DUE TECNICHE DI VOCODING PER LA SINTESI DI PARLATO EMOTIVO MEDIANTE TRASFORMAZIONE DEL TIMBRO VOCALE

Fabio Tesser[1], Enrico Zovato[2], Mauro Nicolao[1,3], Piero Cosi[1]
[1]Istituto di Scienze e Tecnologie della Cognizione, Consiglio Nazionale delle Ricerche, Padova, Italia
[2]Loquendo S.p.A., Torino, Italia
[3]Speech and Hearing Research Group, University of Sheffield, United Kingdom
fabio.tesser@gmail.com, enrico.zovato@loquendo.com, mauro.nicolao@gmail.com, piero.cosi@pd.istc.cnr.it

1. ABSTRACT

In questo articolo vengono descritte due tecniche di modifica del timbro vocale utilizzate in un esperimento di trasformazione della voce con l'obiettivo di riprodurre alcune caratteristiche del parlato emotivo.

Il segnale vocale emesso da un parlatore con stile di lettura neutro viene convertito in modo da riprodurre l'inviluppo spettrale utilizzato dallo stesso parlatore in una situazione emotiva non neutra.

La funzione di conversione tra gli inviluppi spettrali è calcolata utilizzando un metodo ricavato con un addestramento su dati reali. Per questo motivo è stato preso in considerazione un database contenente la voce di un parlatore registrato durante la lettura/recitazione di un corpus di testi con diversi stili emozionali: allegro, triste e uno stile neutro di riferimento.

Le due tecniche di generazione della forma d'onda (vocoding) prese in considerazione sono il *Phase Vocoder* e il *filtro MLSA* (Mel Log Spectrum Approximation). I due prototipi implementati sono stati valutati con test di tipo percettivo, mentre valutazioni oggettive hanno convalidato l'efficacia della funzione di conversione.

2. INTRODUZIONE

Lo studio delle emozioni e della loro comunicazione attraverso la voce ha suscitato un crescente interesse nel mondo della ricerca negli ultimi anni (Scherer, 2003). Molti lavori recenti hanno focalizzato la loro attenzione sullo studio dei correlati psico-acustici delle emozioni nella voce. Tra questi si possono identificare due diversi gruppi di parametri: quelli prosodici e quelli legati al timbro vocale. Ritmo, velocità di eloquio, intonazione e intensità appartengono al primo gruppo, mentre posizione delle formanti e distribuzione dell'energia spettrale appartengono al secondo.

In precedenti lavori (Tesser et al., 2005), è stato proposto e analizzato un sistema per la sintesi del parlato emozionale che si basa su un modello ricavato con addestramento su dati reali (*data-driven*) per i moduli prosodici e l'applicazione di alcune regole euristiche per quello che riguarda la modifica della *voice quality*.

Anche se queste regole hanno fornito dei buoni risultati, il metodo non consente di seguire le rapide variazioni di timbro vocale che possono occorrere all'interno del parlato emozionale e non è un metodo sufficientemente flessibile per essere adattato facilmente a nuovi stili emozionali.

Per questa ragione è stato deciso di apprendere statisticamente, basandosi sull'analisi di dati reali, le caratteristiche spettrali della voce.

Questo lavoro si focalizza, infatti, sull'analisi del timbro vocale, facendo parte di un più ampio progetto riguardante anche l'analisi e la manipolazione delle caratteristiche prosodiche del parlato. Nello specifico, sono state valutate due diverse tecniche di vocoding nel contesto di esperimenti di trasformazione del timbro vocale volti a simulare alcune caratteristiche del parlato emozionale.

Il sistema di trasformazione del timbro vocale utilizzato prende come riferimento i lavori sviluppati nell'ambito delle tecniche di *voice conversion* (Stylianou et al., 1998). Solitamente queste tecniche sono utilizzate per convertire l'identità del parlatore, ma vari esperimenti sono stati eseguiti anche nell'ambito della trasformazione della voce in senso emozionale (Inanoglu & Young, 2009; Kawanami et al., 2003)

L'algoritmo di conversione può essere applicato a qualsiasi segnale vocale, ma per ottenere risultati coerenti, devono essere riprodotte le stesse condizioni della fase di apprendimento. Nel nostro caso è stato utilizzato un database di un unico parlatore e quindi i modelli ricavati sono *speaker dependent*. Un modello di questo tipo può essere applicato per convertire il segnale audio in uscita da un sistema di sintesi vocale a concatenazione di unità variabili in un segnale audio che rappresenti una voce più espressiva.

Un sistema di *voice conversion* è composto di due parti fondamentali: stima della funzione di trasformazione e sistema di risintesi del segnale.

La prima parte si occupa di predire i parametri psico-acustici della voce da simulare (*target*) partendo da quelli della voce originale (*source*).

Per stimare la funzione di conversione, ricavandola da dati reali, è stato necessario registrare un corpus parallelo: la voce dello stesso parlatore legge lo stesso testo con diversi stili emozionali. Le emozioni prese in considerazione sono state la tristezza e l'allegria, con l'aggiunta di uno stile neutro di riferimento.

Il secondo modulo di un sistema di voice conversion deve essere in grado di analizzare e re-sintetizzare la voce modificando i parametri psico-acustici sui quali si è intervenuto. Questi sistemi di elaborazione dei segnali sono appunto i vocoder e si basano su modelli di diversa complessità della voce e dei suoi parametri fondamentali.

Le due tecniche di vocoding prese in considerazione sono il *Phase Vocoder* (Portnoff, 1976) e il filtro MLSA*, Mel Log Spectrum Approximation* (Imai, 1983). In questo articolo sono proposte due diverse implementazioni per la risintesi del segnale vocale usando questi due vocoder.

3. INVILUPPO SPETTRALE

L'utilizzo di tecniche data-driven richiede un metodo compatto e robusto per descrivere il timbro vocale. In questo lavoro è stato scelto di utilizzare l'inviluppo spettrale ricavato dall'analisi mel-cepstrale (Imai, 1983; Fukada et al., 1992).

Questo metodo descrive lo spettro di una porzione (frame) di segnale audio nel dominio della trasformata Z utilizzando M+1 coefficienti e un modello acustico percettivo per valutare le diverse frequenze in accordo con il modello umano:

$$H(z) = \exp \sum_{m=0}^{M} \tilde{c}(m)\tilde{z}^{-m} \quad (1)$$

dove z̃ rappresenta la coordinata nello spazio nella trasformata Z modificata in modo da approssimare la scala Mel (Imai, 1983) .

Per calcolare i coefficienti c̃ l'algoritmo adotta un metodo di ottimizzazione che minimizza la distanza direttamente nel dominio percettivamente rilevante della scala Mel per le frequenze e logaritmica per le ampiezze.

Il timbro vocale di un frame di segnale vocale è quindi rappresentato dal corrispondente vettore di coefficienti mel-cepstrali; un esempio di inviluppo spettrale ottenuto dall'analisi mel-cepstrale di un frame di segnale vocale è mostrata in Figura 1.

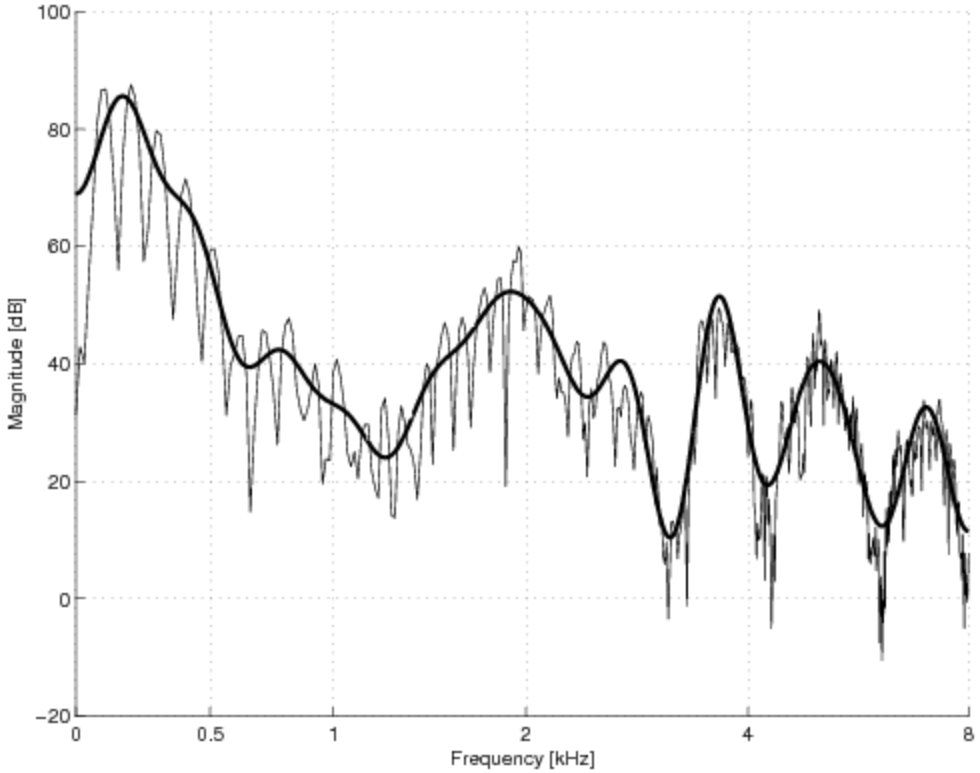

Figura 1: Spettro DFT (linea sottile) e inviluppo spettrale (linea grossa) relativa ad un frame di segnale vocale. L'inviluppo spettrale è ricavato direttamente dai coefficienti mel-cepstrali. La scala delle ascisse è proporzionale alla scala Mel.

Negli esperimenti i file audio sono stati campionati ad una frequenza di 16 kHz, sono stati analizzati utilizzando una finestra di analisi di 40 ms e 10 ms di overlap, infine i primi 26 coefficienti mel-cepstrali sono stati presi in considerazione. Questi parametri sono stati estratti dal segnale mediante il toolkit SPTK[1].

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

4. STIMA DELLA FUNZIONE DI CONVERSIONE

La funzione di conversione è una rappresentazione parametrica della funzione che mappa tra loro coppie coerenti di inviluppi spettrali appartenenti rispettivamente all'insieme neutro ed emozionale del database.

Con il termine coppie coerenti s'intende che i due elementi della coppia sono stati estratti da frame audio appartenenti a frasi parallele (stesso parlatore, stesso testo pronunciato) di due diversi insiemi (uno neutro ed uno emozionale) e corrispondenti alla stessa parte segmentale all'interno dei due file audio. Ad esempio: è stato associato il frame centrale del secondo fonema della registrazione appartenente all'insieme neutro al frame centrale del secondo fonema della registrazione appartenente all'insieme "triste". Per ottenere l'associazione tra coppie di vettori mel-cepstrali coerenti è stata utilizzata una procedura di allineamento automatico tramite DTW (Dynamic Time Warping).

Per stimare la funzione di trasformazione è stata seguita la procedura suggerita da (Stylianou et al., 1998). L'algoritmo utilizza una rappresentazione dello spazio acustico "neutro" tramite un modello statistico, detto GMM (Gaussian Mixture Model), che utilizza una funzione distribuzione di probabilità a mistura di gaussiane che è esplicitata nella seguente formula:

$$p(\boldsymbol{x}_t) = \sum_{i=1}^{Q} \alpha_i \mathcal{N}(\boldsymbol{x}_t; \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i) \tag{2}$$

dove $\mathbf{x}_t$ rappresenta il vettore mel-cepstrale, $\boldsymbol{\alpha}_i$ il peso della i-esima mistura, Q il numero di gaussiane mentre i parametri $\boldsymbol{\mu}_i$ e $\boldsymbol{\Sigma}_i$ sono ricavati dai dati "neutri" utilizzando l'algoritmo di Expectation Maximization.

La funzione di trasformazione ha invece la seguente forma parametrica:

$$\boldsymbol{y}'_t = \mathcal{F}(\boldsymbol{x}_t) = \sum_{i=1}^{Q} P(C_i|\boldsymbol{x}_t) \left[\boldsymbol{\nu}_i + \boldsymbol{\Gamma}_i \boldsymbol{\Sigma}_i^{-1} (\boldsymbol{x}_t - \boldsymbol{\mu}_i)\right] \tag{3}$$

dove il primo termine nella sommatoria è la probabilità condizionata che $\mathbf{x}_t$ appartenga alla mistura i-esima, mentre i parametri $\boldsymbol{\nu}_i$ e $\boldsymbol{\Gamma}_i$ sono calcolati attraverso la soluzione di un sistema di equazioni con il metodo *least square*.

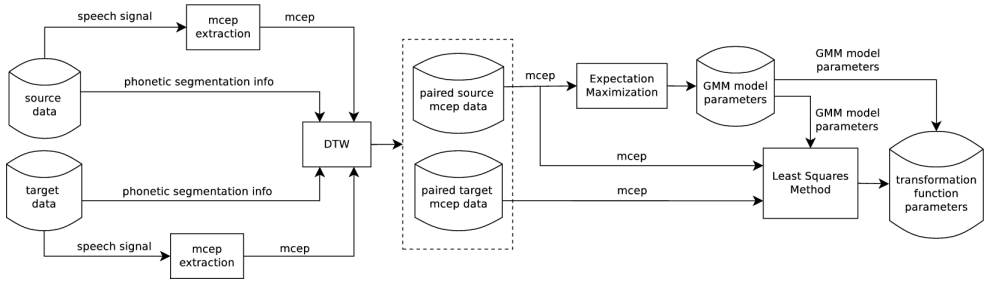

Figura 2: Diagramma funzionale relativo alla stima della funzione di conversione.

Il diagramma di Figura 2 riassume la procedura adottata per il calcolo dei parametri.

Utilizzando la funzione di conversione, quindi è possibile calcolare l'inviluppo spettrale relativo a un'emozione a partire da quello neutro dello stesso frame.

Uno svantaggio di questa tecnica è che la predizione è "frame per frame", e quindi manca di coerenza dinamica. Una prima soluzione a questo problema consiste nell'aggiungere le derivate prima e seconda dei coefficienti mel-cepstrali all'interno della procedura di training. Nella fase di addestramento sono stati fatti degli esperimenti per verificare questa ipotesi.

Nella Figura 3 è mostrato l'errore di predizione (distanza mel-cepstrale) a seconda del numero di componenti gaussiane, e si può confermare che la presenza dei coefficienti dinamici riduce l'errore di predizione.

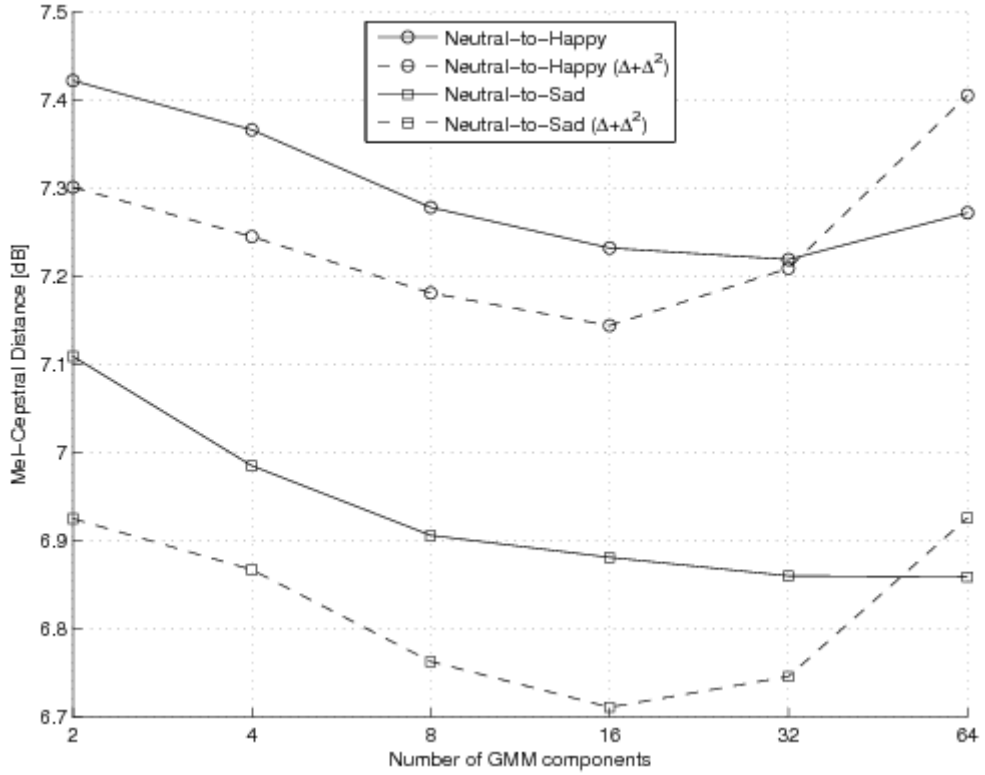

Figura 3: Errore di predizione in funzione del numero di componenti gaussiane. L'errore è stato calcolato utilizzando i dati presenti nel test set. Il test set è costituito dal 20 frasi scelte in modo casuale tra quelle nel corpus (200 frasi).

Come si vede dalla Figura 3, le performance migliori sono ottenute con un numero di gaussiane pari a 16 e con il calcolo dei coefficienti dinamici; questi modelli sono stati quindi utilizzati nella predizione degli inviluppi spettrali del parlato emotivo.

5. I VOCODER

Partendo dagli inviluppi spettrali predetti, sono state sperimentate due tecniche per risintesi del segnale con il timbro vocale modificato: Phase Vocoder implementato tramite FFT (Portnoff, 1976) e filtro MLSA (Imai, 1983).

5.1. Phase Vocoder

La prima tecnica si basa sulla rappresentazione spettrale dei frame ricavati dal segnale vocale (STFT) e permette di manipolare il segnale audio sia nel dominio temporale sia in quello spettrale (Portnoff, 1976). Il Phase Vocoder è principalmente utilizzato per operazioni di *pitch shifting* o *time stretching*, ma può essere utilizzato anche per modificare le caratteristiche timbriche della voce.

La Figura 4 mostra lo schema implementato: i coefficienti mel-cepstrali $\mathbf{x}_t$ sono calcolati tramite l'analisi mel-cepstrale sul segnale "neutro", il vettore *target* $\mathbf{y}_t'$ è calcolato partendo da $\mathbf{x}_t$ con la funzione di conversione (3). Questi due vettori sono utilizzati per calcolare il rapporto tra gli inviluppi spettrali nel dominio della FFT:

$$R(f) = \frac{Y_t'(f)}{X_t(f)} \tag{4}$$

Questo rapporto, dipendente dalla frequenza, è utilizzato come fattore moltiplicativo (*gain*) nell'implementazione del Phase Vocoder che elabora il modulo FFT del segnale neutro in ingresso. Come conseguenza, al segnale in uscita è stato imposto l'inviluppo spettrale predetto $\mathbf{y}_t'$.

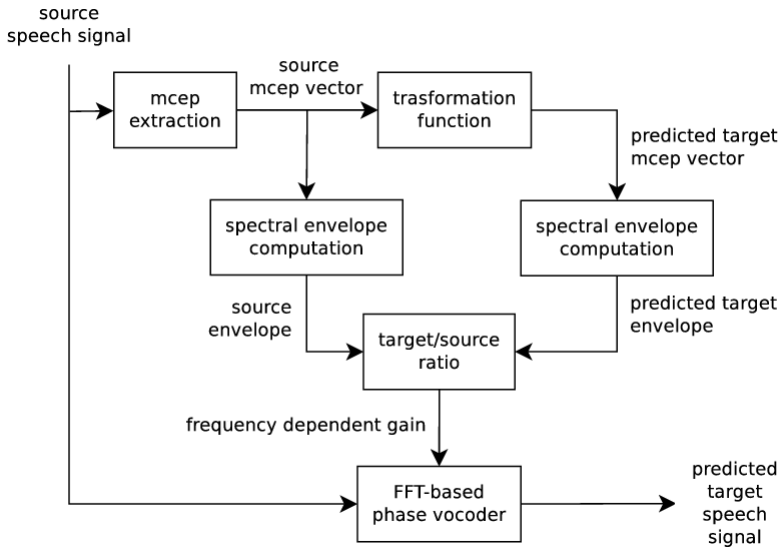

Figura 4: Diagramma funzionale della trasformazione di inviluppo spettrale utilizzando il Phase Vocoder.

5.2. Il filtro MLSA

La seconda tecnica si basa sul modello sorgente-filtro della produzione vocale: la risposta in frequenza del filtro MLSA (Mel Log Spectrum Aproximation) (Imai, 1983) può essere controllata direttamente dai coefficienti mel-cepstrali.

Per modificare il timbro vocale due filtri MLSA sono utilizzati in uno schema di *sbiancamento* e successiva rimodellazione dello spettro, la Figura 5 mostra lo schema utilizzato.

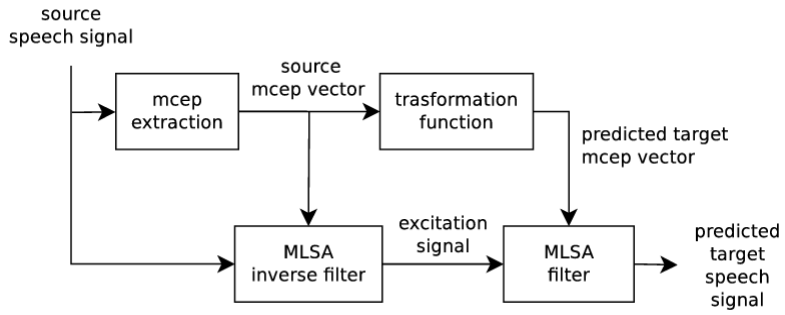

Figura 5: Diagramma funzionale della trasformazione di inviluppo spettrale utilizzando il filtro MLSA.

I vettori source ($\mathbf{x}_t$) e target ($\mathbf{y}_t'$) sono stati ricavati come nel caso del Phase Vocoder. Lo spettro del segnale audio source è stato sbiancato utilizzando un filtro inverso al suo stesso inviluppo. Questa operazione ha permesso di ottenere il segnale di eccitazione simile alla sorgente glottale. Infine, questo segnale è stato inviato in ingresso ad un filtro con risposta in frequenza pari all'inviluppo spettrale target.

6. VALUTAZIONE SOGGETTIVA

Le due tecniche e le due trasformazioni (neutro-triste, neutro-allegro) sono state valutate tramite un test soggettivo di ascolto che prende in considerazione sia la naturalezza della trasformazione, sia la percezione delle emozioni. Il test è stato proposto tramite un'interfaccia web suggerendo ai partecipanti di utilizzare delle cuffie audio. Il test permetteva di assegnare a ognuna delle frasi sintetiche un punteggio in una scala di gradimento MOS (Mean Opinion Score) a 5 punti. I trenta partecipanti hanno prodotto i risultati illustrati nelle Figure 6 e 7.

L'analisi dei risultati significativamente differenti dei vari metodi comparati con lo stile neutro è stata effettuata con un analisi di tipo t-test binario, ed i risultati sono riportati nelle Tabelle 1 e 2.

I dati visualizzati in Figura 6 mostrano che le elaborazioni effettuate conservano un buon grado di naturalezza in entrambi i casi. Analizzando però i dati relativi alla di Tabella 1, si vede che le differenze di naturalezza a confronto con le frasi originali sono significative, e che quindi la naturalezza delle frasi originali è significativamente migliore rispetto alle frasi modificate.

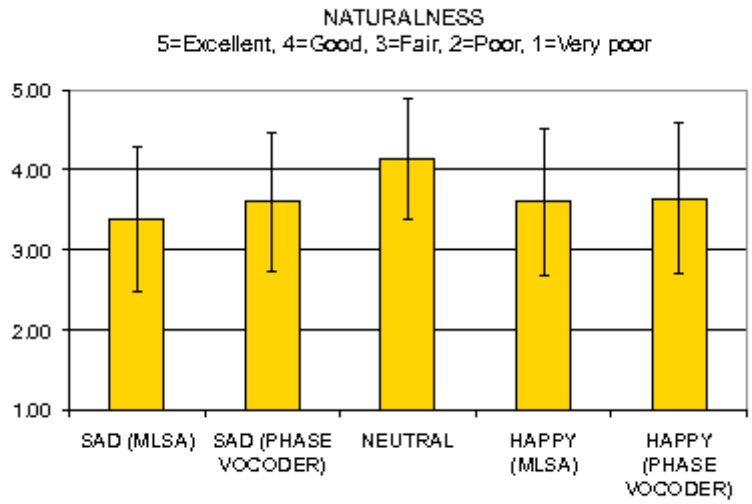

Figura 6: Risultati del test percettivo sulla naturalezza delle frasi originali (NEUTRAL) e modificate con le due tecniche di vocoding per le due emozioni.

METODO	t	P	DIFFERENZA SIGNIFICATIVA
HAPPY (MLSA)	2,8935	0,0054	SI
HAPPY (PHASE VOCODER)	2,3348	0,0230	SI
SAD (MLSA)	3,9111	0,0002	SI
SAD (PHASE VOCODER)	3,4951	0,0009	SI

Tabella 1: Analisi t-test sulla percezione sulla naturalezza delle frasi modificate a confronto con le frasi originali (NEUTRAL).

Per quello che riguarda la percezione delle emozioni si evince da Figura 7 e Tabella 2 che la trasformazione neutro-triste ha risultati migliori (e statisticamente differenti) rispetto a quella tra neutro e allegro. Inoltre, la tristezza è percepita meglio quando la conversione è effettuata con il filtro MLSA. Questo risultato potrebbe derivare dalla proprietà del filtro MLSA di modellare molto bene la distribuzione spettrale in bassa frequenza, visto che tale banda ha un ruolo molto importante nel parlato con stile triste. Nel caso della trasformazione neutro-allegro le due tecniche hanno ottenuto risultati comparabili.

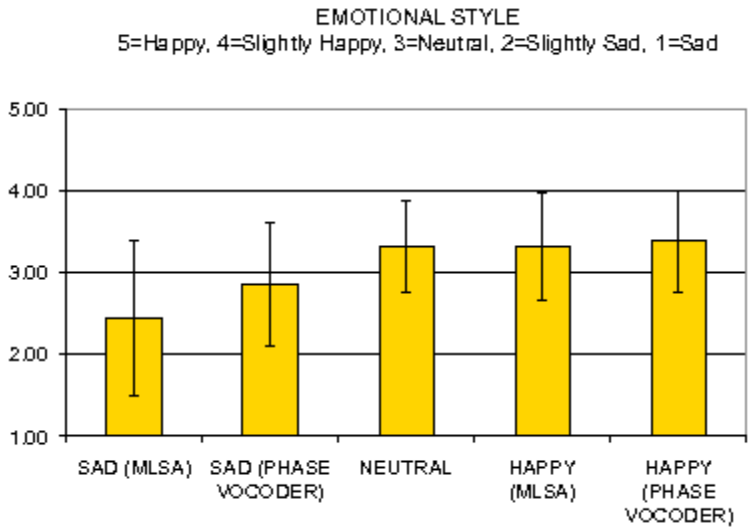

Figura 7: Risultati del test sulla percezione delle emozioni delle frasi originali (NEUTRAL) e modificate con le due tecniche di vocoding per le due emozioni.

METODO	t	P	DIFFERENZA SIGNIFICATIVA
HAPPY (MLSA)	1,6328	0,1079	NO
HAPPY (PHASE VOCODER)	0,8868	0,3788	NO
SAD (MLSA)	3,8915	0,0003	SI
SAD (PHASE VOCODER)	2,8595	0,0059	SI

Tabella 2: Analisi t-test sulla percezione delle emozioni delle frasi modificate a confronto con le frasi originali (NEUTRAL).

7. CONCLUSIONI

In questo lavoro sono stati proposti, implementati e testati due schemi di risintesi del segnale vocale utili per la conversione degli inviluppi spettrali della voce. In particolare tali schemi sono stati testati nell'ambito di un sistema di *voice conversion* adatto a modificare il timbro vocale di una voce emotivamente neutra in una voce con timbro vocale allegro o triste.

La valutazione oggettiva relativa alla funzione di trasformazione mostra l'effettiva riduzione della distanza spettrale tra gli inviluppi neutri modificati e gli inviluppi target emozionali. Dai risultati, diagrammati in Figura 3, si vede che la distanza spettrale è minore nel caso di trasformazione da neutro a triste, piuttosto che neutro-allegro.

I risultati dei test percettivi mostrano che entrambi i metodi proposti hanno modificato lo spettro in modo da risultare percettivamente più vicini al timbro “allegro” o “triste”.

In particolare il filtro MLSA ha ricevuto maggiori preferenze per quello che riguarda la trasformazione neutro-triste.

Un risultato del test soggettivo riguarda la naturalezza, che è stata giudicata buona anche per le frasi modificate con i due sistemi: questo conferma che la conversione non ha introdotto artefatti.

Infine si sottolinea che queste frasi sono state ottenute dalla sola modifica del timbro vocale e che la percezione delle emozioni sollecitate potrà essere sicuramente migliorata aggiungendo ai vari schemi l'adeguata trasformazione prosodica.

RINGRAZIAMENTI

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). L'autore F.T. ringrazia il dipartimento "Speech, Music and Hearing" KTH, Stockholm, Sweden, per l'opportunità di visitare il loro laboratorio. Gli autori infine ringraziano tutti i partecipanti al test percettivo.

BIBLIOGRAFIA

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.

Inanoglu, Z. & Young, S. (2009), Data-driven emotion conversion in spoken English, *Speech Communication*, 51, 3, 268-283.

Kawanami, H., Iwami, Y., Toda, T., Saruwatari, H. & Shikano, K. (2003), GMM-based voice conversion applied to emotional speech synthesis, in *Proceedings of EuroSpeech*, Geneva, Switzerland, 2401-2404.

Portnoff, M. (1976), Implementation of the digital phase vocoder using the fast Fourier transform, *IEEE Transactions on acoustics, speech and signal processing*, 24, 3, 243-248.

Scherer, K. R. (2003), Vocal communication of emotion: A review of research paradigms, *Speech communication*, 40, 1-2, 227–256.

Stylianou, Y., Cappé, O. & Moulines, E. (1998), Continuous probabilistic transform for voice conversion, *IEEE Transactions on Speech and Audio Processing*, 6, 2, 131–142.

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