believed that the benefits derived from time spent speaking to the computer outweighed the problems of occasional erroneous feedback. At that time, over 15 years ago, few language educators shared this opinion.

Rapid advances in ASR technology have led to a new generation of recognizers, capable of processing connected speech as opposed to merely isolated words, and have opened up new possibilities in language learning courseware. During the past few years, several computer-assisted language learning (CALL) systems using ASR have appeared on the commercial market. In this article we describe selected products and comment on strengths and weaknesses from our perspectives as language teacher and applications developer (respectively). We then draw a set of desiderata for speech-interactive CALL based on current theories of language learning, illustrating with reference to speech-interactive courseware developed not commercially but as part of research programs in language learning and instruction. We also look at the climate of acceptance of this technology among today's language educators. Based on our own experience designing and using ASR-based courseware in a variety of settings, we assume that the question is no longer whether automatic speech recognition should be used in language learning applications, but rather how it should be used. We therefore consider the barriers to implementing ASR programs in classrooms and language labs and how these might be overcome.

A SYNOPSIS OF SPEECH-INTERACTIVE CALL

Effective Learning Activities: It is Not the Speech Recognizer, but How It Is Used

To take advantage of ASR capabilities, developers of CALL software have applied speech recognition to a range of language learning activities. It seems reasonable that the kinds of activities selected for CALL be determined to a great extent by the technical capabilities of available recognizers. However, commercial courseware routinely includes activities that assume errorless recognition, even though state-of-the-art speech recognition is never 100% accurate, regardless of the ASR engine applied. This mismatch turns out to be critical in our evaluation of commercial CALL and key in discriminating between commercial software and software developed in a research environment. In the course of testing a variety of software packages, it became apparent to us that the effectiveness of speech-interactive CALL is determined less by the capabilities of the specific recognizer than by (a) the choice of the language learning activity and feedback, and (b) the inclusion of repair strategies to safeguard against recognizer error.



Speech Recognizers Employed: Some Basic Distinctions

Speech-interactive CALL employs a variety of speech recognizers. Four basic distinctions are discussed below.

Speaker-independent ASR systems are intended to recognize almost anyone speaking English (or the selected language), without special training of the recognizer in the tone or speech pattern of an individual speaker. These are more appropriate in language learning applications, especially for non-native speakers. Typically, these systems have been trained using a range of voice types and regional accents. Consequently, they are supposed to accept highly heterogeneous native, and in many instances, comprehensible non-native input.

Speaker-dependent ASR systems, on the other hand, first need to be trained on the speaker, making them less suitable for language learning purposes, although adequate for the use of voice in commands for navigating a personal computer.

Discrete word ASR systems, which aim at recognition of individual words and whose input must be separated by pauses, obviously have more limited applications in language learning. They can be used for word-based language activities, as in the beginning stages of the language learning process, but, in principle, are not suitable for tasks at sentence level. However, in limited contexts, discrete speech recognizers can be used to process short sentence-length utterances of fixed nature.

Continuous ASR systems are intended to process longer phrases and sentence-length (or longer) utterances. Thus, in principle, they are suitable for a variety of language learning activities beyond word level.

Macro-Skills Addressed and Pedagogical Approaches Followed

While CALL has traditionally been applied to listening, reading, and writing, the unique focus of ASR-based CALL is speaking. Three aspects of speaking have received attention:

- Vocabulary development;
- Conversational skills, from practice of more difficult grammatical constructions such as question formation to realistic communication games, where learners speak with a computer-simulated agent to solve a problem;
- Pronunciation, including minimal pairs practice and pronunciation scoring.

The methodological approaches applied to these aspects of speaking are eclectic in the courseware we reviewed, ranging from traditional strict



memorization techniques, to audiolingual practice, to communicative approaches, as seen in dialogue games and simulations with “intelligent agents.” In addition to speaking, listening skills are fostered in the simulated oral dialogues.

Programs employing continuous recognizers, capable of speech recognition at the sentence level, typically listen to a whole conversational turn (a line of a dialogue), whereas discrete word recognizers listen to individual words. Because ASR systems yield higher accuracy when the perplexity (i.e., the number of expected alternative inputs) is low, activities for all three levels listed above typically limit the learner to a small set of multiple-choice responses, whether lexical items or dialogue responses.

A REVIEW OF SPEECH-INTERACTIVE ACTIVITIES IN CALL

Selected Commercial Products

We will review CALL activities focusing on each of the three aspects of speaking listed above. We draw these activities from a sample of commercially available speech-interactive CALL programs for the personal computer, listed below with brief descriptions.

TriplePlayPlus (now updated to Your Way) by Syracuse Language Systems (SLS) uses Dragon Systems' Dragon Dictate as the recognition engine, adapted by SLS to support language learning activities for French, Spanish, and other Western European languages. Dragon Dictate is a speaker-dependent, isolated-word recognizer that can be applied to short sentence-length utterances, as in TriplePlayPlus and its successor, Your Way.

Learn to Speak English 6.0 Auralang is a program for learning English developed by the pioneering French company Auralog. Auralog originally utilized technology from Scott Instruments, which required that actual speech samples from hundreds of individuals be collected for every word and phrase used in the lessons. The major problem with the Scott technology (Scott, 1995) was that it was developed for telephony use and was, therefore, limited in bandwidth, making certain phonetic contrasts impossible to discriminate. The program now uses the Aria Listener.

Vocabulary Builder by Hyperglot, The Learning Company, uses ASR from Lernout & Hauspie, a continuous recognition technology, to teach German (to English speakers) and English (to German speakers).

Dynamic English International by DynEd uses Lernout & Hauspie continuous ASR technology to teach English.

TraciTalk and See it, Hear it, Say it! by Courseware Publishing International (CPI) are two English-teaching programs, developed by CPI in collaboration with the Department of Linguistics at Stanford University



(Hubbard, cited in Scott, 1995). They use IBM's VoiceType Application Factory, an ASR application developer's toolkit that is commercially available and draws on a version of the speaker-independent, continuous recognition engine Sphinx, developed at Carnegie Mellon University. CPI applies this tool to both pronunciation practice and conversational practice, including a set of oral commands to the computer through "Intelligent Agent Traci," whose image appears on the screen. The system runs on SoundBlaster or a compatible sound card. The CPI products represent a bridge between commercial and research-based courseware development.

We will refer to these products as we review ASR-based activities addressing vocabulary development, conversational skills, and pronunciation. Occasionally, we will draw comparisons between these products and the research-based systems reported on in this volume, such as the Voice Interactive Language Training System (VILTS) developed by Rypa and Price at SRI, International, which uses speaker-independent, continuous speech recognition. Created in an academic environment, these systems have paid more attention than commercial products typically do to pedagogical design and to training and tailoring ASR technology to meet pedagogical needs.

Vocabulary Development in Selected Products for Speech-Interactive CALL

Vocabulary development is an area traditionally addressed by CALL and was the first area to which speech recognition was applied (Wohlert, 1984). Vocabulary practice allows the use of discrete speech recognizers, which have been available for applications since the early 80's. Moreover, given that games are a well accepted domain of computers, speech-interactive CALL carries on the tradition with vocabulary games like hangman, crossword puzzles, bingo, etc. We will look at two vocabulary games from two different language software products: Syracuse Language Systems' TriplePlayPlus and Hyperglot's Vocabulary Builder.

We tested each program briefly by inputting a series of correct and incorrect utterances to see how the program would respond. Our observations showed that the games work as intended much of the time, providing a satisfying and fun experience in learning to speak a new language. But both products have drawbacks owing to their overly optimistic assumptions about the performance of ASR technology—a pattern common to all the commercial software we looked at.

TRIPLEPLAYPLUS BINGO VOCABULARY GAME

The TriplePlayPlus CD-ROMs for teaching Spanish and French contain several topically arranged activities divided into three proficiency levels.



One activity is the vocabulary game bingo, modeled after the games used frequently in the second language classroom. Our brief observations of this game found that whereas the software often works well, it also falsely accepts incorrect spoken input and falsely rejects correct spoken input a significant portion of the time.

ACCEPTANCE LEVEL: TriplePlayPlus sets up three acceptance levels for pronunciation, reflecting expected increases in learners' level of proficiency. Learners are instructed that their pronunciation must be very careful at the highest level. Arguably, the system should accept a broader range of pronunciations at the first, beginning level, and get progressively more restrictive as the learner moves up. However, such differences were not detectable in our brief tests. We found the system accepting erroneous input—and the same erroneous input—at each of the three levels. Thus, for Spanish *quatro* 'four' the system accepted *quando* 'when.' For Spanish *gracias* 'thank you,' the system accepted the similar sounding English "no grass here." Similarly, it accepted "my niece" instead of *mais* 'corn.' Other reviewers we spoke to, moreover, noted that true native pronunciation of correct choices was sometimes rejected, at all proficiency levels (M. Rypa, personal communication, June 25, 1996).

These observations demonstrate the pitfalls and limitations of any ASR-equipped games, due to the variable accuracy rates inherent in speech recognition. Even with the most sophisticated recognizers (which are not used in commercial products), analysis of acoustic input is subject to the statistical fluctuations of pattern recognition and thus yields variable accuracy.

In a beginning level game, the recognizer should, in our opinion, be relatively flexible with respect to the learner's pronunciation. How much should the learner's input be allowed to vary from what is acceptable? According to the communicative approach, the input must be comprehensible in order to be accepted. According to standards of the Government Interagency Roundtable, beginning learners' pronunciation can often be comprehensible only to native speakers accustomed to dealing with non-native speakers. Even so, it is questionable whether, for example, Spanish speakers would find *quatro* used interchangeably with *quando* comprehensible in a nonnative.

FEEDBACK: Like many commercial language software programs, the feedback in TriplePlayPlus bingo is of the "correct/incorrect" type. This feedback is given non-verbally. A cheerful sound indicates success and a booing sound, failure. The reactions to this kind of feedback were mixed in the students we interviewed. While some younger college students re-



sponded quite enthusiastically to this kind of feedback, older learners we talked to were annoyed, and sometimes insulted, and suggested a more neutral type of feedback.

VOCABULARY BUILDER, A PICTURE-WORD GAME

Hyperglot's Vocabulary Builder can be described as a traditional word list memorization activity. Its glossary contains over 2000 words and numerous accompanying photographs of excellent quality. The program allows the learner to click on an individual word, which makes a frame appear in a photograph, highlighting the word's visual referent (e.g., for a learner of English, the word tree is associated with a picture of a tree). The program also pronounces the word. Learners can choose the speed with which the program pronounces. Finally, the built-in speech recognizer allows the speaker to see the picture and try to recall the word orally, and it listens for both the correct lexical item and its pronunciation.

ACCEPTANCE LEVEL AND PEDAGOGICAL PRINCIPLES: We observed the occasional false positives and false negatives characteristic of ASR. In addition, we found that the pedagogy employed in this game hinders its potential effectiveness. The well-selected, high-quality photographic images seem at odds with the somewhat stilted teaching style. That is, the learner is expected to memorize a fixed list of words with their corresponding pictures. No synonyms are allowed.

A definite asset of the program is that it provides a thematic organization of the vocabulary. For example, a family photo is shown and the user gets to click on a selected image, such as one representing a mother, and hear its equivalent either in German or in English, depending on which track the learner has selected. The ASR system allows the learner to practice his or her pronunciation in the same context. Practicing words in context is not only a key feature of the communicative approach (Nunan, 1991), but also, presenting vocabulary within semantic domains is consistent with research on how human lexical memory is organized and acquired (Miller et al., 1988).

FEEDBACK: Like bingo, Vocabulary Builder employs feedback of the "correct/incorrect" type (although using more culturally appropriate expressions). It typically contained praise for successful performance and encouragement to try again for unacceptable input.



DISCUSSION

ASR VOCABULARY GAMES AND PEDAGOGICAL THEORY

The philosophy behind vocabulary games agrees with current pedagogical thinking, where vocabulary skills have received renewed attention. This attention follows years of neglect of vocabulary skills when the audiolingual method, and its focus on drills of grammatical patterns, was dominant. Superseding the audiolingual method, the communicative approach places an emphasis on vocabulary development because it aims at using language for authentic communication from the very outset of learning. Rich vocabulary is deemed fundamental for communication (Nunan, 1995, 1991). Akin to the communicative approach, the natural approach (Krashen & Terrell, 1989) considers games an essential part of the learning process. Involvement in games results in spontaneous communication.

LIMITATIONS OF RECOGNIZER ACCURACY

A number of vocabulary building activities are feasible in speech-interactive CALL but, unfortunately, not fail-safe. Commercial courseware such as that we have reviewed often assumes fail-safe performance. Such an assumption is unwarranted, and in our view, detrimental to the cause of wider implementation of speech-interactive CALL. Communication mismatches are a necessary part of any software with ASR capability and, in fact, part of any interaction, even between human interlocutors. To increase the effectiveness of ASR-based learning software, provisions for such mismatches must be built in.

Real-life applications provide a model. Successful real-world applications like the ASR-equipped telephone operator systems (e.g., the AT&T automatic operator) have a built-in recovery from communication failure. If the automatic operator misrecognizes you when you say "Information," "Credit," or "Operator," or when you dictate digits, you have another option. You can resort to dial mode, and the system informs you of this option up front. Although automatic operator systems have been carefully designed to be speaker-independent, native and non-native speakers alike sometimes experience problems. These inevitable technological failures are gracefully handled by providing the user with the option of using the telephone keypad or calling a human operator.

Current ASR-based courseware packages do not sufficiently address the issues of accuracy and error margin, although they do give instructions on how to achieve optimal results (including choice of microphone, noise level, etc.). As a step in the right direction, the Web page of TriplePlayPlus informs its users that their ASR system can be "fooled." The system will



accept words and phrases which are not a part of the language learning activity (as it did in our trials of the software). The learner, according to the instructions provided by TriplePlayPlus, should disregard this fact and focus on the advantages of having computer-based speech interaction.

INPUT VERIFICATION PROCEDURES AS A SOLUTION TO THE LIMITED ACCURACY OF ASR

Many popular language games adapted from textbooks and classrooms are effective in CALL if the learner types the input. However, these games appear not so well suited to ASR-based CALL (especially commercial applications, with less sophisticated recognizers and less time for courseware design and ASR training); this is because winning or losing the game becomes more a function of the recognizer's variable performance than of the learner's knowledge of vocabulary. By fooling the recognizer, one can win; by honestly trying, one can lose.

Thus, the process of adapting games and other vocabulary activities requires measures to safeguard against recognizer errors. One such measure is input verification. Feedback can include input verification as an intermediate step, for example, "I heard 'X.' Is that what you said?" Computer spell checkers with dialogue boxes are examples of using verification to safeguard against a false correction. In vocabulary games the verification feedback could be given in spoken form or in various types of response boxes, including cartoon-type bubbles.

Role-playing games that involve some measure of interaction with real or fictional characters often lend themselves naturally to feedback verification and may be more suitable for ASR-based CALL than are word games. Role-playing games allow for input verification in the form of spoken or written feedback. The Intelligent Agent Traci proposed by CPI, discussed below, is an example.

FEEDBACK OPTIONS

Many experts in language teaching (e.g., Robinson, 1991) recommend against feedback of the "correct/incorrect" type seen in the products we reviewed. Positive rather than negative feedback is recommended because it encourages language learners to make more subsequent attempts at communication. Moreover, implicit rather than explicit feedback is often found more effective as long as the correct form is supplied. Implicit feedback reflects what a native speaker might say in the situation. Instead of saying "correct" or "incorrect," the respondent answers the question or follows the direction. In a restaurant context, for example, if one's utterance is incomprehensible or misunderstood, either the waiter asks for clarifica-



tion or brings the wrong food item. If one's utterance is comprehensible, then one's order is filled correctly.

It should be noted that learners brought up in different educational traditions might expect authoritative feedback of the "correct/incorrect" type (Swan & Smith, 1992). However, this is not the case for the courseware reviewed, which assumes U.S. and Western European audiences.

Whereas praise is consistent with best teaching practice (Nunan, 1991), the emphatic praise seen in Vocabulary Builder—and in general, the "right/wrong/try again" type of feedback seen in both programs—sometimes had an unintended humorous effect because of recognizer error. At times, for example, Vocabulary Builder enthusiastically praises the learner for wrong input.

Conversational Practice in Speech-Interactive CALL

We next review commercial products in terms of language learning activities aimed at conversational practice, a cornerstone of the communicative approach to pedagogy. Many commercial programs apply discrete ASR technology to support practice of conversations of fixed nature. In addition, the new generation of continuous recognizers that are phonetically based promise more flexible practice of conversational skills. However, at the present state of ASR-based technology, it is impossible to have conversations with one's computer that fully comply with the communicative approach: authentic, spontaneous, highly personalized dialogues, with a progressively greater emphasis on the culture of the interlocutors. Our review will consider ways that commercial programs attempt to overcome the inherent limitations of ASR technology for conversational practice while trying to preserve some of the basic principles of communication. The programs selected for review are those created by SLS, DynED, and CPI.

TRIPLEPLAYPLUS BUBBLE DIALOGUES

In activities for ASR-based conversational practice, SLS is guided by a communicative approach. As with vocabulary development, the learner is invited to play a game. The organization of the conversational activity material is thematic and convenient. A variation of a comic strip with bubble dialogues of illustrated characters is presented to the learner. The learner's task is to fill in the dialogue bubbles by choosing the correct response from the alternatives provided. If the user makes the correct choice, then the bubble is filled with the correct alternative. Otherwise, the learner receives feedback in the form SLS uses in other activities: a



disapproving booing sound.

In our trials of the courseware, we did get some feeling of real interaction with a person speaking the target language. Variations of the technique of fill-in-the-bubble dialogues are, in our judgment, a step in the right direction. However, we found this activity not as successful as it could be because of factors noted in our review of vocabulary games: the nature of the “right/wrong/try again” type of feedback in combination with errors made by the ASR.

DYNED QUESTION FORMATION WITH INPUT VERIFICATION

The DynEd activity introduced yet another type of ASR-based activity-question formation. It is commonly accepted that the syntax of English questions presents considerable difficulty to non-native speakers, especially in speaking. A jigsaw type activity at the high-beginner level presents the learner with an interrogative sentence with the words in random order. The learner’s task is to speak the words in the right order.

We found this activity fun and, in our judgment, reflecting a sound pedagogy. By drawing aspects of word order to the learner’s attention, it follows the recent pedagogical trend of fostering awareness of language form (Doughty, 1991; Nunan, 1991; Van Lier, 1996). It also corresponds to the communicative approach with its game-like nature. In our brief tests, we again encountered the problem of misrecognition. Presumably, the recognizer is sensitized to typical errors and anticipates them, but our input was misrecognized a significant number of times. However, this program institutes an input verification procedure to help with recognition problems.

This procedure uses verification-by-dictation. The continuous recognizer processes the whole interrogative sentence that the learner is supposed to utter and returns a typed version of what it heard. Then, the system requests confirmation by the learner: “Is ‘X’ what you said?” This gives the learner an opportunity to verify the recognizer’s output. At times, the typed version reflects the learner’s utterance, but, because the recognizer “mishears” a significant number of times, it returns various kinds of ill-formed input that does not correspond with what the learner has said. While this can potentially become annoying, the DynEd program is nevertheless the first in our review of commercial products to address the issue of possible misunderstandings on the part of the recognizer with realistic input verification.



CPI CANNED CONVERSATIONS AND CONVERSATIONS WITH TRACI

The CPI courseware uses a continuous recognizer for two kinds of conversational exchanges. The first kind is canned conversations in which the learner is supposed to choose the correct line of dialogue among several alternatives. There is often no single correct alternative, and if the recognizer finds the line supplied by the learner acceptable, then the next line of dialogue follows and the story branches accordingly. Thus, the feedback for what the learner says is implicit. The learner is allowed several attempts to produce a legitimate conversational alternative. After a few errors, the system supplies a proper line of dialogue. The graphic design of this conversational exchange includes an animated character whose moving lips tell her story.

The second type of conversational exchange is conversations with Traci, an “intelligent agent” whose live video image appears on the screen. This activity resembles the adventure games common in computer entertainment, but with spoken input by a learner of the language. Instead of navigating the program with a mouse or keys, the learner speaks to Traci, who carries out commands to search and find objects. Thus, from the point of view of the communicative approach, the learner is called upon to solve meaningful problems by speaking with Traci. As a task-driven use of language, this game reflects recent approaches to language teaching.

Like other ASR-based programs, this one is a bit hearing-impaired and does not always respond to the navigation commands uttered by the user. But because of an extremely clever interface, recognition errors are less disturbing. Traci appears to be a bit absent-minded; she simply does not always listen. Putting ourselves in the role of learner, we understood this and did not mind repeating the command an additional time. Thus, the personality of the interface matches the capabilities of the recognizer.

DISCUSSION: ASR-BASED CONVERSATIONS AND THE COMMUNICATIVE APPROACH

A review of speech-interactive conversations in CALL products finds that the following conditions apply:

- The learner can use full-sentence utterances.
- What the learner says is limited to a small number of options.
- Most of the time the learner must say exactly the wording given in the multiple choice alternatives.
- A significant proportion of legitimate utterances can be misunderstood.



Some might say that these conditions are reminiscent of interactions with a language teacher trained in the audiolingual method. Under that method, the input was strictly controlled, the errors were considered fatal, and if one's pronunciation or grammar were not perfect, a rejection definitely followed. Although these conditions are clearly not consistent with best classroom practice, they are consistent with the acknowledged limitations of CALL for language production (Holland et al., 1993): CALL's strength is in exercising routine aspects of production skills.

Our review also indicates various strategies that go toward "exploiting strengths and avoiding weaknesses" of current recognizers, in LaRocca's (1994) words. Four strategies have emerged in the CALL we reviewed:

- Input verification (as in DynEd's question formation game);
- Personality of a conversational character consistent with occasional misrecognition (as in Traci, the intelligent agent);
- Authenticity enhancement (as when CPI's canned conversations let the learner advance the story line);
- Task-based language learning (as in Traci's navigation), where the learner must speak to the computer in order to solve a real or fictional problem.

Task-based language learning is what Nunan (1995) calls learning by doing and is an expression of the communicative approach. In Conversations with Traci, the learner is performing an authentic task in an authentic context.

Our review makes it abundantly clear that the design of the interface is of foremost importance in the success of these strategies. We can further illustrate these strategies with speech-interactive CALL systems developed in research environments.

For example, task-based language learning is illustrated in the microworld of MILT, the Military Language Tutor (Holland, Kaplan, & Sabol, this issue; Kaplan & Holland, 1995), a research prototype that requires the learner to use a second language to solve a problem. In MILT's Arabic microworld, the learner must direct a graphically drawn agent with oral commands in Arabic to search for a critical document hidden somewhere in a series of rooms.

Authenticity enhancement is represented by SRI's VILTS-Echoes, a system for learning French (Rypa & Price, this issue). Underlying the system is the Nuance speech recognizer—speaker-independent, continuous ASR technology originally developed by SRI. Echoes makes use of several features to create an atmosphere of authentic conversation: high-quality photographs, live video, very powerful graphics, and listening activities that draw on spontaneous, unscripted conversations. Over 100 native French voices participate in these conversations, collected in France, exposing



the learner to some of the variety of speech they would hear in the target language country. VILTS-Echoes is the only program among the software packages reviewed here that supplies authentic listening input rather than scripts read by actors.

Pronunciation Practice in Speech-Interactive CALL

Pronunciation is a neglected skill in many classrooms, although learners themselves attach great importance to it (Willing, 1988; Lane, 1993) and it draws significant commercial attention (e.g., Meis, 1995). According to speech-interactive courseware designers, computers with ASR capability provide new opportunities to improve a learner's pronunciation. Discrete recognizers are capable of dealing with pronunciation at the word level. Continuous recognizers, which are phonetically-based, can take phrases and sentences as input. Several CALL products take advantage of these capabilities. Pronunciation activities in the products reviewed fall into three categories:

- Minimal pair exercises;
- Pronunciation scoring as part of vocabulary games and conversational practice;
- Word boundary and phrase segmentation practice.

We will look at these categories of pronunciation activities in selected products: Hyperglot's Vocabulary Builder, CPI's conversation games, and Auralog's Auralang. We will also draw comparisons to several of the research-based systems described in this volume.

MINIMAL PAIR EXERCISES

During the 1950s and 1960s, a new approach to teaching pronunciation emerged, minimal pair exercises. In minimal pair practice, students are told to focus on the contrast between two sounds leading to a difference in meaning, for example, "thin" and "sin" in English. This type of exercise is still a principal way of teaching pronunciation (Baker & Goldstein, 1990; Morley, 1992, 1994; Pennington, 1996).

CPI provides a minimal pair-based pronunciation activity that we found challenging, well designed, and convenient. The background graphical interface, a landscape of pastel colors, follows the recommendations of graphics designers for lessening the presumed psychological strain of pronunciation practice. Recognizer yield was relatively high on our trials. However, it is a question, in our opinion, whether even a small margin of



error can frustrate the learner in a potentially stressful exercise such as minimal pairs, especially with feedback of the “right/wrong/try again” variety. Implicit feedback, such as presenting visual referents of the minimal pair alternatives (someone thinking as opposed to sinking), may be more helpful than judgmental feedback given occasional recognizer errors.

Certain of the research-based systems discussed in this volume—such as those of Aist and Mostow, Dalby and Kewley-Port, and LaRocca—demonstrate advantages over commercial systems. They have minimal pair designs drawn from intensive experimentation with design alternatives, careful selection of minimal pairs, and thorough training of recognizers on the likely mispronunciations of closely defined populations, such as speech therapy clients, as opposed to the general audience assumed by commercial products. In LaRocca’s work, moreover, we see the added advantage of presenting a graphical display of relevant speech articulators, as well as a video of a native speaker pronouncing targeted sounds prior to the student’s being asked to pronounce. The full face and body video depicts the contribution of facial and other gestures as well as articulators to making speech. Although the audio-lingual method emphasized aural perception and instructed teachers to hide their lips (Nunan, 1991), more recent evidence is that seeing as well as hearing sounds articulated helps shape pronunciation and perception (Massaro, 1987). Most of the commercial systems did not provide displays of this sort.

PRONUNCIATION SCORING

Pronunciation scoring is a bonus feature of some commercial programs. It may be added to vocabulary practice as well as to conversation and minimal pairs practice. The commercial programs we reviewed provided scales of pronunciation proficiency that range from three to seven levels.

Vocabulary Builder illustrates the incorporation of pronunciation scoring into a vocabulary game. As described earlier, Vocabulary Builder asks learners to recall the target-language name for each of a series of pictures. The program measures the learner’s pronunciation of the word and provides feedback according to a three-level pronunciation scale. The basic level is called “tourist” and is marked in red; the second, intermediate level is marked yellow; and the third, highest level, is called “native speaker” and marked green.

Because the program expects a particular word and only that word in response to each question, pronunciation can be scored with respect to the anticipated response. We did not specifically test how well the output scores actually discriminate levels of pronunciation proficiency; but the inherent errors in word recognition noted earlier in this program lessen



our confidence in the pronunciation scoring. Moreover, the basis of the categorization at the three levels appears arbitrary. By comparison, the experimental VILTS-Echoes system (Rypa & Price, this issue) uses a five-level system of pronunciation ratings that follows the American Council on the Teaching of Foreign Languages (ACTFL) and the Interagency Language Roundtable (ILR) specifications, and the developers validated the scoring methodically against expert teacher judgments.

Based on the performance of ASR in commercial CALL products, it is our opinion that it is too early for recognizers to grade pronunciation as a background process secondary to word-learning games and other response-selection activities in CALL. A more desirable option may be to give an overall pronunciation score for an entire conversational exchange rather than scoring individual items. It is this procedure that VILTS-Echoes adopts. However, while sophisticated statistical approaches might produce desired results, many experts in the field emphasize the complexities of scoring pronunciation. Among these complexities is the unpredictability of pauses, typical for non-native speech and well described by language teachers. Rypa and Price (this issue) discuss some of these complexities for speech interactive language learning courseware.

WORD BOUNDARIES

One of the greatest difficulties in listening comprehension stems from the learners' inability to recognize where one word ends and another one begins in connected speech—that is, defining word boundaries. Pauses often do not coincide with the lexical items as given in dictionaries. Auralang offers activities focusing on word boundaries and pauses in a sentence (Auralog, 1995). The activities are of a “listen and imitate” type, in the style of the audiolingual language laboratory. As opposed to the language laboratory, however, speech-interactive courseware allows the learner to focus precisely on a desired portion of a sentence. The system also allows the learner to “see” his or her voice in the form of a sound wave and compare it with the wave form of a model native speaker.

We found this kind of activity interesting and potentially instructive. It is not subject to the errors made by ASR when attempting to categorize utterances as this or that word or phrase. However, we did not test the Auralang activity with students, nor did we find teachers familiar with the software who could give an impressionistic assessment. Many educators have questioned whether second language learners can benefit from seeing acoustic features of their utterances compared with acoustic features of a native utterance. They doubt whether learners can be educated in the intricacies of the wave form. In the late nineties, the wave form has been rediscovered, and there is a new audience of technologically sophisticated



learners. Researchers are beginning to see benefits in carefully designed, experimental systems focused on particular spectral features, such as described by Eskenazi (this issue). Eskenazi provides a critique of CALL products that incorporate acoustic wave visualization and offers desiderata for effective use of this method in CALL.

DISCUSSION

We are tempted to say that the approach seen in many pronunciation activities—“Repeat, imitate, get corrected by your teacher”—reflects dated pedagogical practices. At times, the computer is visualized as a high-quality tape recorder. One clear advantage computers can offer is to show the multimodal aspect of pronunciation. In most of the reviewed CALL packages, however, neither video nor photographs of native speakers articulating target sounds accompany ASR-based pronunciation activities. In the pronunciation scoring products, moreover, in which scoring is presented secondary to a vocabulary game or dialogue exercise, we judge that the technology is not up to the rather superficial integration of ASR that appears typical of commercial products. Rather, to exploit ASR effectively for shaping pronunciation, CALL needs the kind of careful internal development and tuning of ASR, together with structuring of exercises, that goes into the experimental systems. Among commercial offerings, minimal pair exercises and acoustic wave form comparisons appear the most promising kinds of activities to help pronunciation.

BARRIERS TO IMPLEMENTATION OF SPEECH-INTERACTIVE CALL

Our review of commercial products has focused on critiquing their use of technology and on recommending improvements. Even so, we have found that most of these products foster a motivation and an opportunity to practice speaking not found in traditional CALL. Many of these products are likely to benefit language learning, we feel, if deployed carefully by language teachers. We have also seen continual improvements across successive versions of commercial products, so that what is available at the time of this article's publication may have advanced considerably beyond what is reviewed here. Moreover, some university-developed systems are ready or nearly ready for distribution to language learning labs.

Yet, it is the home market that provides the largest audience for speech-interactive CALL. Despite progress in the technology, ASR for a variety of reasons is still not widely used in language labs in schools, universities, and government. Our experience in these settings finds that many teach-



ers who already use computer-aided instruction are nevertheless reluctant to employ ASR-based technology. They describe the following barriers:

- Many teachers are overwhelmed by the technical detail of many available reports on the subject.
- They are skeptical of the enthusiasm of commercial product advertisements.
- Although researchers are reporting accuracy exceeding 95% in some simple question-answer tasks, teachers' own casual reviews of available speech interactive courseware seem to point to lower accuracy rates. They fear that a less-than-perfect system may frustrate or mislead the learner.
- Teachers see many of the commercial packages as at odds with the generally accepted communicative approach to language teaching.

How can such barriers be addressed by speech-interactive CALL developers? First, developers can foster a realistic model of today's learners. These learners tend to be technologically sophisticated, and most are already familiar with ASR capability from computer games, automotive applications, the telephone, and voice-controlled computer navigation. They are aware of the limitations of ASR and have realistic expectations.

Second, developers can promote a truthful analysis of the drawbacks and advantages of speech-interactive CALL. Lacking formal assessments of ASR for language instruction, this analysis must be largely intuitive. An obvious advantage is that many non-native speakers find talking to a computer less intimidating than talking to a person and are thus more likely to spend time in speaking practice. This may be especially true in cultures where losing one's face in public is stigmatized. Moreover, as opposed to practicing with tapes in a language lab, CALL is more interactive and can better simulate a conversational situation. To put recognizer errors in perspective, we would note that even human teachers cannot guarantee 100% accurate recognition of students' utterances—particularly in noisy, crowded classrooms, as anyone with language teaching experience knows. This is especially true if the teacher is a non-native speaker, as is often the case in countries where the target language is not the national language.

Third, developers can spotlight the best of speech-interactive CALL designs, those that present authentic communicative contexts by exploiting the computer's multimodal features and branching capabilities, as well as those that use verification procedures to safeguard against ASR errors.

Fourth, it can be emphasized that CALL is well suited to addressing several aspects of form, consonant with a renewed interest in language form in pedagogical theory, as expressed by a number of recent trends in language teaching: the language awareness approach, grammatical con-



sciousness raising, and the organic approach (e.g., Doughty, 1991; Lightbown & Spada, 1990; Pennington, 1996). These trends extend the communicative approach, viewing communicative practice as a series of interrelated tasks, several of which focus on formal aspects of language. For example, Pennington suggests as many as eight related activities for teaching the voiced/voiceless stop contrast that is a common difficulty for students whose first language is Arabic or Spanish. The first activity is presentation (voiceless: p, t, k, versus voiced b, d, g). Other activities cover minimal pair practice, repetition, and discrimination. The activity sequence ends with communicative pair practice. The activities for the p/f contrast (e.g., with Korean students) contain a language awareness part consisting of a sentence: "Is it true that Koreans prefer eating pork with a fork?" Thus, the computer can selectively address some form-focused activities that enable authentic communication.

DESIRABLE FEATURES IN SPEECH-INTERACTIVE CALL

A review of commercial products in comparison with experimental prototypes allows us to begin a checklist of desirable features in speech-interactive CALL. Some of these features are common to all ASR-equipped software, some are common to all multimedia-based computer assisted instruction, and some are specific to multimedia-based CALL. The best products from our review have these characteristics:

- Task-based instruction with an emphasis on communicative authenticity;
- Implicit as opposed to corrective feedback in those tasks;
- Multimodal formats (video, drawings, photos) to enhance authenticity;
- Focus on schematized, relatively predictable conversations;
- Verification procedures and repair strategies to counter speech recognizer errors.

In these features, the role of the computer in the interaction is not the traditional role of tutor but rather of partner in the conversation or agent in the game. Other desirable features include

- Giving learners a chance to correct their own errors;
- Providing visual support for pronunciation activities (e.g., native speakers articulating target phonemic distinctions).

Desirable features that we have not seen in commercial products but have seen in some experimental systems include:



- Adaptive sequencing of items in an exercise to accommodate individual learners' performance;
- Use of authentic texts (e.g., The conversational exchanges should involve oral, non-scripted texts rather than texts read by actors);
- Focus on listening activities as a complement to speaking;
- Attention to the sociolinguistic variability of speech.

Some of these features bear expansion. First, adaptive sequencing has been discussed at length in the literature on computer-assisted instruction (Park & Tennyson, 1983) and intelligent tutoring systems (Anderson et al., 1990). It requires tracking the individual learner's performance throughout a lesson and tailoring the lesson to the individual's needs. It can result, for example, in dropping items the learner gives evidence of having mastered and presenting additional practice for items on which the learner makes errors. This process results in more efficient learning and is demonstrated in this volume by the prototype systems of Dalby and Kewley-Port and Holland, Kaplan, and Sabol (see also Kaplan & Holland, 1995).

Second, verification procedures and repair strategies have been a recurrent theme of this review. Input verification is a key repair strategy for potential misunderstandings between humans or between human and machine. It can take the form of negotiation of meaning by conversational exchange: "Did you say 'X' or 'Y'?" It can also take the form of written feedback or pictorial feedback. Planning for misunderstandings is a critical dimension to take into account when developing or evaluating speech interactive language learning courseware.

The importance of listening as a complement to speaking is evident to many language teachers. Measuring second language performance in speaking is invariably tied to listening challenges and the learners' ability to interact in a given situation (McNamara, 1996). The statement "I can speak, but I don't understand" is not uncommon among language learners, and difficulties with listening comprehension among non-natives are a well documented source of problems in the workplace (Wachowicz, 1996). These difficulties often lie not in ignorance of vocabulary or grammar but in lack of familiarity with the diversity of speech registers and voices representing individual, geographical, and sociological variation within a culture. Yet, although multimedia courseware for listening is commonplace, ASR-based CALL that includes focused listening activities is conspicuously lacking (Wachowicz, 1996). It is well represented, however, in the rich listening exercises that are included in VILTS-Echoes (Rypa & Price, this issue).

To portray the sociolinguistic variability of speech, systems can present a range of voices. For example, we infer that role playing with the possibility of choosing a voice for one's game partner is more effective than



being presented with a fixed voice. Research on second language acquisition and bilingualism points to the importance of finding one's own voice in the target on voice and memory (Sheffert & Fowler, 1995) provides new evidence that voice type affects the memory process.

CONCLUSION

Not surprisingly, a review of speech-interactive systems provides a new perspective on the whole endeavor of creating CALL. We judge that courseware that listens can perform best when it is part of communicative, culturally sensitive activities, fashioned after some real-life, uncomplicated, interactive task. Courseware that listens can also address certain form-focused activities, such as minimal pair practice, if the ASR is carefully tuned, as in the experimental programs we mentioned. The current trends in CALL advocate a learner-centered approach (Brown, 1994; Oller, 1996). Learning is viewed as a collaborative venture in which learners construct knowledge from materials provided by CALL. In this view, computers do not serve as authoritative teachers but provide a context in which learners can use the target language.

Progress in computer-based language learning and speech recognition is rapid and inevitable. Now that computers are an everyday tool for many professions, they can and will be used in language learning. Their role will continue to be redefined, and they may help to elucidate the communication process itself. The ever-changing technological scene will continue to have an impact on how we view learning, and these views in turn will shape the instructional interface.

ACKNOWLEDGEMENTS

We have benefited greatly from discussions with Dr. Jared Bernstein, Professor Stanley Peters, Professor Leo Van Lier, Dr. Rick Jackson, Alexandra Marinov, Renee Robinson, and Dr. Melissa Holland. None of these persons necessarily agrees with all the conclusions reached in this paper.

REFERENCES

- Anderson, J. R., Boyle, C. F., Corbett, A. A., & Lewis, M. W. (1990). Cognitive modeling and intelligent tutoring. *Artificial Intelligence*, 42, 7-49.
- Auralog (1995). *Talk to me: User manual*. Voisins le Bretonneux, France: Author.
- Baker, A., & Goldstein, S. (1990). *Pronunciation pairs: An introductory course for the students of English*. Melbourne: Cambridge University Press.



- Bialystok, E., & Hakuta, K. (1994). *In other words: The science and psychology of second language acquisition*. New York: Basic Books.
- Brown, D. (1994). *Teaching by principles*. Englewood Cliffs, NJ: Prentice Hall.
- Doughty, C. (1991). Second language instruction does make a difference: Evidence from an empirical study of SL relativization. *Studies in Second Language Acquisition*, 13, 431-469.
- Holland, V. M., Maisano, R., Alderks, C., & Martin, J. (1993). Parsers in tutors: What are they good for? *CALICO Journal*, 11 (1), 28-46.
- Kaplan, J., & Holland, V. M. (1995). The application of learning theory to the design of a second language tutor. In V. M. Holland, J. Kaplan, & M. Sams (Eds.), *Intelligent language tutors: Theory shaping technology*. Mahwah, NJ: Lawrence Erlbaum.
- Krashen, S. D., & Terrell, T. (1983). *The natural approach*. Oxford: Pergamon.
- Lane, L. (1993). *Focus on pronunciation: Principles and practice for effective communication*. White Plains, NY: Longman.
- LaRocca, S. (1994). Exploiting strengths and avoiding weaknesses in the use of speech recognition for language learning. *CALICO Journal*, 12 (1), 102-105.
- Lightbown, P., & Spada, M. (1990). Focus on form and corrective feedback in communicative language teaching: Effects on second language learning. *Studies in Second Language Acquisition*, 12 (4), 167-191.
- Massaro, D. W. (1987). *Speech perception by ear and eye: A paradigm for psychological inquiry*. Hillsdale, NJ: Lawrence Erlbaum.
- McNamara, T.F. (1996). *Measuring second language performance*. London: Longman.
- Meis, M. L. (1995). *Kiss your accent good-bye*. Colombia: Cosmopolitan Business.
- Miller, G. A., Fellbaum, C., Kegl, J., & Miller, J. (1988). WordNet: An electronic lexical reference system based on theories of lexical memory. *Revue Québécoise de Linguistique*, 17, 181-213.
- Morley, J. (1992). *Rapid review of vowel & prosodic context*. Ann Arbor, MI: The University of Michigan Press.
- Morley, J. (Ed.). (1994). *Pronunciation pedagogy and theory: New views, new directions*. Alexandria, VA: TESOL.
- Mostow, J., et al. (1994, August). A prototype reading coach that listens. In *Proceedings of the twelfth national conference on artificial intelligence (AAAI)*. Seattle, WA.
- Nunan, D. (1991). *Language teaching methodology*. New York: Prentice Hall.
- Nunan, D. (1995). *Atlas: learning-centered communication*. Boston: Heinle & Heinle.
- Oller, J. W. (1996). Toward a theory of technologically assisted language learning/instruction. *CALICO Journal*, 13 (4), 19-43.
- Park, O. K., & Tennyson, R. D. (1983). Computer-based instructional systems for adaptive education. *Contemporary Education Review*, 2, 121-134.



- Pennington, M. C. (1991). Computer assisted analysis of English dialect and interlanguage: Prosodics. Applications to research and training. In P. Dunkel (Ed.), *Computer-assisted language learning and testing: Research issues and practice* (pp. 133-154). New York: Newbury House.
- Pennington, M. C. (1996). *Phonology in English language teaching*. London: Longman.
- Richards, J., & Rodgers, T. (1986). *Approaches and methods in language teaching*. Cambridge: Cambridge University Press.
- Scott, B. (1995). *To humanize speech technology*. (Technical report). Stanford, CA: Applied Speech Technology Laboratory, Center for the Study of Language and Information, Stanford University.
- Sheffert, S. M., & Fowler, C. A. (1995). The effects of voice and visible speaker change on memory for spoken words. *Journal of Memory and Language*, *34*, 665-685.
- Siegal, M. (1996). The role of learner subjectivity in second language sociolinguistic competency: Western women learning Japanese. *Applied Linguistics*, *17*, 356-381.
- Swan, M., & Smith, B. (Eds.). (1992). *Learner English: A teacher's guide to interference and other problems*. Cambridge: Cambridge University Press.
- Wachowicz, K. A. (1996). Surfing the sound wave. In F. L. Borchardt & E. Johnson (Eds.), *Proceedings of the 1996 CALICO annual symposium: Distance learning* (pp. 262-264). Durham, NC: CALICO.
- Willing, K. (1988). *Learning strategies in adult migrant population*. Adelaide, Australia: NCRC.
- Wohlert, H. (1991, March). German by satellite. *Annals of the American Academy of Political and Social Sciences*.
- Wohlert, H. (1984). Voice Input/Output: Speech technologies for German language learning. *Unterrichtspraxis*, *17* (1), 76-84.

AUTHORS' BIODATA

Krystyna Wachowicz is Program Developer at the University of Arizona, Sierra Vista campus, where she develops instructional models for distance learning and is evaluating the effectiveness of a web-based sustanment course in Russian language and culture. She holds a Ph.D. in linguistics from the University of Texas and has taught graduate level courses in linguistics, second language acquisition, and CALL. She has consulted with the Defense Language Institute Foreign Language Center on teaching methods and CALL.



Brian L. Scott has worked in the speech industry for twenty-six years and holds a Ph.D. in perceptual psychology from the University of Waterloo. He founded Scott Instruments in 1978, a pioneering company in speech recognition technology, and in 1994 established the Applied Speech Technology Laboratory at Stanford University. Widely published in speech technology, signal processing, and human speech interfaces, he holds eight patents—five in speech recognition methods and three in devices for the hearing-impaired.

AUTHORS' ADDRESSES

Krystyna Wachowicz
Office of Language Programs and Distance Learning Initiatives
University of Arizona, Sierra Vista Campus
1140 North Colombo
Sierra Vista, AZ 85635
Phone: 520/626-2422, ext. 154
E-Mail: wach@uasv.arizona.edu

Brian Scott
Universal Interface
1500 Sandy Creek
Denton, TX 76205
E-Mail: blscott@gte.net