# CSLU Toolkit

# A Technological Survey of a Multimodal Library

## Seminar document

Lawrence Michel & Ridha Zarrougui
MSc Students, University of Fribourg

May 2006

## 1 Introduction

The CSLU Toolkit, standing for "Center for Spoken Language Understanding Toolkit", was created to provide the basic framework and tools for people to build, investigate and use interactive language systems. The toolkit incorporates leading-edge speech recognition, natural language understanding, speech synthesis and facial animation technologies. An environment has been designed in a way that these listed technologies may be used in a comprehensive manner with the use of a user-friendly graphical interface. The audience of this toolkit may begin from primary school use to high level academic research purposes.
We will briefly describe the toolkits architecture, have a closer look on a couple of modules which requires our specific attention. Finally, we will explain how we did install the toolkit, built our every first pizza order system and conclude.

## 2 The Toolkit Architecture

The CSLU Toolkit has been developed with modularity requirements in mind. The application developer, named RAD is the main part of the toolkit. It builds all necessary connection with all specific modules, and its main purpose is to enable an user-friendly interface to help building end-application in a matter of time.
Each modules are developed to solve specific problems, such as speech/text recognition (Festival), animation of an anatomically correct 3D head (Baldi) and last but not least, a platform for research in perception and cognition (PSL).

### 2.1 RAD

RAD, standing for Rapid Application developer, is a graphical tool for creating structured dialog-orientated application to enable interaction between the user and the computer. It implements all necessary connection between its modules. RAD offers a user-friendly graphical interface which leads the human-user to quickly build simple applications. Furthermore, it offers the experimented user the possibility to enhance his application by adding specific TCL/TK to extend the scope of problems it addresses.
The use of TCL/TK as a middle layered language makes RAD become a very powerful software. This let's us break through the limitation of what the RAD user interface offers us. For example, we can easily start a stand-alone TCL/TK process after having ordered to it by the use of our voice, such as opening a web browser.

The graphical user interface proposes 3 distinct group of objects :

- Base Objects : Basic object that ship as part of the CSLU Toolkit.

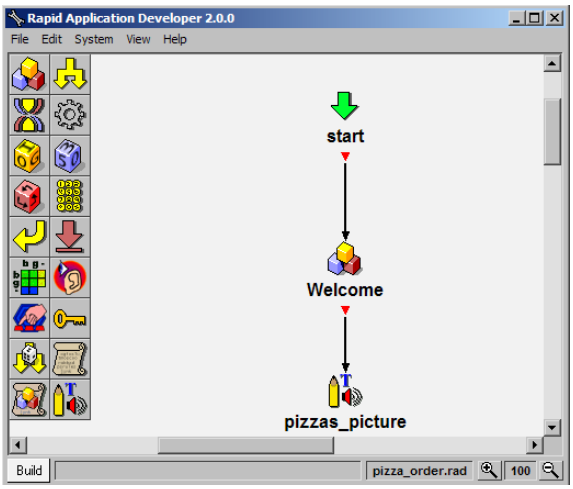


Figure 1: The RAD user interface

- Tucker-Maxon Objects : As part of the Tucker-Maxon plug-in, they were developed for use in a classroom, and enables some multimedia application.

- PSL Objects : Set of objects for conducting experiment (perception and cognition).

## 2.2 Festival

The text-to-speech component of the Toolkit. In our tested version (2.0.0), it could generate speech in English and Spanish. Festival is commonly used through other programs, rather than being interfaced to directly by the user. Unfortunately, at the time of writing this article, no elaborated documentation has been written about it.

## 2.3 Baldi/-Sync

The next module, named Baldi (fig.2), is designed to create and view a facial animation that is aligned with recorded speech audio. It is able to read speech signal and write down into phonem transcription the recognized portions, and vice-versa. In the actual state of its development, it still requires to know in literal text what it is said, as it could perform the phonem-to-signal alignement.
Baldi-Sync (fig.3) is an extended component of the Baldi module. It is intended to enhance the interaction between the application and the user. It enables the use of a 3D humanoid face animation, where movements and facial expression are synchronized with the speech. Extended features offered by the user-interface are handling of complex facial expressions through expression switches. That is, external modules can fix the correct expression according to the way some speech text should be said (angry vs happy, sad, etc.). This component incorporates a very important aspect of computer/user interaction by adding perceptual and cognitive aspects.

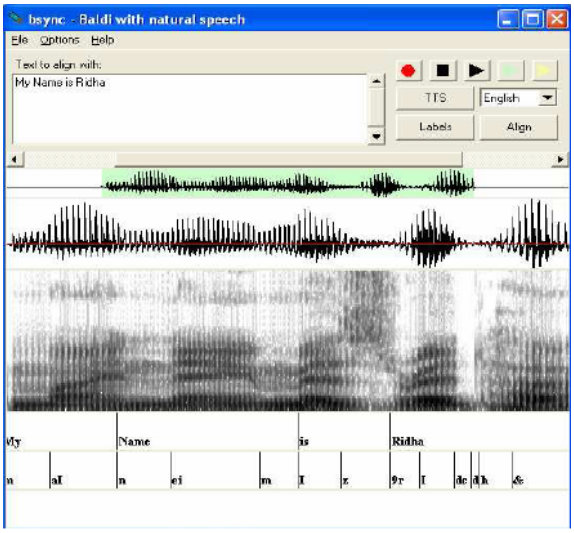


Figure 2: The BALDI module user interface

## 2.4 PSL Tools

The PSL Tools extension adds objects to the RAD user interface for designing and conducting perceptual experiments. There are three objects : *Expcontrol, Stimulus and Response.*
The *Expcontrol*, for Experimenter Control, object is intended to be connected with other objects in a loop. At each call, it will assign values to a user-defined list of variables, by use of stimulus and response variable.
The *Stimulus* object provides a facility for presenting a stimulus of audiovisual speech or a recorded sound. It has two main purposes : it presents the desired stimulus, and stores the time of the stimulus presentation.
The *Response* object provides an easy way to

collect the subject's response through keyboard, mouse or voice. It waits for the response to occur and then sets one or more variables to the received response and the time.


Figure 3: The BALDI-Sync module 3D model user interface

# 3 Installing the CSLU Toolkit

The toolkit is delivered in a single windows executable file. This version of installation contains only the required stuff to let the end user to choose over several possibilities of installation. All modules exposed in chapter 2. are proposed as a standard, but demanding users will benefit of the possibility to download all useful module libraries.
Our application has been exclusively built within RAD. We were especially focusing on RAD's particularity to quickly develop functional speech application. We experienced several objects disposed in it's graphical user interface, and added some extended TCL/TK code.
We finally built our first pizza ordering system using exclusively the speech recognition module as interaction mode. We tested our application in several situations, and we did conclude that it's efficiency is quite suitable for real situations. It might be important to notice that normally people might hesitate before giving an answer, or even add "noisy" words within it's sentence. Baldi module did correctly detect the portion of the speech sample where he would expect a possible correct response. But sometimes, some given words with very close pronunciation but strictly different in it's semantic might make the application be confused. The issue might be fixed if a dictionary is included in the process.

# 4 Possible CSLU applications

In the actual state of its development, the CSLU toolkit is already fully suitable for supporting small to middle size business applications. We are able to list some of them:

- Drive-In Fast-Food customer oriented service : potential customer may address directly to a computer service and may list what food he wants to order.
- Business Process chaining application : Industrial activities may be conducted by voice from an human operator located in a central area, or even decentralized. He will, for example, interact with the computer to bring the business process to its goal.
- PSL Objects : Navigational and diagnostic purposes : a pilot may interact with his machine through a board computer interface which might inform him of the actual state of the drive or flight, and may be ordered to change driving or flying parameters.

# 5 CSLU Strength and weaknesses

The CSLU toolkit is a speech processing oriented application development platform. It is mainly focusing on having RAD application's behaviour relying on effective speech recognition. Becauses such applications might be designed for interacting in real conditions, the effectiveness of speech recognition is strongly depending on the quality of all sound samples that are given for processing.

This might be one of its weakness (which is a well known issue in all actual speech recognition techniques). We tested the toolkit in several environmental sound conditions : The quality of the speech recognition can be affected by the amount of noise captured within the recording device (such as a city environmental noise, or people talking loud behind us). The sampling rate of the recorded speech could as well possibly affect it, due to the fact that less information might be transmitted (telephony applications).
Another weakness we have detailed is that Baldi module requires user input to help it aligning it's detected phonems to the signal.

CSLU toolkit has the strong advantage to have the platform run using clean separate modules. Interacting with them can be done either using the RAD user interface, or by accessing them directly within their own interfaces. Using RAD has the benefit of quickly developing application by describing all specific states in it. This way of doing it makes the understanding of the program much more intuitive. RAD gives as well the opportunity to add more complex functionality to our program by letting us the ability to add custom made TCL/TK coding in it. This enlarges the scope of possibilities the developer may address.
Modules used in this platform are developed using state-of-the-art techniques, such as the PSL tool, created and maintained from the Perceptual Science Laboratory at the University of California, US. That is, all modules within this platform are still under development, and newer version of them will probably address common known issues.

# 6 Conclusion

The CSLU toolkit is definitively a comprehensive environment to build, investigate and use interactive language systems, as it is proposed to be. It addresses various type of users, from the beginner, who can easily make his first steps on designing a simple speech oriented application, to the researcher, who might be having more specific interest on the speech processing capabilities of the proposed modules. Developers have the ability to access the modules separately and benefit of the overall methodology designed to address his requirements.
The actual version we tested (2.0.0) is in a very functional state. All important functionality are there and efficiency is good. Some room is still available for fine tuning and improvement, such as the speech-to-text alignment module.

# References

[1] CSLU toolkit, a comprehensive suite of tools to enable exploration, learning, and research into speech and human-computer interaction. http://www.cslu.ogi.edu/toolkit/