October 2009, Volume 13, Number 3
pp. 4–11

EMERGING TECHNOLOGIES
SPEECH TOOLS AND TECHNOLOGIES

Robert Godwin-Jones
Virginia Commonwealth University

Using computers to recognize and analyze human speech goes back at least to the 1970's. Developed initially to help the hearing or speech impaired, speech recognition was also used early on experimentally in language learning. Since the 1990's, advances in the scientific understanding of speech as well as significant enhancements in software and hardware have allowed speech technologies to be robust enough to be deployed in commercial enterprises. At the same time, developments in the fields of linguistics and second language acquisition have led to greater interest in using computers to help in developing speaking and listening skills. As a result, there are now powerful tools and technologies available for speech analysis. There has been significant progress in projects using speech technologies for teaching pronunciation and speaking. Some major commercial language learning software also now features these technologies. New standards are being developed in this area, which should allow advances and innovations to be shared more easily and implemented more widely.

Visualization of Speech

Since the earliest incarnations of computer-assisted language learning, it has been possible to display human speech in a graphic representation on a computer screen. The visual display generally has shown a waveform or pitch contour. Programs such as Visi-Pitch, available since the MS-DOS days, have been widely used to show a representation of a learner's utterance alongside that of a native speaker. Early pedagogical applications left it up to the learner to determine to what extent the generated contour matched the model. Little guidance was given on how to improve, other than the suggestion to try again. While the actual graphics display has remained the same in many applications of speech visualization, language learning implementations today tend to provide much more useful feedback and assistance to the learner in correcting pronunciation problems. Moreover, there is recognition today that the past practice of having students work with individual sounds and sentences out of context only goes part of the way towards helping with pronunciation, as it leaves out intonation at the sentence and discourse levels. Using artificially generated sentences does not necessarily put learners on the path to communicative ability with natural speech.

Most language learning projects using speech analysis follow the same basic structure: students listen closely to model speech, paying attention to aspects of the native speaker's pronunciation, then are asked to generate the utterance themselves. They receive feedback, often both visual and auditory. In addition to spectrograms and waveforms, the visual feedback may include a representation of the mouth showing physically how the sound is to be produced. The audio feedback may be a recast of the utterance, some kind of encouragement/critique, or possibly the item produced in a different context. Students then repeat (possibly multiple times) and/or proceed to the next item. Some applications take the next step of having the learners practice the pronunciation skills in communicative exercises, whether in peer-to-peer role-playing or in real or simulated social interactions.

One of the issues often raised with the use of traditional approaches to visualizing speech patterns is the difficulty users may have in understanding or interpreting the significance of the displays. Some speech software developers have experimented with game-like interfaces for providing visual feedback, such as using speech input from the student to control movements in a game (replacing joystick controls) or using a racing car interface in which adherence to the road is determined by how far separated the user's utterance is from that of the model native speaker's (Gòmez et al., 2008). One ESL (English as a second language) pronunciation program (Hearsay) used pronunciation accuracy in a bowling environment to determine the number of pins knocked down (Dalby & Kewley-Port, 2008).

There are clear ways in which computer-based pronunciation training can be more efficient than classroom-based practice. The computer provides one-on-one individualized attention with patience enough to allow unlimited tries. Individuals proceed at a pace with which they are comfortable, processing feedback and using available tools as they are needed, or moving quickly through an exercise. The interaction with the computer is private, thus likely to be less stressful to the learner than repeated teacher corrections in a classroom environment. Computer-based training can supply many more native speaker voices as models, a recognized benefit to learning pronunciation. A software program can also be programmed to adapt to individual learner's progress by customizing practice to that student's needs, testing for transference of skills to other contexts or speakers, and doing interval checks to see if knowledge has been retained. The accumulated learner data can be valuable to researchers looking to improve software and learn more about optimal approaches to pronunciation training.

Automatic Speech Recognition

While a valuable tool for pronunciation training, computer-based speech recognition can be used as well to analyze learner speech more generally and to serve as the foundation for creating auditory interactions between the learner and the computer. This is possible through automatic speech recognition (ASR), which allows software to interpret the meaning of a speaker's utterance. This technology has come so far that in recent years it has been deployed widely in commercial systems such as travel reservations, stock price quotes, weather reports, or sports score reporting. Many consumers have had frequent dialogues with disembodied voices through automated phone ASR systems in a variety of help desk or customer service environments. ASR is also the basis for commercial dictation systems such as Dragon NaturallySpeaking. Such systems tend to work well with native speakers and within controlled vocabulary domains. The dictation systems increase accuracy further by allowing the software to be individually trained to recognize particular voices.

Despite their use commercially, there are a number of concerns regarding ASR use in computer-assisted language learning (CALL). The most salient issue is the fact that systems built around the recognition of native speaker speech may not recognize the quite different accent of a language learner. If an ASR system is not reliable enough to understand a correct utterance from a learner, this can be a devastatingly frustrating experience. If, on the other hand, the tolerance is set so low as to recognize a wild approximation of the targeted sound, the learner is not likely to take away much from the interaction. ASR systems are likely to have problems as well with non-grammatical constructions and broken sentences. They tend to give accurate results when used with a pre-defined lexical domain and therefore have difficulty with non-standard vocabulary and off-topic comments. This means that they are less easily deployable (i.e., effective) in environments using free-flowing natural speech. Unfortunately, that is just the environment we want our language learners to be working in. Furthermore, such systems are complex, difficult and expensive to develop, and not easy to integrate into language learning programs.

Nevertheless, there is immense interest in ASR for CALL, with its use steadily increasing in software for language learning. Despite limitations in accuracy and range, the promise of having a computer recognize spoken language holds so much potential for helping students improve speaking skills, that ASR represents an irresistible attraction to CALL developers. For an ASR system to be used with language learners, the speech recognition and analysis system needs to include components not necessary in a system targeted for use by native speakers. The language models, which together with the defined grammar form the database for speech recognition, need to include samples of non-native language. This is important both for the phonetic analysis and for the system's language grammar, which should be able to recognize common grammatical misconstructions. Unfortunately, collecting learner speech is not easy, due to logistical and legal issues. If the system is to be used with specific first language (L1) learners, that simplifies the process, in that common characteristics of learner meta-language for that group can be included. This kind of targeting makes it more likely that frequent non-standard usage will be recognized and that specific feedback for that particular error pattern can be given.

Early ASR systems used template-based recognition based on pattern matching. Today's systems are more sophisticated, complex, and "intelligent," typically using a probability-based system known as the Hidden Markov Model (HMM). They are built around a very large collection of speech samples. The speech analysis works in the following way. The new input from the learner is received, digitized, and analyzed. Then the utterance is compared to the stored information in the database, together with the grammar model and lexicon. A statistical process generates a confidence score for possible matches and delivers an evaluation and feedback based on the results of that process. One of the variables that can be adjusted in that system is where to set the bar in terms of pronunciation accuracy. Setting the bar higher requires closer matches for the learner; too high a setting could be discouraging and counter-productive. That decision should be based on the kind of desired outcome for the learner using the program, namely whether it is simply comprehensibility (by the typical native speaker) or the higher goal of closeness to a native accent. In practice, this distinction is not always made. But as discussed in a seminal article, this should be a fundamental component of how a system is designed, as both the analysis and feedback could be quite different depending on the end goal (Neri, Cucchiarini, Strik, & Boves, 2002).

Speech Analysis and Language Learning Software

There are a number of tools for speech analysis, many of which can be and have been adapted for language learning. KayPentax (formerly Kay Elemetrics) markets the widely used Visi-Pitch, now at version four. The latest release features a waveform editor, auditory feedback and voice games. The games are quite basic, for example, an animated graphic based on pitch and amplitude of the sound input. The company also sells the popular Computerized Speech Lab (CSL), a powerful hardware/software speech analysis system. Its hardware is used in conjunction with a PC and offers high-quality input and output as well as a variety of software add-ons. The latter includes a Video Phonetics Program, which features a synchronized display of video and acoustic data. The most widely used option with CSL is Real-Time Pitch, which provides extensive analysis capabilities to compare two speech samples. KayPentax also sells Multi-Speech, a software only speech analysis program, with similar software options to CSL. The EduSpeak speech recognition system from SRI International supports 9 languages and includes program interfaces for use in Macromedia Director and Microsoft ActiveX. For oral language testing, the Versant suite of tests (available for English, Spanish and Arabic) incorporate speech processing and can be taken over the phone or on a computer. A free on-line demo for English illustrates how the 15-minute test works.

There are as well non-commercial tools available. The Speech Analyzer from SIL International (formerly the Summer Institute of Linguistics) performs frequency and spectral analysis and can be used to annotate phonetic transcriptions. Of particular interest to language learning, it also allows for slowed playback and looping of audio. SIL International makes available other tools for language analysis and recording including Phonology Assistant and WinCECIL, both of which can be used together with Speech Analyzer. A widely used authoring tool is WinPitch LTL, a Windows desktop application. Teachers create lessons consisting of a sequence of speech models, to be repeated or imitated by the learner, with the model speech displayed in graphic form on the left and the learner's input on the right. The program features Unicode word processing and Web linking. As with similar tools, WinPitch LTL offers powerful capabilities but it is likely the rare language teacher who can find the time to create pronunciation lessons themselves from scratch. Teachers who do, like Marjorie Chan for teaching Chinese (2003), find that the flexibility and customizability of such tools, as well, of course, as evidence of student improvement, compensate for the time and effort involved.

Other speech analysis tools include RTSPECT, WASP (both from the University College London), WaveSurfer (from the Swedish Royal Institute of Technology), and the CSLU Toolkit (from Oregon Health and Science University). COLEA is a tool for speech analysis that works within Matlab, a well-known tool in mathematics education. The Sphinx Group at Carnegie-Mellon University (CMU) also makes a set of open source speech recognition engines available. One of the few speech analysis programs currently available for the Macintosh and Linux platforms as well as for Windows is Praat (Dutch for 'talk'), from the Institute of Phonetic Sciences, University of Amsterdam. This is a widely used tool, perhaps in part due to the availability of an extensive tutorial, something often missing from open source tools. An interesting software tool in this area is the Hidden Markov Model Toolkit (HTK), from Cambridge University, a portable toolkit for building and manipulating hidden Markov models, used primarily in speech recognition projects. A widely-used free tool for audio recording, Audacity, also features spectrogram visualizations. It is available for Winows, Mac, and Linux.

These open source tools can be used to build quite sophisticated language learning software. A free video annotation tool, Anvil, for example, allows for import of data from Praat, so as show voice patterns in a multi-layered video annotation. An interesting application built with Praat is SpeakGoodChinese, a Web-based tool for beginning Mandarin learners for help with recognizing and producing tones correctly. Instead of basing analysis of feedback on data from a collected database, a system was developed synthetically based on intelligent algorithms, to calculate possible erroneous pitch tracks. A similar program for learning Mandarin is the Pinyin Tutor, which adds an interesting twist. It asks the learner to type in the pinyin (Romanization of Mandarin) of a speech segment. If the pinyin is incorrect, a synthesized voice reads the pronunciation of the pinyin as the learner entered it. Other programs using open source tools include Fluency from CMU, an English language speech-based learner-computer conversation program, and CandleTalk, a similar English conversation program developed in Taiwan.

While these programs are for the most part demonstration or research projects, there are high-profile commercial language learning programs that feature speech technologies. One of these is Tell Me More from Auralog. As can be seen from an on-line demo, the software engages the learner in a dialogue, asking a question (in both spoken and written form), which invites a spoken response from the learner. The answer is evaluated globally as correct or not and given a score between 1 and 7, shown as a series of rectangles. The learner sees a pitch curve and waveform display of both the learner's utterance and that of a model native speaker. In its specific pronunciation exercises, the program uses speech recognition for analysis and feedback, while showing simulated images of the mouth. It also uses voice recognition in language games, such as finding pronounced items in a word puzzle. Tell Me More allows role-play through the learner taking on a speaking role in a simulated TV program, with the learner practicing first in a "pronunciation workshop" with the lines to be delivered, then speaking those lines in the video playback together with the other characters being voiced by the native speaker actors. Rosetta Stone also uses voice recognition. The feedback for Rosetta Stone gives a correct/incorrect evaluation but also highlights individual sounds mispronounced, which are shown grayed out in the feedback. Several high-profile ESL/EFL programs such as DynEd's Intelligent Tutor (formerly Dynamic English), also makes extensive use of speech technologies. These programs are high-end and high-cost and, in terms of their use of voice recognition, have gotten mixed reviews in professional journals (Hinck, 2003), but nevertheless they enjoy wide use. Given their popularity, it is unfortunate that there are not studies that go beyond reviewing these products and analyze and evaluate their use in controlled language learning environments, including when used as a supplement in traditional classroom environments.

Standards and Outlook

Creating software that incorporates speech technologies is not an easy task and is made more difficult from the fact that there have not been commonly accepted standards in this area. Fortunately, that has been changing recently, with the W3C's (World Wide Web Consortium) efforts in the areas of the Speech

Recognition Grammar Specification (SRGS) and the Semantic Interpretation for Speech Recognition (SISR). These standards are beginning to be implemented. A SRGS-XML editor is available which uses a graphical user interface to create new grammars for voice recognition. The Java format often used in this area, the Java Speech Grammar Format (JSGF), can be converted to SRGS. The widely-used commercial Loquendo ASR uses SISR. It also incorporates a subset of ECMAScript, the official moniker for JavaScript. This use of the main Web browser scripting language, along with the fact that the main Web standards organization (W3C) is involved in standards setting, points to the fact that the future speech technology interface of choice is likely to be a Web browser, rather than the desktop application mostly used today.

A speech recognition program which makes an interesting use of the Web is SPICE, another project out of CMU. It is a tool for developing speech processing for different languages and features both a tool for harvesting Web texts and a browser-based voice recorder for direct user recordings. The idea is to crowd-source development of a speech analyzer for languages for which such systems are not likely to be available. SPICE has been used to generate speech recognition tools for Afrikaans, Bulgarian, and Vietnamese. The system learns sound rules from analysis of user input, updating after each new word is added. Efforts such as this to encourage development of open source speech recognition capabilities for less-commonly taught languages are important, as it is unlikely that the small market share instruction in these languages represents will lead to the development of commercial products. The ability of software to learn and improve is, in fact, another growing trend in this area. Sagawa, Mitamura, and Nyberg (2004) describe the implementation of a system of correction grammars that are dynamically generated based on analysis of dialogues entered into the system. This is clearly bound to become an ever more important component of such systems, as the drive for more accuracy continues.

Another likely direction for the use of voice analysis in language learning software is increased multimedia. Debra Hardison (2005) has shown how beneficial inclusion of video can be for pronunciation training. As shown in the Tell Me More software, video can also be valuable in extending the knowledge gained to real world situations. Indeed, the incorporation of real-world natural speech rather than scripted sentences is another direction that is important. As part of that development, it will be important for software to be able to deal effectively with discourse-level input. In one view, “the use of computer technology has furthered the dominance of sentence-level practice rather than promoting the use of discourse intonation” (Levis & Pickering, 2004). This is clearly not an easy task to take on, however it is particularly important if the software is to address not only prosody but social aspects of speech as well. Levis and Pickering demonstrate, for example, how important changes in tone can be in English to signal attitudes toward the content. Dealing with socio-linguistic aspects of language is a challenging proposition for language technology but intriguing as well, and one that invites new, innovative approaches.

Most of the evaluations of speech recognition projects come from the researchers themselves. It would be useful to have more reviews and comparisons of different approaches. Also helpful would be more of the kind of study done by Engwall and Bälter (2007) that compares pronunciation training in the classroom and on the computer. Their study also explores different options for feedback than those typically given. For example, they put forward the suggestion of adding a third response for feedback, supplementing the traditional “correct” and “incorrect” with something along the lines of “satisfactory for the time being.” This is part of a strategy to encourage students yet provide honest feedback. It is important to be accurate, but also to be as positive as possible, a challenge when evaluating learner pronunciation. Feedback to users is a difficult issue due in part to the multiple sources of error in learner speech, including environmental variables. Feedback ideally should take into account the user's language level, L1 typical interference patterns, and the user's personal learning history (i.e., consideration of issues such as whether an error may be a persistent problem that needs special treatment). For most purposes, feedback should focus on errors that adversely affect comprehension, which may in fact be intonation or rhythm issues rather than individual sounds. I agree with Engwall and Bälter that, given both the frequency of erroneous analysis and the variations in personal learning styles, the best practice in this area generally is to provide minimal feedback as a default, with much richer feedback available on request. It would be helpful to have more studies similar to that of Precoda and Bratt (2008) which deal with telltale indicators of non-native versus native accents, including identifying particular phonemes or intonational patterns which act as signals of non-nativeness. Such studies, conducted on specific L1 (first language) and L2 (second language) combinations, would enable more specifically tailored feedback in speech applications.

Finding confidence-building approaches to computer-based pronunciation training can be particularly important in an area so fraught with emotional baggage. Games can help; Engwall and Bälter (2007) suggest exploring a karaoke style game. An early example of using speech recognition in a game was TraciTalk, a detective story for learning ESL. It used IBM VoiceType to allow users to select from a list of choices on how to proceed to resolve the case. A game element is also introduced in Microworld, part of the Military Language Trainer (Holland, Kaplan & Sabol, 1999). It asks users to give commands, which are then carried out virtually on the computer. In fact, an incorporation of speech recognition into virtual reality is a compelling concept. An example of that combination is Zengo Sayu, an experimental program for learning Japanese. Incorporation of voice analysis is planned as a next step in a Mexican project for teaching English to engineering students. The U.S. military is also continuing in this direction in developing language-learning software. Maatman, Gratch, and Marsella (2005) describe a prototype of a “listening agent” which reacts to user's speech using gestures and posture shifts in an attempt to simulate what happens in actual human conversation.

Finally, a direction for the near future is the appearance of voice recognition in mobile learning programs. Efforts in this area have been underway for some time. A high profile program has been used by the U.S. Army in Iraq to facilitate communication between American soldiers and Iraqi civilians. There are a number of projects underway that target mobile devices specifically. At CMU, for example, there is a good deal of interest in new smart phones, since they feature more powerful processors as well as consistent access to the Internet. This allows both the higher demands of voice processing as well as the possibility of bypassing the limited storage capacities of phones by accessing on demand server-based data. PocketSphinx is a lightweight speech recognition engine for mobile devices developed at CMU. The most recent version of the Apple iPhone has voice recognition built in (in 6 languages) but only for dialing numbers in the address book or playing music. Recently a voice API (application programming interface) for the iPhone, CeedVocal SDK (for English, French, German), has been released, which allows developers to build speech recognition into their iPhone applications. Google's Android phones also have some voice recognition capabilities built in. We are likely to see this trend continue and accelerate, and one would hope that future implementations look beyond catering exclusively to English language speakers, and that standards will develop in this area as well. Considering the enormity of the process involved in developing well functioning speech tools, the availability of accepted standards and agreement on common basic approaches would enable sharing when feasible and would ensure an easier path for developers, who would not need to learn to work with an array of different proprietary approaches.

REFERENCES

Chan, M. (2003). The digital age and speech technology for Chinese language teaching and learning. *Journal of the Chinese Language Teachers Association, 38*(2), 49-86.

Dalby, J. & Kewley-Port, D. (2008). Design features of three computer-based speech training systems. In V. Holland & F. Fisher (Eds.), *The path of speech technologies in computer assisted language learning* (155-173). New York: Routledge.

Engwall, O., & Bälter, O. (2007). Pronunciation feedback from real and virtual language teachers. *Journal of Computer Assisted Language Learning, 20*(3), 235-262.

Gòmez, P., Álvarez, A., Martínez, R., Bobadilla, J., Bernal, J., Rodellar, V., & Nieto, V. (2008). Applications of formant detection in language learning. In V. Holland & F. Fisher (Eds.), *The path of speech technologies in computer assisted language learning* (44-66). New York: Routledge.

Hardison, D. (2005). Contextualized computer-based L2 prosody training: Evaluating the effects of discourse context and video input. *CALICO Journal, 22*(2), 175-190.

Hincks, R. (2003). Speech technologies for pronunciation feedback and evaluation. *ReCALL 15*(1), 3-20.

Holland, V., Kaplan, J. & Sabol, M. (1999). Preliminary tests of language learning in a speech-interactive graphics microworld. *CALICO Journal, 16*(3), 339-359

Levis, J. & Pickering, L. (2004). Teaching intonation in discourse using speech visualization technology. *System, 32*(4) 505-524.

Maatman, R., Gratch, J. & Marsella, S. (2005). Natural Behavior of a Listening Agent. In T. Panayiotopoulos, J. Gratch, R. Aylett, D. Ballin, P. Olivier, & T. Rist (Eds.), *Intelligent virtual agents* (25-36). Berlin: Springer.

Neri, A., Cucchiarini, C., Strik, H., & Boves L. (2002). The pedagogy-technology interface in computer assisted pronunciation training. *Computer Assisted Language Learning, 15*(5), 441-467.

Precoda, K., Bratt, H. (2008). Perceptual underpinnings of automatic pronunciation assessment. In V. Holland & F. Fisher (Eds.), *The path of speech technologies in computer assisted language learning* (71-84). New York: Routledge.

Sagawa, H., Mitamura, T. & Nyberg E. (2004). Correction grammars for error handling in a speech dialog system. In S. Dumais, D. Marcu, & S. Roukos (Eds.), *HLT-NAACL 2004: Short papers* (61-64), Boston: Association for Computational Linguistics.

RESOURCE LIST

Articles on speech technologies

- Interactive translation of conversational speech - IEEE article
- Learning by doing: Space-associate language learning using a sensorized environment
- Star Trek's universal translator: Coming soon to an iPhone near you? - Mobile voice analysis
- Talking paperclip inspires less irksome virtual assistant - Article on CALO (Cognitive Assistant that Learns and Organizes)
- Virtual reality breathes second life into language teaching - Project in Mexico for teaching English to engineering students

Speech analysis tools

- Anvil - Video annotation tool
- Audacity - Spectrogram visualization and audio recording and editing
- CeedVocal - Speech recognition for the iPhone
- COLEA - Speech analysis freeware
- Computerized Speech Lab (CSL) - From KayPentax
- CSLU Toolkit - Multiple tools for speech analysis, recognition, generation, and display

- Janus - Speech translation system
- Julius - Open-source large vocabulary continuous speech decoder
- Let's Go - Info on the mobile speech analysis project from CMU
- Open Mind Speech - Open-source crowd-sourcing project for speech analysis project
- PocketSphinx - Sphinx for handhelds
- Phonology Assistant - Tool using IPA (international phonetic alphabet) characters to index and display data
- Praat - Open source speech analysis program (multi-OS)
- SFS/RTSPECT Version 2.4 - Windows tool for real-time waveforms & spectra
- Speech Analyzer - From SIL International
- TalkBank CMU - Speech technology project
- The CMU Sphinx Group - Open source speech recognition engines
- Video Phonetics Database and Program - From KayPentax
- WaveSurfer - Open source tool for sound visualization and manipulation
- WinCECIL - Tool for viewing speech recordings, automatic pitch contours, and spectrograms
- WinPitch LTL - Voice processing software

Language learning software and projects

- BetterAcent Tutor - Speech analysis program for ESL
- CAMMIA (A Conversational Agent for Multilingual Mobile Information Access) - from Language Technologies Institute, CMU
- CandleTalk - Conversation tool for English
- Fluency - Automatic foreign language pronunciation training (CMU)
- Intelligent dialog overcomes speech technology limitations: The SENECa example
- MyET-MyCT 3 - Uses speech analysis software for tutoring English or Chinese
- Review of Tell Me More Chinese - From Calico Review
- Sakhr Software - Mobile voice applications
- SpeakGoodChinese - Article
- SpeakGoodChinese - Chinese pronunciation program
- SPICE (Speech Processing Interactive Creation and Evaluation) - Speech processing models
- Tactical Language and Culture - US Army language learning software
- TransTac - Spoken language communication and translation system for tactical use
- Zengo Sayu - An immersive educational environment for learning Japanese