ANALISI GERARCHICA DEGLI INVILUPPI SPETTRALI DIFFERENZIALI DI UNA VOCE EMOTIVA

Giampiero Salvi[1], Fabio Tesser[2], Enrico Zovato[3], Piero Cosi[2]
[1]KTH, School of Computer Science and Communication, Dept. of Speech, Music and Hearing, Stockholm, Sweden
[2]Istituto di Scienze e Tecnologie della Cognizione, Consiglio Nazionale delle Ricerche, Padova, Italia
[3]Loquendo S.p.A., Torino, Italia
giampi@kth.se, fabio.tesser@gmail.com, enrico.zovato@loquendo.com, piero.cosi@pd.istc.cnr.it

1. ABSTRACT

In questo articolo viene descritto un nuovo metodo di analisi del timbro vocale tramite lo studio delle variazioni di inviluppo spettrale utilizzato da uno stesso parlatore in situazioni emotiva neutra o espressiva.

Il contesto dell'analisi riguarda un corpus di un solo parlatore istruito a leggere una serie di frasi utilizzando uno stile di lettura neutro e successivamente utilizzando due modalità emotive: uno stile allegro e uno stile triste.

Gli inviluppi spettrali relativi alle versioni allineate delle realizzazioni vocali neutre e espressive (allegra e triste) sono confrontati utilizzando un metodo differenziale. Le differenze sono state calcolate tra lo stato emotivo e quello neutro, di conseguenza le due categorie messe a confronto sono neutro-allegro e neutro-triste.

La statistica degli inviluppi differenziali è stata calcolata per ogni fono. I dati sono stati esaminati utilizzando un metodo di *clustering* gerarchico di tipo agglomerativo. I cluster risultanti sono avvalorati con diverse misure di distanza tra le distribuzioni statistiche ed esplorati visivamente per trovare similitudini e differenze tra le due categorie.

I risultati mettono in evidenza sistematiche variazioni nel timbro vocale relative ai due insiemi di differenze di inviluppi spettrali. Questi tratti dipendono dalla valenza dell'emozione presa in considerazione (positiva, negativa) come dalle proprietà fonetiche del particolare fono come ad esempio sonorità e luogo di articolazione.

2. INTRODUZIONE

La comunicazione delle emozioni tramite la voce si manifesta attraverso variazioni di vari parametri acustici. I parametri prosodici come intonazione, durata e intensità sono tra i più significativi. Altri correlati delle emozioni nella voce sono quelli che riguardano l'analisi spettrale come la posizione delle formanti, la distribuzione della energia spettrale e il rumore spettrale.

I parametri legati al timbro vocale appartengono a questa categoria. Questi parametri sono importanti per sistemi di *voice conversion* (Kain & Macon, 2001) o per i nuovi sistemi di sintesi vocale statistica parametrica HMM (Zen et al., 2009), nei quali il timbro vocale può essere modificato e modellato tramite metodi statistici che analizzano dati reali presenti in corpora vocali. Tuttavia, i modelli utilizzati in questi sistemi sono spesso molto complessi e difficili da interpretare.

L'obiettivo di questo lavoro è di analizzare le variazioni spettrali di timbro vocale di due opposti stili emotivi (allegro e triste) rispetto a uno stile neutro di riferimento. Lo scopo non

è solo quello di studiare questo fenomeno nella voce di un parlante reale, ma anche quello di suggerire nuove strategie per migliorare la parte di analisi e predizione timbrica dei sistemi di *voice conversion* e sintesi statistica parametrica.

Il nuovo metodo proposto per questo studio si compone di due fasi: analisi mel-cepstrale differenziale e *clustering* gerarchico agglomerativo.

L'analisi differenziale è già stata implementata con successo per la predire la prosodia nell'ambito di un sistema Text To Speech (TTS) emotivo (Tesser et al., 2005); in questo lavoro si cerca invece di impiegare la stessa idea di base nel contesto del timbro vocale per identificare in maniera compatta le variazioni di timbro tra stile neutro ed emotivo.

Il metodo del *clustering* gerarchico, utilizzato precedentemente per analizzare grandi corpora con variazioni regionali di accento (Salvi, 2003) (Salvi, 2005), è impiegato qui per permettere di visualizzare similitudini e divergenze tra le due categorie di inviluppi differenziali, tenendo conto delle diversi classi fonetiche.

Per poter analizzare la variazione di timbro tra stile neutro ed emotivo è stato registrato un corpus di tipo "parallelo", nel quale lo stesso testo è letto con diversi stili emotivi (neutro, allegro e triste) da parte di un parlante madrelingua italiano.

Nel prossimo paragrafo sono spiegati l'analisi mel-cepstrale differenziale ed il clustering gerarchico agglomerativo. I dati utilizzati e gli esperimenti sono descritti nel paragrafo 4 ed infine il paragrafo 5 conclude l'articolo.

3. METODO

3.1 Analisi mel-cepstrale differenziale

Il timbro vocale è rappresentato dai coefficienti mel-cepstrali estratti dai corrispondenti frame del segnale vocale. L'analisi mel-cepstrale (Imai, 1983) (Fukada et al., 1992) è stata scelta in quanto estrae i coefficienti minimizzando l'errore di rappresentazione dell'inviluppo spettrale direttamente in un dominio percettivamente significativo. Il vettore di coefficienti mel-cepstrali ricavati rappresenta l'inviluppo spettrale di un frame di segnale vocale.

Per effettuare un confronto tra frame coerenti il database è stato allineato tramite una procedura di DTW (Dynamic Time Warping). La Figura 1 mostra schematicamente il risultato della procedura.

La diversità del timbro tra lo stile neutro ed emotivo è stata valutata tramite l'approccio differenziale: la differenza tra i due vettori mel-cepstrali corrispondenti rappresenta la variazione dell'inviluppo spettrale tra timbro emotivo e neutro. La Figura 2 mostra un esempio di inviluppi spettrali e della loro differenza. La differenza tra i relativi coefficienti mel-cepstrali costituisce il vettore differenziale oggetto dell'analisi successiva.

La proprietà principale dell'analisi differenziale deriva dalla teoria dei sistemi omomorfici (Oppenheim & Schafer, 1968). La trasformazione mel-cepstrale trasforma le convoluzioni in somme. L'inviluppo spettrale derivato dal mel-cepstrum differenziale corrisponde quindi alla risposta in frequenza di un filtro che trasforma il timbro vocale neutro in timbro vocale emotivo. In questo senso questa metodologia di analisi risulta utile per analizzare i dati presenti nei corpora dedicati al training di modelli per *voice conversion*.

Un ulteriore vantaggio del metodo differenziale nel dominio mel-cepstrale è quello di eliminare fattori costanti relativi allo speaker o al canale di trasmissione che non vogliono essere presi in considerazione in questa analisi.

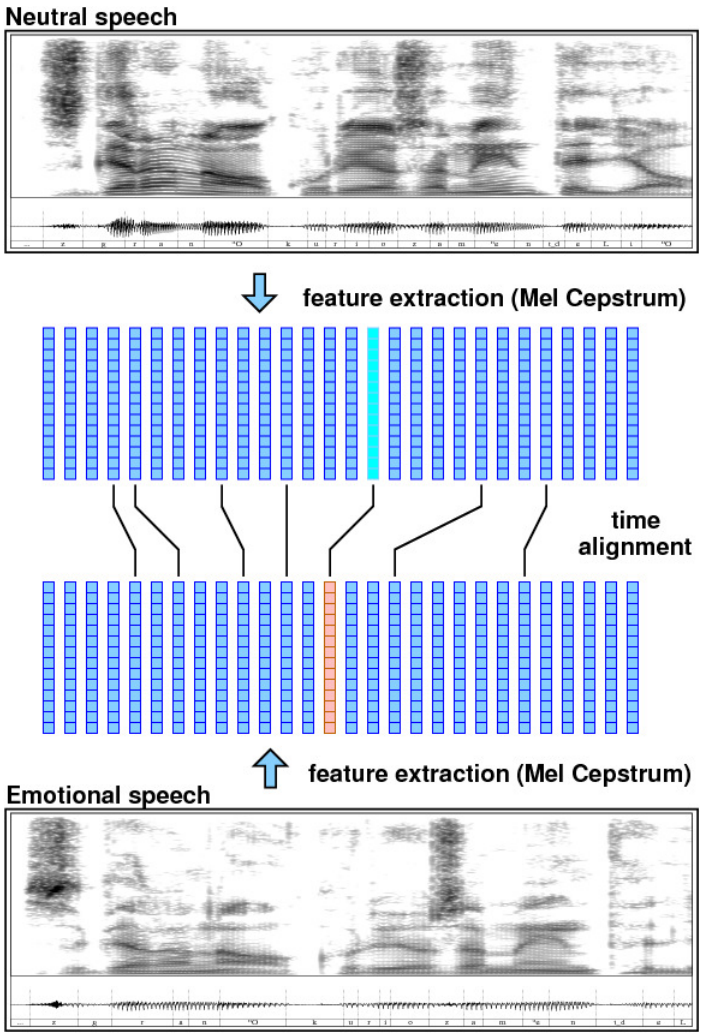

Figura 1: Rappresentazione schematica del Dynamic Time Warping effettuato tra due frasi (una neutra e una emotiva) del corpus parallelo.

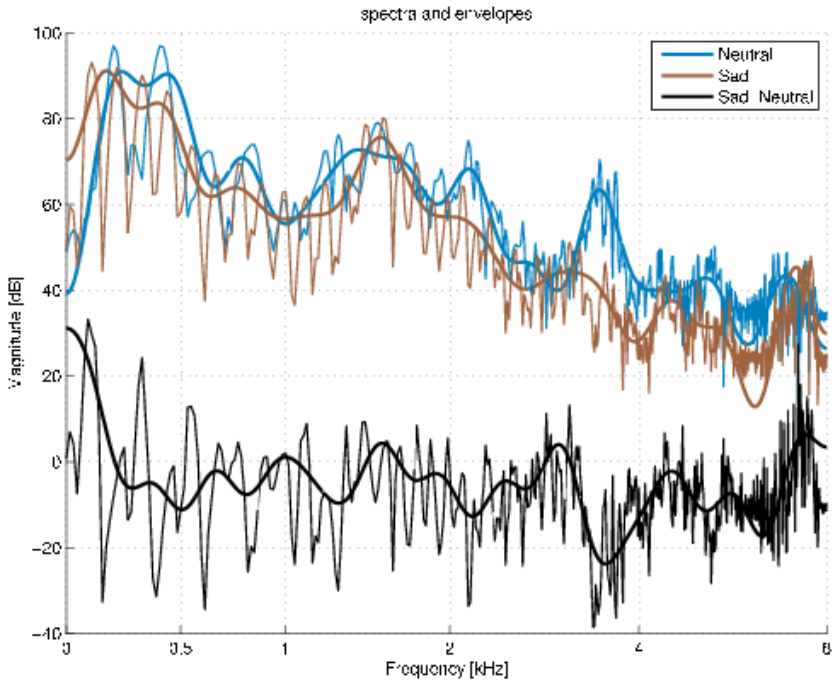

Figura 2: Inviluppi spettrali (linea spessa) e DFT (linea fine) di una coppia di frame coerenti (neutro e triste) e la loro differenza (triste-neutro). La scala delle ascisse è proporzionale alla scala mel.

3.2 Clustering

Per eseguire l'analisi sui dati è stata utilizzata una analisi di tipo cluster. Per ogni fonema è stata calcolata la statistica di primo ($\boldsymbol{\mu}$) e secondo ordine ($\boldsymbol{\Sigma}$) dei vettori rappresentanti le differenze mel-cepstrali.

Questi dati sono successivamente analizzati tramite un metodo di raggruppamento gerarchico (Johnson, 1967). Il vantaggio della analisi di tipo gerarchico rispetto ad altri metodi (ad esempio basati su partizionamento o densità) è la sua proprietà di poter visualizzare i parametri in diversi livelli di dettaglio. Questa proprietà è idonea all'analisi esplorativa dei dati che è tra gli scopi di questo articolo.

Il *clustering* è basato sulla distanza Bhattacharyya (Mak & Barnard, 2006), che tiene conto della diversità tra due distribuzioni prendendo in considerazione le statistiche di primo e secondo ordine.

La distanza d(i,j) tra due distribuzioni i e j è definita come:

$$d(i,j) = \frac{1}{8}(\boldsymbol{\mu}_i - \boldsymbol{\mu}_j)^T \bar{\boldsymbol{\Sigma}}^{-1}(\boldsymbol{\mu}_i - \boldsymbol{\mu}_j) + \frac{1}{2}\ln\frac{|\bar{\boldsymbol{\Sigma}}|}{\sqrt{|\boldsymbol{\Sigma}_i||\boldsymbol{\Sigma}_j|}} \quad (1)$$

con:

$$\bar{\mathbf{\Sigma}} = \frac{\mathbf{\Sigma}_i + \mathbf{\Sigma}_j}{2} \tag{2}$$

Tale distanza è calcolata tra tutte le coppie di fonemi presi in considerazione; il risultato di questa operazione è una matrice come quella visualizzata in Figura 3.

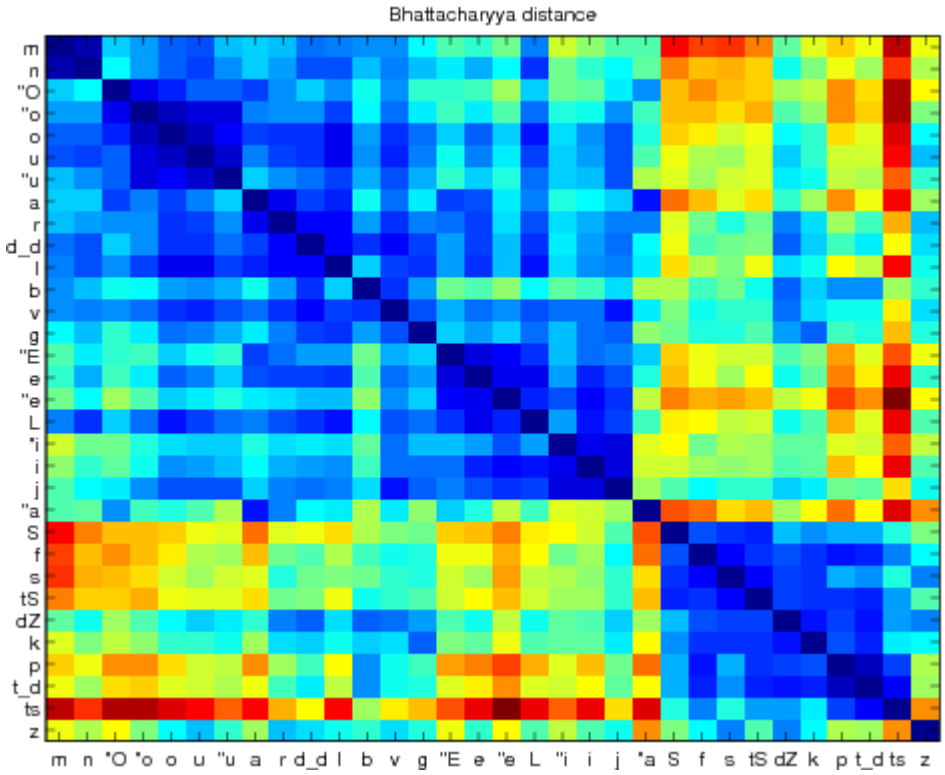

Figura 3: Matrice delle distanze Bhattacharyya tra i cluster calcolata utilizzando i dati degli inviluppi spettrali differenziali nel caso neutro-triste. Sono stati utilizzati i simboli fonetici X-SAMPA (Wells, 1995).

I risultati del raggruppamento gerarchico sono due dendogrammi, uno per la transizione neutro-allegro (Figura 4) e uno per neutro-triste (Figura 5), che identificano la vicinanza tra i vari cluster fonetici.

Per valutare se un dendogramma descrive bene la matrice delle distanze d(i,j) è stato usato il Cophenetic Correlation Coefficient (Sokal & Rohlf, 1962), definito da:

$$\text{coph} = \frac{\sum_{i<j}(d(i,j)-d)(t(i,j)-t)}{\sqrt{\left[\sum_{i<j}(d(i,j)-d)^2\right]\left[\sum_{i<j}(t(i,j)-t)^2\right]}} \tag{3}$$

Valori di questo coefficiente vicino a 1 indicano che il dendogramma rappresenta bene la matrice delle distanze.

Per misurare la similarità tra due dendogrammi si può invece utilizzare il parametro Variation of Information (Meila, 2007):

$$VI(X;Y) = H(X) + H(Y) - 2I(X,Y) \tag{4}$$

dove X e Y sono due partizioni dei dati, H(·) rappresenta l'entropia e I(·,·) il valore di informazione mutua. Il valore di VI è pari a 0 solo se le due partizioni sono equivalenti.

4. ESPERIMENTI

Il metodo è stato applicato ai dati registrati da uno speaker maschile al quale è stato chiesto di leggere lo stesso testo utilizzando i tre diversi stili espressivi.

I coefficienti mel-cepstrali di ordine 26 sono stati calcolati partendo dall'audio campionato a 16 KHz, utilizzando il toolkit SPTK[1], utilizzando finestre di analisi di 40 ms e 10 ms di overlap.

Il metodo differenziale e il *clustering*, come spiegati nel paragrafo 3, sono stati applicati ai dati neutro-allegro e neutro-triste ottenendo i due dendogrammi di Figura 4 e 5.

Ogni fonema, rappresentato in X-SAMPA (Wells, 1995), è visualizzato in basso nei dendogrammi dove forma un unico cluster. Spostandosi verso l'alto, i cluster si uniscono iterativamente fino a diventare un unico gruppo.

Calcolando i valori di Cophenetic Correlation Coefficient, si evince che i dendogrammi risultanti sono una buona rappresentazione della matrice delle distanze tra coppie di cluster in quanto tali valori sono di 0.78 per l'albero neutro-triste e 0.76 per quello neutro-allegro.

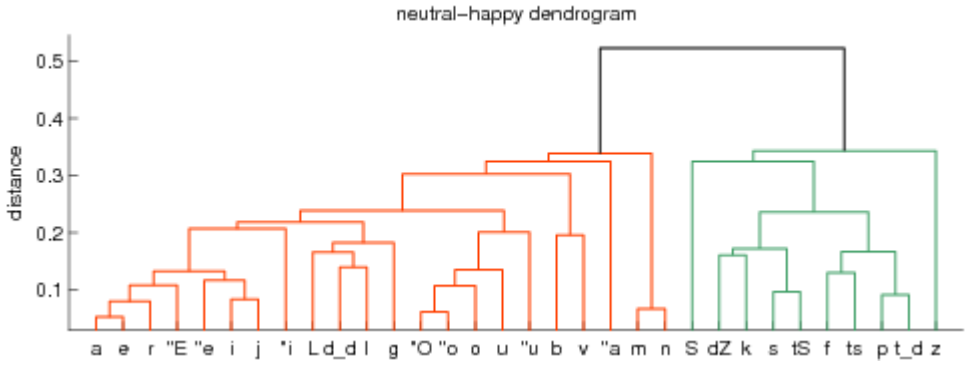

Figura 4: Dendogramma nel caso neutro-allegro.

[1] SPTK Working Group, "Speech Signal Processing Toolkit (SPTK) version 3.3, http://www.sp-tk.sourceforge.net/, December 2009.

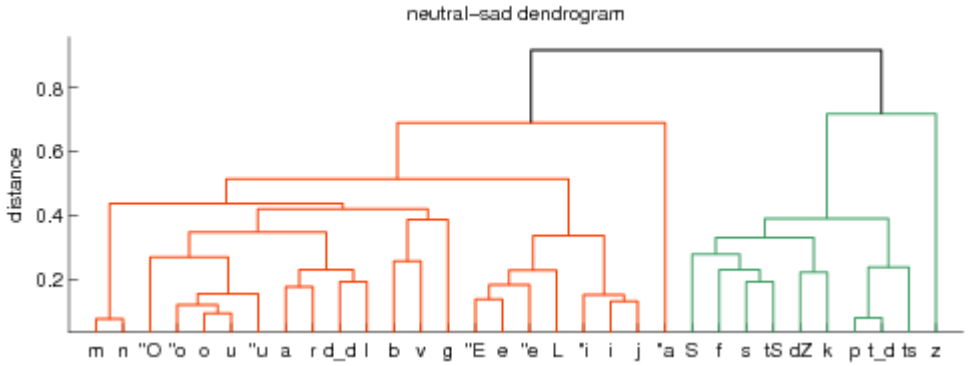

Figura 5: Dendogramma nel caso neutro-triste.

I due dendogrammi mostrano che al vertice di entrambi (partizione di ordine 2) c'è la separazione tra fonemi sonori e sordi[2]. Le successive partizioni sono diverse tra i due alberi ma si possono ritrovare alcune similitudini. Ad esempio i cluster che raggruppano le nasali (/m/ e /n/) oppure le vocali posteriori (/"O/, /"o/, /o/, /u/) sono presenti in entrambi gli alberi.

Il coefficiente Variation of Information, mostrato in Figura 6, è stato utilizzato per quantificare la differenza tra i due dendogrammi.

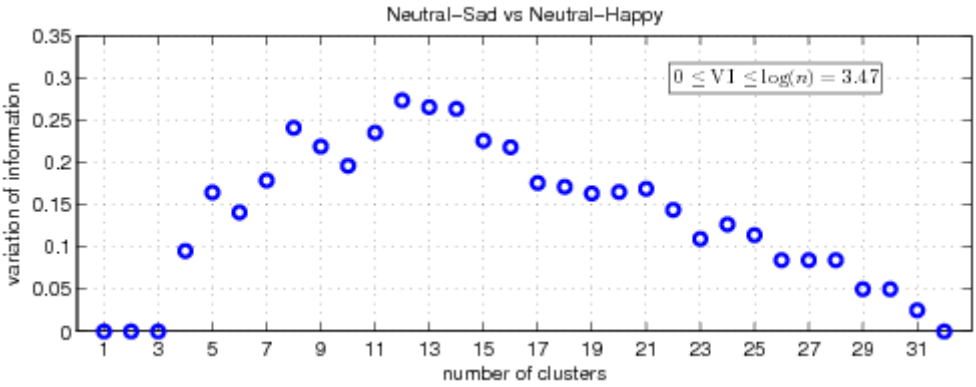

Figura 6: Variation of Information per le partizioni (ordine da 1 a 32) per i dendogrammi neutro-allegro e neutro-triste.

Come ci si aspetta, VI è pari a zero per la prima e l'ultima partizione, ma anche per la partizione 2 (foni sonori, foni sordi) e quella successiva. La distanza successivamente aumenta fino a raggiungere un massimo per poi ritornare a zero nell'ultima partizione.

Per analizzare le differenze anche dal punto di vista acustico, sono stati esaminati gli inviluppi spettrali medi di ogni fonema. Tali inviluppi sono mostrati nelle Figure 7 e 8 e raffigurati con due colori diversi a seconda della partizione di ordine 2 (C1: foni sonori, C2 foni sordi).

[2] Da un controllo informale è risultato che /dZ/ e /z/ erano a volte pronunciati come sordi.

Da questa analisi si può osservare il diverso e opposto comportamento in bassa frequenza delle due categorie emotive: gli inviluppi neutro-triste presentano una amplificazione in bassa frequenza, mentre al contrario quelli neutro-allegro presentano una attenuazione. Questo fenomeno è confermato da studi sulla distribuzione spettrale nella voce emotiva (Banse & Scherer, 1996) anche se per i foni sonori tale fatto può essere influenzato delle variazioni di pitch che si riscontrano nel parlato emotivo. Ad esempio la voce triste solitamente è caratterizzata da un pitch più basso, per cui la corrispondente deriva verso il basso delle armoniche aumenta l'energia spettrale in bassa frequenza.

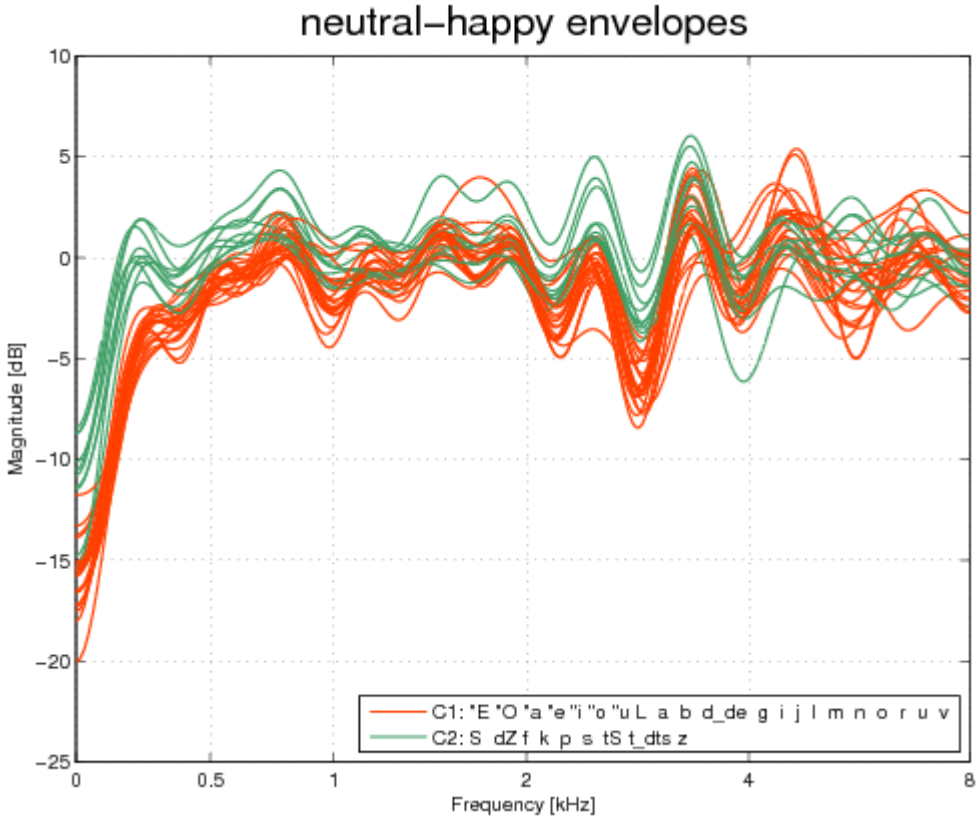

Figura 7: Inviluppi differenziali medi per fono nel caso neutro-allegro.

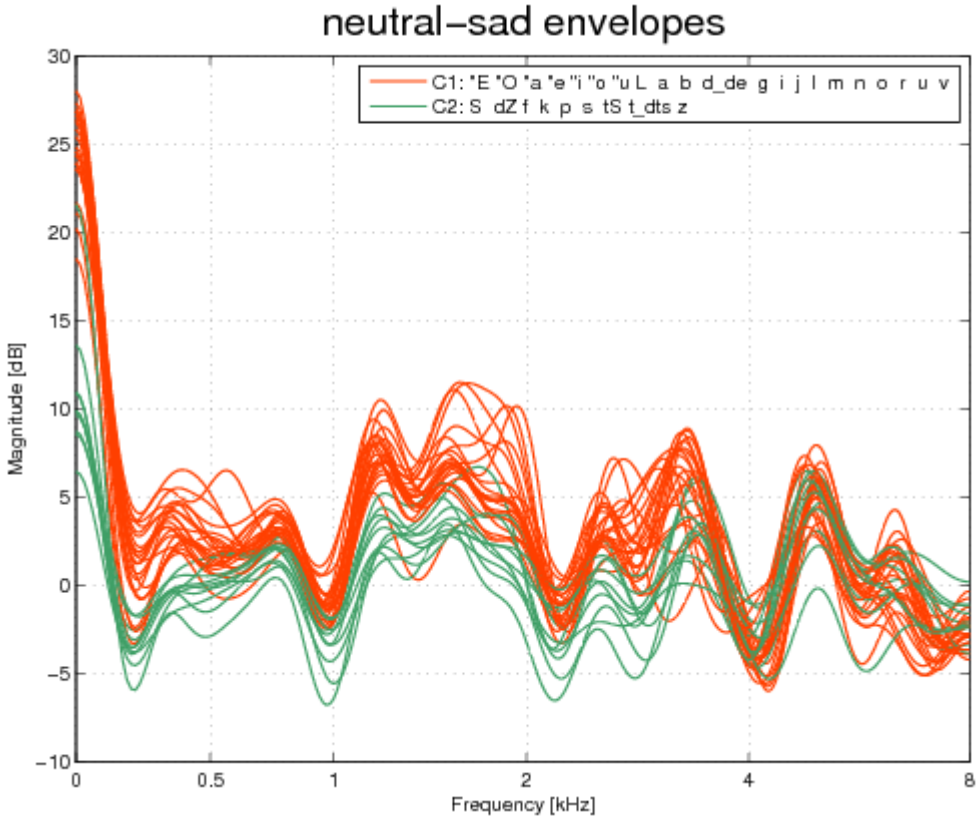

Figura 8: Inviluppi differenziali medi per fono nel caso neutro-triste.

Altri dettagli si possono vedere attorno ai 4 kHz; al di sotto di questa frequenza c'è una divisione abbastanza netta tra foni sordi e sonori, mentre al di sopra i comportamenti dei foni sembrano avere lo stesso comportamento. E interessante notare che il valore di 4000 Hz corrisponde alla *maximum voiced frequency* utilizzata in altri lavori (Pantazis & Stylianou, 2008).

5. CONCLUSIONI

In questo lavoro è stato presentato un nuovo metodo per analizzare corpus di tipo parallelo utilizzando l'analisi mel-cepstrale differenziale in combinazione con il *clustering* gerarchico agglomerativo. Il metodo è stato utilizzato in un database emotivo mostrando come gli inviluppi spettrali variano tra diverse emozioni. Le risultanti partizioni dei due dendogrammi neutro-allegro e neutro-triste mettono in evidenza come la variazione di timbro tra le varie emozioni dipenda dal fono preso in considerazione e quanto la sonorità sia un fattore rilevante.

Le feature utilizzate, derivate dall'analisi mel-cepstrale differenziale, descrivono il filtro necessario per trasformare un inviluppo spettrale neutro nella sua controparte emotiva;

l'osservazione dei dendogrammi suggerisce gruppi di trasformazioni omogenee da applicare nell'ambito della *voice conversion*. Inoltre analizzando in dettaglio gli inviluppi spettrali differenziali si possono suggerire trattamenti diversi per le regioni frequenziali sopra o sotto i 4 kHz.

E' tuttavia importante notare che l'espressione delle emozioni è dipendente dal parlatore e per generalizzare i risultati qui ottenuti sarà necessario estendere questo studio considerando più parlatori.

RINGRAZIAMENTI

L'autore G. S. ringrazia il Swedish Research Council (Vetenskapsrådet grant 2009-4599). L'autore F. T. ringrazia il dipartimento "Speech, Music and Hearing" KTH, Stockholm, Sweden, per l'opportunità di visitare il loro laboratorio. Questo lavoro è stato parzialmente supportato dal progetto europeo FP6 COMPANIONS, "www.companions-project.org", IST 034434 e dal progetto europeo FP7 "ALIZ-E" (grant number 248116).

BIBLIOGRAFIA

Banse, R. & Scherer, K. R. (1996), Acoustic profiles in vocal emotion expression, *Journal of personality and social psychology*, 70(3), 614–36.

Fukada, T., Tokuda, K., Kobayashi, T. & Imai, S. (1992), An adaptive algorithm for mel-cepstral analysis of speech, in *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing*, 92(1), 137–140.

Imai, S. (1983), Cepstral analysis synthesis on the mel frequency scale, in *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing*, 8, 93–96.

Johnson, S. C. (1967), Hierarchical clustering schemes, *Psychometrika*, 32(3), 241–254.

Kain, A. & Macon, M. (2001), Design and evaluation of a voice conversion algorithm based on spectral envelope mapping and residual prediction, in *Proceedings of IEEE ICASSP*, 2, 813–816.

Mak, B. & Barnard, E. (2008), Phone clustering using the Bhattacharyya distance, in *Proceedings of ICSLP96*, 4, 2005–2008.

Meila, M. (2007), Comparing clusterings—an information based distance, *Journal of Multivariate Analysis*, 98(5), 873–895.

Oppenheim, A. & Schafer, R. (1968), Homomorphic analysis of speech, *IEEE Transactions on Audio and Electroacoustics*, 16(2), 221–226.

Pantazis, Y. & Stylianou, Y. (2008), Improving the modeling of the noise part in the harmonic plus noise model of speech, *in Proceedings of IEEE ICASSP*, 4609–4612.

Salvi, G. (2003), Accent clustering in Swedish using the Bhattacharyya distance, in *Proceedings of 15th ICPhS*, 1-4.

Salvi, G. (2005), Advances in regional accent clustering in Swedish, in *Proceedings of Interspeech*, Lisbon, September 4-8 2005, 2841–2844.

Sokal,R. R. & Rohlf, F. J. (1962), The comparison of dendrograms by objective methods, *Taxon*, 11, 33–40.

Tesser, F., Cosi, P., Drioli, C. & Tisato, G. (2005), Emotional Festival-Mbrola TTS Synthesis, in *Proceedings of Interspeech*, Lisbon, September 4-8 2005, 505–508.

Wells, J. (1995), *Computer-coding the IPA: a proposed extension of SAMPA 1995*, ms. http://www.phon.ucl.ac.uk/home/sampa/x-sampa.htm.

Zen, H., Tokuda, K. & Black,A. W. (2009), Statistical parametric speech synthesis, *Speech Communication*, 51 (11), 1039–1064.