FESTIVAL PARLA ITALIANO!

Piero Cosi*, Roberto Gretter** and Fabio Tesser**

*IFD-CNR
Istituto di Fonetica e Dialettologia – Consiglio Nazionale delle Ricerche
e-mail: cosi@csrf.pd.cnr.it www: http://nts.csrf.pd.cnr.it/Ifd

**ITC-IRST
Istituto Trentino di Cultura
Istituto per la Ricerca Scientifica e Tecnologica
e-mail: {gretter,tesser}@irst.itc.it www: http://www.itc.it/enIRST

1. SOMMARIO

L'argomento di questo lavoro riguarda l'implementazione della versione italiana del sintetizzatore vocale da testo scritto basato sulla tecnica della concatenazione di difoni, denominato FESTIVAL[1]. Il sistema è stato interamente realizzato utilizzando l'ambiente di sviluppo denominato EDINBURGH-SPEECH-TOOLS (EST) e un sistema di allineamento/segmentazione automatico sviluppato all'IFD sulla base di un sistema di riconoscimento fonetico per l'Italiano di elevate prestazioni. Sono descritti i vari moduli, Linguistico-Prosodici e Fonetico-Acustici, e le relative procedure utilizzate per la generazione di una nuova voce maschile adulta per l'italiano.

2. INTRODUZIONE

Il sintetizzatore, vocale basato sulla tecnica della concatenazione di unità vocali, denominato ***FESTIVAL*** [1], sviluppato dal "*Centre for Speech Technology Research*" (*CSTR*) dell'Università di Edinburgo, e in generale un qualsiasi sistema di sintesi automatica da testo scritto (Text-To-Speech: *TTS*), può essere efficacemente rappresentato dalle sue tre principali funzioni: l' "*analisi del testo*", che consente di identificare le parole o, in generale, alcune semplici sotto-frasi più comuni; l' "*analisi linguistica*", che consente di individuare la pronuncia corretta delle parole e delle frasi da sintetizzare assieme alla corrispondente struttura prosodica in termini di intonazione e di durata; la "*generazione del segnale*", che consente, quale ultimo passo, di generare una forma d'onda a partire dalle informazioni linguistiche sopra specificate. Ovviamente questa divisione non è "assoluta", ma sembra in ogni caso molto utile per suddividere efficacemente il problema della sintesi da testo scritto, che può essere, infatti, affrontato e risolto mediante numerosi metodi alternativi. Ad esempio vi sono varie tecniche di generazione di forma d'onda e ognuna può richiedere diversi tipi di informazione. La pronuncia delle parole può, non sempre richiedere i fonemi standard, quali unità di base da cui partire, ma unità diverse quali ad esempio i cosiddetti "difoni", oppure altre unità di durata variabile. L'intonazione non vuol dire necessariamente solo il contorno della frequenza fondamentale (f0). Essenzialmente le

[1] Parte di questo lavoro è stato svolto nell'ambito del progetto europeo
MPIRO: Multilingual Personalized Information Objects. European Project IST-1999-10982

tre sezioni sopra elencate ben rappresentano, piuttosto che descrivere le più aggiornate tecniche di sintesi, le principali procedure più comunemente utilizzate per generare una voce sintetizzata funzionante.

Un'altra caratteristica importante di un sistema TTS, che non è frequentemente menzionata, riguarda l' "*architettura del sistema*". In FESTIVAL, l'architettura del sistema è, senza dubbio l'aspetto funzionale più importante che rende possibile e soprattutto semplice tutto il processo di costruzione di nuove voci in varie lingue. FESTIVAL fornisce all'utente una semplice struttura di frase e, cosa ben più importante un semplice ed intuitivo linguaggio, basato sul linguaggio di programmazione ***SCHEME*** [2], diretto discendente del più famoso LISP utilizzato in applicazioni di Intelligenza Artificiale, per manipolarla; inoltre interagisce con il sistema audio mediante un'efficace modalità di *spooling* che consente di elaborare i segnali audio mentre il resto del processo linguistico di sintesi può continuare in parallelo. Mediante gli "***EDINBURGH-SPEECH-TOOLS***" (EST), incorporati in FESTIVAL, sono forniti all'utente o al ricercatore alcuni strumenti di base necessari per la generazione di un sistema TTS quali: estrattori di *f0/pitch*, costruttori di alberi di classificazione o regressione, procedure di analisi del segnale, moduli di I/O, ecc., assieme ad un semplicissimo linguaggio di interfaccia ("*scripting*"). Tutte queste funzioni sono state progettate con lo scopo di far concentrare, l'utilizzatore del sistema, sull'obiettivo di costruire nuove voci, senza dover invece preoccuparsi direttamente del software alla base di tutto il procedimento necessario per la costruzione di un sistema TTS.

Parallelamente a FESTIVAL, soprattutto per merito di Alan W. Black dello "*Speech Group*" della "*Carnegie Mellon University*", è stato creato ***FESTVOX*** [3], un progetto il cui scopo primario è quello di rendere la costruzione di nuove voci sintetiche più sistematico e più documentato, al fine di rendere chiunque, e non solo alcuni specialisti, indipendenti nel processo di creazione di una "propria voce". FESTVOX, infatti, non è nient'altro che un modulo parallelo a FESTIVAL con cui creare una nuova voce, distribuibile ed installabile separatamente al software originario principale, con il quale non interferisce direttamente. Una nuova voce creata con FESTVOX consiste essenzialmente: di un insieme di unità vocali primarie, frequentemente *difoni* anche se vi sono già esempi di voci basate su sintesi ad unità varabili, e di un insieme di *script* specifici per l'analisi acustica, l'analisi del testo, l'analisi linguistica e l'analisi prosodica relativi alla nuova voce stessa ed eventualmente alla nuova lingua sviluppata. In pratica, FESTVOX, mediante un'adeguata documentazione e un completo insieme di *script* in linguaggio SCHEME, offre il supporto per: la progettazione, la registrazione e la segmentazione/etichettatura di corpora vocali di riferimento da cui estrarre le unità di base per la sintesi (fonemi, difoni, unità variabili, ecc.); la costruzione di motori di sintesi in domini di applicazione limitati; la costruzione di modelli prosodici di durata e di intonazione basati su regole o derivati automaticamente dall'analisi di alcuni dati campione; la costruzione di moduli di analisi del testo, anch'essi basati direttamente su regole o automaticamente appresi dai dati stessi; la generazione di lessici di riferimento e di moduli specifici per l'accentazione e la trascrizione grafema-fonema. Sebbene la costruzione di una nuova voce rimanga ancora un procedimento non completamente automatizzato, mediante FESTIVAL e FESTVOX sono già state create nuove voci per l'inglese, l'inglese americano, lo spagnolo casigliano e messicano, il tedesco, il polacco, il greco, il gallese gaelico, il basco, il portoghese brasiliano, e finalmente ora anche l'italiano. Nonostante la qualità delle singole voci nelle varie lingue vari notevolmente e risenta pesantemente della conoscenza specifica e degli strumenti linguistici preesistenti a disposizione dei loro creatori, tutti questi tentativi hanno

comunque prodotto dei sistemi di sintesi TTS in grado di leggere un qualsiasi testo scritto, quale ad esempio quello di un articolo di un quotidiano, con un notevole livello di intelligibilità.

FESTIVAL e FESTVOX sono liberamente utilizzabili, senza scopo di lucro, da chiunque voglia utilizzarli senza alcuna restrizione e la distribuzione e la responsabilità delle varie lingue rimane interamente a carico dei loro creatori originari.

3. FESTIVAL IN ITALIANO

Utilizzando FESTIVAL cioè un sintetizzatore basato sulla concatenazione di unità vocali primarie, si è deciso per il momento di utilizzare la sintesi per concatenazione di difoni invece della più efficace e recente sintesi basata sulla selezione di unità di durata variabile.Questo è stato fatto, sia per velocizzare il processo di creazione di nuove voci per l'italiano, sia per verificare la qualità dello stato dell'arte della sintesi TTS con questa tecnica.

Come regola generale si indica con difono quella porzione di segnale vocale relativa alla sequenza di due fonemi che va da ½ del primo fonema a ½ del successivo. Ci sono però alcune eccezioni come ad esempio nelle geminate dove ai difoni destro e sinistro si aggiunge un difono centrale che va da ¼ a ¾ del fonema geminato.

3.1. Corpus dei difoni

E stato registrato un corpus di frasi "foneticamente ricche" (grammaticalmente e sintatticamente corrette, ma senza senso) pronunciate con intonazione "piatta" da un parlante italiano maschile (P.C.) in una camera silente (semi-anecoica). Come indicato in Figura 1, per la registrazione sono stati utilizzati un microfono Sennheiser MKH 40 P48, connesso sia ad una scheda di acquisizione di un Personal Computer che ad un canale (DX) di un DAT Sony DTC 1000 ES, che un Elettroglottogtrafo collegato soltanto ad un canale (SX) del DAT.

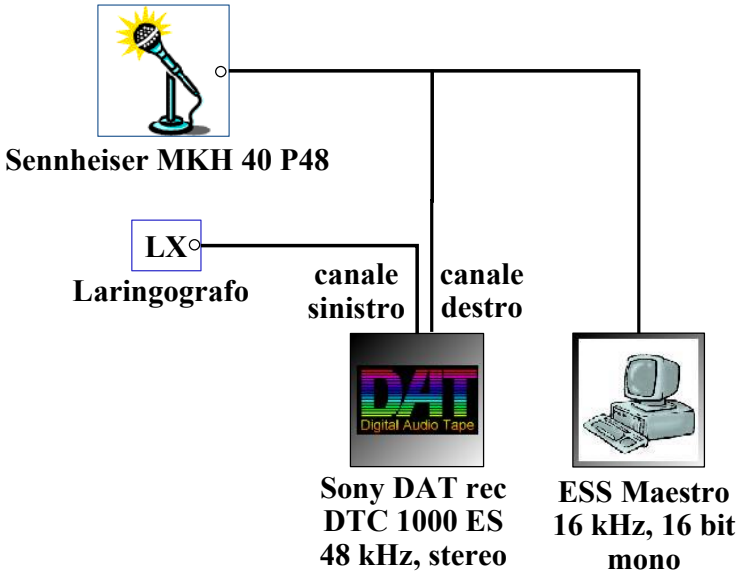


Figura 1. Illustrazione grafica dell'ambiente di registrazione del corpus dei difoni.

La procedura di registrazione è stata interamente automatizzata ed è pilotata da uno script RAD (***CSLU SPEECH TOOLKIT*** [4]) che consente l'acquisizione su file di tipo Windows PCM *.wav (16kHz, 16bit). I "*difoni target*" sono presenti in "*parole target*" all'interno di semplici frasi senza senso, ma grammaticalmente e sintatticamente corrette. Come indicato nei seguenti esempi:

"i ve**nti** **ciapàpa** **zapàpa** sono sta**nch**i"
"i gatti **òpapa** **àpapa** sono belli"
"il tu**ngst**eno **papapò** **papòccia** è gh**ia**cciato"

Vi sono 2 parole target per ogni frase e nelle frasi contorno sono stati inclusi anche alcuni cluster consonantici ed alcune sequenze vocaliche nel caso fosse sorta la necessità di utilizzare anche queste unità più complesse dei semplici difoni per ottimizzare la sintesi. Per rendere le forme d'onda dei singoli difoni ragionevolmente omogenee in ampiezza si è operata una normalizzazione a livello di ogni singola parola. Attualmente i difoni creati sono circa 1300 (CV, CC, VV...) e si pensa di raggiungere i 2000, alla fine del progetto, una volta inclusi anche tutti i casi relativi alle sequenze vocaliche e consonantiche (cluster, rV, VVV,…).

I difoni sono stati codificati in termini di coefficienti LPC e del loro corrispettivo segnale residuo. I coefficienti LPC di ordine 16 sono stati calcolati mediante gli Edinburgh Speech Tools di Festival e la codifica è sincrona con il pitch: nelle parti vocalizzate sono calcolati 16 coefficienti per ogni periodo di pitch, mentre per le parti non vocalizzate l'analisi è effettuata immaginando un pitch virtuale che interpola le parti vocalizzate adiacenti. Il pitch è stato estratto direttamente dal segnale voce utilizzando un software chiamato ***PRAAT*** [5]. Una schermata di PRAAT è illustrata in Figura 2 in cui sono visibili i singoli *pitchmarks* calcolati e la curva di f0 finale su una frase campione. Per automatizzare tutte le fasi di analisi sono stati creati degli appositi script ***TCL*** [6] per interfacciare PRAAT e EST. Il lavoro è stato svolto pensando di automatizzare e velocizzare il più possibile la creazione di una nuova voce: infatti, una volta registrato il corpus di frasi da un nuovo speaker, la segmentazione in difoni viene effettuata da un sistema di riconoscimento fonetico automatico di elevate prestazioni [7], sviluppato all'IFD e allenato su APASCI [8] mediante i CSLU Speech Toolkit. Riconosciuti i fonemi, un'altro script TCL estrae automaticamente la posizione temporale dei difoni dalle parole target per poi analizzarli e codificarli.

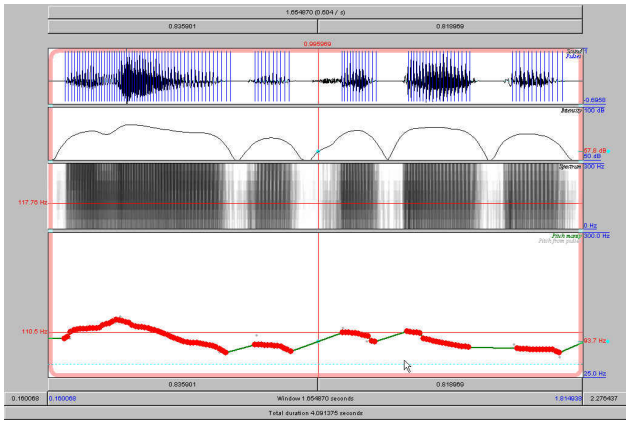


Figura 2. Esempio di un'analisi di f0/pitch con PRAAT [5] su una frase campione. Oltre alla curva di energia ed allo spettrogramma sono visibili i pitchmarck e la curva finale di f0.

3.2. Analisi del testo

L'architettura dei moduli per l'analisi del testo e dei moduli linguistici per l'accentazione la trascrizione grafema-fonema e per la sillabificazione è quella illustrata in Figura 3.

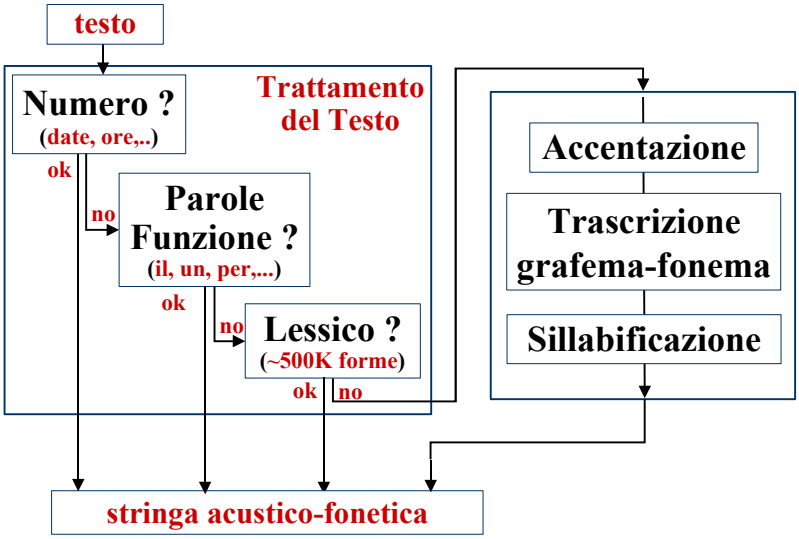


Figura 3. Architettura dei moduli per l'analisi del testo e dei moduli linguistici per l'accentazione la trascrizione grafema-fonema e per la sillabificazione.

La stringa in ingresso è esaminata da un primo modulo che riconosce i dati numerici che sono direttamente trascritti a livello fonemico. I dati non numerici sono distinti a seconda che la parola sia una parola "*funzione*" o una parola "*contesto*". A questo punto le parole sono trascritte nella loro forma fonemica, o passando attraverso il lessico o applicando le regole esplicite di accentazione, trascrizione e sillabificazione. In particolare il modulo numerico espande a livello di parola e poi di fonemi i dati numerici distinguendo tra ore, date, numeri di telefono, ecc. Le parole funzione sono distinte dalle altre perché nei moduli successivi sono trattate in modo diverso. Una volta identificate vengono infatti suddivise a seconda del loro gruppo grammaticale e del sotto-gruppo funzionale (es: ARTICOLI definiti: il lo la i gli le; ARTICOLI, indefiniti: un, uno, una; AVVERBI, tempo: ieri, oggi, dopo, poi, ecc.) La fase di trascrizione grafema-fonema è implementata tramite la ricerca della trascrizione all''interno di un lessico di 500000 forme accentate, trascritte fonemicamente, suddivise in sillabe ed etichettate secondo la loro classe grammaticale (*POS*: *part of speech*) come ad esempio:

("accertare" V (((a tS) 0) ((tS e r) 0) ((t a1) 1) ((r e) 0))).

Il lessico è stato compilato nel formato letto da FESTIVAL per velocizzare la procedure di ricerca. Se la parola da sintetizzare è trovata nel lessico, la sua trascrizione è immediata altrimenti si applicano le regole d'accentazione e di trascrizione grafema-fonema, scritte in

linguaggio SCHEME ed elaborate statisticamente sulla base di un grosso corpus testuale. Viene rimosso l'accento da tutte le parole funzione che vengono poi congiunte con la parola successiva, mentre per tutte le altre forme si applica dapprima una serie di regole statistiche ed euristiche per l'inserimento dell'accento e successivamente una serie di regole per la vera e propria trascrizione grafema-fonema. La simbologia utilizzata è quella SAMPA [9] con le vocali accentate fatte seguire dalla cifra numerica "1".

3.3. **Analisi Prosodica**

A partire dalla stringa fonetica sin qui ottenuta sono selezionate le corrispondenti unità acustiche, nel nostro caso, i difoni, e per ognuno di esso è aggiunta l'informazione riguardante la durata e il suo pitch. Questi dati sono poi inviati al modulo di generazione vera e propria della forma d'onda che utilizza la sintesi LPC eccitata dai residui. ("*Residual Excited Linear Prediction")*.

Attualmente, non disponendo di un analizzatore sintattico complesso, l'analisi prosodico-intonativa, intesa come determinazione della durata e della frequenza fondamentale in corrispondenza dei difoni da sintetizzare, è sicuramente la parte più debole dell'attuale sistema di sintesi e verrà sicuramente migliorata in futuro.

I moduli prosodici sin qui sviluppati sono molto semplici, anche perché attualmente le uniche informazioni estratte dal testo sono quelle riguardanti le parole funzione e i marker d'inizio e fine frase assieme agli altri simboli di punteggiatura.

Ai singoli fonemi viene assegnata una durata media, che nel nostro caso è stata elaborata analizzando statisticamente un corpus di frasi pronunciate da alcuni annunciatori RAI [10].

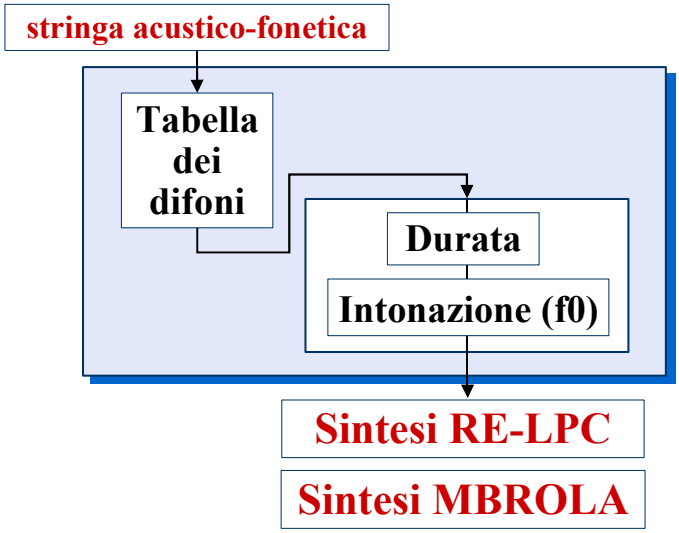
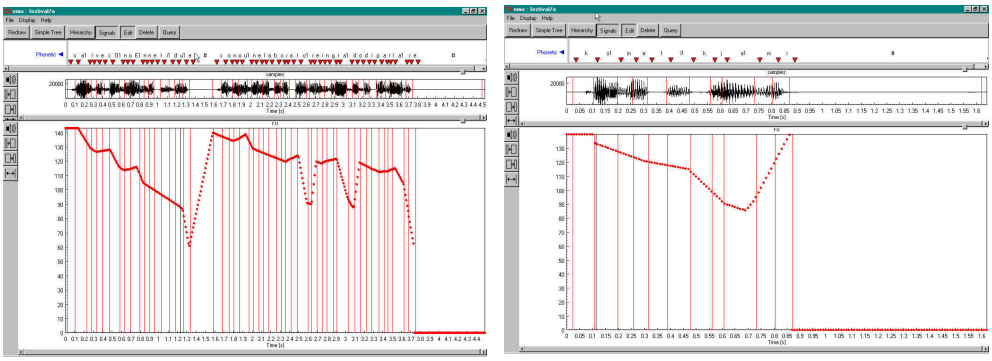


Figura 4. Architettura dei moduli per l'analisi prosodica e per la generazione della forma d'onda mediante RE_LPC e MBROLA [11].

Le "pause" fra le parole sono divise in tre categorie: pause brevi di 250ms caratteristiche ad esempio di alcuni simboli di interpunzione: " ' \ , ;" ; pause medie di 500ms prodotte a volte prima delle parole funzione; pause lunghe di 750ms in corrispondenza dei simboli di punteggiatura conclusivi di frase: " ? . : !". Inoltre, se la

vocale da sintetizzare è accentata la sua durata viene aumentata del 20% rispetto a quella standard.

Per le frasi "*dichiarative*" (si veda l'esempio di Figura 5a), per ogni "*gruppo intonativo*", viene generata per f0 una "*linea di base*" che parte da 140Hz e arriva a 60Hz. Se una sillaba è accentata il contorno di f0 viene elevato approssimativamente di ~10Hz rispetto alla "linea di base". L'ultima sillaba della frase ha un'inclinazione maggiore rispetto all'andamento di base e viene effettuato un "reset" della "linea di base" sulle "parole funzione".



(a) **(b)**

Figura 5. Esempio dell'andamento di f0 per due frasi dichiarative: "Salve, sono NT2. Sono un elaboratore in grado di parlare"(a) e per la frase interrogativa: "Come ti chiami?" (b).

Per quanto riguarda le frasi "interrogative" (si veda l'esempio di Figura 5b), il "Target Point" (TP) corrisponde ai ¾ della vocale dell'ultima sillaba tonica ed in questo punto f0 assume il valore corrispondente all'80% del valore della linea di base (Baseline*0.8), mentre, alla fine della vocale precedente, f0 assume il valore corrispondente al 110% della linea di base (Baseline*1.1). A partire da TP, f0 sale fino a f0max con inclinazione graduale e salita più ripida sull'ultima sillaba.

4. CONCLUSIONI

Finalmente FESTIVAL parla italiano. Le procedure sviluppate per automatizzare la creazione di una nuova voce per l'italiano si sono dimostrate affidabili ed efficienti. In particolare, lo script per la segmentazione e l'estrazione automatica dei difoni, basato sull'allineamento forzato automatico che sfrutta un sistema di riconoscimento fonetico (IFD) sviluppato su APASCI (IRST) mediante CSLU Speech Toolkit (OGI) si è rivelato essenziale per velocizzare al massimo le procedure. Un nuovo parlante, infatti, può registrare tutte le frasi per la generazione dei difoni in circa due settimane e mediante le procedure automatiche, si può preparare una nuova voce in 48 ore, anche se, ovviamente, è sempre necessario un lavoro di controllo di qualità finale per ottimizzare le prestazioni del sistema.

5. SVILUPPI FUTURI

Oltre alla registrazione di nuove voci (femminile adulta, bambino, bambina), saranno provati altri metodi di analisi e codifica del segnale, in particolare per quanto riguarda la sintesi MBROLA [11]. Si cercherà di sviluppare o interfacciarsi a nuovi moduli linguistici quali un analizzatore morfologico e grammaticale, al fine di migliorare la prosodia con nuove e più specifiche regole di durata e intonazione. Inoltre, nell'ambito del progetto europeo MPIRO ([12]), sarà implementata l'interazione con un modulo di generazione automatica del testo, capace di fornire un testo da sintetizzare arricchito con direttive linguistico - prosodiche. Obiettivo di tale progetto è la realizzazione di un chiosco informativo virtuale in grado di fornire informazioni ai visitatori di un museo.

BIBLIOGRAFIA

[1] **FESTIVAL**: Alan W. Black (awb@cs.cmu.edu), Paul Taylor (Paul.Taylor@ed.ac.uk), Richard Caley, Rob Clark (robert@cstr.ed.ac.uk)
CSTR - Centre for Speech Technology - University of Edinburgh.
WWW page: http://www.cstr.ed.ac.uk/projects/festival/.

[2] **SCHEME**, Computer Programming Language.
WWW page: http://www-swiss.ai.mit.edu/~jaffer/Scheme.html

[3] **FESTVOX**: Alan W Black (awb@cs.cmu.edu), Kevin A. Lenzo (lenzo@cs.cmu.edu)
Speech Group at Carnegie Mellon University.
WWW page: http://www.festvox.org/.

[4] M. Fanty, J. Pochmara e R.A. Cole, "An Interactive Environment for Speech Recognition Research", *Proceedings of International Conference on Spoken Language Processing* (ICSLP-92), Banff, Alberta, October 1992, 1543-1546.
CSLU SPEECH TOOLKIT: WWW page: http://cslu.cse.ogi.edu/tools.htm.

[5] **PRAAT**: Paul Boersma (Paul.Boersma@hum.uva.nl),
David Weenik (David.Weenink@hum.uva.nl), Institute of Phonetic Sciences, University of Amsterdam.
WWW page: http://www.fon.hum.uva.nl/praat.

[6] **Tcl/Tk**: K. Ousterhout - ouster@sprite.berkeley.edu
WWW page: http://sol.brunel.ac.uk/tcl/Tcl.html.

[7] P. Cosi e J.P. Hosom, "High Performance "General Purpose" Phonetic Recognition for Italian", *Proceedings of International Conference on Spoken Language Processing* (ICSLP-2000), Beijing, Cina, 16-20 October, 2000, Vol. II, pp. 527-530.

[8] B. Angelini, F. Brugnara, D. Falavigna, D. Giuliani, R. Gretter e M. Omologo, "A Baseline of a Speaker Independent Continuous Speech Recognizer of Italian", *Proceedings of EUROSPEECH 93*, Berlin, Germany, 1993.

[9] A.J. Fourcin, G. Harland, W. Barry e V. Hazan, Eds., *Speech Input and Output Assessment, Multilingual Methods and Standards*, Ellis Horwood Books in Information Technology, 1989.

[10] M. Federico, D. Giordani, and P. Coletti, "Development and evaluation of an Italian broadcast news corpus", *Proceedings of 2nd International Conference on Language Resources and Evaluation*, Athens, Greece, 2000.

[11] **MBROLA**: The MBROLA Project.
WWW page: http://tcts.fpms.ac.be/synthesis/.

[12] **MPIRO**: *Multilingual Personalized Information Objects*. European Project IST-1999-10982 Version : 5.
WWW page: http://www.ltg.ed.ac.uk/mpiro/