UNCLASSIFIED

Defense Technical Information Center
Compilation Part Notice

ADP010384

TITLE: A Platform for Multilingual Research in Spoken Dialogue Systems

DISTRIBUTION: Approved for public release, distribution unlimited

This paper is part of the following report:

TITLE: Multi-Lingual Interoperability in Speech Technology [l'Interoperabilite multilinguistique dans la technologie de la parole]

To order the complete compilation report, use: ADA387529

The component part is provided here to allow users access to individually authored sections of proceedings, annals, symposia, ect. However, the component should be considered within the context of the overall compilation report and not as a stand-alone technical report.

The following component part numbers comprise the compilation report:
ADP010378 thru ADP010397

UNCLASSIFIED

A PLATFORM FOR MULTILINGUAL RESEARCH IN SPOKEN DIALOGUE SYSTEMS

Ronald A. Cole, Ben Serridge[§], John-Paul Hosom[‡], Andrew Cronk[‡], and Ed Kaiser[‡]*

*Center for Spoken Language Understanding; University of Colorado – Boulder; Boulder, CO, 80309, USA
[§]Universidad de las Americas; 72820 Santa Catarina Martir; Puebla, Mexico
[‡]Center for Spoken Language Understanding (CSLU), Oregon Graduate Institute; Beaverton, OR, 97006, USA

*cole@cslu.colorado.edu; [§]serridge@alum.mit.edu; [‡]{hosom,cronk,kaiser}@cse.ogi.edu

ABSTRACT

Multilingual speech technology research would be greatly facilitated by an integrated and comprehensive set of software tools that enable research and development of core language technologies and interactive language systems in any language. Such a multilingual platform has been one of our goals in developing the CSLU Toolkit. The Toolkit is composed of components that are essentially language-independent, and support research and development of recognition, understanding, text-to-speech synthesis, facial animation, and spoken dialogue systems. Portions of the Toolkit have already been ported to Italian, German, and Vietnamese. In addition, a complete Mexican-Spanish version of the Toolkit has been created, and is in daily use at the Universidad de las Americas in Puebla (UDLA). In this paper we outline some of the issues involved in porting the Toolkit to a new language, and describe why the Toolkit is well suited to multilingual adaptation.

1. INTRODUCTION

Speech communication occurs within social and cultural contexts, and is influenced by the perceptions, beliefs, attitudes, and backgrounds of the speakers. Research in spoken language systems requires participation by native speakers who understand not only the language, but also the subtle social conventions and cultural factors that enable natural communication. As a result, the best way to understand and model linguistically-related differences, create natural spoken-dialogue systems, and achieve acceptable machine translation between languages is through multinational collaborative research.

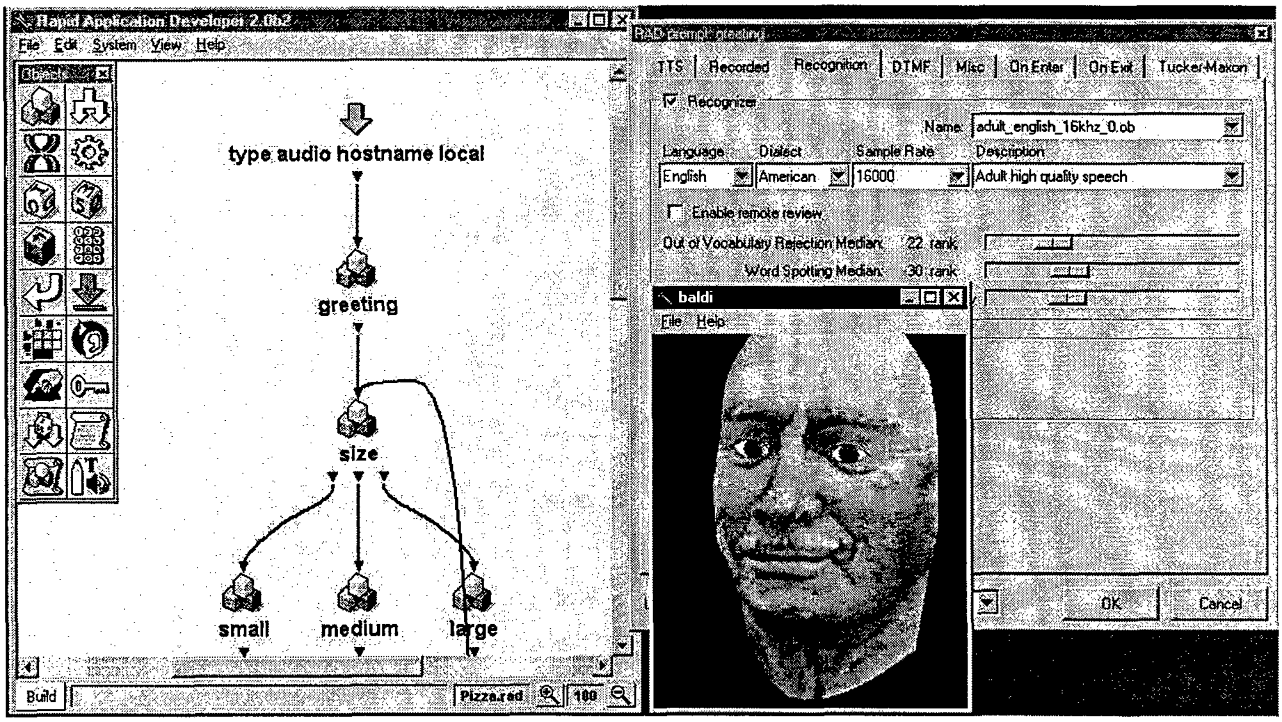
One of the main factors preventing more intensive multinational research is the "knowledge engineering bottleneck" — the massive costs associated with developing and deploying spoken language systems for each additional language and new application. These costs present formidable barriers to progress in human language technology.

Currently, spoken language systems development and research are limited to a few specialized laboratories because of the infrastructure and expertise required. The systems that are created in these laboratories are generally not portable; each new language and application requires collection of speech data, application-specific system development, and human engineering to create a graceful user interface. Data collection for training recognizers and for building language and dialogue models is costly and often must be done via "Wizard-of-Oz" simulation, with humans attempting to mimic the performance of a spoken language system. Such experiments are notoriously expensive and time-consuming. Consequently, all but the most fortunate students and researchers are denied the opportunity to explore this interesting frontier of human-computer interaction.

To break the knowledge engineering bottleneck and realize the potential of international, multilingual spoken-language research, it is necessary to develop available, usable, and powerful tools and corpora to engage and enable a generation of students to study, use, research, and develop language technologies and systems. These tools must be readily applicable to all languages of interest, so that there can be a common research framework. In general, most tools available today were designed by experts for use by other experts, and are not sufficiently tutorial to be used to train new researchers in undergraduate and graduate programs.

Our research efforts are aimed at overcoming these barriers, and the platform we use to integrate our advances is called the CSLU Toolkit. The Toolkit is freely available for research use from the CSLU Web site, and integrates speech recognition, natural language understanding, text-to-speech synthesis, facial animation, dialogue modeling, and spoken-language interface design in one package. The Toolkit is essentially language-independent; we have successfully ported the Toolkit to Mexican Spanish, and we are now developing a Brazilian Portuguese version of the Toolkit with colleagues at the Universidade Federal do Rio Grande do Sul, in Porto Allegre, Brazil.

The remainder of this article describes the CSLU Toolkit and its use a platform for multilingual research We hope that release of the Toolkit will remove entry barriers to research and education in human language technology, and enable researchers and students around the world to participate in creating the multilingual spoken-language systems of the future.



☑ Recognizer

Name: adult_english_16khz_0.ob

Language	Dialect	Sample Rate	Description
English	American	16000	Adult high quality speech

□ Enable remote review

Out of Vocabulary Rejection Median: 22 rank

Word Spotting Median: 30 rank

Figure 1. Screen shot of the Rapid Application Developer (RAD), the animated face Baldi (with texture map), and a parameter window for setting properties of the recognizer.

2. THE CSLU TOOLKIT

The CSLU Toolkit is a comprehensive set of tools and technologies for learning about, researching and developing interactive language systems and their underlying technologies [1, 2, 3, 4]. It is available, free of charge, from the CSLU OGI Web site [5], along with CSLU's multi-language phonetically hand-labeled speech corpora. The Toolkit supports real-time interactive dialogues on standard off-the-shelf PC platforms running Windows (Solaris and Linux will be available soon). It provides a modular, open architecture supporting distributed, cross-platform, client/server-based networking. This flexible environment makes it possible to easily integrate new components and to develop scalable, portable speech-related applications.

The components of the Toolkit include both neural-network and HMM-based speech recognition systems, a natural-language semantic parser called PROFER, the Festival text-to-speech system, an anatomically accurate talking face called Baldi, and software for recording, displaying, labeling, and manipulating speech. The Toolkit also includes a GUI-based application developer called RAD and the documentation required to train HMM and neural-network based recognizers.

The tools are designed to enable inexperienced users to rapidly design, test and deploy spoken language systems. In addition to the pre-existing components, users can write their own C-level or script code for integration into the Toolkit. The recognition, synthesis, and natural language systems (and their tutorials) also support basic research and system development. Research advances can then be evaluated in real-world applications designed with the Toolkit's dialogue design tools.

Because the Toolkit is portable, runs on affordable off-the-shelf computing platforms, and provides both the knowledge (tutorials) and resources needed to conduct research, it removes some of the main entry barriers that currently prevent universities and research laboratories from establishing new programs in human language technology.

3. A PLATFORM FOR RESEARCH AND DEVELOPMENT OF MULTILINGUAL SPOKEN LANGUAGE SYSTEMS

The CSLU Toolkit is a proven platform enabling international collaboration in multilingual research and system development. Between 1996 and 1998, a joint NSF/CONACyT program supported collaboration between OGI and UDLA, the Universidad de las Americas, Puebla. The collaboration aimed to establish UDLA as a center of excellence in speech technology in Mexico, capable of educating students, of developing state-of-the-art Mexican Spanish spoken language systems, and of supporting research and education in language technologies throughout Mexico. These goals were accomplished.

As a result of our collaboration, the Tlatoa speech group has developed a complete Mexican-Spanish version of the Toolkit by collecting and transcribing corpora, implementing Mexican-Spanish recognition and text-to-speech systems, and converting Toolkit documentation to Spanish. UDLA now develops and distributes Mexican Spanish language resources, trains undergraduate and graduate students in language technology [6], publishes articles on speech research [7], and transfers knowledge and technology to other Mexican universities [8]. The collaboration has also produced industrial investment — a U.S. speech technology company has hired two UDLA students, has established a subsidiary in Puebla, and has made a substantial investment in the Tlatoa speech group (more than recovering the original investment from CONACyT).

The Toolkit has also been used successfully in European labs. Michael McTear has used the Toolkit to train undergraduate students at the University of Ulster to develop interactive language systems [9, 10]. Piero Cosi at the Istituto di Fonetica e Dialettologia Consiglio Nazionale delle Ricerche (Institute of Phonetics and Dialectology - National Research Council) has used the Toolkit to develop English and Italian speech recognition systems and compare Hidden Markov Model and neural network approaches [11, 12].

4. COMPONENTS OF THE TOOLKIT

In this section, the main components of the Toolkit are described. Issues in developing each component in a new language will be discussed, with examples from previous multilingual efforts where applicable.

4.1 Rapid Application Developer (RAD)

RAD is the Toolkit's high-level application developer. RAD's easy-to-use graphical authoring environment enables users to rapidly design and test spoken dialogue systems. It 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 enables users to design interactive dialogues by specifying prompts, recognition vocabularies, and actions. Prompts can be either recorded or typed in as text, in which case they are produced as speech using the Toolkit's text-to-speech system. Both recorded and synthesized prompts are produced automatically by Baldi, the animated talking face [13]. Words or phrases to be recognized at any dialogue state are simply typed in by the system builder. Arbitrary actions can be associated with recognized utterances, such as producing a new prompt, displaying an image or retrieving and displaying information from a Web site. RAD contains many useful objects for retrieving, organizing and presenting information. In addition, users can develop new objects using the Tcl/Tk programming language. By connecting RAD objects, dialogues of arbitrary complexity can be designed. A sample screen shot of a RAD dialogue using Baldi with texture mapping is shown in Figure 1.

RAD currently includes both English and Mexican-Spanish recognizers and TTS voices. Addition of new recognizers and voices is easily done by creating new containers for the relevant objects (including the dictionary, if applicable) and storing them in the appropriate directories. Once these steps have been accomplished, RAD functions in the target language.

4.2 Facial animation

Baldi can be programmed within RAD to produce synthetic or recorded speech with different emotions. The face can be made transparent during speech production, revealing the movements of the teeth and tongue, and the orientation of the face can be changed while speaking to view it from different perspectives. Recently, a more complex and accurate tongue (consistent with electropalatography and imaging data), a hard palate, and three-dimensional teeth have been incorporated in Baldi. These features offer unique capabilities for language instruction — features that cannot be easily controlled in real faces.

Baldi is totally language-independent, in that he is controlled entirely by phoneme-level input. The input to Baldi consists of Worldbet [14] phonetic symbols, which are ASCII representations of the IPA and can represent all phonemes available in that alphabet. (The Worldbet system is used throughout the Toolkit, giving multi-language implementations a consistent phonetic representation.)

4.3 Speech Recognition

The Toolkit includes (a) English and Spanish digit and alpha-digit recognizers for recognizing sequences of digits and/or letters; (b) general-purpose English and Spanish recognizers for recognizing arbitrary words or phrases specified as text; and (c) a medium vocabulary English speech recognition system (MVCSR) that supports training of acoustic and language models for real time recognition of continuous speech with vocabularies up to 5000 words. The Toolkit supports research and education using several approaches to computer speech recognition, including artificial neural network (ANN) classifiers, hidden Markov models (HMM), and segmental systems. The Toolkit also includes step-by-step tutorials for training and testing new ANN and HMM recognizers.

The methods for training speech recognizers in the Toolkit are essentially language independent, with the selection of phonetic symbols and training corpora the only language-dependent parts. Pitch information is not currently used in the default feature set, but for tonal languages such as Mandarin or Vietnamese, the default feature set can be easily modified to include such information.

The Toolkit appears to be gaining acceptance as a platform for recognition research. In addition to English and Mexican Spanish recognizers, we are aware of

Toolkit recognizers developed for digit recognition in Italian, Vietnamese, and Korean. Consistent results have been observed across languages; for digit recognition, recognizers trained on telephone-band speech have word-accuracy levels of about 97% to 98%, and recognizers trained on microphone-quality speech have word-accuracy levels of about 99% [12, 15].

4.4 Natural Language Understanding

People do not always speak grammatically, and they often make false starts, or correct themselves as they are speaking. To parse this kind of spontaneous input requires a robust parser — that is, a parser that when confronted with such ill-formed input doesn't break, but finds the best allowable partial parse. Robust parsers, like Carnegie Mellon's Phoenix parser [16], are based on semantic *case-frame* architectures. They allow *slots* within a particular case-frame to be filled in any order, and allow out-of-grammar words to be skipped over. Thus, partial parses can be returned as *frames* in which only some of the slots have been filled. Typically, semantic case-frame parsers are implemented as chart parsers, and accept a transcript from the speech recognizer as their input. This requires separate grammars for the recognizer and the semantic parser, and limits the possibility of feedback between the two.

We have developed a semantic case-frame parser that runs as a finite-state machine rather than as a chart parser [17]. We believe this makes it more amenable to being tightly integrated into a speech recognizer, in such a way that the recognizer and semantic parser can share grammars and provide immediate feedback to each other. This tight integration is the aim of our current research. However, our initial version of Profer (which stands for Predictive, RObust, Finite-state parsER) can be used as a standard robust parser in a second-pass system, accepting the transcript produced by a recognizer. For example, using a grammar that defines sequences of numbers (each of which is less than ten thousand, greater than ninety-nine, and contains the word "hundred"), inputs like the following string of three numbers, which is rife with false starts and on-line corrections, can be robustly parsed by Profer [18]:

Input:

first I've got twenty ahhh thirty yaaaaaa thirty ohh wait no twenty twenty nine hundred two errr three ahhh four and then two hundred ninety uhhhhh let me be sure here yaaaa ninety seven and last is five oh seven uhhh I mean six

Parse tree:

```
[fsType:number_type,
 hundred_fs:
   [decade:[twenty,nine],hundred,four],
 hundred_fs:
   [two,hundred,decade:[ninety,seven]],
 hundred_fs:
   [five,hundred,six]]
```

Profer is essentially a regular grammar parser. It allows the grammar writer to specify patterns in the input that should be "tagged" in the output parse tree as belonging to certain slots in a particular frame. The names of slots and frames are arbitrary — they can describe standard syntactic elements or task-specific semantic categories. Both tag-names and the patterns that define them are language independent. The grammar writer has free reign in this regard. Thus Profer is a language-independent tool, and has been used to define both English and Spanish grammars. A step-by-step tutorial has been developed for Profer to develop a conversational system for retrieving movie times and locations from a Web site.

4.5 Festival Speech Synthesis System

The Toolkit integrates the Festival text-to-speech synthesis system [19], a complete environment for learning, researching, developing, and using 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 (pitch and amplitude) to utterances, and generating speech using either diphone or unit-selection concatenative synthesis. In addition, a graphical user interface enables users to "mark up" a text string to control many features of the resulting synthesized speech (e.g., rate, pitch, and amplitude) and to insert pauses, filled pauses, coughs, and sneezes.

During the summers of 1997 and 1998, researchers in the Speech Synthesis Research Group at OGI developed Spanish and German voices for use in the CSLU Toolkit. Students from the University of the Americas Puebla (UDLA), the University of Stuttgart, and the University of Bonn collaborated in these efforts. More information on these projects is available at http://cslu.cse.ogi.edu/tts.

While details vary, the overall process of developing a new voice is consistent between languages. As Festival is a concatenation-based synthesizer, a speech corpus must be designed and collected that optimally covers the target linguistic space. For example, a sample target linguistic space might be the phonemes of a language. In practice, such simple speech units are not used because they do not capture the coarticulatory effects between phonemes. Both the Spanish and German voices developed at OGI use the diphone as the basic unit of concatenation.

A promising technique known as unit selection is the focus of much ongoing research at OGI and other speech labs. In unit selection, longer — possibly non-uniform — "chunks" of speech may be extracted from a large, continuous-speech corpus. The goal of unit selection is to reduce the number of concatenation points in an utterance and increase the number of coarticulatory events captured in the speech in order to improve naturalness.

The process of developing text-to-speech corpora for waveform synthesis of new voices in new languages requires a series of steps. A protocol or script must be designed that contains at least one instance of each speech unit. Often the protocol is comprised of nonsense words or word pairs from which the diphones may be extracted. As not all phoneme-to-phoneme transitions exist in a given language, the advice of a native speaker or a trained linguist is exceedingly useful in keeping the protocol to a manageable size.

Once the protocol is optimally designed and the speaker selected, recording may proceed. High-quality recording is vital to the successful deployment of a new voice. The recording studio should be as anechoic as possible and possess high-quality microphones. A laryngograph is used to measure the impedance across the glottis during the session. These data are used to determine pitch marks, which are needed for smoothing concatenation points and altering prosody. The bulk of time invested in the development of the voices was spent separating and labeling these data.

For any language, rules must be developed to transform text into a sequence of tokens. For instance, the English text, "Dr. Suess spent $2.01 on Lorax Dr." may be represented by the tokens "doctor suess spent two dollars and one cent on lorax drive". Festival allows these rules to be easily scripted in Scheme, a dialect of the LISP programming language. In addition, mechanisms for determining the pronunciation of a token must be prepared. Since Spanish is a very consistent language, a set of letter-to-sound rules suffices. However, as the English and German languages are not particularly consistent, a lexicon must be found or created.

Finally, modules for the prediction of phoneme duration and pitch must be devised. These can be as simple as averages or they may be trained from data. Festival provides a number of tools for training prosodic modules from data.

Once all the above steps are complete, the voice may be defined within Festival. While this may be a challenging task for the first time Festival developer, once achieved for a particular language, the file formats and configuration files for each additional language are quite similar and readily created. In fact, the German synthesizer was speaking "guten tag" after only one day's work. The remaining month was spent collecting and preparing the speech data.

4.6 SpeechView

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 at CSLU for research, corpus development activities, and forms the basis for an interactive spectrogram reading class [20]. SpeechView supports simultaneous recording and subsequent annotation of auditory and visual speech data, and was recently used to collect bimodal speech data from over 250 children. SpeechView is entirely language-independent.

4.7 Perceptual Science Laboratory (PSL)

PSL provides a user-friendly research environment for designing and conducting multimodal experiments in speech perception, psycholinguistics, and memory. It enables users to manipulate auditory and visual stimuli; design interactive protocols for multi-media data presentation and multi-modal data capture; transcribe and analyze subjects' responses; perform statistical analyses; and summarize and display results. We plan to use PSL in our research to evaluate auditory visual synthesis for new languages. PSL, like SpeechView, is language-independent.

4.8 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. 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.

5. CONCLUSION

The Toolkit has proven itself to be well suited for multilingual research in several areas. It is in use in over 300 laboratories worldwide, and has enabled research leading to over 200 publications.

In recognition, both English and Mexican-Spanish general-purpose recognizers have been created and are incorporated within the rapid application developer (RAD). Furthermore, the tutorial for training a digits recognizer has been used successfully by others in languages as diverse as Italian and Vietnamese. The semantic parsing tools in Profer are essentially language independent and are being used in both English and Mexican-Spanish applications.

In text-to-speech, we have developed Mexican-Spanish and German voices, the implementations of which were performed in one month, including the time required to collect and hand-label the diphone databases. In addition, we are currently refining a unit selection approach which is easily applicable to other languages and promises to improve naturalness.

Once the recognition and TTS components have been implemented in a given language, the graphical authoring tools enable rapid development of structured dialogue applications in that language. Finally, the components of the Toolkit can easily be interchanged, allowing quick substitution of an English recognizer with an Italian one, or German TTS with English. These factors all contribute to making the CSLU Toolkit powerful and easy to use in a multilingual environment.

6. ACKNOWLEDGEMENTS

This work was made possible by NSF Challenge award NSF/CDA-9726363, NSF CARE award UCB/NSF/153-0584, NSF grant NSF/IIS-9614217, ONR/DARPA grant ONR/N00014-94-1-1154, NSF Career award IIS-9875950, the support of Intel Corporation, and the CSLU Member companies.

REFERENCES

[1] S. Sutton, D. G. Novick, R. A. Cole, and M. Fanty. Building 10,000 spoken-dialogue systems. In Proceedings of the International Conference on Spoken Language Processing (ICSLP), Philadelphia, PA, October 1996.

[2] S. Sutton, R. Cole, J. de Villiers, J Schalkwyk, P. Vermeulen, M Macon, Y Yan, E. Kaiser, B. Rundle, K Shobaki, P. Hosom, A. Kain, J Wouters, M Massaro, and M Cohen. "Universal Speech Tools: the CSLU Toolkit." In Proceedings of the International Conference on Spoken Language Processing (ICSLP), pages 3221-3224, Sydney, Australia, November 1998.

[3] R. Cole, S. Sutton, Y. Yan, P. Vermeulen, and M. Fanty. Accessible technology for interactive systems: A new approach to spoken language research. In Proceedings of the International Conference on Acoustics, Speech and Signal Processing, Seattle, Washington, May 1998.

[4] R. Cole. "Tools for research and education in speech science." In Proceedings of the International Conference of Phonetic Sciences, San Francisco, CA, Aug 1999.

[5] http://cslu.cse.ogi.edu/toolkit

[6] A. Barbosa, "A new Mexican Spanish voice for the Festival text to speech system." Masters Thesis, May 1997, UDLA-Puebla.

[7] B. Serridge, R. Cole, A. Barbosa, A. Vargas, and N. Munive. "Creating a Mexican Spanish Version of the CSLU Toolkit" Proceedings of the International Conference in Spoken Language Processing, Sydney, Australia, November 1998.

[8] B. Serridge, "An Undergraduate Course on Speech Recognition Based on the CSLU Toolkit." Proceedings of the International Conference in Spoken Language Processing, Sydney, Australia, November 1998.

[9] M. McTear, "Modelling spoken dialogues with state transition diagrams: experiences with the CSLU toolkit." Proc 5th International Conference on Spoken Language Processing, Dec 1998, Sydney, Australia, 1223-1226.

[10] M. McTear, "Using the CSLU toolkit for practicals in spoken dialogue technology." M.A.T.I.S.S.E. workshop on Method and Tool Innovations for Speech Science Education, April 1999, University College London, April 16 – 17.

[11] 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", 4th IEEE Workshop on Interactive Voice Technology for Telecommunications Applications (IVTTA-ETWR98), Turin, Sep. 1998, pp. 135-140.

[12] P. Cosi and J.P. Hosom "HMM/Neural Network-Based System for Italian Continuous Digit Recognition", In Proceedings of the International Conference of Phonetic Sciences (ICPhS), San Francisco, CA, August 1999.

[13] D. Massaro, *Perceiving Talking Faces: From Speech Perception to a Behavioral Principle*, MIT Press, Cambridge, 1998.

[14] J. Hieronymus, *ASCII phonetic symbols for the world's languages: Worldbet.* AT&T Bell Laboratories, Technical Memo, 1994.

[15] J.P. Hosom, R.A. Cole, P. Cosi "Improvements in Neural-Network Training and Search Techniques for Continuous Digit Recognition", Australian Journal of Intelligent Information Processing Systems, accepted for publication.

[16] W.H. Ward, "The Phoenix System: Understanding Spontaneous Speech", IEEE ICASSP, April 1991.

[17] E. Kaiser, M. Johnston, and P. Heeman, "Profer: Predictive, Robust Finite-State Parsing for Spoken Language." In Proceedings of ICASSP, Phoenix, Arizona, March 1999.

[18] E. Kaiser, "Robust, Finite-State Parsing for Spoken Language Understanding", in the 37th Annual Meeting of the Association for Computational Linguistics (ACL99), June 1999.

[19] A. W. Black, P. Taylor, and R. Caley, "The Festival Speech Synthesis System," http://www.cstr.ed.ac.uk/projects/festival.html, 1998.

[20] 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.