Atti delle XIV Giornate del GFS
Viterbo, 4-6 dicembre 2003

MODELLO PROSODICO "DATA-DRIVEN" DI FESTIVAL PER L'ITALIANO

Fabio Tesser*, Piero Cosi**, Nadia Mana*,
Cinzia Avesani**, Roberto Gretter*, Fabio Pianesi*

ITC-IRST *: Istituto Trentino di Cultura, Centro per la Ricerca Scientifica e Tecnologica
e-mail: {tesser, mana, gretter, pianesi}@ itc.it

ISTC-SPFD-CNR **: Istituto di Scienze e Tecnologie della Cognizione Sezione di Padova "Fonetica e Dialettologia" e-mail: {cosi, avesani}@csrf.pd.cnr.it

1. SOMMARIO

Il modello prosodico implementato nell'attuale versione di FESTIVAL in italiano è basato interamente su semplici regole fonetico/acustiche.
In questo lavoro vengono descritte le metodologie e gli esperimenti realizzati per la creazione del nuovo modello prosodico "data-driven" della nuova versione di FESTIVAL. I parametri di questo nuovo modello sono stati appresi mediante tecniche automatiche di modellizzazione prosodica che si basano totalmente sulle informazioni contenute nei corpora vocali utilizzati in fase di sviluppo. Al fine di valutare le proprietà e l'efficacia di questo modello, è stato realizzato un primo esperimento di stima dello stesso, messo a confronto con il modello originale "rule-based", mediante un test percettivo. I risultati di questa prima valutazione vengono qui presentati brevemente.

2. INTRODUZIONE

Il compito dei moduli prosodici di un sistema TTS (Text To Speech) è quello di predire l'andamento dei parametri prosodici (durata dei fonemi, f0, …) di una frase attraverso le informazioni che si possono ricavare dal testo in ingresso al sintetizzatore. Usualmente due approcci sono utilizzati per la creazione di questi moduli: la tecnica "rule-based" oppure la tecnica "data-driven". Nei prossimi paragrafi verranno messe a confronto queste due diverse possibilità e verrà spiegata in dettaglio la metodologia utilizzata per la creazione dei modelli prosodici "data-driven".

3. "RULE-BASED" (REGOLE MANUALI O ESPLICITE)

Usando la tecnica "rule-based", i fenomeni prosodici vengono trattati attraverso un set di regole definite manualmente. Questo approccio è stato

usato per la prima versione di Italian Festival (Cosi et alii, 2001): inizialmente è stato definito un set di regole riguardanti l'andamento della durata dei fonemi e del pitch in frasi dichiarative e interrogative. Per fare un esempio, una regola per la durata è: se la vocale da sintetizzare è accentata, la sua durata viene aumentata del 20% rispetto a quella standard.
L'approccio "rule-based" ha alcuni limiti:

- è richiesta una conoscenza dettagliata della prosodia italiana
- per una buona copertura di tutti i fenomeni, sarebbe necessario inserire molte regole prosodiche ma non tutte le possibili regole sono codificate in letteratura; solo quelle più standard lo sono.
- solitamente il numero di regole scritte non è sufficiente ad ottenere una prosodia naturale: ogni frase è pronunciata alla stessa maniera, mentre nella realtà ci sono molte varianti che rendono più naturale la sintesi.

4. "DATA-DRIVEN"

Le tecniche "data-driven" utilizzano algoritmi di "machine learning" basati sulla classificazione statistica per predire l'andamento dei parametri prosodici, cercando di "copiare" quello presente in un corpus vocale. La creazione di un modello prosodico per la sintesi da testo consiste nel far apprendere ad un modello statistico in che modo i parametri prosodici da predire di un corpus vocale sono in relazione con le informazioni disponibili e ricavabili all'ingresso del TTS ("features").
Alcuni vantaggi di questo metodo sono:

- permette di ricavare in maniera automatica molte e varie regole prosodiche utilizzando un corpus vocale;
- non richiede di conoscere in dettaglio le regole della prosodia italiana;
- utilizzando corpora specifici, è possibile "catturare" la prosodia di una specifica persona, di uno stile di lettura, o addirittura di una emozione.

Nel caso specifico l'algoritmo statistico utilizzato è quello dei CART "Classification And Regression Trees" (Breiman et alii 1984).

4.1 Corpora prosodici

Per creare un CART servono dei corpora che contengano il parametro da predire e le "features" da usare per predire tale parametro.
Il corpus Carini (Avesani et alii, 2003) è stato progettato appositamente per la modellizzazione automatica di moduli prosodici; esso contiene tre racconti di Dino Buzzati e un set di frasi interrogative lette da uno speaker professionista.

Il corpus è stato segmentato automaticamente a livello di fonemi, con una procedura di allineamento automatico basata su un sistema ASR (Automatic Speech Recognition) (Angelini et alii, 1993), ed annotato dal punto di vista linguistico (Part Of Speech tagging - POS - e annotazione sintattica a costituenti) e prosodico (ToBIt) (Avesani, 1994).
Due moduli prosodici sono stati presi in considerazione:

- Durata: la variabile da predire è la lunghezza dei fonemi; tale informazione è ricavabile automaticamente e in maniera molto affidabile con un sistema ASR (Automatic Speech Recognition).
- Pitch: il parametro da predire è la frequenza fondamentale f0; tale informazione è ricavabile in modo automatico utilizzando algoritmi di pitch extractor (Boersma, 2001), i quali purtroppo non sono così affidabili come i sistemi ASR.

Per avere un riferimento numerico oggettivo di quanto un modulo prosodico data-driven riesca a "seguire" l'andamento prosodico del corpus si utilizzano le misure di RMSE (Root Mean-Square Error) e Correlazione tra il segnale prosodico originale e quello predetto.
Il 90% del corpus è stato utilizzato per l'addestramento e il 10% per il test. Tutti i valori di RMSE e Correlazione mostrati in seguito sono calcolati sulla parte di test del corpus.

4.2 "features"

Le informazioni strutturali usate per il training sono state estratte automaticamente dal testo, usando la tipica struttura HRG (Heterogeneous Relation Graph) di Festival (Taylor et alii, 1998; Cosi et alii, 2002).
Per quanto riguarda le informazioni linguistiche e prosodiche, l'attuale modulo NLP (Natural Language Processing) di Festival è limitato al riconoscimento di Function Word e POS della parola, e non contiene nessun predittore di tipo prosodico categoriale (ToBI).
Per verificare se sarebbe opportuno integrare questi moduli in una futura versione di Festival, sono in corso degli esperimenti per indagare se e quanto potrebbero aumentare le performance dei moduli prosodici utilizzando informazioni linguistiche di alto livello (analisi Sintattica e a costituenti) e informazioni prosodiche (ToBIt).

4.3 Il problema dello data-sparseness

La maggior difficoltà che si affronta quando si utilizzano algoritmi di apprendimento statistico è il problema della poca densità dei dati (data-sparseness). Due possibili soluzioni possono essere adottate per risolvere questo problema: la prima è quella di avere più dati; la seconda invece è quella di trovare delle trasformazioni "intelligenti" della variabile da predire in modo da raggruppare dati omogenei che prima erano "distanti" tra di loro.
Due esempi di questo tipo di trasformazione sono la trasformazione z-score (Campbell & Isard, 1991) per la durata e il modello PaIntE (Möhler, 1998)

per l'andamento del pitch. Nelle prossime due sezioni verranno illustrati tali metodi e i risultati ottenuti.

4.4 Durata

Z-score è un modello molto efficace per predire la durata dei singoli fonemi di una frase: la durata di un fonema d_i è calcolata stimando il numero k di deviazioni standard σ_i dalla sua media μ_i, utilizzando la relazione: $d_i=\mu_i+k\sigma_i$. In tal modo si elimina la dipendenza della durata dal tipo di fonema e il CART può generare una serie di regole indipendenti dal tipo di fonema. Utilizzando tale modello sul corpus Carini si sono ottenuti valori di errore RMSE di 0.7775 (in unità z-score) e una correlazione di 0.6046.

4.5 Pitch

La predizione automatica di f0 è molto difficile per vari motivi: a) i pitch-extractor non sono del tutto affidabili; b) il problema dello data-sparseness è presente in maniera considerevole per il fatto che il pitch range può variare molto anche tra una frase e un'altra di uno stesso speaker e inoltre realizzazioni di accenti equivalenti dal punto vista percettivo possono avere valori di f0 diversi; c) i singoli valori di f0 non sono molto significativi. Lo sono di più i movimenti di f0 ed è percettivamente importante rilevare tali andamenti solo in determinate sillabe che portano informazione.

Dati questi motivi, per seguire i movimenti di f0 è stato usato un modello apposito e percettivamente significativo: PaIntE.

Alla sua implementazione originale sono state aggiunte due modifiche per limitare il problema dello data-sparseness: la normalizzazione in semitoni e la quantizzazione vettoriale dei parametri. La normalizzazione in semitoni è una semplice trasformazione dell'asse della frequenza usata per trasformare i valori di f0 in una misura percettivamente più vicina all'orecchio dell'uomo.

La quantizzazione vettoriale (VQ) invece cerca di raggruppare le traiettorie omogenee di f0 in uno stesso insieme, consentendo al CART di predire solamente l'insieme di appartenenza della traiettoria invece di predire i 6 parametri di PaIntE. In questo modo si limita il numero di possibili casi da predire a N, definito come il numero di gruppi omogenei delle traiettorie, il quale è fissato dall'algoritmo di VQ. Sono stati fatti diversi esperimenti di apprendimento al variare di N.

N	RMSE(Hz)	Correlazione
32	39.6764	0.2622
64	38.5798	0.3275
80	36.3755	0.4340
128	41.6729	0.2666

Tabella 1: RMSE (Hz) e Correlazione delle traiettorie di f0 predette

Come è mostrato nella *Tabella 1*, il risultato migliore è stato ottenuto con N=80. Per N inferiori le performance diminuiscono perché l'algoritmo VQ crea degli insiemi meno omogenei; per N maggiori l'algoritmo di apprendimento del CART si comporta peggio perché deve distinguere tra un numero maggiore di casi.

5. VALUTAZIONE SOGGETTIVA PRELIMINARE

Per avere una valutazione preliminare della bontà dei moduli prosodici creati, sono stati realizzati dei test d'ascolto soggettivi. I soggetti del test erano 10 persone con competenze diverse nell'ambito della prosodia. L'esperimento consisteva nell'ascoltare una coppia di frasi sintetizzate partendo dallo stesso testo ma con due diversi moduli prosodici; l'ascoltatore doveva indicare quale stile prosodico era più adatto per raccontare una storia, con la possibilità anche di valutare egualmente adatti i due stimoli (**Nessuna Preferenza**).
I modelli prosodici messi a confronto erano:

- **Regole:** modulo prosodico "rule-based"
- **Rai-news:** modulo prosodico data-driven il cui apprendimento è stato fatto su di un corpus di tipo telegiornalistico
- **Carini:** modulo prosodico data-driven appreso sul corpus Carini

Il corpus di questo test era formato da 6 frasi, per un numero totale di 18 confronti. Le coppie di stimoli erano presentate in maniera casuale e dopo il primo ascolto era possibile riascoltare la coppia di stimoli.
Siccome il corpus Carini contiene la registrazione della lettura di novelle, l'aspettativa era che, se il modello ha imparato in maniera efficace dai dati, il modulo **Carini** risulti il preferito dagli ascoltatori.
I risultati del test (*Tabella 3*) confermano questa aspettativa, mostrando chiaramente la preferenza netta accordata al modulo Carini sia rispetto al modulo **Regole** che a quello **Rai-news**.
Interessante poi il fatto che sia stato preferito **Regole** a **Rai-news**. Ciò può essere spiegato dal fatto che lo stile prosodico di tipo giornalistico è più distante da quello narrativo di un semplice modulo "rule-based".

Modulo 1 vs **Modulo 2**	**Preferenza Modulo 1**	**Preferenza Modulo 2**	**Nessuna Preferenza**
Regole vs **Rai-news**	52.8 %	34.7 %	12.5 %
Carini vs **Rai-news**	83.3 %	12.5 %	4.2 %
Carini vs **Regole**	88.9 %	9.7 %	1.4 %

Tabella 3: Risultati del test soggettivo sullo stile di lettura

6. CONCLUSIONI E SVILUPPI FUTURI

Le metodologie utilizzate per la creazione automatica di modelli prosodici per TTS hanno subìto una notevole influenza dal mondo del "machine-learning". La rappresentazione dei dati prosodici e la creazione di appositi modelli aiutano in maniera considerevole la fase di apprendimento degli algoritmi statistici. Creando migliori modelli che diminuiscano il problema del data-sparseness si potranno migliorare ancora le prestazioni.
Dai primi test di valutazione soggettiva si deduce che è possibile "catturare" la prosodia di uno stile di lettura. Sono previste ulteriori e più complete sessioni di valutazione per verificare la naturalezza dei moduli prosodici e per confrontare diverse strategie di modellizzazione automatica.
Inoltre si cercherà in futuro di modellare la prosodia relativa ad una emozione, ovvero di una frase pronunciata con espressione emotiva (ad es. con tono felice oppure arrabbiato oppure triste, ecc.).

Bibliografia

Angelini B., Brugnara F., Falavigna D., Giuliani D., Gretter R., and Omologo M., "A baseline of a speaker independent continuous speech recognizer of italian,", *Proceedings of EUROSPEECH 93*, Berlin, Germany, 1993, pp. 847–850.

Avesani C., "ToBIt: un sistema di trascrizione per l'intonazione italiana", *Atti delle V giornate di Studio del Gruppo di Fonetica Sperimentale*, Trento, Novembre 1994.

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

Boersma P., "PRAAT, a system for doing phonetics by computer", *Glot International*, vol. 5, pp. 341–345, 2001.

Breiman L., Friedman J., Olshen R., and Stone, *Classification and regression trees*, Wadsworth and Brooks, 1984.

Campbell N. and Isard S., "Segment durations in a syllable frame," *Journal of Phonetics*, pp. 37–47, 1991.

Cosi P., Tesser F., Gretter R., Avesani C., Macon M. : "Festival Speaks Italian!", *EUROSPEECH 2001*, Aalborg, Denmark, September 2001.
Cosi P., Avesani C., Tesser F., Gretter R., and Pianesi F.,"On the use of Cart-Tree for prosodic predictions in the Italian Festival TTS,*" Voce, Canto, Parlato - Studi in onore di Franco Ferrero*, pp. 73–81, 2002.

Möhler G., "Describing intonation with a parametric model," in *Proceedings of ICSLP98*, Sydney, 1998, pp. 2581–2584.

Taylor P., Black A., and Caley R., "The architecture of the Festival speech synthesis system," *3rd ESCA Workshop on Speech Synthesis*, pp. 147–151, 1998.