VILTS: THE VOICE INTERACTIVE LANGUAGE TRAINING SYSTEM

Marikka Elizabeth Rypa
SRI International, Speech Technology and Research Laboratory, Menlo Park, CA, 94025
Email: marikka@speech.sri.com

ABSTRACT

ECHOS is a voice interactive language training system being developed to foster improvement in French comprehension and speaking skills, incorporating speech recognition and pronunciation evaluation. Speech recognition allows students to navigate through units using oral communication with various types of system feedback. The pronunciation scoring being developed is validated by expert human raters. We will discuss the motivation for the program, the nature of the interdisciplinary effort, and the resulting system architecture. Challenges and trade-offs of designing activities using unscripted material and aspect of speech research as related to this application will be described. Finally, we discuss opportunities to use speech recognition on other platforms.

1. INTRODUCTION

In this presentation we describe the current implementation of the Voice Interactive Language Training System. We discuss the program vision and motivation, the development goals, the lesson unit architecture incorporating authentic data, the interdisciplinary nature of the work. We end with a brief overview of future directions.

The French flagship version of the VILTS, called ECHOS, is being developed to foster improvement in listening and speaking skills using state-of-the-art speech recognition technology. Designed to support language learning and skills maintenance at beginning, intermediate, and advanced levels, the VILTS lesson architecture stresses learner-centered navigation through listening and speaking activities. The system currently runs on an SGI Indy* with 32 megabytes of RAM and 2 gigabytes of disk space.

2. DEVELOPMENT GOALS

One of the development goals of the VILTS project is to mine a range of resources and expertise, from speech research to pedagogical design, in service to language education. Another goal is to incorporate authentic, unscripted materials in an engaging, interactive, flexible lesson architecture. A third goal is to exploit state-of-the-art advances in speech research in two related but distinct speech technologies: speech recognition, which guides user interaction in lesson unit activities, and automatic pronunciation scoring of student speech, which provides feedback validated by human expert raters.

3. LESSON UNIT ARCHITECTURE

Using spontaneous, unscripted French conversations on various topics, supplemented by excerpts from the French newspaper LeMonde, the VILTS offers the student authentic, unrehearsed French speech as might be heard in an interview. Conversations were collected at various levels on ten separate topics, including domains such as travel, health, education, and politics, from a pool of 100 native speakers in France. The conversations and test form the basis for lesson unit activities. A read version of the conversations was also recorded by the same speakers so that both spontaneous speech and a clearer and (generally) slower version of each conversation is available to the student in the lesson units.

The system architecture is designed to afford the student a high degree of flexibility that will accommodate different learner styles, from more structured, incremental learning to exploratory navigation of the system resources. Students are given choices at several levels: after system logon, the student can elect to move directly to a lesson unit, or review unit completed with pronunciation scores, or engage in pronunciation exercises based on areas of weakness. After electing to move on to another lesson unit, the student chooses a level (beginning, intermediate, or advanced), and then chooses a unit topic from the topics available at that level. Once a unit topic is chosen, the activity menu allows the student to navigate among various activities that are clustered around listening, speaking, and reading aloud, as illustrated in Figure 1.

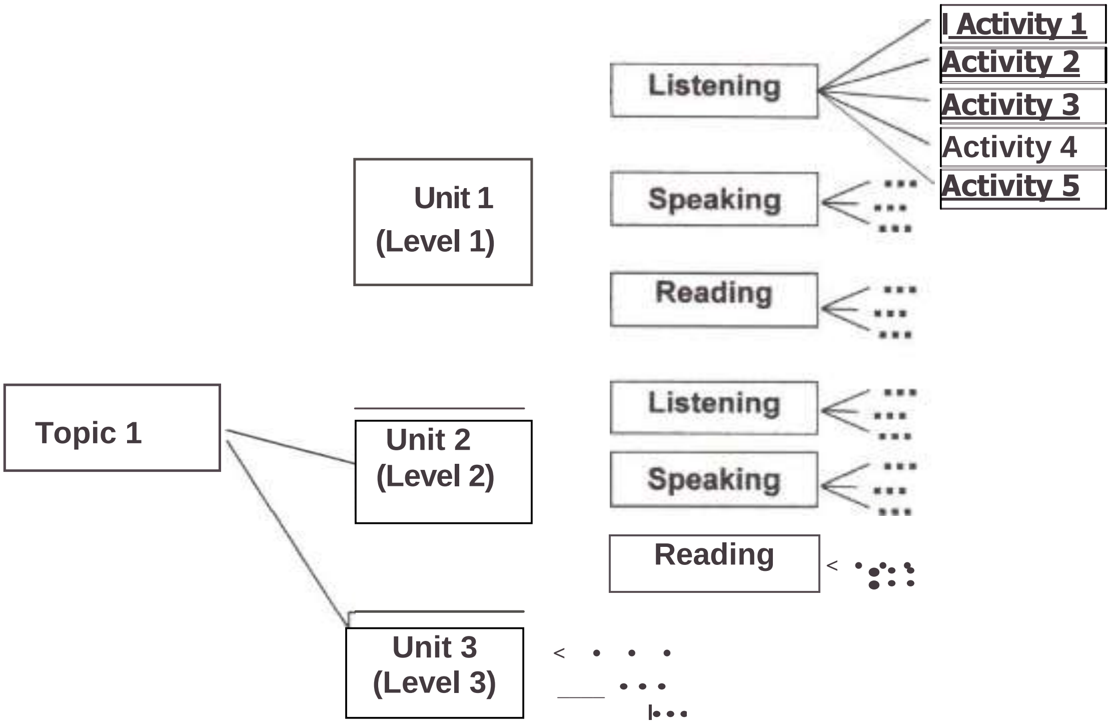

Figure 1

4. SPEECH TECHNOLOGY AND PEDAGOGICAL DESIGN

4.1. Speech Recognition and Pronunciation Evaluation

As student speech is elicited through a variety of lesson activities, the French speech recognizer listens for oral input that will guide the system response. Each student utterance elicits a response from the system within the context of a particular lesson activity. As the student completes a lesson unit and enough speech has been collected, pronunciation-searing algorithms will be employed to compare the student speech with that of native speakers. The system has been designed to elicit a body of student speech sufficient to assure a measure of confidence that the returned pronunciation scores are meaningful. The scores returned to the students will be validated by calibrating the results of the automatic scoring algorithm with the results of human expert ratings of nonnative speech.

4.2. Data Collection

Speech data collection represents a major part of the VILTS project as support for both native and nonnative speech modeling while providing core lesson material. Both unscripted conversations and read versions of these conversations were collected from 100 native speakers over a period of six months in France. Nonnative speech from another 100 speakers at various levels of fluency was collected in the form of read and imitated speech (where the reader could hear the sentence spoken by a nonnative first). This data was then rated on a five-point scale by five expert French instructor raters. The automatic pronunciation algorithms being developed at SRI are mapped to these expert teacher ratings to provide the student with automatic system feedback validated by the teacher scores. The correlation of human scores with the automatic system scoring provides validation to the system feedback given to the student.

4.3. Using Unscripted Materials

The use of unscripted interviews as core material presented us with both a rich repository of resources and a range of challenges in creating lesson units.

To accommodate a focus on comprehension of spoken interactions without textual support, as well as to meet the requirement that all user input be elicited by means of text (to ensure that the pronunciation evaluation is scoring the correct utterance), a range of activities was developed to encompass some activities based on spoken material alone, some activities combining both spoken utterances and text, and some activities where student interactions are guided largely through text. It is widely accepted that the ability to discriminate target sounds is a necessary condition for near-native production of these sounds. The suggested activity order in the VILTS presents comprehension and discrimination activities before speech-eliciting activities.

The goal of incorporating three levels of difficulty in the conversational material presented a further challenge. Conversations do not naturally fall neatly into these categories, nor do most conversations take place at the most beginning levels. Interviewers were trained in guiding conversations from advanced beginning levels on to advanced levels.

Natural, unscripted exchanges are also replete with components that interrupt the smooth flow of speech, such as false starts, repeats, stammering, and other disfluencies as well as deviations from grammatically correct 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 models for student speech. Pruning and extracting appropriate units of speech for interactions represented an additional constraint in using the raw conversations.

4.4. Speech And Pedagogy

A major challenge in the VILTS project was to develop a system that was both technically feasible from a speech recognition standpoint and pedagogically viable. Automatic pronunciation scoring imposed requirements that all input speech be supported by text so that the scorer would return accurate scores and not penalize the student for producing an unexpected utterance. However, since comprehension activities with no text as a crutch were also envisioned, the system was designed to begin with comprehension and discrimination activities and flow into activities eliciting speech input by the student.

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, false acceptance, and false rejection. When the rejection weight is high, the rate of false rejections is higher, but the rate of false acceptances and misrecognition is lower, for example. On the other hand, lowering the rejection weight results in more frequent misrecognitions and false acceptances, but less frequent false rejections. Research is being conducted on the optimum strategy for pedagogical purposes.

5. FUTURE DIRECTIONS

The vision for the VILTS program is to port the software to a more widely used platform such as the Pentium PC. This will make the system more accessible to a broader audience and admit a greater range of authoring tools to more easily add new content and graphic material with a greater range of activities. Another direction is to make learning interactions and pronunciation scoring available via the World Wide Web. Users will be able to access the Web from their personal computers and interact with language learning activities, using the telephone for speech input and receiving scores on their pronunciation.

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

Biodata: Dr. Marikka Rypa is project manager of the Voice Interactive Language Training System in the Speech Technology and Research Laboratory 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 instruction. This work has resulted in experimental language learning systems that exploit emerging linguistic technologies to promote reading, writing, listening, and speaking skills.