VILTS: A Tale of Two Technologies

Marikka Elizabeth Rypa
Patti Price
SRI International

ABSTRACT. The Voice Interactive Training System (VILTS) is a language-training prototype developed to help improve comprehension and speaking skills. The system incorporates two related technologies: speech recognition and pronunciation scoring. Speech recognition allows students to navigate through units by using oral communication skills. Pronunciation scoring, validated through correlation with expert raters, provides assessment of speaking skills. We discuss the motivation for the program, the interdisciplinary efforts involved, and the resulting system architecture. We also describe challenges and trade-offs in designing activities using unscripted material and in integrating new speech technology. Finally, we discuss system evaluation and opportunities for future directions.

1. INTRODUCTION

ECHOS, the French version of the VILTS, was developed to help improve listening and speaking skills by using state-of-the-art speech recognition technology. Designed to support language learning and maintenance of skill at beginning, intermediate, and advanced conversation levels, the VILTS lesson
architecture stresses learner-centered navigation through listening and speaking activities. Two related technologies underlie the system design: speech recognition and pronunciation evaluation. In the following sections we motivate the program, describe the system architecture, explain the use of speech technology in the language pedagogy, and discuss future directions.

2. MOTIVATION AND GOALS

Active speaking skills are central to the needs of most language learners, and people learn best through contextual, culturally valid, interactive listening and speaking (e.g., Lee & VanPatten, 1995). In the last few years, great strides have been made in multimedia educational tools; features such as sophisticated animation, graphics, and audio input and output capabilities are increasingly used to develop engaging interfaces. Until recently, however, user interaction has been limited by the lack of robust speech recognition technology, and oral user input was largely confined to recording and playback. Recent advances in speech technology have yielded high-performance, continuous-speech, speaker-independent recognition. This technology supports the development of more sophisticated interactive language learning software and enables users to navigate using spoken utterances in active exchanges with the system. We explore several uses of *speech recognition* technology in interactive language learning. We also consider another speech technology, *pronunciation scoring*: the evaluation of nonnative speech as compared to native pronunciation calibrated through correlation with expert human raters. These two technologies, speech recognition and pronunciation scoring, place different requirements on the lesson interface; the motivation for the VILTS program was to bring together advances in each technology to support language education in a way that was most technically feasible and pedagogically valid.

We have used technology to present a broad range of natural, authentic speech from a diverse set of talkers in an immersive, learner-centered pedagogical experience. As core material for lessons, we created a structured corpus balanced for age, sex, topic of discussion, and complexity of language use. These materials were incorporated into an engaging, interactive, flexible architecture to support learner-centered navigation through various levels, topics, and activity types in an environment that could foster individual learning styles. Our interdisciplinary collaboration linking speech technology with language pedagogy has resulted in a prototype system with interesting new possibilities for language learning. Our work has focused on French, as seen in ECHOS, although we have also explored the teaching of Spanish and English, and all our algorithms are portable to other languages. Inherent in an interdisciplinary effort are both benefits and challenges. The advantages include a broadening of ideas and input to the lesson activities, which contributes greatly to the richness and range of learning interactions developed for the project. The challenges include the incorporation of diverse viewpoints in a cohesive, engaging, pedagogically valid system that uses the technology appropriately.

3. THE VILTS ARCHITECTURE

The modular VILTS architecture was designed to be extensible to other languages and to a wide range of activity types. Dialogues or other types of exchanges can easily be incorporated as core materials, and activity types are developed for existing and new languages. The VILTS can complement classroom instruction; it does not include a full language learning program, but it enables this extension. VILTS uses the speaker-independent, continuous speech recognizer from Nuance Communications, based on SRI's speech recognition technology. The acoustic models used are custom-developed by SRI for language education: the models are wide-band (as opposed to telephone bandwidth) and are based on a homogeneous pool of French speakers (Parisians) as well as English learners of French at various levels of experience. SRI developed the pronunciation scoring described in this paper.

3.1 A Conversational Core

The ECHOS version of the VILTS uses spontaneous, unscripted French conversations on various topics, supplemented by read excerpts from the French newspaper *Le Monde.* Conversations were collected on ten common topics, including travel, health, education, environment, and politics. Linguistic complexity of the conversations was controlled by the interviewer's questions. "Beginning" level conversations contain relatively simple vocabulary and constructions and were elicited by simple (usually "yes/no") questions. "Intermediate" and "advanced" conversations contain progressively longer speech segments and more complex and idiomatic expressions. These levels were elicited, respectively, through questions that are more complex and questions aimed at eliciting a monologue from the conversant (who claimed to be an expert in the topic of the interview). A pool of 100 native speakers in Paris recorded the conversations,
representing a variety of speaker characteristics and speaking styles.

3.2 Instructional Flexibility and Learning Styles

The VILTS lesson architecture is shown in Figure 1. The content is conversational, thematically-based activities based on the topics of the French interviews we collected. Although the architecture suggests
order implicitly in the layout of the interface, users can also navigate freely though levels, topics, and activity types. The left-to-right screen layout of activities, beginning with listening comprehension activities, encourages the student to proceed from listening to speaking, that is, from passive to more active skills. The suggested path proceeds from comprehension practice through speaking to reading longer segments aloud. Students can, however, choose how to access the materials and activities based on individual interests and needs. The flexibility of the architecture accommodates different learner styles, from more structured, incremental learning to free exploration. For example, after system logon, the student can move directly to a lesson unit, or the student can review units completed. This flexibility was motivated by research in second language learning that shows the primacy of individual styles of learning (e.g., Oxford, 1995) and suggests that no single sequence of instruction or modalities is optimal for all learners.

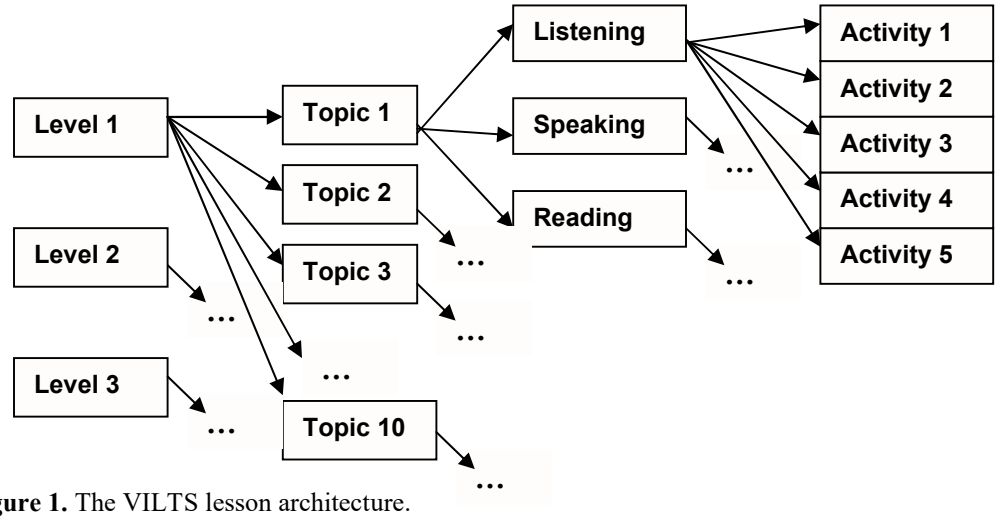

Figure 1. The VILTS lesson architecture.

The student may choose one of three levels (beginning, intermediate, or advanced) and then choose a topic that interests them from those available at that level (e.g., travel). Users can remain at one level longer to become more facile and confident with the material, or they can forge ahead in a more exploratory style. Although the system structure is designed to foster initial practice of listening comprehension skills with an emphasis on speech and minimum exposure to text, users can choose activities that offer slower versions of the spoken samples as well as text support (such as transcriptions). This approach accommodates both oral- and text-based learning styles. The structure of the lesson activities is illustrated in Figure 1. Each of the five activity types increases in complexity from top to bottom and consists of several instances with different content based on the conversations or the related newspaper text. The architecture also allows the student to engage in pronunciation exercises based on areas of diagnosed weakness.

Because our audience is adult learners who are assumed to be capable of self-monitoring, we have provided a learning environment in which the user makes decisions and controls the sequence of instruction (with reference, of course, to feedback from the system). This architecture contrasts with an intelligent tutoring system (ITS) approach, which seeks to adapt content and sequence automatically to the student. The ITS vision is cited in LaRocca, Morgan, and Bellinger (this issue), and a related adaptive sequencing argument is put forth in Holland, Kaplan, and Sabol (this issue). These approaches incur heavy computational overhead, whereas we sought to focus investment in speech recognition and interface development.

3.3 Student Resources: Browsers and Talking Notebooks

Browsing capabilities supplement the choices offered by the VILTS. Keywords with translations and pronunciation of individual words by a native and in conversational context are available in all activities. In addition, as the last lesson activity (in the left-to-right sequence), a "talking notebook" is available. In the talking notebook, the student can review a list of lesson vocabulary items (words or phrases), see translations, hear the items spoken in isolation by a native speaker, and hear the items spoken in the various utterance

contexts in which they appear in the lessons. Users can also add words with translations to this list, and the audio features automatically apply so that the word can be heard in isolation or in context.

3.4 A Communicative Approach: Listening Exercises and Dialogue Interactions

Pedagogical research in the VILTS project focused on how best to map the two speech technologies--speech recognition and automatic pronunciation scoring--into useful and valid lesson activities. The lesson interactions involving speech recognition were informed by the communicative approach to language teaching, under which cluster most of the popular methods that succeeded the audiolingual method. This approach is advocated in most government language teaching institutions as well as in public schools and universities. The communicative approach stresses the interactive use of a second language in a meaningful context with a high degree of comprehensible input, for example, exposure to the target language at a slightly higher level than the level at which a student is completely comfortable (Krashen, 1981). To carry out this approach, our design had to emphasize the meaningful use of language for communication.

Following communicative considerations, the VILTS lesson architecture was designed to take advantage of the rich repository of authentic conversations we had collected by incorporating them into a strong communicative framework. The emphasis in the listening mode is on (a) listening for the gist of a conversation to determine the main ideas and issues and (b) spotting key words and phrases. For example, in a phrase-spotting activity at the beginner level in ECHOS, the learner hears a short conversation and then is asked "What did you hear?" and sees a list of French phrases:

- *tout le monde*
- *se poser*
- *peut-être*

The learner clicks on the phrases heard in the order they occurred in the spoken segment.

The emphasis in the speaking and reading-aloud modes is on appropriate dialogue-like responses to specific cues that suggest conversational situations, given through photographs, graphics, and voice. For example, a typical speaking activity presents a situation and then has the student respond by speaking a response from a menu of utterance choices. In a discussion about health habits, learners are asked about their own opinions and practices. To the question, "Do you smoke?" learners can choose to say one of three responses written on the screen:

- I smoke,
- I don't smoke,
- I used to, but I don't smoke anymore.

Any one of these phrases is a valid response, and the system highlights what the student is recognized to have said and issues a dialogue-like, nonjudgmental response. Alternatively, some choices are less appropriate to the question or cue. For example, in a discussion of health habits in ECHOS, the learner of French is asked about going to the doctor. The choices available are *Je préfère aller voir ma famille* 'I prefer to go see my family,' *Je ne vais jamais chez le médecin* 'I never go to the doctor,' and *J'en ai un*, 'I have one of them.' If the last response is recognized as the one spoken, the system highlights it and issues an expression of incomprehension. However, all response choices are structurally and grammatically properly formed because the focus of the program was not the diagnosis and repair of structural errors.

At the same time that we stressed communication, we wanted to incorporate pronunciation evaluation. This is an aspect of language instruction not stressed in the communicative approach (Hammond, 1995) but one that we believe is essential to complete language learning (see, in this issue, Dalby & Kewley Port and Eskenazi).

3.5 Listening Before Speaking

Studies in second language learning suggest that the differences between production and comprehension are not as great as might be thought. For example, the beneficial effect of comprehension training on production is presented in Postovsky (1977). Krashen (1985), in discussing the benefits of the "silent period" during which second language learners produce very little during initial exposure to language, argues that the competence of the learner is strengthened by the act of understanding. In the VILTS activities screen layout, the user is implicitly encouraged to begin with comprehension and discrimination activities. Speaking activities that involve shorter interchanges follow, and the last set of activities concentrates on longer segments of read speech to supplement the lesson with examples of prosody in longer text. The user is also encouraged to proceed from the top activity in each cluster to the bottom, moving from easier to more difficult activities. The top activity in each cluster presents the material used as a basis for the remaining activities in that cluster. This sequence, schematized in Figure 2, is suggested by the layout of the activities. This sequence is not imposed, however, and students are free to explore in various ways, as outlined above.

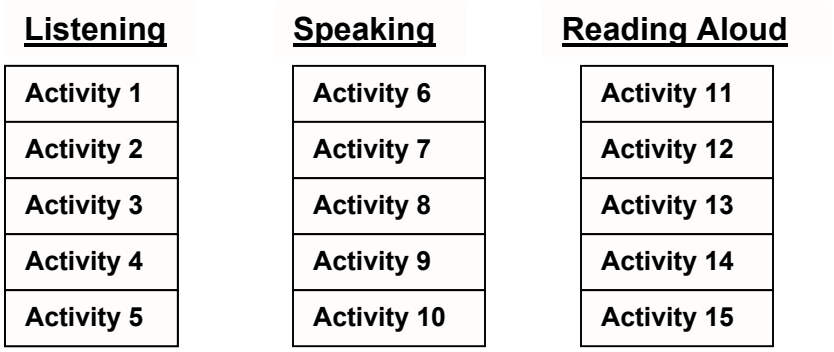

Figure 2. VILTS Screen layout suggesting, but not imposing, a lesson activity sequence.

4. SPEECH TECHNOLOGY AND PEDAGOGICAL DESIGN

Speech is the first and most used linguistic medium, and speech technology enables computer-aided instruction to focus not just on text but also on speech. Students often have difficulty generalizing from the one, typically very careful, style of speaking that a teacher may represent to the many casual styles observed "in the street." We therefore thought it was important both to use spontaneous speech from a variety of talkers and to include more careful read versions to help bridge this gap. The use of unscripted, spontaneous materials, however, poses some challenges. Other major challenges have been the appropriate insertion of the technology into a pedagogical plan, the validation of the pronunciation scoring, and user evaluation. These challenges are discussed, respectively, in the sections below.

4.1. The Use of Unscripted Materials

Speech data collection represents a major task in our development of speech technology for language learning. We collect speech data from both native and nonnative speakers in order to train recognition models and to create scoring algorithms. In developing the ECHOS version of the VILTS, we also collected natural conversations from the same native speakers to serve as the basis for the lesson activities. Thus, the same speech samples serve to shape the recognizer and, inserted in the lessonware, to give listening practice. The conversations we collected were supplemented with read samples of authentic text on similar topics in the French newspaper *Le Monde*.

Our initial challenge came in gathering conversations at the three levels of difficulty called for in the VILTS architecture (Figure 1). Conversations do not naturally fall neatly into these categories, nor do most conversations take place at the most beginning level. To address this issue, interviewers were trained in guiding conversations from beginning levels, with simple constructions and common vocabulary, to advanced levels incorporating greater linguistic sophistication. SRI collaborated with U.S. government instructors to gauge levels of difficulty and to provide appropriate sample questions to ensure that the conversations would be suitable for the leveled teaching framework standard in government instruction (e.g., we drew on "yes/no" questions to elicit beginning level conversations). The materials from *Le Monde* were simplified as appropriate for use in the lower levels.

How and where to incorporate these materials in the lesson architecture presented a further challenge. Natural, unscripted exchanges contain interruptions such as false starts, repeats, stammering, and other disfluencies, as well as deviations from standard linguistic patterns. While this speech is very useful for training the student's ear to real interactions, the more carefully spoken, read speech is used in the system as the model for student speech. In the VILTS, the unscripted conversations were mapped to the listening comprehension exercises so that the user would become accustomed to the flow of authentic conversation with all its natural disfluencies and its linguistic and prosodic variations and irregularities. However, in the subsequent modes of speaking and reading aloud, the clearer, read version of the conversation was used as a model when the user was asked to produce speech. In some activities, such as role playing or question posing, a combination of the spontaneous and the read versions was used to mimic more closely a natural interaction. In these exercises, the user's turn was modeled by the careful speech and the system response was the conversational version, to more closely mimic a learner's conversation with a native speaker. Finally, the text from *Le Monde* was used as a model in the reading-aloud mode to incorporate prosody models for longer segments of speech.

4.2. Speech Recognition in a Pedagogical Plan

A major challenge in the VILTS project was to develop a system that was both technically feasible and pedagogically valid. The communicative framework we chose to follow emphasizes meaningful interactions in the target language. However, as Hammond (1995) points out, communicative theory does not make explicit claims about the teaching of pronunciation, and by its omission, accords pronunciation a less important status. Spoken interaction with a minimum of textual support is a major feature of the communicative approach. However, a system in which student speech was to be evaluated and a valid score returned seemed to require text on the screen, which would be read. Speech recognition technology, although it has made remarkable strides in recent years, is still far from being able to understand arbitrary speech. Spontaneous speech from native speakers is quite a challenge, and

nonnative speech is still more challenging. We decided that the desire for score validity should take precedence over the desire for spontaneity when pronunciation was to be scored. Therefore, we developed initial lesson activities in which all pronunciations being scored were based on some text support. Lesson activities in which pronunciation is not scored can be more flexible, as will be discussed later.

Although communicative theory relegates pronunciation practice to a relatively minor role, this view has been controversial (Hammond, 1995), and even some proponents agree that evaluation and instruction would be beneficial to many learners. A central issue in pronunciation training is to determine the possible role of explicit instructions (Terrell, 1989). There is also a body of research showing that adult learners are capable of perceiving, imitating, and learning new phonetic distinctions (e.g., Flege & Hammond, 1982; Hammond & Flege, 1988; Rochet, 1993). A study by Catford and Pisoni (1970) examined auditory versus articulatory training in English-speaking students learning "exotic" sounds such as the glottal stop and a glottalized "k." Based on their test results, they suggest that ear-training and mimicking alone, while effective for some students, are less effective in general than articulatory training in teaching both auditory discrimination and the production of exotic sounds. Other studies with Japanese students learning to identify English /r/ and /l/ offer additional evidence indicating that the knowledge gained during perceptual learning may be transferred to production (Bradlow et al., 1995).

We developed an architecture for the VILTS that would admit a communicative approach in which student responses were elicited as meaningful in context. The architecture includes separate modules to review student progress as they go through lessons; these modules consist of pronunciation evaluations after each lesson on overall pronunciation and individual sounds. While students are engaged in communicative activities in the lessons, the recognizer guides the interactions with appropriate, conversation-like responses from the system. Simultaneously, student input is logged and stored as a basis for pronunciation evaluation, but a score is established only after a lesson is completed and sufficient speech has been collected to return a reliable score. Since it is desirable to collect multiple student utterances to ensure consistent and reliable pronunciation evaluation, the background collection of input during communicative activities--before returning a score at the conclusion of a lesson--supports both the communicative component of learning activities and the robustness required of the scoring technology.

Because we also envisioned comprehension activities with no text as a crutch, the system was designed to begin with listening comprehension and discrimination activities and to flow into activities eliciting speech from the student. A range of activities was developed: those based on spoken material alone, those combining speech and text, and those based largely on text. This approach is seen in the trimodal architecture of the lessons containing sets of listening, speaking, and reading-aloud activities (see Figure 2).

Another issue in research and implementation of the French speech recognizer in the VILTS was that of trade-offs in weighting types of possible recognition errors. The main types of potential recognition errors are misrecognition (or false acceptance) and false rejection. When the rejection weight is high, the rate of false rejections is high, but the rate of false acceptances/misrecognitions is low. On the other hand, lowering the rejection weight results in more frequent misrecognitions/false acceptances but less frequent false rejections. Research was conducted on the optimum strategy for pedagogical purposes. We tried to minimize both false acceptance and rejection. However, we biased the system toward rejection because we felt that forcing a student to repeat a marginal utterance was preferable to the risk of falsely accepting an error and potentially confusing the student.

4.3. Validation of Automatic Pronunciation Scoring

In our scoring paradigm, both native and nonnative speech data are collected, and a database is created of ratings from human experts (e.g., language instructors) to enable the development of machine scores. We use the data to assess the reliability of human ratings of pronunciation and to develop and assess pronunciation scoring that correlates well with humans. We treat pronunciation evaluation as a prediction problem: Can we predict the score a human expert would assign to a particular speech segment? Using the speech and the expert-ratings data, we build statistical models and assess various machine scores as predictor variables.

For the VILTS and related systems, we have developed and tested our algorithms on data collected from native speakers of American English speaking French and Spanish. In the ECHOS project the human experts were senior French instructors at a major government language school. In the Spanish project the raters were a panel of five native Spanish speakers. We showed that, when a sufficient amount of speech data are available (many student utterances over the course of a lesson), certain machine scores (e.g., the log-posterior and the normalized duration) achieve a correlation with the human scores comparable to the correlation between human raters. Thus, the agreement between human raters serves as a benchmark for the degree of agreement desired between automatic scores and human judgements. Note that our scoring is intended to operate on multiple-word utterances containing arbitrary words without the need for collecting additional data. This is a far different situation from systems trained on specific minimal pair input such as the pronunciation trainer described by Dalby and Kewley-Port (this issue), where pronunciation scoring has been found to proceed reliably with smaller samples of speech.

Speech recognition technology is key to the automatic evaluation of pronunciation quality. However, standard speech recognition algorithms were not designed with the goal of pronunciation scoring. Therefore, new methods and algorithms had to be devised to match the perceptual capabilities of human listeners in scoring speech quality. The factors affecting human scores (and therefore our automatic scores) include duration, syllabic timing, spectral, and posterior scores. Best results are obtained by combining several

scores. We combined scores using neural networks and classification trees, and we also used Bayesian approaches. Further improvements were achieved by modeling intraword pauses in nonnative speech. By using these scores, we were able to obtain human-machine score correlations comparable to human-human correlations on the same data set.

For example, for our Spanish data, the panel of native speakers rated the overall pronunciation of each nonnative sentence on a scale of 1 to 5, ranging from "strongly nonnative" to "almost native." These human scores were used to evaluate the effectiveness of the machine scores and to calibrate the mappings from the machine scores to the predicted pronunciation scores. To assess the consistency of these human scores, two types of correlation were computed. At the sentence level, pairs of corresponding ratings for all the individual sentences were correlated. At the speaker level, the scores for all the sentences from each speaker were averaged, and then the sequence of pairs of corresponding average scores for each of the speakers was correlated. The correlation between raters was computed on a subset of 2800 sentences rated by all raters. The average sentence/speaker-level interrater correlation was r=0.68/0.91. Our best sentence-level result using automatic pronunciation scores is a correlation of r=0.609, very close to the average .68 correlation found between human raters. Details of these results appear in Neumeyer et al. (1998).

4.4. User Evaluation of the VILTS

Initial user evaluation of the VILTS prototype was begun after the French recognizer was judged sufficiently robust to support smooth interactions. The first lesson was used as the testbed for the evaluations. A questionnaire was developed, focusing on a qualitative analysis. Five subjects were interviewed in three sessions, each lasting approximately two hours. Table 1 summarizes the demographics of the subject population.

Subject	Age	Education	French Experience	Level
Female 1	30	M.S.	4 years high school, 1 year college	High
Female 2	33	B.S.	4 years high school, ~5 semesters post-college	Mid-high
Female 3	37	Ph.D.	3 years high school, 2 years college, 6 months in France	Low-mid
Male 1	31	B.S.	No formal, 2 years living in France	Low
Male 2	33	Ph.D.	4 years high school, 1 year living in France	High

Table 1. VILTS Evaluation subject data.

Subjects were provided with written instructions describing the goals of the project and the current state of development. They were informed that the project was not targeted to structured grammar learning but rather to the elicitation of speech in an engaging and interactive environment to support pronunciation scoring and feedback. The questionnaire was presented orally, and the interviewers encouraged dialogue and comments throughout lesson use. Subjects were asked about their likes and dislikes with respect to each activity and the program overall. They were assisted in working through the lessons only when they seemed stalled or frustrated. Sessions were recorded and transcribed for analysis. Although project resources permitted evaluation with only one lesson, the results were illuminating as a guide to future refinements as well as to interface design for language learning in general. In some areas all subjects agreed, and in others they had divergent opinions, often seeming to depend on their language skill levels. We focused on documenting system strengths and areas needing future work.

The interviewees noted the following strengths of the system:

- They all reacted very positively to interactions with authentic, unscripted materials, because they felt that the point of learning a language was to converse in real-life situations. The unscripted nature of the conversations made the activity resemble a real-life situation.
- They liked hearing native French without much knowledge of the content and trying to figure out what was being said. Better speakers enjoyed the challenge of trying to comprehend without seeing text; weaker speakers wanted an option to display the text.
- The incorporation of high-quality speech recognition proved to be a major strength for all subjects; an interactive activity that elicited French from them was an important factor in their enthusiasm.
- Most important to all subjects was the high quality and real-time nature of the interactions that mimicked real-life interactions.
- Very important to all subjects was the ability of the system to readily recognize nonnative French input and to reject poor pronunciations. They found that when their pronunciation was rejected, they were able to consult key words, practice and compare with a native speaker until they were more confident, and then return to the program and continue with improved pronunciation. The subjects also noted that, through practice with the key words, they learned the vocabulary. They found that they used the key words less as they learned more, which afforded them a high level of satisfaction.
- The key word and browsing capabilities were also well received, particularly the ability to hear the words in context. The subjects with weaker French ability requested the addition of more such key words in the system.
- The self-paced nature of the program was another highly appreciated system component. Subjects liked the ability to navigate freely through the activities and to repeat or review as needed. Since loudspeaker buttons were available to review native pronunciations of all system material as often as necessary, subjects took advantage of this feature to suit their level of comfort

and to practice until they felt confident enough to move on. "You could work with it until you understood it" was a comment often heard in the feedback.

- Interviewers noted that four of the five subjects asked if they could return on their own time to use the system. The fifth, the weakest in French ability, felt that the system was more challenging than suitable to his level.

The interviewees made the following suggestions for improvements:

- More support mechanisms are needed at various levels. The subjects wanted to see the VILTS include video clips, an on-line dictionary, and expanded hypertext capabilities. Since only one lesson was available for evaluation, they felt that a greater range of levels should be accommodated. Concomitant with this was the request for shorter segments of material for users with lower French ability, along with more textual support. The subjects also asked for more help in repairing pronunciation.
- Although the system is highly navigable, subjects would have liked even more flexibility, for example, to move backwards within an activity. They would also have liked the ability to determine the level of pronunciation skill required. Future research is needed to support this function robustly.
- Subjects suggested tracking mechanisms, practice spaces, and a timing feature. They wanted to see indications of their progress through the lessons and also to see how much of an individual activity they had completed. Some expressed a desire to practice both responses and pronunciation before embarking on a scoring session. Finally, some of the more advanced subjects thought that proceeding through the activities against a timer would encourage them to improve their fluency in French.

We observed that, in general, the subjects needed some exploration and ramp-up time with more than one activity to become accustomed to any interface.

5. FUTURE DIRECTIONS

The initial user evaluations suggest several future directions. Clearly, more research is required to understand how technology can improve language learning under different circumstances and for different types and levels of learners. We need to understand differences in short-term versus long-term improvements and how these are affected by different practice and feedback methods. Future directions are also informed by the ongoing research in speech technologies. We believe that the future will continue to require multidisciplinary interaction between those representing language pedagogy and those representing speech and language technology. Below we outline directions in terms of additional lesson development, followed by pronunciation scoring and feedback.

5.1. Additional Lesson Development: Toward Free-form Utterances

The unscripted French conversations we collected for ECHOS could support additional lesson activities. Our research suggests that the most expeditious next step might be to develop new lessons with a narrower scope (e.g., fewer than 15 activities per lesson). The activities would be selected on the basis of pedagogical impact relative to development effort. In addition, excerpts from conversations not yet used could provide a greater breadth of exposure without the overhead of developing coherent, linked materials around full conversations, some of which are more interesting or easier to work with than others. Students could select lessons according to a topic of interest, or lessons could be indexed for pronunciation or linguistic form. This approach would allow rapid development of new lessons, would support additional French speech production by students, and would provide a rich environment for interactions using a variety of French speakers and speaking styles. The greater the number of lessons available to support student speech and practice, the greater the opportunity to foster improvement and track progress.

We have investigated some of these ideas in a new language learning application that leverages the components and architecture of the VILTS. This program, called Special Operations Language Voice Interactive Training (SOLVIT), was developed in just a few months by reusing VILTS components and adding new activities. In SOLVIT the student is coached through successively more independent spoken interactions in French to a level of free-form utterances bounded only by the types of constructions and vocabulary introduced in the lesson. Students produce these utterances without textual support and without reading utterance choices from the screen. To foster the predictable utterances needed by speech recognition, SOLVIT coaches the student on target sentence elements in early speaking activities, then in later activities encourages the student to recombine these elements to create new utterances. SOLVIT also uses the principle of graphically displayed artifacts to bound students' utterances. Normal artifacts of human interaction--desk calendars, restaurant menus, office checklists, road maps--constrain what we talk about in ordinary conversations; in a language lesson they serve to limit what the learner says. Figure 3 shows this principle at play in a sample SOLVIT screen, which displays a map and a checklist for assessing the conditions of supply routes from an airfield. The student must ask about the routes shown on the map and assess the features (shown as icons) on the checklist.

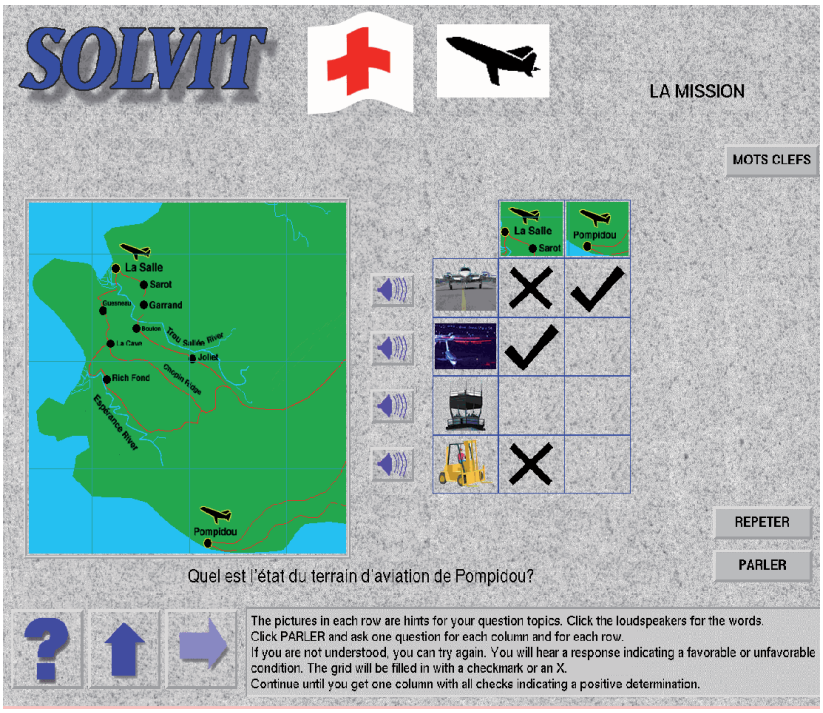

Figure 3. Sample SOLVIT screen showing a task to evaluate road conditions using spoken French.

The artifacts shown in Figure 3 appear as a natural part of the lesson scenario: to provide relief to a hurricane-stricken Caribbean island where the inhabitants speak only French.

5.2. Pronunciation Scoring and Feedback

Pronunciation scoring, as we saw in the initial user studies, provides useful feedback to the learner. It has the potential of being even more useful if we can devise ways to provide more detailed feedback, diagnosis, and repair strategies. In selecting feedback, it is important to understand both technical challenges and pedagogical validity. Future directions include the following:

- **Sentence/Sounds Imitation.** Increased exposure to French native speech, together with imitation and repetition, has been a basic technique used in language classrooms. Although commercial systems sometimes return pronunciation scores (see Wachowicz & Scott, this issue), it is unclear what those scores mean. Based on our research to date, we estimate the need for about 30 sentences from the learner to provide a valid score, that is, one that correlates well with human experts. As research on pronunciation scoring and feedback progresses, it may be possible to return a valid score with fewer utterances. Then we envision exercises where the student practices on first short and then incrementally longer utterances. With such a system, students could practice and compare their speech with that of a native speaker as often as they wished, at their own pace, and receive feedback on the effectiveness of this method in improving pronunciation.
- **Formative Feedback**. Formative feedback to improve student pronunciation could be provided through various graphical representations. Representations of the speech production apparatus (lips and vocal tract) and/or the speech waveform could be used to compare the student's speech with that of a native speaker or to illustrate a student's pronunciation problems. Many displays are possible and several are now in use. However, little data are available showing whether such displays actually improve pronunciation, and, if so, whether students can generalize to conditions of normal interactions when such feedback is not available. We recommend research aimed at assessing whether feedback improves pronunciation and whether any gains carry over to conditions where no feedback is available.
- **Speech Representation**. There are several ways of representing the speech segment relevant to pronunciation training, including speech waveforms and spectrograms (see Eskenazi, this issue). These can be displayed along with corresponding phonetic transcriptions and word alignments. An interactive system that allows users to click on certain regions of the waveform and hear them could be developed. Similar displays of native speakers could be presented to illustrate differences between student and native speech. Scores might also be color-coded to represent the different phonetic segments so that students could determine where their areas of difficulty lie. Since many of these displays (especially spectrograms and phonetic symbols) incur significant training costs before students can make use of them, research is needed to understand the costs and benefits of teaching these skills.

ACKNOWLEDGEMENTS

We thank Leo Neumeyer and the SRI algorithms team for their work on the speech technology components of VILTS, as well as the U.S. Government language teachers and developers who collaborated on graphical design and language pedagogy and who also provided many of the ratings of pronunciation. We also thank Harry Bratt of SRI who led the Solvit effort and the Defense Advanced Research Projects Agency for its support of SOLVIT, and researchers at the U.S. Army Research Institute and staff at the Special Operations Forces Language Office for collaborating on the content of the SOLVIT lesson. We especially thank Melissa Holland of the U.S. Army Research Institute for extensive useful comments on this paper.

NOTE: All product and company names mentioned in this paper are the trademarks of their holders.

BIBLIOGRAPHY

Bradlow, A., Pisoni, D., Akahane-Yamada, R., & Tohkura, Y. (1995). Training Japanese listeners to identify English /r/ and /l/ IV: Some effects of perceptual learning on speech production. *Progress Report No. 20*, *Research on Spoken Language Processing*. Bloomington, IN: Indiana University.

Catford, J.C., & Pisoni, D. (1970). Auditory vs. articulatory training in exotic sounds. *Modern Language Journal*, LIV (7), p. 477-481.

Flege, J., & Hammond, R. (1982). Mimicry of non-distinctive phonetic differences between language varieties. *Studies in Second Language Acquisition,* 5(1), 1-17.

Hammond, R. (1995). Foreign accent and phonetic interference. In Fred R. Eckman (ed.) *Second language acquisition theory and pedagogy*. Mahwah, NJ: Lawrence Erlbaum Associates.

Hammond, R , & Flege, J. (1988). *Attitudes, experience, and the mimicry of sounds: Implications for second language acquisition*. Paper presented at the Seventh International Symposium on International Perspectives on Language, Literature, and Culture, George Mason University, Fairfax, VA.

Krashen, S. (1981). *Second language acquisition and second language learning*. Oxford, England: Pergamon.

Krashen, S. (1985). *The input hypothesis*. London: Longman.

Lee, J., & VanPatten, B. (1995). *Making communicative language teaching happen*. New York: McGraw-Hill.

Neumeyer, L., Franco, H., Abrash, V., Julia, L., Ronen, O., Bratt, H., Bing, J., Digalakis, V., & Rypa, M. (1998). WebGrader(TM): A multilingual pronunciation practice tool. In *Proceedings of the Workshop on Speech Technology in Language Learning*, Stockholm, Sweden.

Oxford, R. (1995). Linking theories of learning with intelligent computer-assisted language learning. In V.M. Holland, J. Kaplan, & M. Sams (Eds*), Intelligent language tutors: Theory shaping technology*. Mahwah, NJ: Erlbaum.

Postovsky, V.A. (1977). Why not start speaking later? In M.K. Burt, H.C. Dulay, & M. Finocchari (Eds.), *English as a Second Language*. New York: Regents.

Rochet, B. (1993). *The role of auditory training in teaching nonnative speech contrasts.* Paper presented at the Third Conference on Second Language Acquisition and Foreign Language Learning, Purdue University, West Lafayette, IN.

Terrell, T. (1989). Teaching Spanish pronunciation in a communicative approach. In P. Bjarkman & R.. Hammond (Eds.), *American Spanish pronunciation - Theoretical and applied perspectives*, Georgetown University Press, Washington DC.

Biographical Data

Marikka Rypa, now at Nuance Communications, was project leader of the Voice Interactive Language Training System at SRI International. She has a Ph.D. in German and Linguistics from Stanford University. Her experience includes several years of teaching languages at Stanford, California State University, and Indiana University. At Xerox Palo Alto Research Center and at SRI, she has conducted applied research as leader of interdisciplinary teams to investigate the role of new linguistic theories and technologies in supporting computer-assisted language instruction. This work has resulted in experimental language learning systems that exploit emerging speech and natural language processing technologies to promote reading, writing, listening, and speaking skills.

Patti Price, Director of the Speech Technology and Research (STAR) Laboratory at SRI International, has been active in speech technology for over 20 years. She has broad, multi-disciplinary expertise in speech analysis, perception, production, recognition, synthesis, prosody, and speech understanding. In addition to her work at SRI since 1988, she has conducted research at the University of Pennsylvania, where she received her Ph.D., as well as at MIT, Haskins Laboratories, the French telecommunications research center (CNET), BBN Laboratories, and the Institute for Perception Research in Eindhoven, the Netherlands. She has also served on numerous advisory, editorial, and program committees.