CSLU SPEECH TOOLKIT: UN EFFICACE "TOOL" PER LO SVILUPPO DI SISTEMI INTERATTIVI MULTIMODALI"

Piero Cosi
Istituto di Fonetica e Dialettologia – C.N.R.
Via G. Anghinoni, 10 - 35121 Padova (ITALY)
Tel: 049 8274421 Fax: 049 8274416
e-mail: cosi@csrf.pd.cnr.it www: http://www.csrf.pd.cnr.it

SOMMARIO

Sono illustrate le principali caratteristiche e funzionalità del software denominato CSLU-Speech-Toolkit che si configura come un insieme integrato di specializzate tecnologie di programmazione che rappresentano lo *stato dell'arte* negli strumenti per la ricerca, lo sviluppo e l'apprendimento dei sistemi di riconoscimento e sintesi del linguaggio naturale.

INTRODUZIONE

I sistemi automatici di sintesi da testo scritto e di riconoscimento del linguaggio naturale rendono possibile all'uomo di interagire con il computer mediante la voce, il metodo di comunicazione umana più naturale e comune. Questi sistemi sono realizzati sfruttando le conoscenze acquisite nel corso degli anni nel campo del riconoscimento automatico, dell'elaborazione del linguaggio naturale e delle tecnologie per l'interfaccia uomo-macchina. Essenzialmente si basano sul riconoscimento delle parole pronunciate, sull'interpretazione della loro sequenza al fine di ottenerne un opportuno significato e sull'attuazione di un'adeguata risposta. Le applicazioni sono numerosissime e, sebbene questi sistemi siano sostanzialmente agli albori, è oltremodo facile intuire la loro enorme potenzialità nel poter rivoluzionare il modo in cui le persone nel futuro si rapporteranno con le macchine. Numerosi e notevoli sono stati i passi avanti compiuti nel campo della ricerca e delle applicazioni. Infatti, non si può più parlare di esclusivi prototipi di ricerca appannaggio di pochi laboratori scientifici, poiché esistono un gran numero di sistemi funzionanti in compiti specifici, quali, la pianificazione di viaggi, l'esplorazione urbana ecc.. Dovendo essere utilizzati in "applicazioni reali" questi sistemi devono essere assai più robusti degli iniziali prototipi di ricerca in quanto devono poter funzionare correttamente in presenza di rumore, sia di canale sia d'ambiente ed indipendentemente dal variare della velocità d'eloquio, dell'accento o del sesso dell'utilizzatore. Devono esibire inoltre un comportamento 'intelligente', devono in pratica essere in grado di saper reagire anche in condizioni di parziale riconoscimento, che può avvenire in seguito all'occorrenza di pronunce scorrette da parte dell'utente o a causa d'altri fenomeni indesiderati. Dovranno esibire inoltre la capacità di integrarsi efficacemente con altri modi di comunicazione, cercando di "capire" in anticipo le intenzioni dell'utente attraverso le sue espressioni facciali, il movimento delle labbra, degli occhi ecc. e sfruttando tutte le molteplici potenzialità multimediali offerte dalla tecnologia per elaborare le proprie azioni come risposta ai quesiti dell'utente rendendo l'interazione oltremodo naturale ed immediata.

Purtroppo lo sviluppo di un sistema di riconoscimento del linguaggio parlato è un'attività molto complessa che generalmente richiede, per la progettazione, la valutazione e la vera e propria implementazione del sistema, un lungo periodo che può facilmente durare parecchi mesi o meglio alcuni anni. Per poter sfruttare efficacemente questa nuova tecnologia, un sempre maggior numero di laboratori deve poter disporre di strutture informative adeguate ed è impensabile che le conoscenze necessarie allo sviluppo di una tale tecnologia siano parcellizzate e non comunemente utilizzabili. E' con questo obiettivo che all'*Oregon Graduate Institute* (*OGI*) di Portland, ed in

particolare presso il *Center for Spoken Language Understanding* (*CLSU*)[1], è stato sviluppato il software denominato *CSLU-Speech-Toolkit* [2] [1] che ha proprio lo scopo di fornire ad un sempre più elevato numero di ricercatori, come anche di non addetti ai lavori, lo strumento necessario per creare e sviluppare personalmente in modo semplice ed interattivo nuovi sistemi di riconoscimento del linguaggio naturale sempre più orientati alle applicazioni [2]. Per gli utenti più esperti in tecnologie vocali i Toolkit rappresentano un vero e proprio banco di prova, efficacissimo per lo sviluppo e la verifica delle proprie ricerche, anche le più avanzate [3-5].

CSLU SPEECH TOOLKIT

Il software denominato CSLU-Speech-Toolkit è stato progettato per facilitare lo sviluppo della ricerca e delle sue possibili applicazioni, nel campo delle tecnologie vocali per un'ampia gamma d'utilizzatori e d'utilizzazioni. Fra le molteplici possibilità si possono elencare:

- l'abilitazione di esperti in specifici domini di applicazione all'utilizzo delle tecnologie vocali per la progettazione di sistemi di riconoscimento del linguaggio naturale, anche multi-lingue, con semplici strumenti di sviluppo;
- la produzione di sistemi di riconoscimento di elevate prestazioni a partire da specifiche di progetto di alto livello;
- l'apprendimento delle tecnologie vocali, con particolare riferimento ai sistemi di dialogo interattivo, mediante opportuni corsi introduttivi incorporati all'interno dei Toolkit;
- la realizzazione di ricerche sull'interazione uomo-macchina con particolare riferimento ai sistemi di dialogo;
- lo sviluppo delle ricerche sulle tecnologie vocali per una loro introduzione in applicazioni reali ed una loro valutazione.

Il software CSLU-Speech-Toolkit è un ambiente di sviluppo che integra un insieme di tecnologie vocali che comprendono il riconoscimento automatico del linguaggio naturale, la sintesi automatica da testo scritto[3], e l'animazione video di facce parlanti[4]. Nel sistema sono inclusi vari strumenti di programmazione, essenziali per una facile implementazione delle applicazioni. L'architettura generale, come illustrato in Figura 1, è composta da tre elementi principali: un insieme di librerie contenenti i moduli di base(***Core*** o ***Kernel***) generamente in C++, un'opportuna ed immediata interfaccia di programmazione (***CSLUsh*** [6]) e un sistema di sviluppo grafico per una più semplice realizzazione delle applicazioni finali (***RAD***, ***Rapid Application Developer***), questi ultimi realizzati mediante l'utilizzazione del sempre più diffuso linguaggio *Tcl/Tk* [7].

Il *cuore* (***Kernel***) del sistema consiste in un insieme di moduli che implementano gran parte degli aspetti fondamentali relativi alle tecnologie vocali. Queste librerie, generalmente realizzate in C++, rappresentano l'interfaccia di programmazione (API, *Application Programming Interface*) indipendente dall'*hardware* e dai diversi sistemi operativi utilizzati. I singoli moduli comprendono specifiche funzioni per l'elaborazione e l'analisi del segnale vocale, per l'apprendimento di reti neurali artificiali (NN, *Neural Networks)* e di catene di Markov Nascoste (HMM, *Hidden Markov Models*), per il riconoscimento automatico mediante algoritmo di ricerca di *Viterbi* [8] e per l'utilizzazione di un'interfaccia telefonica. Inoltre sono inclusi un robusto *parser* del linguaggio

---

1 Oregon Graduate Institute of Science and Technology (OGI) - Center for Spoken Language Understanding (CSLU), P.O. Box 91000, Portland Oregon 97291-1000 USA, http://cslu.cse.ogi.edu.

2 Per coloro che volessero utilizzare il software descritto in questo lavoro, il software CSLU-Toolkit può essere facilmente recuperato (gratuitamente per scopi accademici e di ricerca) all'indirizzo Internet: http://cslu.cse.ogi.edu/.

3 Il software CSLU-Toolkit è strettamente integrato con il sistema software di sintesi da testo FESTIVAL, sviluppato presso l'Università di Edimburgo: http://www.cstr.ed.ac.uk/projects/festival.html

4 Il software CSLU-Toolkit è strettamente integrato con il sistema software di animazione di volti parlanti BALDI, sviluppato presso l'Università di California, Santa Cruz: http://mambo.ucsc.edu/

naturale [9], una versione aggiornata del sistema di sintesi da testo scritto denominata *Festival* [10], sviluppata presso l'Università di Edimburgo e un sistema di animazione di volti parlanti denominato *Baldi* [11] sviluppato presso l'Università di California, Santa Cruz, oltre a dizionari testuali di pubblico dominio. L'insieme di questi moduli è molto flessibile poiché i singoli moduli possono essere facilmente compilati all'interno di un singolo programma o richiamati individualmente all'interno dell'interfaccia di programmazione, secondo quanto richiesto dalle specifiche applicazioni.

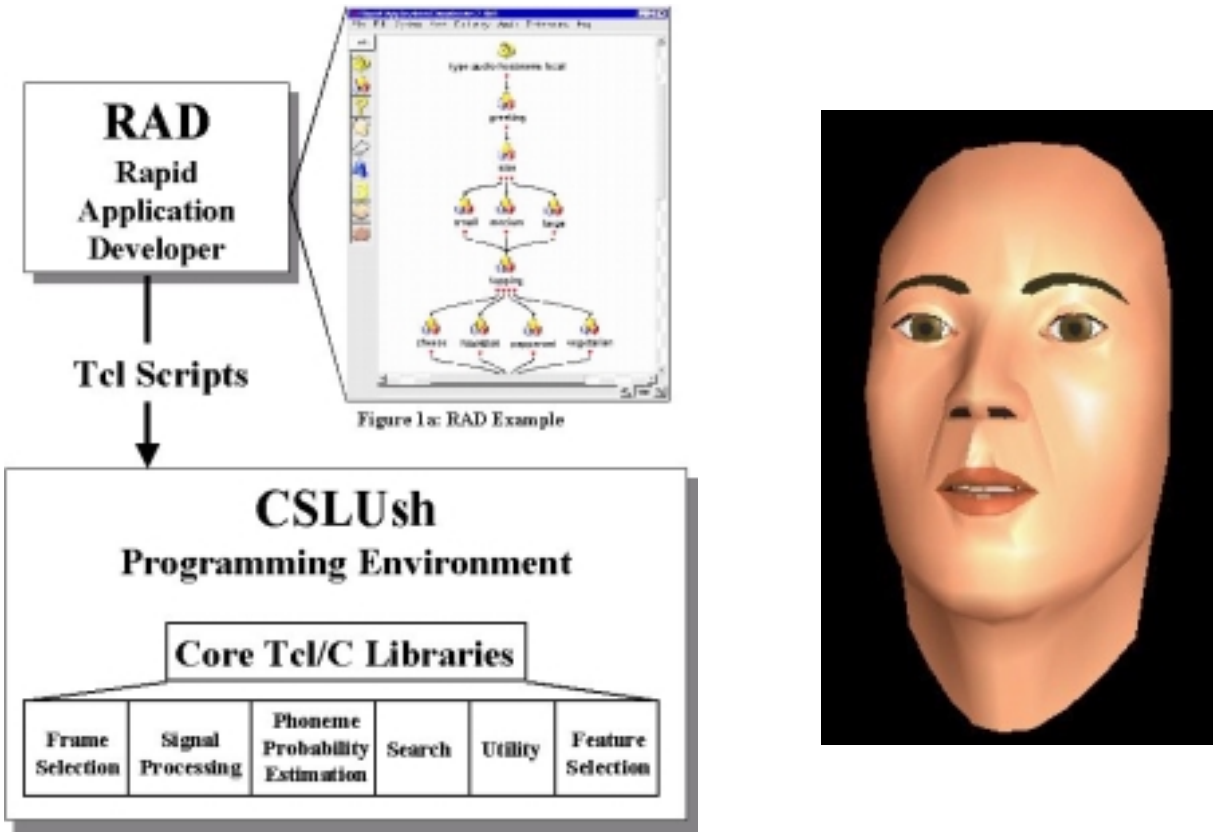

Figura 1. Architettura del software CSLU-Speech-Toolkit e illustrazione dell'agente parlante Baldi.

Il livello principale per lo sviluppo delle applicazioni è costituito dall'interfaccia di programmazione denominata ***CSLUsh***, pronunciato come "*slush*", ed è interamente basata sul diffusissimo e portabilissimo linguaggio di programmazione interattivo denominato Tcl/Tk [7]. CSLUsh incorpora, infatti, i moduli di base API precedentemente descritti mediante specifiche funzioni interamente realizzate in Tcl/Tk. Un'applicazione è quindi costruita fondendo insieme i moduli di base e utilizzando funzioni aggiuntive d'interfaccia, quali, ad esempio, la gestione degli eventi relativi all'organizzazione dei vari *file*, degli eventi collegati ai problemi della rete, oppure degli eventi relativi all'interfaccia utente grafica. Le varie applicazioni possono essere eseguite anche su piattaforme *hardware-software* differenti e connesse ad una rete locale (LAN), oppure in Internet utilizzando le specifiche funzionalità *client-server* fornite dal linguaggio Tcl, quali i protocolli TCP e UDP, la capacità di funzionare come *daemon*, l'esecuzione remota di comandi Tcl ed il trasferimento dei dati-oggetti. I moduli di base API sono raggruppati in librerie funzionali che sono richiamate, dinamicamente, in fase d'esecuzione, rendendo le applicazioni scalabili in termini di risorse computazionali. Questo consente anche un'estensione delle funzionalità del software senza che questo debba essere ricompilato in seguito ad eventuali aggiunte o modifiche.

Il terzo componente, a livello più "*alto*", è costituito dall'ambiente di sviluppo grafico denominato ***RAD* (Rapid Application Developer)**. RAD integra i moduli per il riconoscimento

vocale, la sintesi, l'animazione d'agenti parlanti e gli strumenti di visualizzazione in un efficace ambiente grafico di sviluppo per la realizzazione e l'esecuzione di semplici applicazioni. RAD include una tavolozza d'oggetti di dialogo grafici ed una semplice interfaccia di tipo *drag-and-drop*. Gli oggetti della tavolozza sono dei veri e propri blocchi di programmazione grafica, che l'utente seleziona ed organizza collegandoli appropriatamente per realizzare un modello di dialogo a stati finiti. Particolare cura è stata riposta nel cercare di semplificare al massimo la procedura di sviluppo delle applicazioni. Ad esempio, infatti, la specifica per l'utilizzazione di un sistema di riconoscimento è semplice al punto di specificare soltanto le parole o le frasi che devono essere riconosciute. Allo stesso modo per le procedure relative alla sintesi vocale, è richiesta esclusivamente l'immissione testuale dei messaggi da produrre, oppure l'indicazione di un eventuale collegamento con messaggi precedentemente registrati. L'agente parlante Baldi (vedi Figura 1) [8] è integrato completamente nel sistema ed è automaticamente sincronizzato, sia con la voce sintetica, sia con quella registrata. Non appena l'utente diventa più esperto con l'ambiente di programmazione grafica, può immediatamente estendere l'insieme iniziale di funzioni fornite da RAD e creare, ad esempio, nuovi front-end di analisi per applicazioni già sviluppate quali la lettura automatica della propria posta elettronica o di pagine *www* testuali.

CONCLUSIONI

Il software CSLU Speech Toolkit è attualmente utilizzato in vari laboratori di ricerca, scuole ed industrie in tutto il mondo ed alcune delle principali attività sviluppate riguardano **il riconoscimento automatico del linguaggio**, la **sintesi automatica da testo scritto**, il **riconoscimento del parlante**, l'**insegnamento delle lingue**, l'**alfabetizzazione informatica** e la **riabilitazione del linguaggio** nel caso di bambini non udenti. L'elevato grado di semplicità rende il sistema utilizzabile da un gran numero di utenti anche non esperti e la sua modularità ed affidabilità lo rendono senza dubbio un insieme integrato di specializzate tecnologie di programmazione, rappresentanti lo stato dell'arte negli strumenti per la ricerca, lo sviluppo e l'apprendimento dei sistemi di riconoscimento del linguaggio naturale.

BIBLIOGRAFIA

[1] Sutton, S., Cole, R.A., de Villiers, J., Schalkwyk, J., Vermeulen, P., Macon, M., Yan, Y., Kaiser, E., Rundle, B., Shobaki, K., Hosom, J.P., Kain, A., Wouters, J., Massaro, D., and Cohen, M., "Universal Speech Tools: The CSLU Toolkit," ICSLP-98, vol. 7, pp. 3221-3224, Sydney, Australia, November 1998.

[2] Cole R., Sutton S., Yan Y., Vermeulen P., Fanty M., Accessible Technology for Interactive Systems: A new approach to spoken language research, Proc. ICASSP-98, II 1037-1040.

[3] Sutton S., Novick D., Cole R., Fanty M., Building 10,000 spoken-dialogue systems, Proc. ICSLP-96, II 709-712.

[4] Hosom J.P., Cole R.A., Cosi P., Evaluation and Integration of Neural-Network Training Techniques for Continuous Digit Recognition, Proc. ICSLP-98, III 731-734.

[5] 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," IVTTA-ETWR-98, pp. 135-140, September 1998.

[6] J. Schalkwyk, J.H. de Villiers, S. van Vuuren, and P.Vermeulen, "Cslush: An Extendible Research Environment," Proc. of Eurospeech 97, Rodhes, September 1997, pp 689-692.

[7] J.K. Ousterhout, Tcl and the Tk Toolkit. Addison Wesley, 1994.

[8] L. Rabiner and B-H Juang, Fundamentals of Speech Recognition, Signal Processing Series, Alan V. Oppenheim, Series Editor, Prentice Hall, Englewood Cliffs, New Jersey,1993.

[9] Kaiser, E.C., Johnston, M., and Heeman, P.A., "PROFER: Predictive, Robust Finite-State Parsing for Spoken Language," ICASSP-99, vol. 2, pp. 629-632, Phoenix, AZ, March 1999.

[10] Black, A. and Taylor, P., "Festival Speech Synthesis System: System Documentation (1.1.1)," Human Communication Research Centre Technical Report HCRC/TR-83, Edinburgh, 1997.

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