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

EMOZIONI E "QUALITÀ VOCALICA": ESPERIMENTI CON MODELLI DI SINTESI SINUSOIDALE

Carlo Drioli †, Graziano Tisato †, Piero Cosi †, Fabio Tesser ‡

†Istituto di Scienze e Tecnologie della Cognizione – CNR
Sezione di Padova Fonetica e Dialettologia

‡ Centro per la Ricerca Scientifica e Tecnologica
Istituto Trentino di Cultura

1. INTRODUZIONE

La qualità dell'emissione vocale, intesa come il risultato delle varie modalità di fonazione possibili, riveste un ruolo importante nella resa delle emozioni nella comunicazione verbale. E' comune osservare, infatti, in corpora di registrazioni di parlato emotivo, casi di fonazione non modale (voce soffiata, sussurrata, laringalizzata, rauca, stridula e così via). La trasmissione delle emozioni nella comunicazione verbale è, inoltre, un aspetto che sta interessando in vario modo il settore delle tecnologie del parlato. Ambiti nei quali l'elaborazione del contenuto emotivo promette di avere un impatto sostanziale sono, ad esempio, quello del riconoscimento automatico del parlato (*automatic speech recognition*) e della sintesi del parlato da testo scritto (*text-to-speech synthesis*).

In questo lavoro si valuta l'efficacia di un sistema di elaborazione della voce basato sulla rappresentazione sinusoidale del segnale e finalizzato all'analisi e alla sintesi del parlato emotivo, ponendo particolare attenzione alla qualità dell'emissione vocale. La qualità fonatoria del segnale vocale su un corpus costituito da sequenze 'VCV', pronunciate ripetutamente da uno stesso parlatore con differenti intenzioni espressive, viene valutata oggettivamente attraverso un insieme di indici acustici opportunamente scelti. Lo stesso viene fatto per delle sequenze ottenute per "trasformazione" di una versione neutra in diverse versioni emotive, e vengono confrontate le versioni emotive originali con le versioni ottenute per trasformazione.

2. CORPUS VOCALE

Per la creazione del corpus vocale, e' stato chiesto ad un soggetto, di sesso maschile e con esperienza di recitazione, di pronunciare due strutture fonologiche 'VCV: "Aba" / 'aba / and "Ava" / 'ava /, simulando, sulla base di uno scenario appropriato, sei stati emotivi: rabbia (A), gioia (J), paura (F), tristezza (SA), disgusto (D), sorpresa (SU), oltre allo stato emotivo "neutro" (N), corrispondenti ad una frase dichiarativa. Questo insieme di 14 parole è

stato ripetuto numerose volte con ordine casuale. Di seguito si riporta una breve descrizione delle caratteristiche del database vocale in termini di parametri acustici comunemente correlati alle emozioni (tabelle dettagliate dei valori di F0, intervallo di variazione di F0 ed Intensità, sono riportati in Drioli et alii (2003)). La rabbia (A) si distingue dalle altre emozioni per il valore più alto di intensità (76.7 dB), per valori intermedi di F0 (178 Hz), per un limitato intervallo di variazione di F0 (18 Hz) e per una durata (195 ms) più breve della neutra (N=231 ms). Il rapporto (circa 7/5) fra la F0 media della rabbia (178 Hz) e della neutra (126 Hz), risulta significativamente dissonante, se espresso in termini musicali (6 semitoni, e cioè un Tritono). Nella rabbia, sia le vocali accentate che le atone presentano una caratteristica ruvidezza sonora, che emerge soprattutto nelle accentate. Il disgusto (D) si caratterizza per la durata più lunga (293 ms), per una intensità intermedia (72.3 dB), per una F0 (139 Hz) poco sopra il valore della neutra, per uno stretto intervallo di variazione di F0 (15 Hz), ed infine per un intervallo di seconda maggiore fra la F0 di D e della neutra. Nella D le vocali accentate come le atone si contraddistinguono per una caratteristica laringalizzazione (*creaky voice*), leggermente più pronunciata nelle atone. Sia la gioia (J) che la sorpresa (SU) presentano una F0 elevata (261 e 266 Hz rispettivamente), intervalli di variazione della F0 piuttosto ampi (68 e 90 Hz) limitatamente alle vocali accentate, valori di intensità medio-alti (74.9 e 72.5 dB) e durate (188 e 179 ms) ridotte rispetto alla neutra e alle altre emozioni, sia nelle accentate che nelle toniche. L'intervallo musicale che formano gioia e sorpresa con la neutra è di circa una ottava aumentata. Queste due emozioni risultano le più difficili da distinguere percettivamente. Una caratteristica distintiva della gioia è una forte componente rumorosa di espirazione nell'attacco del suono, a differenza dell'attacco netto della sorpresa. La paura (F) si distingue per una durata (211 ms) confrontabile con la neutra, il pitch con il valore più alto (289 Hz), l'intervallo di variazione ridotto (27 Hz) e un valore di intensità (70.8 dB) confrontabili con N. L'intervallo che lega la paura con la neutra è di una ottava e una seconda maggiore. Questa emozione è caratterizzata da un attacco rumoroso nelle vocali toniche e in generale dalla presenza di soffio. La tristezza (SA) presenta un valore di F0 medio-alto (209 Hz), un ampio intervallo di variazione della F0 (57 Hz), un valore di intensità (70.4 dB) simile a N, e una durata fra le più alte (269 ms). Non ci sono caratteristiche qualitative che distinguano la tristezza dalle altre emozioni eccetto una generale "colorazione" scura del suono.

2.1 Indici acustici correlati alla qualità vocale e analisi statistiche

Il segnale vocale è stato segmentato manualmente e analizzato mediante il software di analisi PRAAT (Boersma, 2001) ed alcuni script Matlab. I seguenti parametri acustici, fra i più usati nelle ricerche sul parlato emotivo, sono stati scelti quali correlati della qualità vocalica nelle emozioni (Banse

& Scherer, 1996; Alter et alii, 2003): *Shimmer* e *Jitter*, definiti come le perturbazioni di ampiezza e lunghezza del periodo della forma d'onda; l'Harmonic-to-Noise ratio (*HNR*), definito come il rapporto tra l'energia della parte armonica e l'energia della rimanente parte del segnale; l'Indice di Hammarberg (*HammI*), definito come la differenza tra l'energia massima nella banda di frequenze 0-2000 Hz e l'energia massima nella banda 2000-5000 Hz; la pendenza dell'energia spettrale sopra i 1000 Hz (*Do1000*), calcolata come il gradiente dell'approssimazione ai minimi quadrati dell'andamento spettrale sopra i 1000 Hz; l'energia relativa nella parte alta dello spettro (sopra i 1000 Hz) rispetto a quella nella parte bassa (fino ai 1000 Hz) per segmenti vocalizzati (*Pe1000*) ; una misura di piattezza dello spettro (*SFM*), calcolata come il rapporto tra la media geometrica e la media aritmetica della distribuzione spettrale di energia. I parametri acustici sono stati calcolati sul database limitatamente ai segmenti relativi alla vocale accentata e a quella non accentata. Per ogni emozione e' stata calcolata la media di ciascun indice e, infine, e' stata calcolata la differenza standardizzata rispetto al caso neutro (in Fig. 1 sono illustrati i risultanti profili dei parametri per le 6 emozioni).

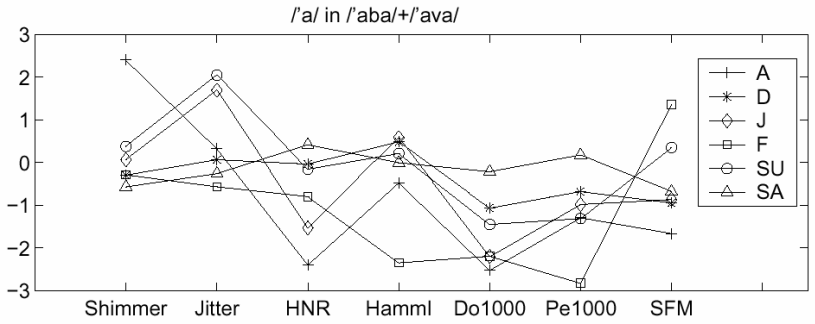

Figura 1: Profili dei parametri acustici per le diverse emozioni (segmento della vocale accentata nelle parole / 'aba/ e / 'ava/).

3. CONVERSIONE DI FRASI NEUTRE IN EMOTIVE

Il metodo di rappresentazione del segnale adottato per questo studio è il noto modello sinusoidale (McAulay & Quatieri, 1986). L'algoritmo di analisi opera attraverso una trasformata FFT su finestre temporali di segnale e produce una rappresentazione in termini di componenti sinusoidali tempo-varianti (qui chiamate parziali). Il numero *H* di parziali è assunto costante per tutte le finestre e, per la i-ma finestra, il risultato della modellazione sinusoidale è un insieme di parametri di frequenza, fase e ampiezza che descrivono ciascuno parziale, più una componente residua di rumore. La risintesi del suono si basa sull'inversione della procedura di analisi, cioè sulla trasformazione inversa dell'analisi sinusoidale seguita da una procedura di sovrapposizione (*overlap-and-add*) di finestre successive. La componente rumorosa non è presa in considerazione nel presente studio.

La rappresentazione sinusoidale permette di controllare alcune delle caratteristiche principali del segnale vocale, come durata, pitch ed intensità, semplicemente interpolando le strutture di analisi, e traslando o scalando la frequenza e l'ampiezza delle parziali. Quando si realizzano variazioni di pitch, è pratica corrente interpolare le ampiezze delle parziali traslate rispetto all'inviluppo spettrale originale, in modo da mantenere la posizione delle formanti. Regole di questo tipo non sono d'altra parte disponibili per la riproduzione delle trasformazioni di qualità vocalica implicate nella produzione delle diverse emozioni nel parlato. Ci si basa dunque su un metodo statistico che consente di "imparare" le trasformazioni spettrali da una base di dati di registrazioni emotive. Il metodo di elaborazione usa una funzione di conversione spettrale basata su Gaussian Mixture Models (*GMMs*) addestrata sui dati spettrali dal corpus di voce (Stylianou & Cappé & Moulines, 1998). I dettagli dell'implementazione della conversione si possono trovare in Drioli et alii (2003), e non vengono riportati qui.

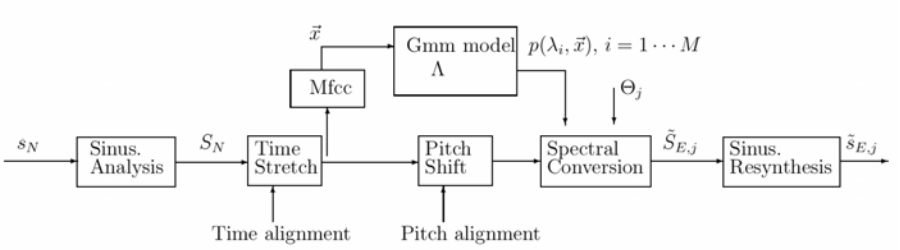

Figura 2: Schema delle trasformazioni operate per trasformare il segnale vocale S_N senza intenzione emotiva, in un segnale S_E con intenzione emotiva.

In Fig. 2 è illustrato lo schema che riassume le trasformazioni operate sul segale di partenza (senza intenzione emotiva) per trasformarlo in un segnale con contenuto emotivo. Sono proposte due diverse procedure: la prima prevede solo l'allineamento temporale (*time stretching*, Ts) e l'allineamento del pitch con mantenimento delle formanti (*formant preserving pitch shifting*, Ps) della frase neutra rispetto alle frasi emotive. A fronte di questa elaborazione viene generata una nuova frase che presenta le caratteristiche prosodiche dell'emozione desiderata e la qualità vocalica derivante dalla frase neutra (essendo la procedura di mantenimento delle formanti l'unica elaborazione spettrale effettuata). La seconda procedura prevede anche una trasformazione spettrale (*spectral modification*, Sm) basata sul modello visto. Questa elaborazione è finalizzata ad allineare anche la qualità vocalica del segnale ottenuto per trasformazione alla qualità del segnale emotivo.

4. RISULTATI SPERIMENTALI

Le trasformazioni descritte nella sezione precedente sono state applicate su una ripetizione delle registrazioni della parola /'ava/. La registrazione neutra in primo luogo è stata trasformata con le procedure di allineamento

temporale e del pitch, così da sincronizzarla e allinearla a ciascuna delle altre sei registrazioni emotive. I segnali vocali risultanti sono stati analizzati per ottenere lo stesso insieme di misurazioni acustiche introdotte prima. Ulteriori sei trasformazioni sono state ottenute operando le trasformazioni spettrali, e per queste è stato calcolato l'insieme corrispondente dei parameri acustici. In Fig. 3 sono riportate, per ogni parametro acustico, le differenze standardizzate con rispetto alla neutra (vengono indicati in figura soltanto i risultati riguardo alla vocale accentata).

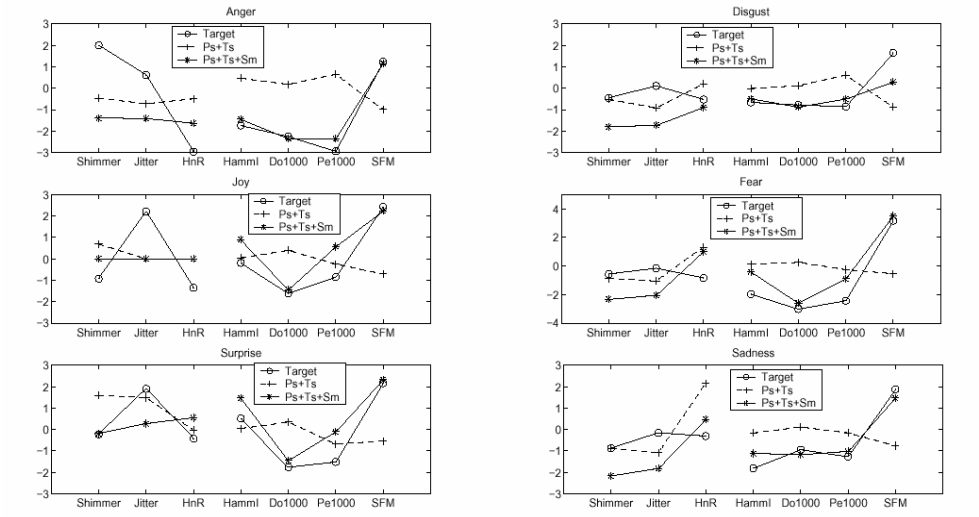

Figura 3: Risultato delle trasformazioni.

Confrontando i parametri delle trasformazioni "Ts+Ps" (linea: -+-) e quelli delle trasformazioni "Ts+Ps+Sm" (linea: -*-) con i parametri relativi ai campioni originali (linea: -o-), possiamo vedere che in generale la procedura Ts+Ps non riesce a riprodurre i pattern acustici delle diverse emozioni. Nella maggior parte dei casi, i nuovi valori dei parametri si trovano intorno allo zero, indicando che la variazione rispetto ai valori dalla neutra è minima. Fanno eccezione il parametro di HNR, poiché la componente rumorosa non è riprodotta in risintesi, e una diminuzione in alcuni parametri spettrali spettrale, probabilmente a causa dell'aumentare del pitch. La successiva introduzione dell'elaborazione spettrale nella procedura (Ts+Ps+Sm) conduce ad effetti positivi per una parte dei parametri, in particolare per HammI, Do1000, Pe1000, SFM, mentre non si riscontra nessun beneficio per i parametri di Shimmer, Jitter, e HNR. Ciò, per Shimmer e Jitter, è motivato dal fatto che il modello di analisi/sintesi sinusoidale, non permettendo trasformazioni con la risoluzione di un singolo periodo, non riesce a fornire un buon modello per variazioni di ampiezza e frequenza pitch-sincrone; per l'HNR, ciò è dovuto alla mancanza di un modello per le

componenti rumorose e per le altre componenti non armoniche del segnale, che dunque non sono riprodotte nella risintesi. I benefici per i restanti parametri possono essere spiegati considerando che tutti rappresentano le caratteristiche dell'inviluppo spettrale e sono calcolati sulla base dallo spettro a breve termine del segnale.

5. CONCLUSIONI

E' stato usato un approccio di sintesi basato su modello sinusoidale per convertire una frase (emotivamente) neutra in frasi con contenuti emotivi diversi. Si sono messi a confronto due diversi livelli di elaborazione: nel primo caso sono stati elaborati la durata dei fonemi e l'andamento di F0, nel secondo caso l'elaborazione e' stata migliorata mediante l'utilizzo di funzioni di conversione spettrale. Questo elemento ha migliorato la resa della qualità vocali caratteristiche delle diverse emozioni. I parametri acustici estratti dalle frasi trasformate sono stati confrontati con quelli estratti dalle frasi emotive originali e i risultati hanno messo in evidenza che il metodo di elaborazione utilizzato risulta essere in grado di riprodurre l'andamento di alcuni parametri per le diverse emozioni.

Si può concludere affermando che il metodo di analisi/sintesi sinusoidale offre delle caratteristiche positive per l'elaborazione della voce nel contesto del parlato emotivo. Sono attualmente allo studio dei criteri per il controllo di parametri che in questo studio non sono stati presi in considerazione, quali le variazioni di F0 ed ampiezza con risoluzione del periodo (*shimmer, jitter*) e le caratteristiche della componente rumorosa.

Ringraziamenti

Questo studio è parte del progetto PF-STAR (Preparing Future multiSensorial inTerAction Research, Progetto Europeo IST- 2001-37599).

Bibliografia

Alter, K. & Rank, E. & Kotz, S. A. & Toepel, U. & Besson, M. & Schirmer, A. & Friederici, A. D. (2003), Affective encoding in the speech signal and in event-related brain potentials, Speech Communication, vol. 40, pp. 61–70.

Banse, R. & Scherer, K. R. (1996), Acoustic profiles in vocal emotion expression. In Journal of Personality and Social Psycology, vol. 70, no. 3, pp. 614–636.

Boersma P. (2001) PRAAT, a system for doing phonetics by computer, Glot International, vol. 5, no. 9/10, pp. 341– 345.

Drioli, C. & Tisato, G. & Cosi, P. & Tesser, F. (2003) Emotions and Voice Quality: Experiments with Sinusoidal Modeling. In Proceedings of VOQUAL'03, pp. 127-132, Geneva, August 27-29.

McAulay R. J. & Quatieri, T. F. (1986), Speech analysis/ synthesis based on a sinusoidal representation, IEEE Transactions on Acoustics, Speech and Signal Processing, vol. ASSP-34, no. 4, pp. 744–754.

Stylianou, Y. & Cappé, O. & Moulines, E. (1998), Continuous probabilistic transform for voice conversion, IEEE Transactions on Speech and Audio Processing, vol. 6, no. 2, pp. 131–142.