RECENTI SVILUPPI DI FESTIVAL PER L'ITALIANO

Piero Cosi*, Roberto Gretter** e 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

Recentemente è stata resa disponibile la prima versione di ***FESTIVAL***[1] per l'italiano e in questo lavoro sono descritti gli ultimi, e più recenti, sviluppi del sistema.

Assieme ad una descrizione riassuntiva dell'architettura del sistema, sono presentate le due voci maschile e femminile, attualmente disponibili in tre differenti motori di sintesi, e le nuove regole di durata, automaticamente apprese, mediante CART-tree, da un corpus di parlato letto (annunci televisivi di notizie o fiabe per bambini). Dopo aver introdotto le linee guida per l'estrazione automatica delle nuove regole di intonazione, che saranno successivamente incluse nel sistema finale, sono illustrati alcuni esempi di sintesi.

2. INTRODUZIONE

Come già descritto in un precedente lavoro [1], il sintetizzatore vocale ***FESTIVAL*** [2] è basato sulla tecnica della concatenazione di unità vocali. L'architettura generale del sistema, illustrata schematicamente in Figura 1, comprende un blocco di Moduli Linguistici responsabili dell' "*analisi testuale e linguistica*" del testo in ingresso e da un blocco di Moduli Fonetico-Acustici responsabili dell' "*analisi prosodica*", intesa come determinazione dell'intonazione e della durata, e della "*generazione del segnale*" che consente, quale ultimo passo, di generare una forma d'onda a partire dalle informazioni linguistiche sopra specificate.

L'architettura dei Moduli Linguistici è illustrata in Figura 2. 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.

[1] Parte di questo lavoro è stato svolto nell'ambito dei progetti di ricerca denominati MPIRO (Multilingual Personalized Information Objects. European Project IST-1999-10982 - WWW page: http://www.ltg.ed.ac.uk/mpiro/), finanziato dalla Comunità Europea, e TICCA (Tecnologie cognitive per l'interazione e la cooperazione con agenti artificiali), finanziato dal CNR e dalla Provincia Autonoma Trentina.

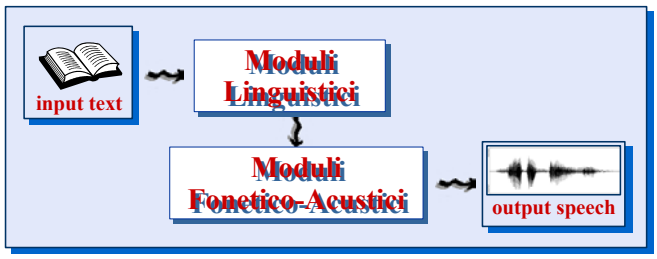

Figura 1. Architettura generale di un sintetizzatore da testo scritto.

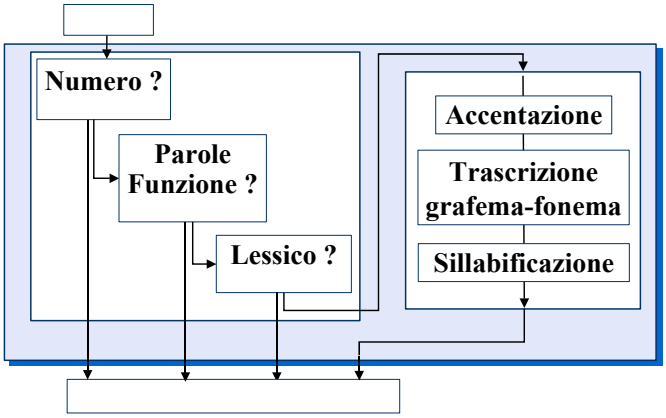

Figura 2. Architettura dei Moduli Linguistici per l'analisi del testo, l'accentazione, la trascrizione grafema-fonema e per la sillabificazione.

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".

L'architettura dei Moduli Fonetico-Acustici è illustrata in Figura 3. A partire dalla stringa fonetica sin qui ottenuta sono selezionate le corrispondenti unità acustiche, nel nostro caso, i difoni, e per ognuna di esse è aggiunta l'informazione riguardante la durata e

la frequenza fondamentale. 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",) o la sintesi MBROLA* [3].

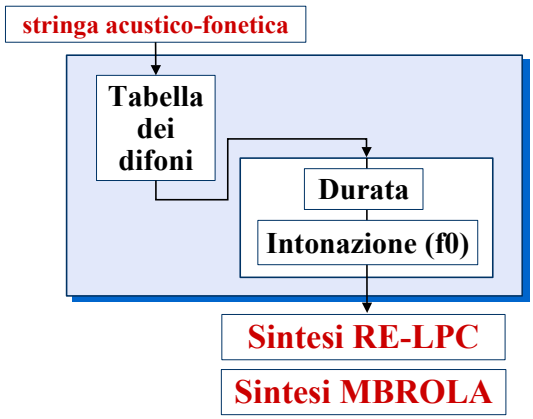

Figura 3. Architettura dei Moduli Fonetico-Acustici per l'assegnazione delle regole di durata e intonazione (f0) e per la generazione della forma d'onda mediante sintesi LPC o MBROLA [3].

3. ANALISI PROSODICA: Durata e Intonazione (f0)

Attualmente, non disponendo di un analizzatore sintattico complesso, l'analisi prosodico-intonativa, intesa come determinazione della durata e dell'intonazione (f0, frequenza fondamentale) in corrispondenza dei difoni da sintetizzare, è sicuramente la parte più debole dell'attuale sistema di sintesi.

3. 1 Metodo 1 (Regole Esplicite)

In una prima fase di sviluppo [1] i moduli prosodici erano molto semplici e basati interamente su regole esplicite sia per la durata che per l'intonazione come schematizzato in Figura 4.

Ai singoli fonemi viene assegnata una durata media elaborata analizzando statisticamente un corpus di frasi pronunciate da alcuni annunciatori RAI [4]. 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 è aumentata del 20% rispetto a quella standard.

Per le frasi "*dichiarative*" per ogni "*gruppo intonativo*", viene generata per f0 una "*linea di base*" che inizia a 140Hz e arriva fino 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". Per quanto riguarda le frasi "interrogative" 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.

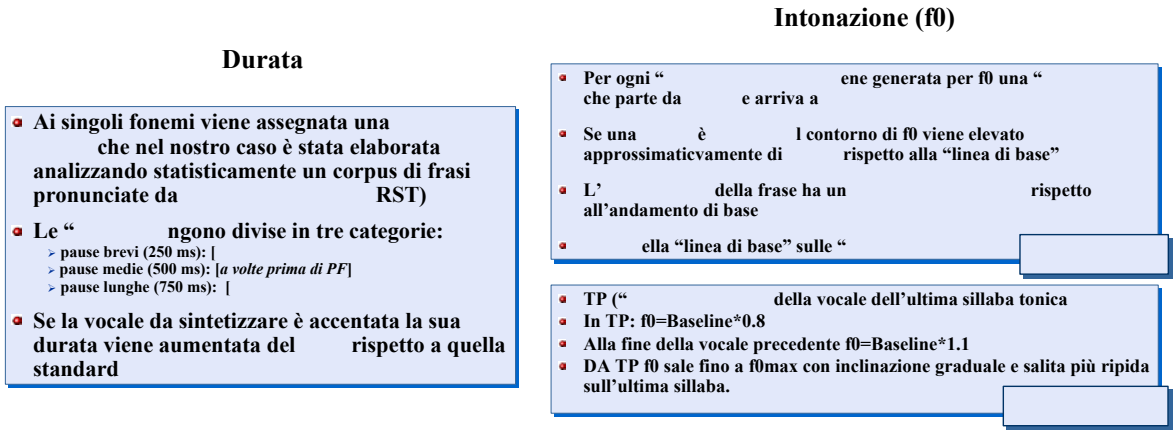

Figura 4. Regole esplicite per la durata e l'intonazione (f0).

3. 2 Metodo 2 (CART-tree)

L'assegnazione della durata è effettuata non più per regole esplicite ma mediante un'approccio statistico basato sulla teoria dei CART (Classification and Regression Trees) [5]. Un CART non è altro che un metodo statistico per ricercare ed isolare sottoinsiemi di dati o relazioni, simili fra loro, all'interno di più complesso insieme di dati. In un CART ad ogni nodo devono essere soddisfatte determinate condizioni e nel nostro caso queste condizioni sono di tipo "***binario***", come illustrato graficamente in Figura 5, e sono relative alle caratteristiche dei fonemi in esame.

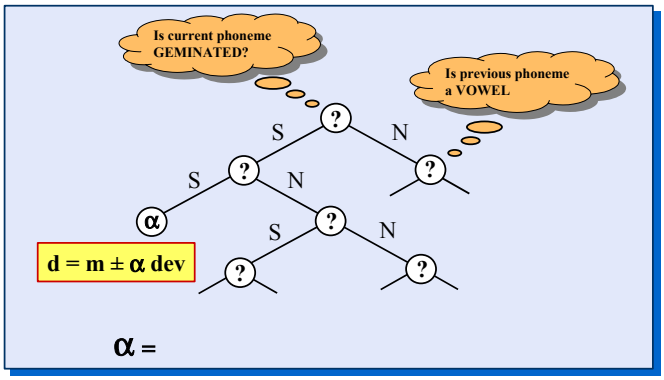

Figura 5. Rappresentazione grafica del metodo statistico denominato CART-tree.

Un insieme di regole CART, sia per la durata che per l'intonazione (f0), sono state apprese automaticamente su un corpus di "parlato naturale" (annunci televisivi di notizie o

fiabe per bambini) mediante il modulo denominato "***wagon***" del software ***Edinburgh Speech Tools*** [6]. Per un trattamento più omogeneo dei dati, per rappresentare i parametri "durata" e "intonazione (f0) sono state utilizzate le rispettive deviazioni standard dalla media per ogni fonema e nel modello il CART predice lo "***zscore***" cioè il numero di deviazioni standard dalla media, piuttosto che i parametri durata o f0 direttamente. Il programma wagon costruisce il CART a partire da un determinato vettore di "***feature***" illustrate in Figura 6 rispettivamente per la durata e l'intonazione (f0).

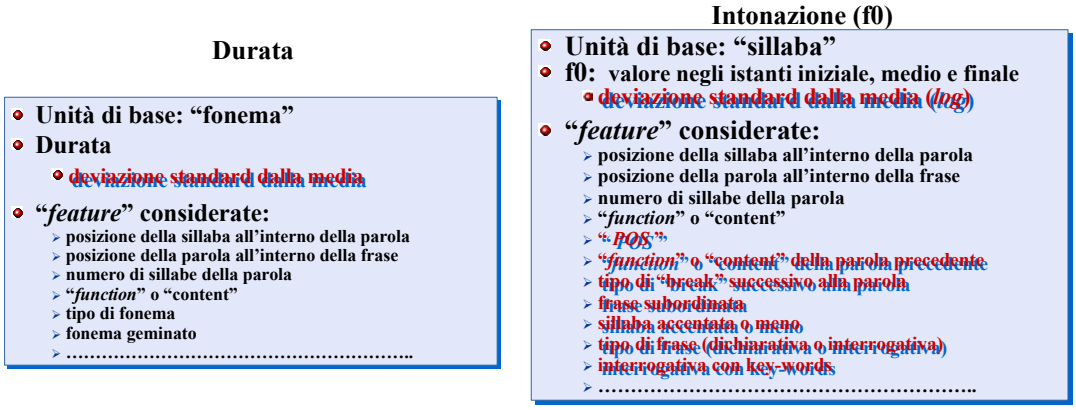

Figura 6. Parametri e "*feature*" utilizzati per l'analisi con il metodo statistico CART-tree per la durata e l'intonazione (f0).

4. CONCLUSIONI

Alcuni test preliminari hanno dimostrato una maggior naturalezza nella sintesi utilizzando il nuovo metodo statistico rispetto al primo metodo deterministico, anche se ben più esaustive prove d'ascolto dovranno essere effettuate per poter affermare con certezza queste conclusioni.

Uditivamente, inoltre, è facilmente riscontrabile una differenza di stile, se in fase di sintesi si utilizzano le regole prosodiche apprese in corrispondenza dei due CART allenati uno sul corpus di materiale vocale relativo ad annunci televisivi, di tipo quindi giornalistico, e l'altro sulle fiabe per bambini.

E' poi inportante notare che il materiale vocale per l'apprendimento dei CART non è stato ancora etichettato "prosodicamente", ad esempio mediante sistemi di tipo ToBi [7]. Si presume quindi che gli eventuali CART allenati includendo anche queste informazioni dovrebbero portare a sempre più precisi andamenti intonativi.

5. SVILUPPI FUTURI

Oltre alla registrazione di nuove voci di bambino e bambina, si cercherà di sviluppare o interfacciarsi a nuovi moduli linguistici quali, ad esempio, un analizzatore morfologico, grammaticale, sintattico e semantico, al fine di migliorare la prosodia con nuove e più specifiche regole di durata e intonazione. Sempre con quest'obiettivo, si cercherà inoltre di identificare le eventuali correlazioni fra le informazioni linguistiche e i contorni intonativi. Inoltre, nell'ambito del progetto europeo MPIRO ([8]), 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] Cosi P., Gretter R., Tesser F., "FESTIVAL parla italiano!". *Atti XI Giornate di Studio del G.F.S.*, Padova, Italy, November 29-30, Dicember 1, 2000, pp. 235-242.

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

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

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

[5] Breiman L., Friedman J., Stone C.J., Olshen R.A., "*Classification and Regression Trees*". Chapman & Hall/CRC, 1984.

[6] *The Edinburgh Speech Tools Library*. http://www.cstr.ed.ac.uk/projects/speech_tools/

[7] Beckman M., Ayers E.G., "*Guidelines for ToBI Labelling*". Version 3. Ohio State University. http://ling.ohio-state.edu/Phonetics/E_ToBI/etobi_homepage.html

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