STIMA DEI MOVIMENTI LABIALI A PARTIRE DAL SEGNALE VERBALE

Piero Cosi, Luca Versini

Istituto di Fonetica e Dialettologia – C.N.R.
Via G. Anghinoni, 10 - 35121 Padova (ITALY)
e-mail: cosi@csrf.pd.cnr.it www: http://www.csrf.pd.cnr.it

SOMMARIO

E' descritto un sistema per la stima automatica dell'andamento dei parametri articolatori, in particolare di quelli che caratterizzano il movimento della bocca di un parlatore, a partire dal segnale vocale co-prodotto, e sono illustrati alcuni dei risultati ottenuti applicando il modello ad un corpus di stimoli VCV.

INTRODUZIONE

Nell'ambito della trasmissione digitale di immagini in movimento è necessario far fronte a problemi di occupazione di banda: si scontrano, infatti, l'esigenza di trasmettere grandi quantità di informazioni con quella di utilizzare porzioni limitate di banda, sia che si faccia uso di canali a banda stretta, sia che si vogliano allocare più trasmissioni distinte su uno stesso canale. La ricerca ha portato alla nascita di tecnologie di compressione con prestazioni soddisfacenti per quanto riguarda la trasmissione di immagini ad alto *bit-rate* (MPEG), mentre per quanto riguarda la trasmissione di immagini a bassissimo bit-rate, buona parte delle problematiche devono ancora trovare risposta. Questo secondo tipo di trasmissioni è utilizzato in genere per il videotelefono o per le videoconferenze, in situazioni nelle quali, in pratica, l'immagine considerata è comunemente un mezzobusto o il primo piano di un parlatore. Si è quindi pensato di cercare di migliorare la qualità delle immagini in ricezione sfruttando anche le informazioni contenute nel segnale vocale, e in particolare, nel caso in esame, la relazione fra il segnale audio e i movimenti articolatori del viso. Si vuole ottenere un miglioramento della qualità delle immagini, al costo di una complessità superiore, sia computazionale che circuitale, del ricevitore, ove sarà necessario eseguire delle elaborazioni sul segnale audio allo scopo di ricavare i parametri articolatori corrispondenti. La complessità di queste elaborazioni ed il tempo necessario per eseguirle via software sembrano attualmente rendere problematico l'utilizzo di questi sistemi in applicazioni real-time, specie in quelle che, sfruttando ad esempio la trasmissione via satellite, implicano intrinsecamente la presenza di un ritardo rilevante

fra invio e ricezione del segnale. Una possibile soluzione ai problemi di complessità computazionale potrebbe essere individuata nella realizzazione prettamente via *hardware* del ricevitore.

METODO

Gli stimoli utilizzati sono stati estratti da un corpus di parole senza senso del tipo VCV (vocale-consonante-vocale), acquisito mediante il sistema ELITE [1]; nel *database* di riferimento sono contenuti l'andamento reale dei movimenti labiali ed il corrispondente segnale verbale co-prodotto. Per le vocali sono state considerate le tre vocali cardinali (a,i,u), mentre per le consonanti sono state utilizzate le occlusive (b,d,g,k,p,t). Il segnale vocale è rappresentato in frequenza mediante 8 coefficienti *Mel Cepstrum* (MFCC) [2], calcolati ogni 10 ms.; l'evoluzione temporale di questi coefficienti è fornita in ingresso ad una rete neurale del tipo *Time Delay Neural Network* (TDNN) [3], con struttura a tre strati opportunamente dimensionati, la quale è stata istruita a generare l'andamento di un particolare parametro articolatorio, ad esempio l'andamento dell'*apertura labiale*. Volendo realizzare una procedura di stima dei parametri articolatori che funzionasse in modalità *speaker idependent*, ovvero che risentisse il meno possibile delle differenze di pronuncia e di articolazione che contraddistinguono i singoli individui, si è ricavato per ogni stimolo un modello sia per l'andamento dei MFCC sia per l'andamento dei parametri articolatori. Questo ha permesso di addestrare la rete neurale, non con le sequenze di MFCC associate ad una particolare pronuncia di un dato stimolo, bensì con una trasformazione di tali sequenze, ottenuta tramite un algoritmo di allineamento temporale (DTW - *Dynamic Time Warping)* [4], sui i modelli prima determinati. Questo modo di procedere è stato adottato allo scopo di limitare la variabilità degli ingressi forniti alla rete, e di agevolarla così nel compito di classificazione.

La costruzione dei modelli per prima ha comportato l'utilizzo della procedura di allineamento temporale, in quanto ad ogni passo della procedura, il modello parziale $M_i(n)$ è stato ottenuto come media pesata del modello prima determinato $M_{i-1}(n)$ (o della sequenza di riferimento $S_1(n)$ nel caso di $M_1(n)$), e della sequenza di MFCC i-esima $S_i(n)$ allineata al modello disponibile $M_{i-1}(n)$, come evidenziato dall'espressione:

$$M_1(n) = S_1(n)$$

$$M_2(n) = \frac{1}{2} \cdot [M_1(n) + S_2(w_2(n))]$$

.....

$$M_L(n) = \frac{1}{L} \cdot [(L-1) \cdot M_{L-1}(n) + S_L(w_L(n))]$$

sono stati cioè mediati in successione tutti gli andamenti relativi allo stimolo in esame dopo un opportuno allineamento temporale (rappresentato in termini di trasformazione del dominio temporale con l'espressione $w_i(n)$), ottenuto in riferimento al modello parziale determinato fino a quel momento.

La procedura di costruzione dei modelli ha implicato preventivamente la necessità di normalizzare la dinamica degli andamenti da mediare, ad un intervallo comune di valori,

essendo significativo in pratica solo l'andamento del modello e non l'ampiezza dei valori ad esso associati, che sarebbero comunque stati scalati per poter essere trattati dalla rete neurale. L'operazione di normalizzazione è stata realizzata in modo che ogni sequenza di coefficienti cepstrali contenesse valori nell'intervallo [0,1], mentre la dinamica dei parametri articolatori è stata riportata all'intervallo [-0.9,0.9], in accordo con quella delle uscite della rete neurale utilizzata. Nella figura 1, è riportato un esempio di modello per un MFCC, confrontato con un andamento reale dello stesso coefficiente.

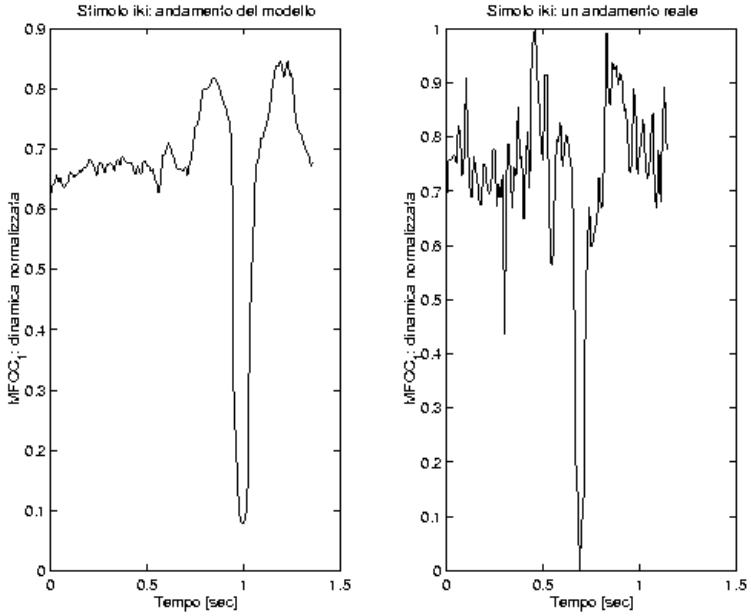


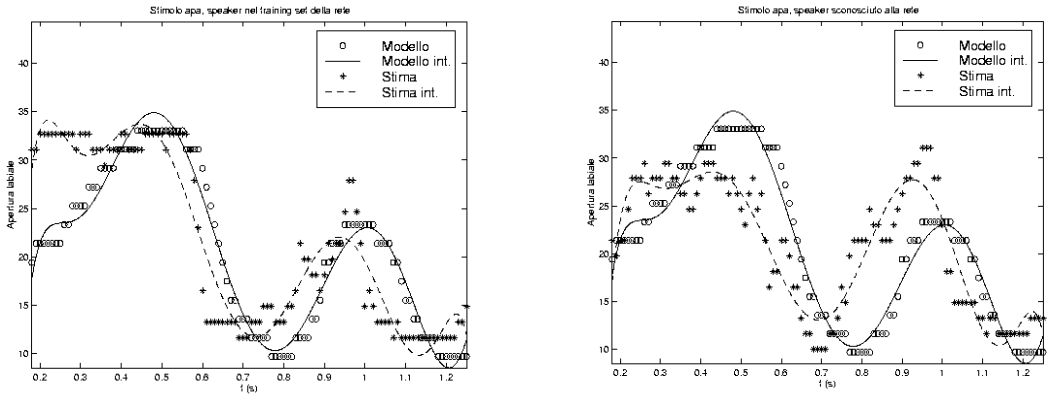
Figura 1. Esempio di modello per un coefficiente MFCC, confrontato con un andamento reale.

Risulta evidente come l'andamento reale sia maggiormente dettagliato (il modello può essere, infatti, interpretato come versione "passa basso" dell'andamento reale), ma anche come il modello ne riproduca fedelmente l'inviluppo. Per quanto riguarda invece i modelli per i parametri articolatori, questi ricalcano fedelmente gli andamenti reali, poiché l'operazione di media ha minor effetto su questi "segnali" che già in origine sono caratterizzati da bassa velocità di variazione.

In fase di verifica del funzionamento della procedura di stima è stato utilizzato l'algoritmo DTW per eseguire la preselezione dello stimolo (con una percentuale di riconoscimento corretto pari al 70%), che ha permesso di allineare la sequenza MFCC considerata al relativo modello, e di utilizzare l'allineamento ottenuto come ingresso alla rete neurale; l'identificazione dello stimolo considerato ha anche permesso di ripristinare la dinamica della stima a quella del modello relativo.

RISULTATI

I risultati sono sicuramente incoraggianti sia in termini di corretta identificazione dello stimolo in ingresso, sia in termini di precisione della stima dei vari parametri articolatori, come risulta dagli esempi illustrati nella Figura 2.



(a) (b)

Figura 2. In questa figura sono riportati rispettivamente un esempio di ricostruzione dell'andamento del parametro articolatorio AL “apertura labiale”' corrispondente allo stimolo “ 'apa ”, in modalità “*speaker-dependent*” (a) e “*speaker independent*” (b).

L'andamento stimato dalla rete neurale è confrontato con quello ideale del modello, ed è possibile notare una buona corrispondenza fra le due sequenze, una volta superata una fase di transitorio dovuta ad effetti di coarticolazione. In Figura 2a, è illustrato il risultato ottenuto per una sequenza di test relativa ad un parlatore del quale sono state utilizzate altre registrazioni (non quella di test) per allenare la rete. Risultati altrettanto confortanti si hanno quando la rete è fatta lavorare con sequenze relative a parlatori ad essa sconosciuti come illustrato in Figura 2b.

CONCLUSIONI

Dal punto di vista delle possibili applicazioni, la stima automatica dell’andamento dei parametri articolatori labiali permette, ad esempio, di costruire immagini sintetiche da inserire tra quelle effettivamente trasmesse nel caso di trasmissioni a basso *bit-rate*, consentendo quindi una visione più “continua”. Un altra possibile applicazione potrebbe riguardare la sintesi di “*talking heads*”, cioè di volti sintetici parlanti, per applicazioni di traduzione da una lingua ad un'altra, ad esempio per non-udenti, oppure per applicazioni didattiche o di intrattenimento.

BIBLIOGRAFIA

[1] F. Ferrigno, A. Pedotti, *ELITE: A Digital Dedicated Hardware System for Movement Analysis via Real-Time TV Signal Processing*, IEEE Transactions on Biomed. Eng., BME-32, 1985, 943-950.

[2] S.B. Davis, P. Mermelstein, *Comparison of Parametric Representations for Monosyllabic Word Recognition in Continuously Spoken Sentences*, IEEE Trans. ASSP, Vol. 28, No. 4, 1990, 357-366.

[3] S. Haykin, *Neural Networks. A comprehensive Foundation – second edition*, Prentice-Hall, Inc., Englewood Cliffs, New Jersey, 1999.

[4] L. R. Rabiner, R.W Shafer, *Digital Processing of Speech Signals*, Prentice-Hall, Inc., Englewood Cliffs, New Jersey, 1978.