MODELLI PROSODICI EMOTIVI PER LA SINTESI DELL'ITALIANO

Fabio Tesser^, Piero Cosi*, Carlo Drioli*, Graziano Tisato*

^Centro per la Ricerca Scientifica e Tecnologica, ITC-IRST, Trento.
* Istituto di Scienze e Tecnologie della Cognizione, Sezione di Fonetica e Dialettologia, ISTC-CNR, Padova.

tesser@itc.it
{cosi, drioli, tisato}@pd.istc.cnr.it

SOMMARIO

È noto come, nel campo della sintesi vocale *Text To Speech* (TTS), si utilizzano due approcci per l'implementazione di regole prosodiche efficaci: la tecnica *rule-based* oppure la tecnica *data-driven*. La prima risulta essere poco naturale e molto laboriosa, poiché le regole devono essere dedotte dalla sola informazione fornita dal testo in ingresso. La generazione *data-driven* della prosodia è un approccio alternativo che ha il vantaggio di essere più espressiva e di facilitare il compito soprattutto se si vogliono creare differenti set di stili prosodici.

In questa comunicazione s'illustrerà la generazione *data-driven* della prosodia basata su alberi di decisione, utilizzando un database non emotivo. In particolare sarà illustrata un'estensione alla procedura degli alberi di classificazione nello spazio degli eventi intonativi quantizzati: VQ-PaIntE (*Vector Quantization - Parametric Intonation Events*).

L'aspetto rilevante di questo lavoro è che si è utilizzato un approccio differenziale nella predizione della prosodia emotiva: il modulo prosodico emotivo implementato all'ISTC cerca di "imparare" le differenze tra la prosodia neutra (senza emozioni) e i dati prosodici corrispondenti alle emozioni. Un'attenzione particolare è stata posta nella scelta del dominio nel quale eseguire questa differenza. Nell'articolo sarà anche mostrato come l'applicazione della PCA (*Principal Components Analysis*) semplifica di molto gli alberi di decisione risultanti per modellare l'intonazione.

1. INTRODUZIONE

Si è assistito negli ultimi anni ad un costante tentativo di migliorare l'efficacia e la naturalezza dell'interazione uomo-macchina con la simulazione di caratteristiche espressive ed emotive tipiche della comunicazione umana.

Mentre le attuali tecnologie di sintesi della voce riescono facilmente a produrre un segnale vocale intelligibile, ad es. (Balestri *et alii*, 1999), difficilmente si può sostenere che la voce prodotta da un TTS sia espressiva.

Una sfida molto ambiziosa in questo campo è dunque quella di aggiungere al sistema TTS la possibilità effettiva di simulare una intenzione espressiva od emotiva.

Per ottenere questo scopo, un sintetizzatore vocale deve agire sui correlati acustici delle emozioni, che sono, a livello sovrasegmentale, i parametri prosodici fondamentali (*pitch*, durata, intensità), e, a livello fonetico, parametri timbrici che si possono riassumere sotto la denominazione di *qualità della voce* (*voice quality*). Quest'ultima proprietà distingue le

modalità con cui viene prodotto il segnale glottale (voce aspirata, soffiata, tesa, ecc.), ed è in relazione con lo stato emotivo ed anche le patologie della voce.

I sistemi di sintesi vocale basati sul modello sorgente-filtro offrono una notevole possibilità di controllo sia dello spettro glottico che delle formanti, ma in compenso ottengono risultati peggiori dei sistemi per concatenazione di difoni o per unità variabili (*corpus-based* o *unit selection*) in termini di naturalezza.

Nel campo della voce emotiva, d'altra parte, i sistemi ad unità variabili devono assicurare la presenza delle unità fonetico-prosodiche che cambiano da emozione ad emozione. A meno di eseguire una re-sintesi (Zovato *et alii*, 2003) delle unità, questa tecnologia richiede un database per ogni emozione (ognuno dei quali può raggiungere una dimensione anche 50 volte superiore a quella dei sistemi per difoni).

L'alternativa è quella di utilizzare un sintetizzatore per difoni che abbia la possibilità di cambiare il timbro e la *voice-quality* dei difoni, e utilizzi dei moduli prosodici appositamente calibrati per le emozioni (Drioli *et alii*, 2003).

Il punto di partenza è un sintetizzatore per la lingua italiana sviluppato sulla piattaforma Festival (Cosi *et alii*, 2000), che utilizza un database di 1300 difoni, sintetizzati con una versione dell'algoritmo MBROLA (Dutoit *et alii*, 1993) esteso per permettere la modifica della *voice-quality* dei difoni.

In questa comunicazione ci si occuperà della prosodia ed in particolare di come ottenere dei buoni moduli prosodici emotivi.

La metodologia *data-driven* per la generazione della prosodia si è dimostrato aderente ai dati reali e adattabile a domini differenti (Tesser *et alii*, 2003) come quello delle emozioni.

Uno dei problemi che bisogna affrontare quando si utilizzano tecniche di *machine learning* è quello del *data-sparseness*: un'accurata scelta dei domini prosodici e la possibilità di raggruppare assieme dati omogenei tra loro è la soluzione utilizzata in questo lavoro.

2. DATABASE

L'approccio *data-driven* cerca di "catturare" la prosodia presente in un database, perciò per ottenere dei buoni risultati con le tecniche *data-driven* è necessario partire da un altrettanto buon database.

Questo implica la necessità che il materiale del database sia coerente con lo stile prosodico che vogliamo riprodurre ed includa una buona varietà di fenomeni linguistici.

Una ulteriore conseguenza è che, al contrario del riconoscimento della voce, i database per l'apprendimento automatico della prosodia contengano la voce di un solo parlatore, che solitamente è un attore o un doppiatore.

3. PROSODIA *DATA-DRIVEN* "NEUTRA"

Nella prima fase si è cercato di generare i moduli prosodici relativi ad uno stile "neutro" ovvero senza emozioni. Il database utilizzato è il CARINI database (Avesani et alii, 2003).

La metodologia e i modelli prosodici sono quelli presentati in (Tesser et alii, 2004), con l'eccezione del modulo intonativo. Infatti, la predizione dei contorni intonativi è il compito più arduo nella generazione automatica della prosodia a causa dell'elevata variabilità della struttura intonativa umana, e quindi è stato necessario migliorare il relativo modello prosodico.

3.1 CARINI database

Il database CARINI è composto dalla registrazione di 3 racconti di Dino Buzzati[1] letti da uno speaker professionista. Il dominio di questo database è di tipo narrativo, e quindi lo stile di lettura scelto dallo speaker è relativamente calmo, rilassato e chiaro. La durata totale del database è di circa un'ora, per un totale di 698 frasi e 7709 parole.

3.2 Quantizzazione vettoriale, classificazione e misura di impurità

L'algoritmo di costruzione dell'albero di decisione dei *cluster* VQ-PaIntE (Möhler et alii, 1998) è stato modificato in modo che la misura di impurità tenga in considerazione la distanza nello spazio dei vettori quantizzati, fornendo maggiori informazioni sulla forma degli eventi intonativi durante la fase di costruzione dell'albero di decisione.

La Figura 1 mostra due *cluster* ricavati dalla quantizzazione vettoriale nello spazio PaIntE.

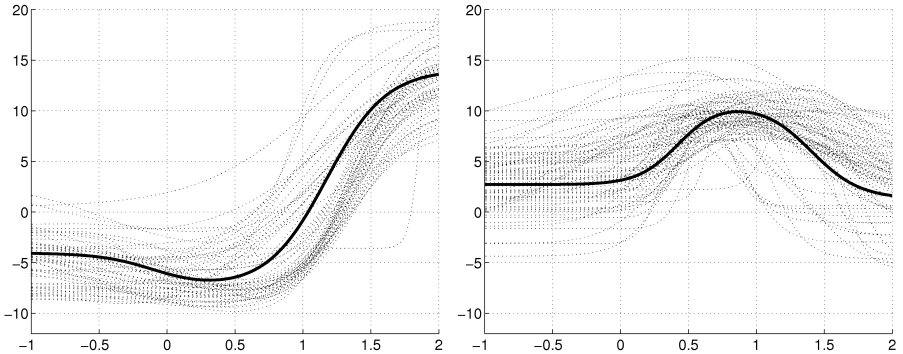

Figura 1: Esempio di quantizzazione vettoriale PaIntE. Le linee tratteggiate rappresentano tutti i pattern PaIntE che sono rappresentati dal centroide PaIntE disegnato con linea continua.

Solitamente l'algoritmo per la creazione dell'albero di classificazione utilizza una misura di impurità[2] del nodo che prende in considerazione l'appartenenza o meno ad una data categoria.

Questo è corretto quando il costo della scelta di una categoria al posto di un'altra è equi-probabile. Nel caso dei vettori PaIntE, le categorie rappresentano un contorno intonativo che può essere più o meno simile ad altri, e quindi differisce dai casi equi-probabili.

Si è quindi utilizzato una misura di impurità che tiene in considerazione la distanza vettoriale tra i vari contorni intonativi: la probabilità della categoria ω_j al nodo t dell'albero viene calcolata utilizzando una misura di similarità (S_{01}) tra i vettori:

$$P_t(\omega_j) = \frac{\sum_{i=1}^{N_t} S_{01}(\mathbf{x}_i, \mathbf{x}_j)}{N_t}$$

[1] I tre racconti sono "Il Colombre", "I sette messaggeri" e "La giacca stregata" di Dino Buzzati.

[2] Solitamente si utilizza l'entropia o la Gini Impurity

Dove N_t è il numero di elementi presenti nel *cluster*. Questa modifica fornisce informazioni più accurate riguardo alla similarità dei contorni di F0 durante la fase di costruzione dell'albero di decisione.

4. EMOZIONI E DATABASE

I dati vocali migliori sui quali studiare le emozioni sono sicuramente quelli registrati durante la naturale occorrenza degli stati emotivi, ma ci sono molti problemi a collezionare una database formato da questo tipo di dati: sono eventi poco frequenti, possono essere registrati male e inoltre molto spesso è difficile determinare l'emozione che è stata espressa. Inoltre nel caso della generazione automatica delle emozioni vi è la necessità di continui e grandi corpus. In questi casi è preferibile ottenere la voce emotiva simulando l'emozione chiedendo a doppiatori o speakers professionisti di produrre l'espressione vocale dell'emozione. In questo lavoro si è utilizzato l' Emotional-CARINI database.

4.1 E-CARINI database

Nel database E-CARINI (Emotional-CARINI) lo speaker legge uno dei racconti usati nel database "neutro" (Il colombre) attuando le sei emozioni di Ekman (Ekman, 1992): disgusto, paura, gioia, rabbia, tristezza, sorpresa.

5. APPROCCIO DIFFERENZIALE

La letteratura che riguarda le emozioni e la voce (Anolli *et alii*, 1997) spesso fa un confronto tra lo stato emozionale e quello "neutro". In questi studi si fa spesso riferimento a espressioni come "bassa velocità di eloquio" o "alto livello di F0" ed è chiaro che questi aggettivi sono paragonati con l'espressione "neutra". Questa è solo una descrizione a livello macro prosodico ma può essere generalizzata a livello segmentale. Prendendo ispirazione da questo si è cercato di trasformare la prosodia "neutra" in quella emotiva, utilizzando sempre le tecniche *data-driven*. Sia per la durata che per l'intonazione è possibile esprimere questa idea con la seguente equazione:

$$\Delta_x = x_E - x_N$$

dove x rappresenta il parametro prosodico (durata o intonazione), x_E rappresenta il valore emotivo, e x_N il valore "neutro". Quello che il modello prosodico deve "imparare" è Δ_x.

In Figura 2 è illustrata la fase di sintesi del parametro prosodico emotivo: le *features* linguistiche strutturali (IDSDATA) sono utilizzate sia dal modulo neutro che da quello differenziale. Per produrre il valore emotivo queste due componenti devono essere sommate.

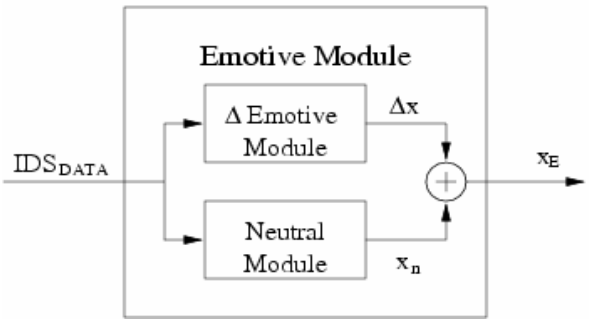

Figura 2: L'approccio differenziale per la generazione della prosodia emotiva.

6. MODULI PROSODICI EMOTIVI

Per utilizzare l'approccio differenziale è necessario distinguere prima di tutto tra macro prosodia e prosodia segmentale. Per un miglior confronto tra i valori prosodici nelle diverse emozioni e nello stato "neutro" è necessario normalizzare questi valori in una stessa scala.

Si sono utilizzati z-score per i dati delle durate e parametri PaIntE normalizzati per l'intonazione.

6.1 Durata

Per ogni emozione sono state calcolate le statistiche di durata dei fonemi, le medie, le deviazioni standard e le differenze con lo stato "neutro" (vedi Tabella 1).

Emotion	μ (s)	σ (s)	μ_Δ (s)	σ_Δ (s)
Neutral	0.094	0.045	-	-
Anger	0.077	0.034	-0.017	-0.010
Disgust	0.103	0.055	0.009	0.011
Fear	0.078	0.036	-0.016	-0.009
Joy	0.076	0.032	-0.018	-0.013
Sadness	0.104	0.052	0.010	0.007
Surprise	0.076	0.033	-0.018	-0.012

Tabella 1: Media e deviazione standard delle durate dei fonemi nelle varie emozioni.

La durata di ogni singolo fonema è stata prima normalizzata con la tecnica z-score, utilizzando le statistiche precedentemente calcolate, ed in seguito è stata effettuata la differenza tra i dati emotivi e quelli "neutri". Questi dati assieme alle *features* linguistiche sono stati utilizzati per creare i moduli emotivi differenziali.

6.2 Intonazione

Per ogni emozione sono stati calcolati i valori di minimo, massimo ed escursione di F0 (vedi Tabella 2).

Emotion	LB(Hz)	μ(Hz)	UB(Hz)	R(Hz)	LB_Δ(Hz)	μ_Δ(Hz)	UB_Δ(Hz)	R_Δ(Hz)
Neutral	62	105	213	169	-	-	-	-
Anger	66	122	258	192	4	17	45	23
Disgust	53	81	238	185	-9	-24	25	16
Fear	66	114	223	157	4	9	10	-12
Joy	63	129	308	245	1	24	95	76
Sadness	53	89	208	155	-9	-16	-5	-14
Surprise	66	136	250	184	4	31	37	15

Tabella 2: Minimo, media, massimo, ed escursione di F0 nelle varie emozioni.

I vettori PaIntE sono stati quindi normalizzati, riportando i valori ad una frequenza normalizzata nell'intervallo [0,1].

A questo punto si è eseguita la differenza tra i vettori emotivi e quelli "neutri". Analizzando i vettori risultanti si è notato che seguono una direzione preferenziale, dipendente da emozione ad emozione (vedi Figura 3).

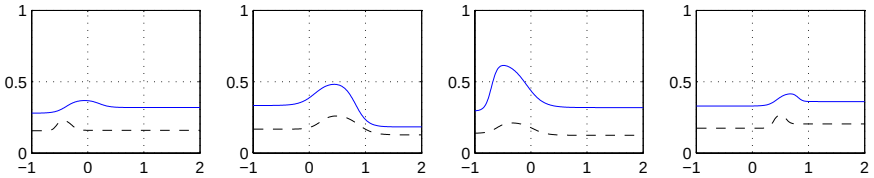

Figura 3: Similarità dello spostamento del pitch tra "neutro" (linea nera tratteggiata) e l'emozione rabbia (linea blue continua) per 4 eventi intonativi.

Si è quindi pensato di compiere l'analisi delle componenti principali (PCA) sull'insieme di questi vettori.

Il risultato dell'analisi ha confermato l'esistenza di queste direzioni preferenziali. Si è pensato quindi di utilizzare un semplice albero di regressione su tale componente per la predizione del Δ emotivo. Tale soluzione porta una notevole semplificazione della procedura di generazione dell'albero di decisione per le emozioni.

7. ESEMPI AUDIO

Gli esempi audio di Tabella 3, sono stati generati utilizzando i moduli prosodici *data-driven*; per ogni emozione è stata scelta una frase (prima colonna della Tabella 3), che è stata sintetizzata utilizzando sia il modulo prosodico "neutro" (seconda colonna della Tabella 3), che il modulo prosodico emotivo (ultima colonna della Tabella 3).

Ti ho ripetuto mille volte di non tirare la coda al gatto.	neutral	anger
Che schifo. Il tuo alito puzza di pesce andato a male.	neutral	disgust
Ho sentito uno sparo. Veniva dalla stanza di Andrea.	neutral	fear
Ohh che bello. Ho appena superato l'esame.	neutral	joy
Sono passati dieci anni da quel giorno e da allora non riesco più a sorridere	neutral	sadness
Veramente incredibile. Tutti avevano l'ombrello, ma la giornata era di sole splendente.	neutral	surprise

Tabella 3: Confronto audio tra prosodia "neutra" ed emotiva.

8. CONCLUSIONI E SVILUPPI FUTURI

Ad un primo ascolto i moduli prosodici generati con le procedure qui descritte, hanno generato una prosodia che si presta ad esprimere efficacemente le emozioni.

Sviluppi futuri di questo lavoro potrebbero prendere in considerazione l'aggiunta di un modulo prosodico *data-driven* sia per l'intensità che per i parametri di *voice-quality*,vista l'importanza di questo parametro acustico nelle emozioni.

I risultati dei precedenti algoritmi sono stati implementati in una nuova versione di Festival per l'italiano, mentre sono in fase di sviluppo alcuni esperimenti soggettivi di accettabilità e riconoscimento delle emozioni.

RINGRAZIAMENTI

Parte di questo lavoro è stato sponsorizzato dal progetto europeo PF-STAR (Preparing Future multi Sensorial inTerAction Research, European Project IST-2001-37599, http://pfstar.itc.it)

BIBLIOGRAFIA

Anolli L. & Ciceri R., *La voce delle emozioni.* Franco Angeli s.r.l, 1997.

Avesani C., Cosi P., Fauri E., Gretter R., Mana N., Rocchi S., Rossi F., & Tesser F., Definizione ed annotazione prosodica di un database di parlato-letto usando il formalismo tobi. In *Il Parlato Italiano*, Napoli, 13-15 Febbraio 2003.

Balestri M., Paechiottia A., Quazza S., Salza P. L. & Sandri S., Choose the best to modify the least: a new generation concatenative synthesis system. In *Proc. of EUROSPEECH*, Budapest, Hungary, Sept. 1999.

Cosi P., Gretter R. & Tesser F., Festival parla italiano! In *XI Giornate di Studio del G.F.S.*, Padova, November 29-30, December 1 2000.

Drioli C., Tisato G., Cosi P. & Tesser F., Emozioni e qualità vocalica: Esperimenti con modelli di sintesi sinusoidale. In *XIV Giornate di Studio del G.F.S.*, Viterbo, 4-6 Dicembre 2003.

Dutoit T. & Leich H., MBR-PSOLA: Text-To-Speech synthesis based on an MBE re-synthesis of the segments database. In *Speech Commun.,* 13(3-4):167-184, November 1993.

Ekman P., *An argument for basic emotions.* In Basic Emotions. N.L. Stein and K. Oatley (eds), hove, uk: lawrence erlbaum. edition, 1992.

Möhler G. & Conkie A., Parametric modeling of intonation using vector quantization. In CDROM proceedings of *Third ESCA International Workshop on Speech Synthesis*, Jenolan Caves, Australia, 1998.

Tesser F., Cosi P., Mana N., Avesani C., Gretter R. & Pianesi F., Modello prosodico "data-driven" di festival per l'italiano. In *XIV Giornate di Studio del G.F.S.*, Viterbo, 4-6 Dicembre 2003.

Tesser F., Cosi P., Drioli C. & Tisato G., Prosodic data driven modelling of a narrative style in Festival TTS. In *CDROM proceedings of 5th ISCA Speech Synthesis Workshop*, Pittsburgh, PA, U.S.A., June 14-16 2004.

Zovato E., Picchiotti A., Quazza S. & Sandri S.: Analisi prosodica di una base dati di parlato emozionale. In *XIV Giornate di Studio del G.F.S.*, Viterbo, 4-6 Dicembre 2003.