Demonstrations of Dialogue Design Tools in the CSLU Toolkit

Ron Cole, Jacques de Villiers, Kal Shobaki
Center for Spoken Language Understanding
http://cslu.colorado.edu http://cse.ogi.edu/cslu

Dominic W. Massaro, Jonas Beskow, Michael M. Cohen
Perceptual Science Laboratory, University of California, Santa Cruz
http://mambo.ucsc.edu

ABSTRACT

The CSLU Toolkit and accompanying tutorials are designed to provide a platform for researching and developing language technologies and systems, and to engage naïve users in using and experimenting with interactive language systems. We provide a set of demonstrations in this special session that illustrate capabilities. They include rapid prototyping of a spoken dialogue system that integrates an animated talking face, speech recognition and text-to-speech synthesis, and a variety of applications created by practitioners using the rapid application developer.

TOOLKIT MODULES

The CSLU Toolkit consists of a number of component systems or modules for researching, developing, and using spoken language systems [1]. More detailed descriptions of the modules described in this section can be found on the CSLU Toolkit Web site [2].

Rapid Application Developer. RAD is the Toolkit's graphical authoring environment. This software makes it possible for people with little or no knowledge of speech technology to learn to develop speech interfaces and applications. RAD seamlessly integrates the core technologies of facial animation, speech recognition and understanding, and speech synthesis with other useful features such as word-spotting, barge-in, dialogue repair, telephone and microphone interfaces, and open-microphone capability. RAD's "drag and drop" interface is easy to use and learn to use—naive users can build simple dialogues (in minutes) for conversing with an animated talking face.

Facial animation. Baldi is an animated three dimensional talking head developed at the Perceptual Science Laboratory at the University of California, Santa Cruz [3]. Baldi's synthesis program controls a wireframe model, with a control strategy for coarticulation, controls for paralinguistic information and affect in the face, text-to-speech synthesis, and synchronization of auditory and visual speech. Basic emotions of surprise, happiness, anger, sadness, disgust, and fear can be communicated through facial expressions.

Baldi produces accurate visual speech that can be understood by skilled speech readers. Recently, a more complex and accurate tongue model (consistent with electropalatography and imaging data) containing a hard palate and three-dimensional teeth have been added [4]. The face can be made transparent during speech, revealing the movements of the teeth and tongue, and the orientation of the face can also be changed while speaking so it can be viewed from different perspectives, such as from the back of the head. These features offer unique capabilities for language instruction.

Speech Recognition. The Toolkit comes complete with "general purpose" speaker- and vocabulary-independent speech recognition engines for both word spotting and continuous speech recognition applications. Recognizers have been trained for telephone speech and microphone speech for adults and for microphone speech for children. In addition, vocabulary-specific recognizers are included for digit recognition and alpha-digits strings [5]. Mexican Spanish recognizers are included for vocabulary-independent word spotting applications and continuous digit recognition [6].

Natural Language Understanding. PROFER (Predictive RObust Finite-state parsER,) is a semantic parser modeled after Carnegie Mellon University's Phoenix system [7]. As a robust semantic parser, PROFER can be used to extract semantic patterns from the output of the Toolkit's recognizers while tolerating many of the features of spontaneous speech, such as false starts, filled pauses and ungrammatical constructions. A step-by-step tutorial has been developed for PROFER. In this tutorial, students learn to develop a conversational system using natural continuous speech (or typed input) for retrieving movie times and locations from a Web site.

Festival Speech Synthesis System. The Toolkit integrates the Festival text-to-speech synthesis system, developed at the University of Edinburgh [8]. 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,

ESCA Workshop and Research workshop. Interactive Dialogue in Multi-Modal Systems. Kloster Irsee, Germany. June, 1999.

assigning prosodic contours (e.g., pitch, amplitude) to utterances, and generating speech using either diphone or unit-selection concatenative synthesis.

SpeechView is the Toolkit's interactive analysis and display tool. It allows users to create new waveform and label files, display data that are associated with a waveform (such as spectrograms or pitch contours), and modify existing waveforms and label files. It is used for conducting laboratory exercises in an interactive spectrogram reading class [9], and provides an excellent tool for education in speech signal processing.

BaldiSync enables users to synchronize any speech waveform with Baldi's facial movements by supplying the waveform and the text of the utterance as a sequence of typed words.

Perceptual Science Laboratory (PSL) provides tools to support research in perception and cognition [10]. PSL provides a user-friendly research environment for designing and conducting multimodal experiments in speech perception, psycholinguistics, and memory. Since PSL tools can be used to teach students to conduct research using the scientific method, it offers them new ways to conceptualize problems and investigate the world.

Speech Performance Assessment and Measurement (SPAM) is a database program designed to capture and analyze all behaviors produced by the user and system during an interactive dialogue.. In conjunction with PSL, SPAM provides an invaluable tool for designing and evaluating user interfaces for conversational interactions. [11]

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 [12]. These libraries serve as basic building blocks for Toolkit programming.

ACKNOWLEDGEMENTS

This work was supported in part by NSF CHALLENGE grant CDA-9726363, ONR/DARPA grant N00014-94-1-1154, NSF CARE grant EIA-9996075, Public Health Service grant PHS R01 DC00236), National Science Foundation grant 23818, Intel Corporation, and the University of California Digital Media Innovation Program. The views expressed in this paper do not necessarily represent the views of NSF, ONR, DARPA or PHS. We gratefully acknowledge the dedicated efforts of the teachers and students at the Tucker Maxon Oral School.

REFERENCES

1. Sutton, S., Cole, R., et al. "Universal Speech Tools: the CSLU Toolkit," In Proceedings of the International Conference on Spoken Language Processing (ICSLP), pages 3221-3224, Sydney, Australia, November 1998.
2. http://cslu.cse.ogi.edu/toolkit/
3. D. W. Massaro. Perceiving Talking Faces: From Speech Perception to a Behavioral Principle, MIT Press, Cambridge, 1998.
4. Cohen, M. M., Beskow, J., & Massaro, D.W. (1998). Recent developments in facial animation: An inside view. Proceedings of the International Conference on Auditory-Visual Speech Processing—AVSP'98 (pp. 201-206). Terrigal, Australia.
5. P. Cosi, J.P. Hosom, J. Schalkwyk, S. Sutton, and R. A. Cole. Connected digit recognition experiments with the ogi toolkit's neural network and hmm-based recognizers. In Proceedings, 4th IEEE Workshop on Interactive Voice Technology for Telecommunications Applications (IVTTA-ETWR98), Turin, Italy, September 1998.
6. B. Serridge, B., Cole, R., Barbosa, A., Vargas, A. and Munive, N. Creating a Mexican Spanish Version of the CSLU Toolkit" Proceedings of the International Conference in Spoken Language Processing, Sydney, Australia, November 1998.
7. Kaiser, E.C., Johnston, M., and Heeman, P. A. Profer: Predictive, Robust Finite-State Parsing for Spoken Language. In Proceedings of ICASSP, Phoenix, Arizona, March 1999.
8. A. Black, and P. Taylor. Festival Speech Synthesis System: System documentation (1.1.1), Human Communication Research Centre Technical Report HCRC/TR-83, Edinburgh, 1997.
9. T. Carmell, J.P. Hosom, and R. Cole. A computer-based course in spectrogram reading. In Proceedings of ESCA/SOCRATES Workshop on Method and Tool Innovations for Speech Science Education, London, UK, Apr 1999.
10. http://mambo.ucsc.edu/psl/tools
11. R.Cole, D. W. Massaro, J. de Villiers, B. Rundle, K. Shobaki, J. Wouters, M. Cohen, J. Beskow, P. Stone, P. Connors, A. Tarachow, and D. Solcher. New tools for interactive speech and language training: Using animated conversational agents in the classrooms of profoundly deaf children. In Proceedings of ESCA/SOCRATES Workshop on Method and Tool Innovations for Speech Science Education, London, UK, Apr 1999.
12. J. Wouters, B. Rundle and M. Macon. Authoring Tools for using the SABLE markup standard. Eurospeech99, Budapest, Hungary, September, 1999.