UNIVERSAL SPEECH TOOLS: THE CSLU TOOLKIT

Stephen Sutton[1,3]*, Ronald Cole, Jacques de Villiers, Johan Schalkwyk*[3]*, Pieter Vermeulen*[1,3]*, Mike Macon, Yonghong Yan, Ed Kaiser, Brian Rundle, Khaldoun Shobaki, Paul Hosom, Alex Kain, Johan Wouters, Dominic Massaro*[2]*, Michael Cohen*[2]

[1]Center for Spoken Language Understanding, Oregon Graduate Institute

[2]Perceptual Science Laboratory, University of California, Santa Cruz

[3]Fluent Speech Technologies, Portland , Oregon

ABSTRACT

A set of freely available, universal speech tools is needed to accelerate progress in the speech technology. The CSLU Toolkit represents an effort to make the core technology and fundamental infrastructure accessible, affordable and easy to use. The CSLU Toolkit has been under development for five years. This paper describes recent improvements, additions and uses of the CSLU Toolkit.

1. INTRODUCTION

Since 1993, the Center for Spoken Language Understanding (CSLU) has focused on incorporating state-of-the-art spoken-language technology into a portable, comprehensive and easy-to-use software environment. The result of these efforts is the CSLU Toolkit. The toolkit integrates learning materials, authoring tools and core technologies such as speech recognition, text-to-speech synthesis, facial animation and speech reading. The toolkit is designed to support basic research, development and education activities related to spoken language systems and human-computer interfaces.

What are our motives and goals for creating and distributing such a toolkit? We are convinced that the best way to accelerate progress in the field of spoken language technology is to get as many people as possible interested and involved, by making it accessible, affordable and easy to use (Sutton et al., 1996; Cole et al., 1998). This is precisely the role that the CSLU Toolkit fulfills. We have developed learning materials, core technologies, infrastructure, and software tools that are truly universal in that they benefit a range of users, from novice to expert, and are applicable to a wide range of tasks.

This paper provides a more detailed overview of the CSLU Toolkit, reports on its latest features and enhancements, and outlines our future plans.

2. TOOLKIT OVERVIEW

The toolkit provides a modular, open architecture supporting distributed, cross-platform, client/server-based networking. It includes interfaces for standard telephony and audio devices, and software interfaces for speech recognition, text-to-speech and animation components. This flexible environment makes it possible to easily integrate new components and to develop scalable, portable speech-related applications.

The major toolkit components are outlined below:

Speech recognition: The toolkit supports several approaches to speech recognition including artificial neural network (ANN) classifiers, hidden Markov models (HMM) and segmental systems. It comes complete with a vocabulary-independent speech recognition engine, plus several vocabulary-specific recognizers (e.g., alpha-digits). In addition, it includes all the necessary tutorials and tools for training new ANN and HMM recognizers.

Speech synthesis: The toolkit integrates the Festival text-to-speech synthesis system, developed at the University of Edinburgh (Black & Taylor, 1997). CSLU has developed a waveform-synthesis “plug-in” component (Macon et al., 1997) and six voices, including male and female versions of American English and Mexican Spanish. Festival provides a complete environment for learning, researching and developing synthetic speech, including modules for normalizing text (e.g., dealing with abbreviations), transforming text into a sequence of phonetic segments with appropriate durations, assigning prosodic contours (e.g., pitch, amplitude) to utterances, and generating speech using either diphone or unit-selection concatenative synthesis.

Facial animation: The toolkit features Baldi, an animated 3D talking head developed at the University of California, Santa Cruz. Baldi, driven by the speech recognition and synthesis components, is capable of automatically synchronizing natural or synthetic speech with realistic lip, tongue, mouth and facial movements. Baldi’s capabilities have recently been extended to provide powerful tools for language training. The face can be made transparent revealing the movements of the teeth and tongue while producing speech. The orientation of the face can be changed so it can be viewed from different perspectives while speaking. Also, the basic emotions of surprise, happiness, anger, sadness, disgust, and fear can be communicated through facial expressions.

Authoring tools: The toolkit includes the Rapid Application Developer (RAD), which makes it possible to quickly design a speech application using a simple drag-and-drop interface. RAD seamlessly integrates the core technologies with other useful features such as word-spotting, barge-in, dialogue repair, telephone and microphone interfaces, and open-microphone capability. This software makes it possible for people with little or no knowledge of speech technology to develop speech interfaces and applications in a matter of minutes.

Waveform analysis tools: The toolkit provides a complete set of tools for recording, representing, displaying and manipulating speech. Signal representations such as spectrograms, pitch contours and formant tracks can be displayed and manipulated in separate windows. The display tools allow recognition results, such as phonetic or word decoding, to be displayed and time-aligned with recognized utterances. Three-dimensional arrays can also be aligned to utterances, showing, for example, the output categories of a neural network phonetic classifier.

Programming environment: The toolkit comes with complete programming environments for both C and Tcl, which incorporate a collection of software libraries and a set of API's (Schalkwyk et al., 1997). These libraries serve as basic building blocks for toolkit programming. They are portable across platforms and provide the speech, language, networking, input, output, and data transport capabilities of the toolkit. Natural language processing modules, developed in Prolog, interface with the toolkit through sockets.

3. TOOLKIT IMPROVEMENTS

In this section, we describe recent and near-term improvements to the Toolkit's components and core technologies, which will be available in upcoming releases of the toolkit.

3.1 Rapid Application Developer

New capabilities have been added to the Rapid Application Developer that expand its scope and make it even easier to build real-world spoken language systems, including:

- ***Barge-In***: Host-based barge-in solution allows users to interrupt the computer when it is speaking and still be recognized. It is intended mainly for telephony applications using Dialogic D/21 analog series hardware, although an early implementation exists for most full-duplex sound cards.

- ***Open Microphone***: An open-microphone feature lets the computer continually listen for a specified keyword to be spoken before triggering a particular dialogue event.

- ***Touch-tone Input***: Full-feature touch-tone (DTMF) allows the integrated use of telephone keypad input along with speech recognition. This includes a "type-ahead" capability allowing expert users to anticipate future input requests.

- ***Wizard-of-Oz:*** A Wizard of Oz mode allows for remote monitoring and control of the dialogue flow in speech applications from a networked computer. This enables, for instance, a teacher to listen to a child's speech production in real time, override the speech recognizer and provide instant human feedback.

- ***Multilingual***: A language selection feature allows the creation of spoken dialogue systems in different languages, including English, Spanish and German.

- ***Natural Speech and Lip Syncing***: Integrated support for natural speech output, including recording tools and automatic generation and synchronization of facial animation with the recorded speech output.

- ***New Dialogue Objects***: Several new dialogue objects have been added including: (a) a media object offering support for displaying images, playing sounds and presenting text; (b) a list-builder object that provides a convenient way of specifying and randomizing the presentation of data; and (c) a login object that requires users to register when using an application. This object is useful for maintaining profiles of individuals and performance data.

3.2 Speech Recognition

The toolkit's recognition capabilities have been improved in several ways. The digit recognizer has also been improved, with reduction in word level error rates of about 50% (Cosi et al., 1998; Hosom et al. 1998). Additionally, the word level error rates of the toolkit's general-purpose (speaker-independent and vocabulary-independent) recognizer have been reduced by about 25%. Furthermore, these advances have led to the development of general-purpose recognizers for both Mexican Spanish and German.

Work at the Universidad de las Americas, in Puebla Mexico has produced Mexican Spanish recognizers for continuous digits and spelled words (Serridge et al., 1998), which is available in a Spanish version of the toolkit.

Our experiences using the CSLU toolkit for learning and language training with profoundly deaf children (Cole et al., 1997) have motivated the development of recognizers trained on children's speech. To improve recognition of children's speech, we collected speech data from over 1100 children in a local school district. There were approximately 100 children recorded from each grade level between kindergarten and 10th grade. The utterances included a set of phonetically balanced words and phrases that provide good coverage of all phonetic contexts. Preliminary work on the children's' speech recognizers has yielded encouraging results, with word error rate reductions greater than 70%, when comparing grade-specific recognizers (i.e., trained and tested on different children from the same grade level), against the general-purpose recognizer.

3.3 Natural Language Understanding

Natural language capabilities are under development and being integrated into the toolkit in the form of a semantic parser named PROFER ("Predictive RObust Finite-state parsER," pronounced *proffer*).

PROFER is modeled after Phoenix (Carnegie Mellon University's *caseframe* robust parser). It accepts the same grammar definitions, and yields the same output as Phoenix. However, internally it is a single Finite-State Machine rather than a Recursive Transition Network as is Phoenix. This means that for more complex grammars and longer input it can function up to several orders of magnitude faster than a chart-based system like Phoenix.

As a robust semantic parser, PROFER can be used to extract semantic patterns from the output of the toolkit's recognizers, cleaning up insertions due to misrecognition and tolerating the extra-grammaticalities of spontaneous speech. But its speed and the fact that it is sequentially predictive, like a Generalized Left-Right parser, make it an attractive candidate for tightly-coupling with a recognizer — in order to augment the standard *n*-gram language models with a layer of longer-range, finite-state predictions.

PROFER's run-time speed comes from pushing most of the bookkeeping which chart-based or Left-Right parsers perform during run-time into a compilation stage. This means slower compilation, but large grammars can be highly componentized, and putting components together can take only a fraction of the time for general top-to-bottom compilation.

In the future we will be incorporating the ability to accept standard feature-structured grammar definitions, and also adding a semantic carrying capability, that could allow some degree of compositional semantics to occur during the parse itself.

3.4 Speech Synthesis

We have continued to make refinements to the diphone synthesizer signal processor to improve smoothness of concatenations. In addition, several new features have been added, including:

- ***Mexican-Spanish Synthesizer***: A Mexican Spanish text-to-speech system has been developed through collaboration with researchers at Universidad de las Americas, in Puebla.
- ***German Synthesizer***: A German text-to-speech system was developed by visiting students from Germany under the direction of Mike Macon (CSLU), and Alan Black (University of Edinburgh).
- ***Voice Conversion***: A voice conversion utility that automatically generates a mapping from the diphone synthesizer voices to the recorded voice of any user (Kain & Macon, 1998). The capabilities of this voice conversion system will be extended in future releases to offer tools for new users to map to their own voice.
- ***Paralinguistics***: A facility for arbitrary embedding of paralinguistic sounds into the Festival output has been developed. This is done in such a way as not to interfere with control the talking face in later processing.
- ***Mark-up:*** A prototype of the text markup language in toolkit has been implemented. This allows insertion of embedded commands for inserting sounds, pauses, speech rate control, etc. A prototype implementation of a "syllabification" mode has been developed for pronunciation practice.
- ***Prosodics***: A prototype "copy synthesis" tool allows transplanting of natural prosodic contours onto Festival outputs.
- ***Transformation***: A speech modification tool allows rate and pitch transformations of natural speech.

3.5 Facial Animation

Facial animation is fully integrated into the toolkit. Baldi, an animated conversational agent, provides accurate visual speech, aligned with either synthetic or recorded auditory speech. During the past year, researchers at the University of California, Santa Cruz have developed a more accurate tongue model for language training, which will be introduced into classrooms at the Tucker Maxon Oral School in September 1998. The goal of this work is to produce an anatomically valid and pedagogically useful display of visible articulators that can be used in language training. This work uses both static and dynamic observations of 3D ultrasound data and electropalatography to train the parameters of the tongue model, using new models of the palate and teeth, shown in Figure 1. In addition, a high-speed algorithm has been developed for detection and correction of collisions, to prevent the tongue from protruding through the palate and teeth.

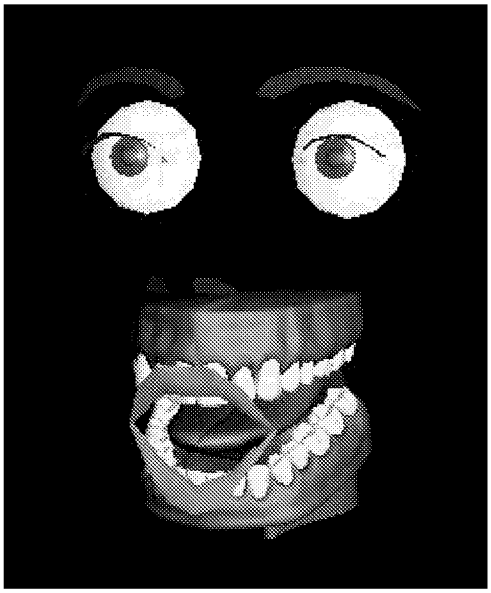

Figure 1: Baldi features accurate, visible articulators.

4. TOOLKIT USE TODAY

The CSLU Toolkit is in daily use in research labs, classrooms and industry. Some of the main toolkit activities are highlighted here:

Research: Research in speech recognition, speaker recognition, natural language processing, speech synthesis, language training and interactive media systems is ongoing in research labs worldwide using the CSLU Toolkit. More than 500 sites have downloaded the research (beta) version in the last five months.

System Development: Several companies in CSLU's industrial consortium are using the toolkit on a daily basis to develop and evaluate application prototypes, and to investigate and improve core technologies.

Education: Toolkit short courses on building spoken dialogue systems have been given to middle school students, high school students, graduate students and professionals (Sutton et al., 1997). Based on the success of these courses, the toolkit is being installed in 10,000 computers in K through 12 classrooms in the state of Oregon. We will be working with teachers and students in the near future to integrate interactive systems into classroom activities.

Language Training With Profoundly Deaf Children Since September, 1997, CSLU has been working with teachers and deaf students, ages 8 to 12, at the Tucker Maxon Oral School, to adapt the toolkit to language training and classroom learning activities (Cole et al., 1997). This project serves as daily testbed for the toolkit technologies and authoring tools, and many of toolkit improvements reported here are motivated by our experiences with the students and teachers.

5. AVAILABILTY

The Toolkit is currently runs on Windows 95/NT. A UNIX version is planned for later in the year. The toolkit can be downloaded free of charge for non-commercial use from the CSLU web site (http://www.cse.ogi.edu/cslu). A commercial license is available from Fluent Speech Technologies (http://www.fluent-speech.com).

6. REFERENCES

Black, A. and Taylor, P. "Festival Speech Synthesis System: system documentation," Human Communication Research Centre Technical Report HCRC/TR-83, 1997.

Cole, R., Carmell, T., Connors, P., Macon, M., Wouters, J., de Villiers, J., Tarachow, A., Massaro, D., Cohen, M., Beskow, J., Yang, J., Meier, U., Waibel, A., Stone, P., Fortier, G., Davis, A., Soland, C., "Intelligent Animated Agents for Interactive Language Training" Proceedings of ESCA-StiLL, Marholmen, Sweden, May 1997.

Cole, R., Sutton, S., Yan, Y., Vermeulen, P., Fanty, F. "Accessible Technology for Interactive Systems: A New Approach to Spoken Language Research," Proceedings of International Conference on Acoustics, Speech and Signal Processing, Seattle, WA, May 1998.

Cosi, P., Hosom, J.P., Shalkwyk, J., Sutton, S., and Cole, R. A., "Connected Digit Recognition Experiments with the OGI Toolkit's Neural Network and HMM-Based Recognizers," 4th IEEE Workshop on Interactive Voice Technology for Telecommunications Applications, Turin, Italy, September 29-30, 1998.

Hosom, J.P., Cole, R. A. "Evaluation and Integration of Neural-Network Training Techniques for Continuous Digit Recognition", Proceedings of 1998 International Conference on Spoken Language Processing, Sydney, Nov.-Dec. 1998.

Kain A., and Macon, M., "Spectral voice conversion for text-to-speech synthesis," Proceedings of the International Conference on Acoustics, Speech, and Signal Processing, vol. 1, pp. 285-288, May 1998.

Macon, M., Cronk, A., Wouters, J., and Kain, A. "OGIresLPC: Diphone synthesizer using residual-excited linear prediction,'" Tech. Rep. CSE-97-007, Department of Computer Science, Oregon Graduate Institute of Science and Technology, Portland, OR, September 1997.

Schalkwyk, J., de Villiers, J., van Vuuren, Sarel and Vermeulen, P. "CSLUsh: An Extendible Research Environment", Proceedings of Eurospeech'97.

Serridge, B., Cole, R., Barbosa, A., Munive, N., and Vargas, A., "Creating a Mexican Spanish version of the CSLU toolkit" To be presented at International Conference of Spoken Language Processing 1998, Sydney, Australia.

Cole, R., Sutton, S., Yan, Y., Vermeulen, P., and Fanty, M., "Accessible Technology for Interactive Systems: A new approach to spoken language research," ICASSP 1998.

Sutton, S., Kaiser, E., Cronk, A., and Cole, R. "Bringing Spoken Language Systems to the Classroom", Proceedings of Eurospeech'97.

Sutton, S., Novick, D., Cole, R., and Fanty, M., "Building 10,000 spoken-dialogue systems." Proceedings of the International Conference on Spoken Language Processing, Philadelphia, PA, 1996.

7. ACKNOWLEDGMENTS

This research was made possible with support from:

NSF Challenge grant (NSF CDA-9726363 Creating Conversational Agents for Language), NSF SGER grant (NSF ECS-9726645 Bringing Spoken Language Systems to the Classroom), DARPA/ONR grant (ONR N00014-94-1-1154 Rapid Prototyping of Spoken Language Systems), NSF/CONACyT grant (NSF IRI-9614217 Advancing Human Language Technology In Mexico), Language Resources grant (NSF IRI-9529006 Human Language Resources/Multilanguage Systems), Fluent Speech Technologies funding, Members of the Center for Spoken Language Understanding industrial consortium.