Language Learning & Technology
http://llt.msu.edu/vol3num2/emerging/index.html

January 2000, Volume 3, Number 2
pp. 8-12
(page numbers in PDF differ and should not be used for reference)

EMERGING TECHNOLOGIES

Speech Technologies for Language Learning

Bob Godwin-Jones
Virginia Commonwealth University

Contents:

- Recorded Speech: From Analog to Digital
- Speech Recognition
- Speech Synthesis
- Multilingual Speech to Speech
- Speech on the Web
- Resource List

Recorded Speech: From Analog to Digital

Using machines to allow students to work with the target language in its spoken form is one of the oldest applications of technology to language learning. Generations of tape players have allowed learners to listen to examples of native speech and to imitate and compare their own utterances. Dual track cassette players remain a staple of many language labs. Variable speed tape players which allow for slower playback while maintaining pitch have also been widely used. Of course, the trend today is away from analog and towards digital formats. Publishers are making their accompanying audio programs available on CD. At some schools, audio programs have been digitized and made available to students through an intranet or local area network (publishers usually require some kind of restricted access). Audio files are typically digitized in wav or aiff formats or in a streaming format like QuickTime or RealAudio, and played back through Web browsers. Jay Kunz (Mississippi State) has been using PureVoice (a freeware audio program bundled with Eudora) to digitize audio for playback at three different speeds, while maintaining the original pitch.

Digital audio provides for random access, variable playback speeds, and incorporation into interactive language learning applications. However, the greatest potential benefit to language learners the digital age is providing in the area of spoken language practice lies in speech recognition and speech synthesis. The greater processing speed of current personal computers, the commercialization of speech technologies, and the tremendous interest in making the World Wide Web voice-accessible have led to interesting developments in these areas. We are not yet at the point that a beginning Spanish student can have a free-ranging conversation with a computer, but parts of what needs to be there for that to be possible are beginning to fall into place.

Speech Recognition

In the last several years programs have begun to appear which allow word processing and other kinds of computing tasks to be accomplished through voice input. Programs like IBM's ViaVoice or Dragon System's Naturally Speaking have become mainstream software products and have been extended to a variety of languages. These are, however, productivity products, not language learning software. In fact, the needs of language learners in respect to speech recognition software are quite different from those of regular consumers. The commercial speech recognition products are typically trained to recognize an

individual user's voice input, with the assumption that there will not be significant changes in that user's speech patterns. Clearly this is not the case for language learners, whose spoken language will change (we hope) as they learn. The programs, in fact, are designed to recognize the speech of native speakers, not of struggling beginners.

Current language learning software programs which incorporate speech recognition do not generally attempt to process continuous speech from users, but rather operate within a controlled environment, limiting the user's vocabulary. Some of the better-known software which incorporates speech recognition includes TriplePlay (Syracuse Language Systems), Echos (Stanford Research Institute), and New Dynamic English (DynaEd). The forthcoming Spanish for Business Professionals (University of Houston-Downtown) makes extensive use of natural language processing in its application of speech technology. Typically, such programs focus on word/phrase discrimination and word order/syntax transformations, incorporating speech recognition in question formation and answering, transformational grammar exercises, and responses to audio/video input. Using voice input rather than the keyboard or mouse enhances active learning and simulates more closely real world communication. Advanced research and development in speech recognition is being done at the US Military Academy through project Santiago, which incorporates Entropic's speech recognizer into the WinCALIS authoring environment.

Speech Synthesis

For a two-way conversation to take place, the computer needs not only to recognize human speech, but also to talk back, to synthesize human speech. The use of speech synthesis has become familiar to consumers through the telephone system. Text-to-speech has been available for some time in personal computers as well. Apple's PlainTalk, for example, allows user to have texts in English or (Mexican) Spanish read back to them in a variety of voices. Most synthesized speech, however, is a far cry from HAL's smooth voice in 2001. The artificial voices sound like what they are, computers pretending to be human beings. The robotic voices, however, have improved considerably in recent years. Some computer products are capable of providing voices which, while not quite mistakable for human, are quite useful for language learning.

Some demos of text-to-speech synthesis, such as Lucent's are available on the Internet. Selecting a language from the pull-down menu of their demo page allows playback of text in the selected language of English, French, German, or Spanish (reading the default text or a text of your choice), using a variety of voices and speeds with adjustment of parameters such as pitch and breathiness. On-line examples are available from Lucent in other languages as well, such as Mandarin Chinese, Italian, Navajo, or Romanian. Demos are available using other text to speech engines as well, such as Elan (English, French, German, Brazilian Portuguese, Russian, Spanish) and Festival (English, French, Welsh). An experimental use of text-to-speech generated on the fly at user request can be seen in a sample German story. Processing and download time make Internet use of synthesized speech less practical in give and take Web interactions (for example, reading back student input, or using natural language processing to generate responses).

Multilingual Speech to Speech

Speech to speech takes speech processing a step further towards a voice-only computer interface. It incorporates both speech recognition and speech synthesis, enabling computing devices to first understand spoken language, then analyze the utterance through natural language processing, and finally formulate and utter a response through synthesized speech. Several urban police departments in the US are currently experimenting with hand-held computers which can perform speech to speech in several different languages.

Pioneering work in speech to speech technology has been done at Carnegie Mellon University beginning in the 1980's through Project JANUS. Several prototype applications have been developed through JANUS including a conversational speech translator for Spanish and a portable travel assistant that provides speech translation, information and navigation assistance to travellers. On the commercial front, Vocal Systems has developed OmniBabel. OmniBabel is software which accepts spoken language input in several different languages, then processes and recognizes in that or another language, making use of artificial intelligence to try to generate a meaningful response.

Speech on the Web

There is considerable interest today in making the Web voice accessible so that users can have Web pages read to them, from a Web browser over the telephone. AudioWeb from Rutgers University offered one of the first implementations of voice enabled Web access. Netphonic (recently renamed MyTalk) uses extended HTML to provide telephone access to Web services. Vocalis SpeeHTML uses a modified version of HTML to provide interactive voice services. IBM recently announced SpeechML, which provides a markup language for speech interfaces to Web pages. HTML, however, is not well suited for speech applications. HTML does not consistently separate formatting and content and lacks tag sets for speech elements such as voice or pitch as well as formatting for dialogs and conditional structures. However, the arrival of XML (extensible markup language) largely overcomes these limitations.

Several recent XML initiatives hold considerable promise for the interactive use of speech technologies (both speech recognition and synthesis) over the Internet, namely VoiceXML and SABLE. These are projects sponsored by a consortium of companies striving to establish a common standard for incorporating speech commands and dialogs into Web-accessible documents. SABLE is an XML-based markup language for Text-to Speech synthesis. It is designed to support applications such as language learning tutorials. SABLE incorporates the use of "aural cascading style sheets," as a standard way to format documents for voice access. In traditional style sheets, level 1 headers (H1 in HTML) might be given a style definition to be rendered in 24 point bold red font. In an aural style, the parameters would deal with voice, pitch, and other voice attributes (as in H1 {voice family: paul, male}).

VoiceXML, or Voice markup language, is a joint project of some 20 companies spearheaded by Motorola, AT&T, and Lucent. It is an ambitious project to support complete, interactive speech services over the Web (and through telephone access to Web pages). Initial specifications (version 0.9) were published recently (9/99). The following example is taken from that document. It asks the user for a choice of drink and then submits it to a server script:

```
<?xml version="1.0"?>
<vxml>
  <form>
    <field name="drink">
      <prompt>Would you like coffee, tea, milk, or
nothing?</prompt>
      <grammar src="drink.gram"/>
    </field>
    <block>
```

```
    <goto next="http://www.drink.example/drink2.asp" submit="drink"
method="get"/>
   </block>
  </form>
</vxml>
```

A *field* is an input field. The user must provide a value for the field before proceeding to the next element in the form. A sample interaction based on this code might be:

> C (computer): Would you like coffee, tea, milk, or nothing?
>
> H (human): Orange juice.
>
> C: I did not understand what you said.
>
> C: Would you like coffee, tea, milk, or nothing?
>
> H: Tea
>
> C: (continues in document drink2.asp)

This is of course an example of limited usefulness for language learning. But the ability to write formatted text to generate meaningful spoken dialogues with conditional statements and responses would provide a powerful new technology tool for language learning.

Resource List

Information and Research on Speech Projects & Related Technologies

- Commercial Speech Recognition Extensive and well-maintained list
- Speech Research Long list (lots of graphics) from UCSC
- Multilingual Speech Processing Article by Alexander Waibel
- inSTIL Integrating Speech Technology in Language Learning SIG (EUROCALL)
- C-Star Consortium for Speech Translation Advanced Research
- SABLE A Synthesis Markup Language (Bell-Labs)
- Voice XML Forum Joint XML project for speech
- Speech in Java Information on Java Speech Markup Language (JSML) & Java Speech Grammar Format (JSGF)
- JANUS Speech to speech project from Carnegie Mellon
- Verbmobil Large-scale German speech project
- Speech Translation Links By Joachim Quantz
- Natural Language Processing Yahoo list

- Natural Language Processing Webopedia site
- Computing Resource Repository Source for papers on computational linguistices, natural language processing (ACM)
- The Association for Computational Linguistics
- Machine Translation Links From Catherine N. Ball (Georgetown)

Voice and Speech Applications/Software

- The Festival Speech Synthesis System From the University of Edinburgh
- PureVoice From Eudora/Qualcomm
- Plaintalk Apple's speech technology in different formats
- L&H Software PowerTranslator, VoiceXpress, RealSpeak
- Lucent Multilingual Text-to-Speech Systems
- IBM ViaVoice
- Philips Speech Processing
- Unisys Natural Language Speech Assistant
- Natually Speaking From Dragon Systems
- Vocalis SpeechWare
- Locus Dialogue
- Elan TTS

Speech Demos on the Web

- Elan English, French, German, Brazilian Portuguese, Russian, Spanish
- Lucent English, French, German, Italian, Spanish
- Festival English, French, Welsh
- RealSpeak English
- Pig Latin Translator (Lucent)