MODELLIZZAZIONE AUTOMATICA DELL'INTONAZIONE PER UN SISTEMA DI SINTESI DA TESTO SCRITTO IN ITALIANO

Piero Cosi, Cinzia Avesani

ISTC-SPFD
Istituto di Scienze e Tecnologie della Cognizione- Sezione di Padova
"Fonetica e Dialettologia"
Consiglio Nazionale delle Ricerche
e-mail: {cosi, avesani}@csrf.pd.cnr.it
www: http://www.csrf.pd.cnr.it/

Fabio Tesser, Roberto Gretter, Nadia Mana, Fabio Pianesi

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

SOMMARIO

In questo lavoro, viene illustrata una versione leggermente modificata del modello denominato PaIntE (*Parametric Representation of Intonation Events*), basato su una speciale tecnica di parametrizzazione di F0. I parametri del modello sono stati automaticamente ottimizzati utilizzando un insieme di frasi italiane "prosodcamente ricche" ed etichettate secondo la convenzione ToBI.

Questo metodo sarà utilizzato in futuro per comandare il modello intonativo della versione italiana di FESTIVAL. L'affidabilità del modello è stata testata mediante alcuni test preliminari che hanno mostrato risultati assai promettenti.

1. INTRODUZIONE

La maggior parte delle teorie intonative ipotizzano che l'intonazione possa essere modellata mediante un insieme di differenti "entità" fonologiche realizzate acusticamente come differenti andamenti di F0.

Il modello "*tone sequence*" (TSM) introdotto da Pierrehumbert (Pierrehumbert, 1980), e la corrispondente convenzione di trascrizione ToBi (Silverman et alii, 1992), mediante la quale il contorno intonativo viene descritto come una sequenza di toni alti (H) e bassi (L) assieme a specifici indicatori di zone di frontiera fra successivi contorni intonativi, rappresentano l'esempio più significativo di una tale descrizione fonologica dell'intonazione. La cosiddetta "*British School*", partendo da un differente punto di vista, enfatizza al contrario i movimenti di F0 invece dei corrispondenti target intonativi, (Halliday, 1967), e l'insieme di questi movimenti consiste di andamenti di salita, di discesa, di salita e discesa e di loro eventuali combinazioni. Anche se descrivono l'intonazione in due modi differenti, entrambe le teorie condividono, però, il

metodo *composizionale* di descrizione dell'intonazione. Esse, infatti, combinano un numero distinto di elementi di base per la costruzione del contorno intonativo globale.

Gli approcci basati direttamente sui dati, al contrario delle teorie classiche sull'intonazione, spesso utilizzano invece per la descrizione dei contorni di F0 dei parametri globali continui. Questo è principalmente dovuto a ragioni pratiche e cioè al fatto che le funzioni utilizzate per l'approssimazione del contorno intonativo sono realizzate mediante parametri continui. I contorni di F0 sono, infatti, approssimati mediante specifiche funzioni ottenute variando un set di *n* parametri continui e il risultato, rappresentato da un vettore *n-dimensionale*, è la caratterizzazione del sottostante andamento intonativo. Il modello TILT (Taylor & Black, 1994; Taylor, 1988), ad esempio, ben rappresenta un tale approccio basato direttamente sui dati intonativi. Questi modelli parametrici sono estremamente validi nel rappresentare l'intonazione poiché simulano in modo efficace i veri movimenti di F0 e possono, inoltre, assegnare specifici contenuti prosodici ai loro parametri. Si può anche dire che questi modelli cercano di catturare solo le informazioni relative all'intonazione cercando di scartare tutte le informazioni ridondanti.

Differentemente dagli altri modelli intonativi, i due principi contraddittori espressi dalle teorie classiche e da quelle basate su dati, sono stati incorporati simultaneamente nel modello denominato *PaIntE* (*Parametric Representation of Intonation Events*) (Dusterhoff & Black, 1997; Mohler & Conkie, 1998). Questo modello, infatti, utilizza, per la corretta rappresentazione del contorno intonativo, 6 parametri continui determinati approssimando la curva di F0 mediante un'apposita funzione estesa a tre sillabe adiacenti e focalizzata in istanti specifici caratterizzati o da sillabe accentate o da opportune etichette ToBI (Möhler, Draft).

2. IL MODELLO "*PaIntE*"

Il modello *PaIntE* (*Parametric Representation of Intonation Events*), come illustrato in Figura 1, approssima F0 mediante una particolare funzione estesa su tre sillabe adiacenti di cui quella centrale è rappresentata o da una sillaba accentata o da un particolare target TOBI. Questa funzione è costituita dalla somma di due sigmoidi caratterizzate da un ritardo temporale fisso, da un limite superiore comune e da un parametro di allineamento temporale costante γ. In pratica è possibile specificare il livello, l'intensità, l'allineamento e la pendenza delle due sigmoidi mediante l'utilizzo di 6 parametri continui ("*intonation event parameters*"):

$$f(x) = d - \frac{c_1}{1+\exp(-a_1(b-x)+\gamma)} - \frac{c_2}{1+\exp(-a_2(x-b)+\gamma)} \qquad (1)$$

con:

- ***a1***, ***a2***: pendenze delle sigmoidi "ascendente" e "discendente"
- ***b***: allineamento della funzione (la lunghezza della sillaba è normalizzata ad 1)
- ***c1***, ***c2***: ampiezze delle sigmoidi "ascendente" e "discendente"
- ***d***: frequenza del picco della funzione.

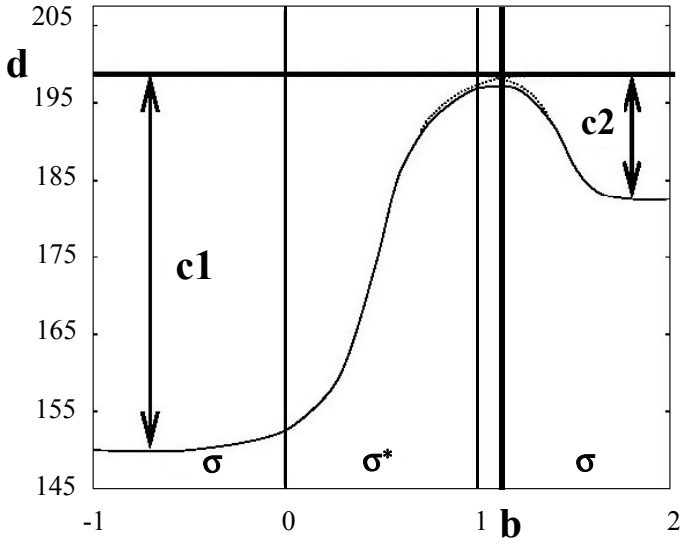

Figura 1: La funzione del modello PaIntE rappresentata dalla soma di una sinusoide ascendente e di una discendente con un ritardo temporale fisso (l'asse temporale è normalizzato ad 1 sulla lunghezza delle sillabe).

Per *decorrelare*, inoltre, la forma della curva di F0 dalla velocità di eloquio e dal particolare contesto sillabico l'asse delle ascisse è normalizzato ad 1 rispetto alla durata delle sillabe. Nella nostra implementazione il modello originale è stato modificato per rappresentare più efficacemente, oltre ai comuni andamenti di forma a "*picco*" tipici della metodologia di etichettatura ToBI, anche andamenti di forma a "*valle*", mediante l'introduzione di valori negativi per le costanti *c1* e *c2* in (1). Questo accorgimento si è rivelato assai utile, soprattutto per l'italiano, nei casi in cui la forma specifica di alcuni toni ToBI, come ad esempio L* o qualche tono di confine (Avesani, 1995), siano meglio rappresentati da una tale tipologia di andamento. I 6 parametri in (1) assumono così significati differenti a seconda della forma scelta per la funzione: "picco" (*c1*, *c2* > 0) o "valle" (*c1*, *c2* < 0) come illustrato in Figura 2.

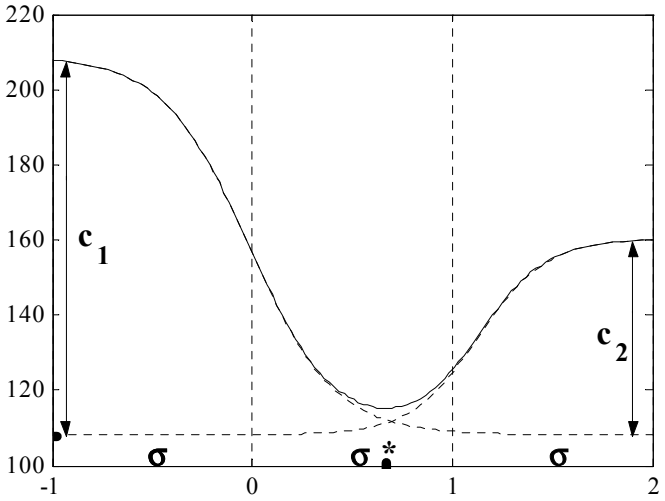

Figura 2: La funzione del modello PaIntE in una tipica configurazione di forma a "*valle*".

Seguendo il suggerimento di Möhler (Möhler, Draft), viene applicata, in oltre, una normalizzazione di F0 ad ogni singola frase, allo scopo di filtrare ed eventualmente eliminare completamente l'influenza del diverso livello del valore della frequenza fondamentale media di ogni parlante sulla forma della funzione. Considerando sia la funzione di forma a *picco* che quella di forma a *valle* i limiti massimi e minimi del livello di F0 sono definiti da:

$$\begin{aligned} UL &= \max [\max d, \max(d-c1), \max(d-c2)] \\ LL &= \min [\min d, \min(d-c1), \min(d-c2)] \end{aligned} \qquad (2)$$

Il modello è stato stimato sia *senza* che *con* le informazioni ToBi. Nel primo caso le sillabe "target" sono tutte quelle accentate di parole *piene*, escluse quindi tutte le parole *funzione*, e quelle in posizione finale di frase.

3. I DATI

Per la stima e l'ottimizzazione dei parametri del modello PaIntE, nel caso in cui non si sono utilizzate le informazioni della trascrizione ToBI (*PaIntE-NT*), quindi considerando esclusivamente le informazioni sull'accento lessicale, sulle parole funzione e sulla punteggiatura, è stato utilizzato un corpus di notizie televisive pronunciate da un annunciatore di una rete televisiva nazionale (*RAI-news*) (Federico et alii, 2000) composto da 558 frasi (~11500 parole), mentre nel caso ToBI (*PaIntE-T*), è stato considerato solo un piccolo sottoinsieme dello stesso corpus.

4. METODO

Tutte le frasi sono state automaticamente allineate alla loro trascrizione fonetica mediante una procedura di allineamento automatico appositamente sviluppata sulla base di un sistema di riconoscimento fonetico "*general-purpose*" per l'italiano (Cosi et alii, 2000) realizzato ed allenato mediante il corpus APASCI (Angelici et alii, 1993) utilizzando l'ambiente software CSLU Speech Toolkit (Sutton , 1998).

La frequenza fondamentale (*F0*), calcolata mediante *PRAAT* (Boersma e Weenink, 2000) è stata interpolata nelle porzioni di segnale non vocalizzato ($F0_I$) e *smussata* mediante un filtro passa-basso a 20Hz ($F0_{IS}$). Per la procedura automatica di ottimizzazione è stata infine considerata una finestra di tre sillabe ($F0_{ISW}$) centrata, su ogni sillaba target corrispondente a tutte le sillabe accentate di parole *piene*, escluse quindi tutte le parole *funzione*, e a quelle in posizione finale di frase, oppure a quelle caratterizzate da un etichetta ToBI, a seconda del caso in esame. E' stata poi effettuata una normalizzazione temporale ad 1 sulla lunghezza di ogni sillaba al fine di eliminare le influenze della diversa velocità di eloquio ($F0_{ISWN}$). Questo segmento di F0 è stato poi approssimato mediante la funzione del modello *PaIntE* introdotta in (1) e per questo è stato utilizzato un metodo iterativo di approssimazione del *gradiente*, secondo la formula:

$$\min_{d,b,a_1,c_1,a_2,c_2} \left\{ \frac{1}{2} \sum_{x\in[-1,2]} p(x)^2 [f(x) - F_{0-ISWN}(x)]^2 \right\}, \qquad (3)$$

per stimare i parametri del modello PaIntE che meglio ricostruiscono l'andamento reale di F0 all'interno della finestra temporale considerata. Come indicato in (3) si è introdotta, inoltre, una funzione peso di forma triangolare $p(x)$, centrata su ogni punto "*critico*" utilizzato nella procedura di ottimizzazione, di ampiezza proporzionale all'importanza del punto stesso e estensione anch'essa proporzionale alla sua corrispondente zona di influenza.

Come graficamente illustrato nell'esempio di Figura 3, in un tipico caso L+H*, sono stati scelti per la procedura di ottimizzazione, in modo euristico, un massimo di 8 punti critici i quali sono stati posizionati su:

- gli eventi ToBI (nei casi di eventi bi-tonali ToBI come ad esempio L+H* sono stati considerati 2 punti critici);
- gli istanti di confine della finestra in esame (-1, e 2 nelle Figure 1,2 e 3);
- la posizione del valore massimo (*F0max*) e minimo (*F0min*) di F0 per ogni sillaba;
- la posizione del valore massimo ($\Delta F0max$) e minimo ($\Delta F0min$) della derivata di F0 per ogni sillaba.

La procedura di ottimizzazione è stata eseguita due volte considerando due diversi insiemi di valori di inizializzazione di alcuni punti critici corrispondenti ad una configurazione di tipo "picco" e ad una di tipo "valle", come indicato in (3).

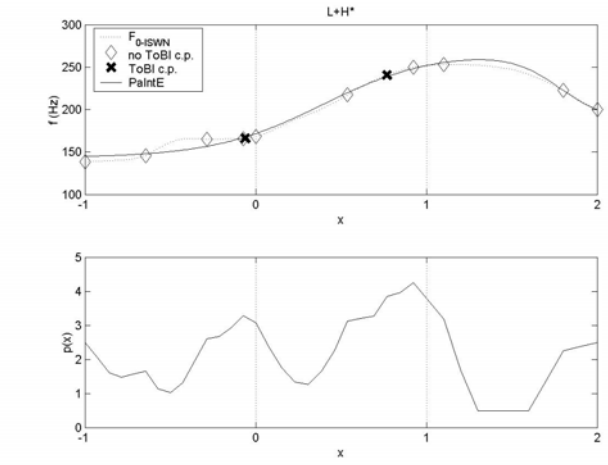

Figura 3: Applicazione della procedura di ottimizzazione del modello PaIntE nel caso di un tono ToBI complesso L+H*; in basso è illustrata la corrispondente funzione peso p(x) (PaIntE: [a.,b,a1,c1,a2,c2] = [268.3668, 1.350618, 3.126952, 125.1043, 7.031559, 81.88443]).

La ricostruzione globale della curva di F0 è stata poi effettuata interconnettendo tutti gli eventi intonativi stimati dal modello PaIntE nediante una semplice interpolazione lineare. Un esempio della ricostruzione di una curva di F0 mediante il modello PaIntE, nel caso di una semplice frase dichiarativa in italiano, è illustrato in Figura 4.

Per valutare la qualità e la precisione del modello (*PaIntE-NT)* nel ricostruire l'andamento di F0, su tutto il corpus *RAI-news* non etichettato mediante ToBI, è stato calcolato l'errore quadratico medio e il corrispondente valore di correlazione fra l'andamento originale di F0 e quello ricostruito dal modello. Come illustrato in Tabella 1, il valore medio di questi due indici per un totale di 558 frasi risulta assai promettente.

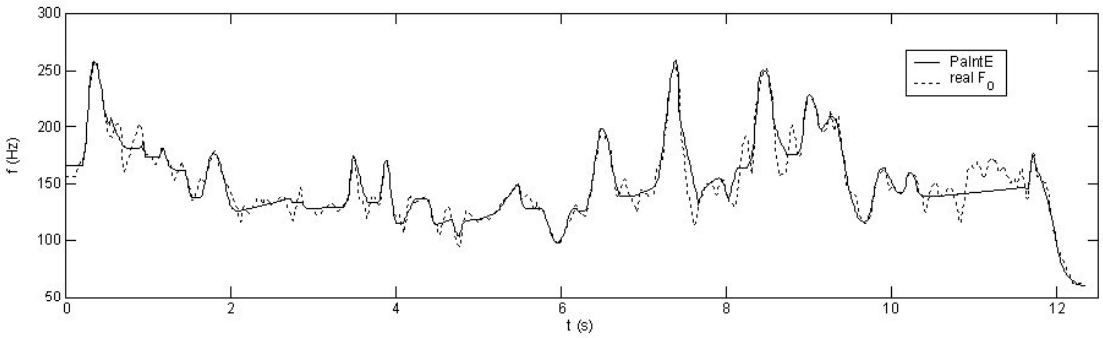

Figura 4. Esempio dell'applicazione del modello PaIntE-T ad una semplice frase dichiarativa in italiano.

Un ulteriore raffronto tra i contorni originali e quelli ricostruiti di F0 è stato effettuato poi solo su tre frasi dello stesso corpus *RAI-news* per entrambe i modelli *PaIntE-NT* e *PaIntE-T* precedentemente descritti. In Tabella 2, si può notare che in questo caso il nodello *PaIntE-T* si comporta notevolmente meglio del corrispondente *PaIntE-NT*.

RAI-news Corpus	*PaIntE-NT* **(no ToBI)**	
	RMSE	Correlazione
(558 frasi)	**12.764**	**0.8845**

Tabella 1: RMSE e correlazione fra andamenti originali e ricostruiti dal modello *PaIntE-NT* su tutto il corpus *RAI-news* senza utilizzare alcuna informazione sulla trascrizione ToBI.

	PaIntE-T **(ToBI)**		*PaIntE-NT* **(No-ToBI)**	
frase	RMSE	Correlazione	RMSE	Correlazione
f0002	9.15	0.964	9.87	0.959
f0004	6.76	0.968	12.96	0.879
f0007	6.46	0.976	11.82	0.917
media	**7.46**	**0.969**	**11.55**	**0.918**

Tabella 2: RMSE, correlazioni e valori medi degli andamenti originali di F0 e di quelli ricostruiti rispettivamente dai modelli *PaIntE-T* (ToBI) e *PaIntE-NT* (No-ToBI) per 3 frasi del corpus *RAI-news*.

Questi valori numerici, tuttavia, non prendono in considerazione le reali differenze percettive fra gli andamenti originali e quelli ricostruiti e quindi in futuro sarà necessario progettare degli specifici test percettivi appositamente studiati per questo scopo. Nei testi preliminari che sono stati eseguiti le differenze percettive fra le frasi originali e quelle ricostruite mediante l'andamento stimato dal modello PaIntE per mezzo di una semplice tecnica di sintesi per sovrapposizione e somma di finestre di segnale denominata PSOLA (Dutoit, 1993) e implementata in *PRAAT* (Boersma e Weenink, 2000), sono realmente molto piccole e si può ritenere che questo risultato sia dovuto al fatto che la parametrizzazione del modello viene eseguita solo nei punti percettivamente più importanti tralasciando quindi le zone meno influenti a livello percettivo.

5. OSSERVAZIONI CONCLUSIVE

Il modello *PaIntE* modificato descritto in questo lavoro sembra qualitativamente appropriato ed efficace nel descrivere e ricostruire gli andamenti di F0 per l'italiano.

Si può ritenere che la tipica configurazione di tipo a "*valle*" e a "*picco*" delle funzioni utilizzate per implementare il modello ricostruiscano in modo efficace gli andamenti di F0 e questo è dovuto principalmente al fatto che, nella procedura di ottimizzazione, sono stati scelti solo alcuni punti percettivamente rilevanti del segnale.

I risultati numerici delle prime valutazioni qualitative sembrano assai promettenti e ci consentono di dire che i risultati migliorano sensibilmente qualora l'informazione prosodica (ToBI) sia considerata nella procedura di stima ed ottimizzazione del modello. Per una valutazione più completa e precisa sono ovviamente necessari altri esperimenti in cui si possano mettere a confronto differenti configurazioni dell'algoritmo di ottimizzazione quali ad esempio quelle corrispondenti ad una scelta diversa dei punti critici da considerare nella procedura.

6. SVILUPPI FUTURI

Per meglio comprendere e paragonare l'importanza delle informazioni prosodiche al fine di ricostruire efficacemente l'andamento di F0 (*PaIntE-NT* vs *PaIntE-T)* sarà utilizzato ulteriore materiale etichettato ToBI e, in particolare, sarà considerato un corpus di alcune novelle di Dino Buzzati, un noto scrittore italiano, pronunciate da un attore professionista (Avesani et alii, 2003).

Inoltre, in seguito all'osservazione dell'esistenza di configurazioni standard degli andamenti di F0 in corrispondenza di toni ToBI simili, come illustrato in Figura 5, e fortemente motivati da alcune teorie intonative che ci suggeriscono che gli accenti e i fenomeni di confine intonativi possono essere descritti mediante un discreto e ridotto numero di pattern, risulta logico ipotizzare che alcuni tipici pattern intonativi possano essere facilmente suddivisi in differenti categorie di riferimento, e questo può senza dubbio essere esplorato mediante l'utilizzazione di alcune tecniche statistiche, come ad esempio la quantizzazione vettoriale che potrebbero rivelarsi assai promettenti in questo campo di ricerche sulla modellizzazione prosodica.

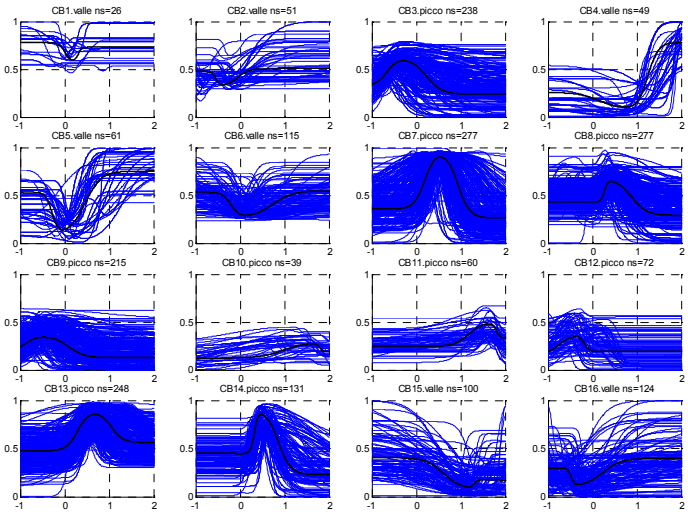

Figure 5. Configurazioni di F0 ricostruite dal modello PaIntE in vari esempi di toni ToBI in italiano.

Inoltre, per uno stesso tono ToBI, gli andamenti di F0 risultano essere assai simili e questo supporta l'idea che questi pattern possano facilmente essere appresi mediante opportune procedure statistiche denominate CART *Classification and Regression Trees* (Breiman et alii, 1984). Per questi motivi sia la quantizzazione vettoriale (VQ) sia i CART saranno considerati negli sviluppi di questo lavoro e qualora le loro funzionalità risultino efficaci saranno inclusi nel modello prosodico della versione finale del sistema di sintesi da testo scritto denominato FESTIVAL (Cosi et alii, 2001), di cui recentemente è stata sviluppata la nuova versione per l'italiano.

7. RINGRAZIAMENTI

Parte di questo lavoro è stata possibile grazie alle attività sviluppate nell'ambito dei progetti: MPIRO (Multilingual Personalized Information Objects, European Project IST-1999-10982, http://www.ltg.ed.ac.uk/mpiro/), TICCA (Tecnologie cognitive per l'interazione e la cooperazione con agenti artificiali, progetto congiunto fra il CNR e la Provincia Autonoma Trentina), e PF-STAR (Preparing Future multiSensorial inTerAction Research, European Project IST- 2001-37599, http://pfstar.itc.it/).

BIBLIOGRAFIA

Angelini B., Brugnara F., Falavigna D., Giuliani D., Gretter R., Omologo M. (1993), A Baseline of a Speaker Independent Continuous Speech Recognizer of Italian, *in Proceedings of EUROSPEECH 93*, Berlin, Germany, 1993.

Avesani, C. (1995), ToBIt: un sistema di trascrizione per l' intonazione italiana, in *Atti delle 5e Giornate di Studio del Gruppo di Fonetica Sperimentale (A.I.A.)*, 1995, Povo (TN), Italy, pages 85-98.

Avesani C., Cosi P., Fauri E., Gretter R., Mana N., Rocchi S., Rossi F., Tesser F (2003), Definizione ed annotazione prosodica di un database di parlato-letto usando il formalismo ToBI, in *Atti de Il Parlato Italiano*, Napoli 13-15, febbraio, 2003 (questo volume).

Boersma, P., Weenik, D. (2000). Praat, a system for doing phonetics by computer, version 3.4. *Technical Report 132*, *Institute of Phonetic Sciences of the University of Amsterdam*, www.praat.org.

Breiman L., Friedman J., Stone C.J., Olshen R.A. (1984), *Classification and Regression Trees*, Chapman & Hall/CRC, 1984.

Cosi P., Hosom J.P., High Performance (2000), "General Purpose" Phonetic Recognition for Italian, in *Proceedings of ICSLP-2000*, Beijing, Cina, 16-20 October, 2000, Vol. II, pages 527-530.

Cosi P., Tesser F., Gretter R., C. Avesani C. (2001), Festival Speaks Italian!, in *Proceedings of EUROSPEECH 2001*, Aalborg, Denmark, Sep 3-7 2001, pages 509-512.

Dusterhoff K., Black A.W. (1997), Generating F0 Contours for Speech Synthesis Using the Tilt Intonation Theory, in *Proceedings of ESCA Workshop on Intonation*, Athens, Greece, 1997.

Dutoit T., Leich H. (1993), MBR-PSOLA: Text-To-Speech Synthesis based on an MBE Re-Synthesis of the Segments Database, *Speech Communication*, Elsevier Publisher, December 1993, vol. 13, n° 3-4, pp. 435-440.

Federico M., Giordani D., Coletti P. (2000), Development and Evaluation of an Italian Broadcast News Corpus, in *Proceedings of LREC-2000*, Athens, Greece, 2000.

Halliday M.A.K. (1967), *Intonation and grammar in British English*, Mouton, The Hague, 1967.

Mohler G., Conkie A. (1998), Parametric Modeling of Intonation Using Vector Quantization, in *Prooceedings of Third International Workshop on Speech Synthesis*, Jenolan Caves, Australia, 1998.

Möhler G. (Draft), Improvements of the PaIntE Model for F0 Parametrization. Research Papers, Draft version, from the Phonetics Lab, AIMS Universität Stuttgart, Institut für Maschinelle Sprachverarbeitung (to appear).

Pierrehumbert J. (1980), *The Phonology and Phonetics of English Intonation*, PhD thesis, MIT, Cambridge, MA,1980.

Silverman K., Beckman M., Pitrelli J., Ostendorf M., Wightman C., Price P., Pierrehumbert J., Hirschberg J. (1992), ToBI: a Standard for Labeling English Prosody, in *Proceedings of ICSLP 1992*, Vol 2, pages 867-870.

Sutton S., Cole R., Villiers J., Schalkwyk J., Vermeulen P., Macon M., Yan Y., Kaiser E., Rundle B., Shobaki K., Hosom P., Kain A., Wouters J., Massaro D., Cohen M. (1998), Universal speech tools: the CSLU toolkit, in *Proceedings of ICSLP-98*, Sydney, Nov 30-Dec 4, 1998, Vol. 7, pp. 3221-3224.

Taylor P. (1988), The Tilt Intonation Model, in *Proceedings of ICSLP 1998*, Sydney Australia, 30th Nov-4th Dec 1998, Paper 827, Vol. IV, pp. 1383-1386.

Taylor P., Black A.W. (1994), Synthesizing conversational intonation from a linguistically rich input, in *Proceedings of ESCA Workshop on Speech Synthesis*, Mohonk, NY, 1994, pages 175-178.