CONFRONTO TRA DIVERSE TECNICHE DI CONVERSIONE PER LA SINTESI TTS DELLE EMOZONI

Mauro Nicolao, Carlo Drioli, Piero Cosi

Istituto di Scienze e Tecnologie della Cognizione - Sede di Padova "Fonetica e Dialettologia"
Consiglio Nazionale delle Ricerche, via Martiri della Libertà, 2 - 35127 Padova, Italia
nicolao@pd.istc.cnr.it, drioli@pd.istc.cnr.it, cosi@pd.istc.cnr.it

SOMMARIO

Nel presente lavoro vengono descritti gli sviluppi delle tecniche per la creazione di una funzione che converte un segnale vocale neutro in uno caratterizzato emotivamente, basate su quanto già sviluppato in precedenti lavori (Nicolao et alii, 2005; Nicolao et alii, 2006).

Sono stati investigati numerosi metodi per creare questa funzione e si è cercato di stabilire l'efficienza delle diverse trasformazioni, attraverso dei criteri oggettivi (distanza di Itakura-Saito) e soggettivi (test percettivi).

Tutte le funzioni sviluppate si basano su un approccio di tipo statistico. Nello specifico, per descrivere lo spazio acustico del segnale vocale neutro vengono utilizzati sia modelli a mistura di funzioni gaussiane (GMM), sia catene di Markov nascoste (HMM).

E' stata valutata anche la possibilità di applicare le funzioni di conversione in vari punti del sistema: o come semplice *post processing* del segnale vocale neutro o agendo direttamente su un database di difoni utilizzato da un sintetizzatore vocale.

I segnali di riferimento per l'allenamento dei modelli statistici sono ricavati da due database di segnali vocali creati *ad hoc*. Uno è stato registrato con lo scopo di raccogliere il materiale per costruire una voce per un sintetizzatore a concatenazione di difoni (MBROLA o SMS). Si ipotizza che questo insieme di file audio sia privo di caratterizzazione emotiva. Lo stesso parlatore, ha inoltre registrato un database di file audio cercando di fornire ad essi una forte componente emotiva (l'emozione utilizzata in questo lavoro è la *collera*).

Da questi insiemi, tramite un processo di *copy synthesis*, si sono ottenuti due *corpora* perfettamente allineati per quanto riguarda durate, intonazione e fonemi pronunciati. Lo studio si è potuto quindi focalizzare sulla modellizzazione delle sole differenze provocate al segnale dall'emozione presente.

Lo spazio acustico del segnale sintetizzato è stato diviso in classi omogenee e, ad ognuna, è stata associata una diversa funzione di trasformazione. Il nostro metodo è, quindi, costituito da 34 funzioni, specializzate per ogni fonema.

In Tabella 1 sono elencati le differenze progettuali che differenziano i segnali analizzati nel progetto.

Modello Statistico	Punto di applicazione	Metodo di conversione
GMM o HMM	Al DB di difoni o al segnale vocale sintetizzato	Conversione completa, semplificata o GMM-smoothed

Tabella 1: Elementi di differenziazione tra i segnali ottenuti.

Nel test di conversione, per ogni segnale neutro considerato, sono stati ottenuti 12 segnali vocali trasformati, uno per ogni strategia di trasformazione considerata.

I migliori punteggi sono stati ottenuti dalle funzioni basate sul modello HMM e dal metodo di *post processing*.

1. SINTESI VOCALE E SINTESI VOCALE EMOTIVA

Il presente lavoro si colloca nell'ambito dello studio della sintesi vocale emotiva e mira a confrontare i risultati di varie tecniche di E-TTS.

Con il termine E-TTS (Emotive Text To Speech) si intende un sistema di sintesi vocale che, partendo da un testo scritto, sia in grado di trasformarlo in un segnale vocale che esprima anche una "emozione" decisa a priori.

Lo stato dell'arte, nell'ambito della sintesi da testo scritto, prevede la possibilità di

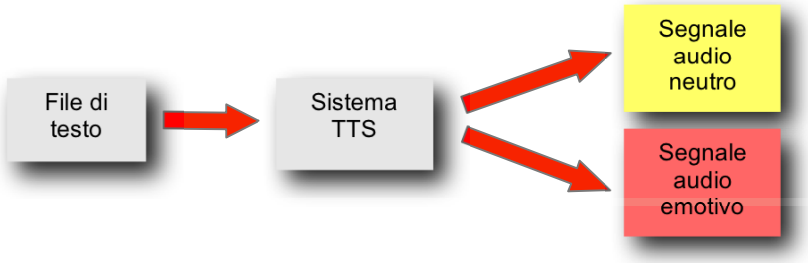

Figura 1: Schema di un sistema di sintesi TTS con la possibilità di sintesi delle emozioni

creare da qualsiasi file di testo un file audio, ma, come illustrato in Figura 2, mentre l'intelligibilità della comunicazione è garantita anche ad ottimi livelli, la naturalezza e la qualità complessiva della voce lasciano ancora insoddisfatti.

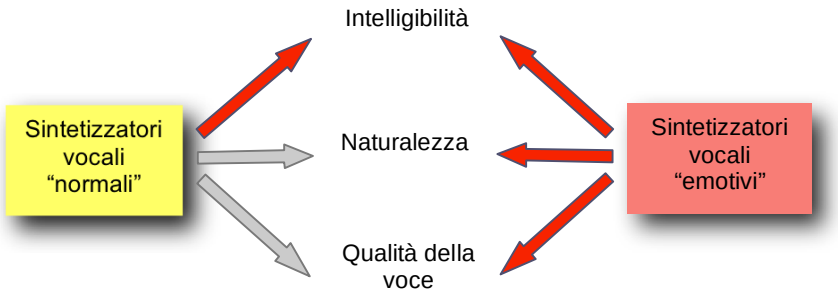

Figura 2: Differenze tra una voce sintetizzata neutra e una voce sintetizzata emotiva.

Si è cercato, quindi, di sviluppare una voce sintetizzata che esprima anche le caratteristiche della voce umana che vengono normalemente riconosciute come emotive.

2. ARCHITETTURA DEL PROGETTO

Il progetto si articola nelle seguenti fasi:

- Reperimento del *materiale audio*
- Copy synthesis

- Scelta delle *funzioni* di conversione dello spettro
- Allenamento dei modelli statistici
- *Esempi* di trasformazione con i modelli
- Analisi delle *performance*

2.1 Il materiale audio

Al fine di creare le varie funzioni di conversione, è necessario allenare un modello statistico, per questo sono necessari dei segnali di riferimento. Questi sono stati ricavati da due database di segnali vocali creati appositamente per l'esperimento.

Uno è stato registrato con lo scopo di raccogliere il materiale per costruire una voce per un sintetizzatore a concatenazione di difoni ed è stato organizzato in modo da ottenere del parlato privo di alcuna componente emotiva.

Lo stesso parlatore successivamente ha registrato un piccolo corpus a forte caratterizzazione emotiva. Come emozione da modellare è stata scelta la *collera* poiché è normalmente riconosciuta come una di quelle più identificabili e facilmente riproducibili.

2.2 La copy synthesis

Partendo dai materiali registrati, sono stati creati due *corpora* su cui allenare i modelli statistici. Questo è uno dei passi più delicati e peculiari del progetto poiché si è tentato di ottenere dei segnali di riferimento che differissero esclusivamente per le caratteristiche "*emotive*" che vogliamo modellare.

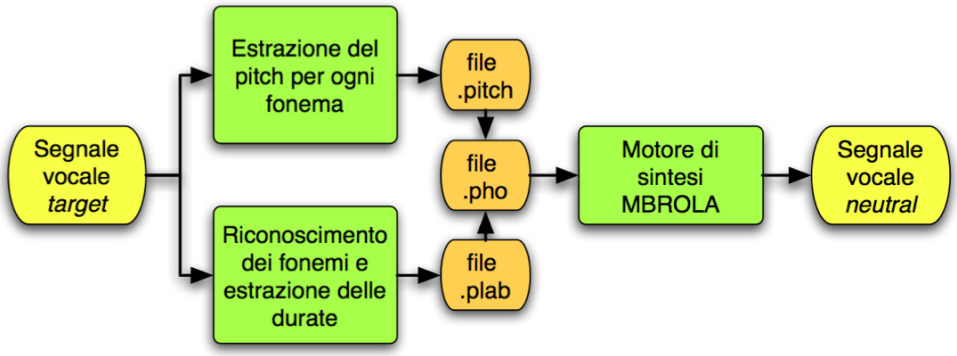

Figura 3: Schema del processo di *copy synthesis*.

I due insiemi di file audio ottenuti sono:

- una voce **emotiva** (*target*) costituita da tutto il materiale registrato, caratterizzato dalla collera
- una voce **neutra** prodotta con un sintetizzatore vocale a difoni (*neutral*)

La voce *neutral* è stata ottenuta attraverso un processo di *copy synthesis*[1], usando come punto di partenza la voce *target*, secondo lo schema illustrato in Figura 3.

Si è così ottenuto un insieme di file audio che, come riassunto in Figura 4, ha la stessa intonazione, gli stessi fonemi pronunciati con la stessa durata e, particolare importante, lo stesso timbro di voce dell'insieme *target*.

[1] Per maggiori dettagli sul processo di *copy synthesis* si può fare riferimento a (Nicolao et alii, 2005)

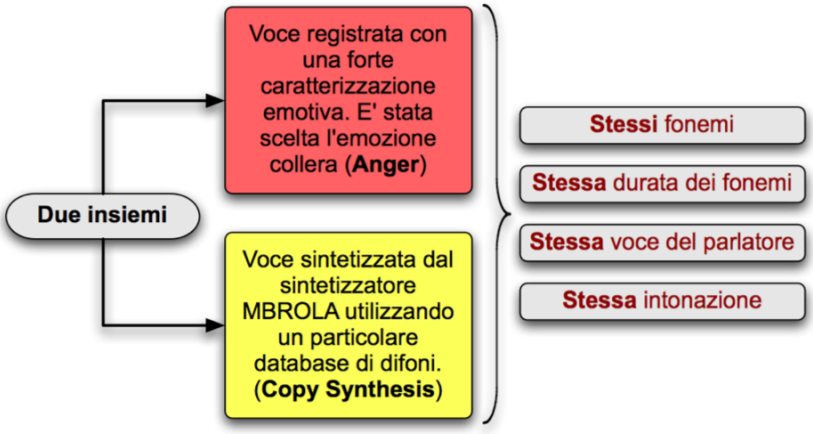

Figura 4: Corrispondenze tra i due corpora per l'allenamento del modello statistico.

3. IL SISTEMA DI CONVERSIONE

Sono state studiate varie metodologie per convertire un segnale neutro in uno "emotivo".

La prima scelta possibile riguarda il sintetizzatore vocale che deve essere utilizzato. Pur dando dei risultati acusticamente migliori, i sintetizzatori ad *unit selection* si prestano meno a manipolazioni del segnale durante il processo di sintesi; per questo motivo si è deciso di prendere in considerazione solo motori di sintesi a concatenazione di difoni.

In particolare sono stati scelti:

- MBROLA[2]
- SMS[3]

Grazie alla scelta di un sistema a concatenazione di difoni completamente *open source* come SMS, si è potuta analizzare anche l'efficacia dello spostamento del punto di applicazione della funzione di conversione:

- prima della generazione del segnale vocale (sui difoni del *database* utilizzato dal sintetizzatore)
- dopo della generazione del segnale vocale (come *post processing* del segnale sintetizzato)

Sono state inoltre analizzate varie funzioni di conversione possibili. Tutte comunque presentano delle caratteristiche comuni:

- agiscono sulla forma dell'inviluppo dello spettro.
- usano il modello statistico come base

[2] Motore di sintesi vocale che, partendo dalle etichettature dei fonemi e le relative durate, utilizzando un database di difoni precedentemente registrato, elabora le forme d'onda e crea un segnale vocale (MBROLA, *http://tcts.fpms.ac.be/synthesis/mbrola*)

[3] Motore di sintesi vocale che si basa sull'analisi e sulla risintesi della parte sinusoidale e residuale di un database di difoni per generare un segnale vocale. (Serra & Smith, 1990)

Si differenziano però, come vedremo più avanti nel dettaglio, per come vengono calcolati i parametri di conversione. Avremo quindi le funzioni:

- Full Conversion
- Vector Quantization
- Smoothed GMM and MAP Adaptation

Un'ulteriore specializzazione si ha per il tipo di modello statistico su cui si basano:

- Modello a mistura di gaussiane (GMM)
- Modello a catene di Markov nascoste (HMM).

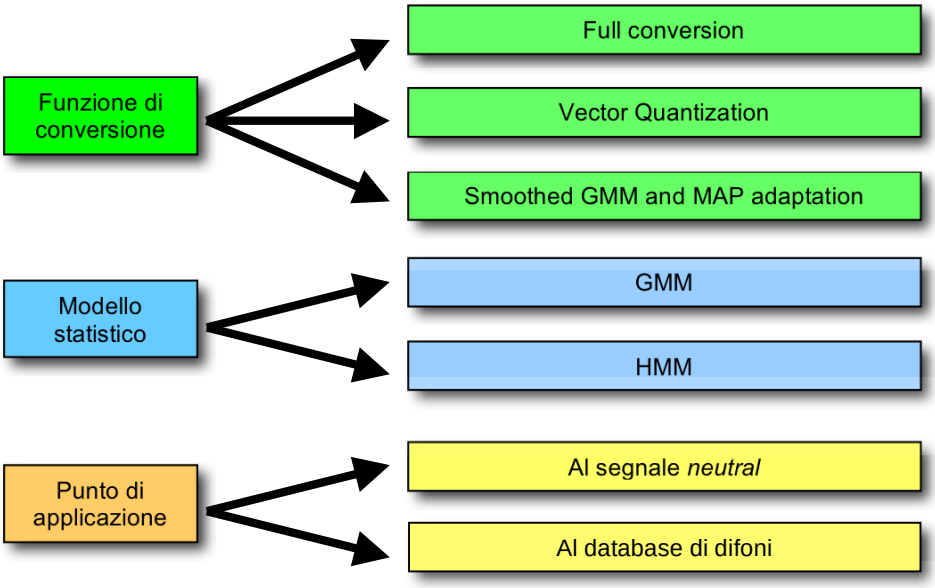

Figura 5: Le possibili architetture del sistema di conversione.

In Figura 5 si può vedere uno schema delle scelte progettuali fatte all'interno di questo progetto.

3.1 Le funzioni di conversione

Come elencato nel capitolo precedente, le funzioni di conversione utilizzate sono tre. Queste hanno caratteristiche diverse che si concretizzano principalmente in una maggiore o minore complessità di calcolo, condizionata dalla quantità di parametri da gestire, e nel numero di disturbi introdotti sul segnale prodotto.

Nella scelta di una funzione di elaborazione di un segnale, quest'ultimo elemento è uno dei criteri più importanti, infatti è necessario introdurre il minor numero possibile di artefatti o discontinuità nel segnale trasformato. Sfortunatamente, questo fenomeno è quasi inevitabile e deve essere quindi controllato attraverso adeguate scelte progettuali. Se lo scopo principale della trasformazione è fornire alla voce sintetizzata maggiore naturalezza, è, quindi, preferibile fornire alla voce una "emozione" di minore intensità, piuttosto che introdurre un disturbo che degradi il segnale.

Alla luce di queste considerazioni, sono state cercate le funzioni seguenti:

Full conversion

In questo caso, si assume che la funzione di conversione dei parametri abbia la seguente forma (Stylianou et alii, 1998).

$$\mathcal{F}(\mathbf{x}_n) = \sum_{i=1}^{M} P(\mathcal{C}_i \mid \mathbf{x}_n)[\nu_i \; + \; \mathbf{\Gamma}_i \mathbf{\Sigma}_i^{-1}(\mathbf{x}_n - \mu_i)] \tag{1}$$

dove $P(\mathcal{C}_i \mid \boldsymbol{x}_n)$, $\boldsymbol{\mu}_i$ e $\boldsymbol{\Sigma}_i$ sono i parametri derivanti dal modello statistico utilizzato. Il primo rappresenta la probabilità che un vettore ($\boldsymbol{x}_n$) dello spazio acustico *neutral* appartenga all'i-esima mistura ($\mathcal{C}_i$), mentre gli altri due elementi sono il valore centrale (il vettore delle medie $\boldsymbol{\mu}_i$) intorno al quale si collocano i vettori $\boldsymbol{x}_n$ e la dispersione caratteristica (la matrice delle covarianze $\boldsymbol{\Sigma}_i$), intorno a questo valore, del modello calcolato sullo spazio acustico del segnale *neutral*.

I parametri che definiscono questa funzione sono il vettore P-dimensionale $\boldsymbol{\nu}_i$ e la matrice di dimensione P x P, $\boldsymbol{\Gamma}_i$, con i=1,...,M (M, il numero di componenti della mistura) e sono calcolati tramite la risoluzione di un sistema lineare che mette in relazione lo spazio dei vettori del segnale *neutral* con quello del segnale *target*.

Anche se la funzione di conversione (1) non è supportata da un adeguato modello statistico teorico, può essere utile interpretare i parametri $\boldsymbol{\nu}$ e $\boldsymbol{\Gamma}$ come vettore delle medie e matrice della covarianza di un modello a mistura di gaussiane dello spazio acustico *target*.

Vector Quantization

Questa funzione di conversione (Stylianou et alii, 1998) è una versione semplificata della precedente. Infatti, si ottiene eliminando il prodotto tra matrici che è a secondo membro dell'addizione di (1) e si ottiene

$$\mathcal{F}(\mathbf{x}_n) = \sum_{i=1}^{M} P(\mathcal{C}_i \mid \mathbf{x}_n)\; \nu_i \tag{2}$$

in cui compare il solo parametro $\boldsymbol{\nu}$ da calcolare e sono scomparsi i riferimenti espliciti ai parametri del modello statistico[4]. Il parametro si ottiene dalla formula

$$\nu^{(k)} = (\mathbf{P^T P})^{-1}\mathbf{P}^T\mathbf{y}^{(k)} \tag{3}$$

che mette in relazione lo spazio *neutral*, i cui riferimenti compaiono all'interno della matrice **P**, con lo spazio *target*, rappresentato dal vettore **y**.

La principale motivazione per la quale viene introdotta questa semplificazione è la minore complessità nel calcolo dei parametri, ed è giustificata dal fatto che i valori, nel prodotto delle matrici semplificato, sono ampiamente trascurabili rispetto ai valori del primo membro.

[4] in realtà i parametri del modello statistico (vettore delle medie $\boldsymbol{\mu}$ e matrice delle covarianze $\boldsymbol{\Sigma}$) entrano ancora nella formula attraverso il calcolo della probabilità di attivazione di una gaussiane $P(\mathcal{C}_i \mid \boldsymbol{x}_n)$ e attraverso la matrice **P** della (3) (Nicolao et alii, 2005; Stylianou et alii, 1998).

Smoothed GMM and MAP adaptation

Questa ultima funzione si discosta dalle precedenti, perché mira a ottenere, non l'intero vettore dello spazio acustico *target* come nelle funzioni precedenti, bensì la differenza tra questo e quello dello spazio *neutral*. La funzione (Chen et alii, 2003) è quindi di tipo additivo,

$$\mathcal{F}(\mathbf{x}_n) = \mathbf{x}_n + \sum_{i=1}^{M} P(\mathcal{C}_i \mid \mathbf{x}_n)\,(\nu_i - \mu_i) \tag{4}$$

anche in questo caso il parametro della funzione è unico, ma viene calcolato in modo decisamente diverso:

$$\nu_i = \frac{r}{r + \sum_{n=1}^{Q} p_i(\mathbf{x}_n)}\mu_i + \frac{\sum_{n=1}^{Q} p_i(\mathbf{x}_n)\mathbf{y}_n}{r + \sum_{n=1}^{Q} p_i(\mathbf{x}_n)} \tag{5}$$

in cui Q è il numero di vettori dello spazio acustico *target* che è possibile mettere in relazione con quelli dello spazio *neutral*, *r* è una costante di adattamento empirica, mentre $p_i(\mathbf{x}_n) = P(\mathcal{C}_i \mid \mathbf{x}_n)$.

Questa funzione è stata introdotta poiché impedisce ai vettori trasformati di assumere valori che si discostino troppo da quello di partenza, si riescono così a ridurre molto disturbi quali i salti di fase o gli scarti di energia. Inoltre, essendo la dinamica da modellare meno ampia, le variazioni sono più precise.

Tutte e tre queste funzioni si basano su informazioni ricavate da modelli statistici, la loro efficacia è pesantemente influenzata dalla bontà di questi che, a sua volta, dipende dalla mole di dati che viene considerata per l'allenamento.

3.2 Applicazione del modello di conversione

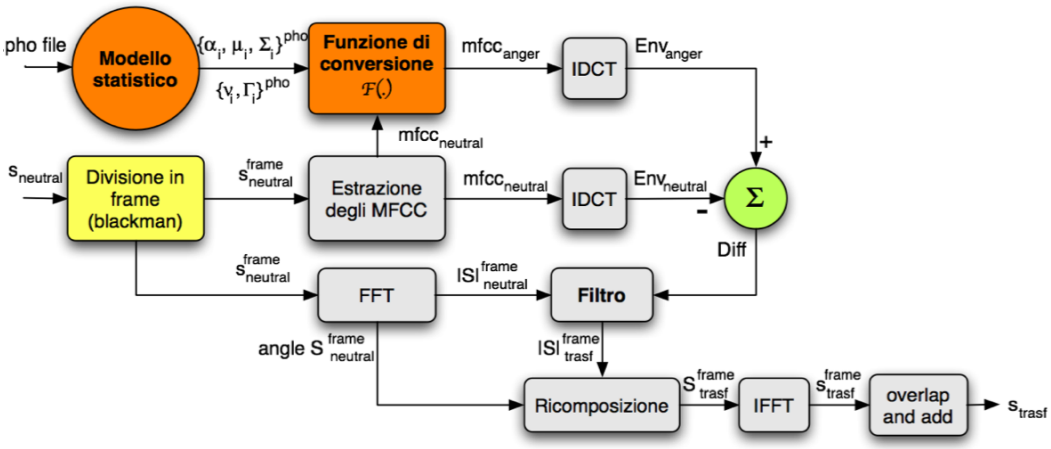

Figura 6: Processo di trasformazione di un segnale audio.

Le funzioni sopradescritte sono applicate a dei vettori di coefficienti MFCC (Mel-Frequency Cepstrum Coefficients) estratti dal segnale audio che deve essere trasformato. I coefficienti MFCC sono una rappresentazione dell'inviluppo spettrale di un segnale. Convertendo questi vettori, si converte di conseguenza quest'ultimo e la variazione calcolata viene quindi applicata allo spettro corrispondente.

In Figura 6 si può vedere lo schema completo del metodo attraverso il quale la funzione di conversione può essere applicata ad un segnale audio, sia esso costituito da una frase completa o da un singolo difono.

Da analisi fatte nel nostro precedente lavoro (Nicolao et alii, 2005) si è visto che questo sistema di applicazione è un ottimo metodo di conversione. Se i parametri della conversione sono buoni anche la voce trasformata che ne deriva è molto simile al *target* desiderato.

4. CALCOLO DEI PARAMETRI DELLE FUNZIONI E DEL MODELLO STATISTICO

Per calcolare i parametri delle funzioni di conversione illustrati in precedenza, è necessario sviluppare un sistema di allenamento. Si dovrà quindi:

- sviluppare un modello statistico che descriva lo spazio acustico neutral,
- risolvere i sistemi di equazioni lineari derivanti dalle formule descritte in precedenza che stabiliscano una relazione tra spazio acustico *neutral* e spazio acustico *target.*

Lo schema utilizzato è descritto in Figura 7.

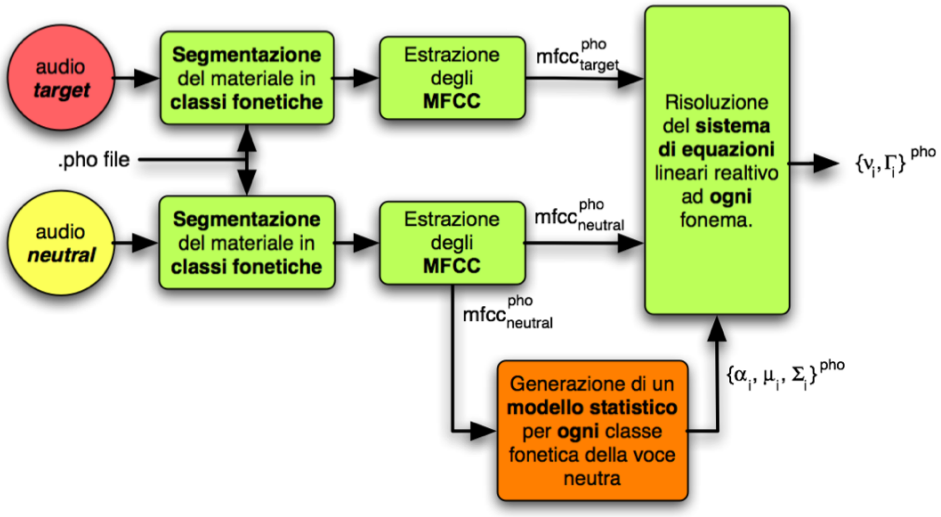

Figura 7: Architettura del metodo per l'allenamento del modello statistico e per il calcolo dei parametri delle funzioni di trasformazione.

4.1 I modelli statistici

Sono stati presi in considerazione due diversi modelli statisti:

- il modello a mistura di gaussiane (GMM – Gaussian Mixture Model)
- il modello a catene di Markov nascoste (HMM – Hidden Markov Model)

Entrambi i modelli permettono, dato un vettore di parametri in ingresso ***x***, di avere la probabilità che questo appartenga ad una specifica classe fonetica. Questo ha permesso, tra l'altro, di specializzare le conversioni, secondo la classe fonetica di appartenenza. Sono stati ottenuti 34 modelli diversi, uno per ogni fonema, e, per ognuno di questi, è stata calcolata la relativa funzione di trasformazione.

Entrambi i modelli statistici servono a descrivere la probabilità con cui si verificano delle occorrenze all'interno di uno spazio vettoriale (nel caso in oggetto, lo spazio acustico del segnale *neutral*) tramite una mistura di funzioni gaussiane multi dimensionali.

Il modello HMM, in aggiunta, fornisce una diversa descrizione dello spazio vettoriale in corrispondenza della parte del fonema (stato) che si sta analizzando (nel caso in oggetto, si hanno: parte iniziale di un fonema, parte centrale e parte finale).

4.2 Calcolo dei parametri delle funzioni

Attraverso le informazioni (contenute in un file *.pho*), relative alle trascrizioni e i fonemi che sono pronunciati, si divide il *training set*. In questo modo, si sono ottenute delle collezioni di materiale audio allineato e omogeneo rispetto al fonema pronunciato. Da queste è possibile partire per allenare contemporaneamente i modelli statistici, uno per ogni fonema pronunciato, e, in seguito, mettendo in corrispondenza il materiale *neutral* e quello *target*, i parametri da inserire nelle funzioni precedentemente illustrate.

Nel sistema di calcolo dei parametri ci sono delle differenze a seconda che il modello statistico, su cui si basa il sistema, sia GMM o HMM.

Nel caso di HMM, infatti, il risultato dell'allenamento del modello statistico dà in realtà una tripletta di parametri.

$$\{\alpha_i, \mu_i, \Sigma_i\}^{pho} \rightarrow \begin{Bmatrix} \{\alpha_i, \mu_i, \Sigma_i\}^{pho1} \\ \{\alpha_i, \mu_i, \Sigma_i\}^{pho2} \\ \{\alpha_i, \mu_i, \Sigma_i\}^{pho3} \end{Bmatrix}$$

In questo caso, è necessario decidere quale insieme di valori utilizzare per il calcolo dei parametri della funzione. Questo si ottiene con una scelta, *a posteriori*, della successione di stati più probabile, fatta attraverso l'algoritmo di Viterbi. Si applica un *buffer*, della durata del fonema che si sta analizzando. Questa analisi si colloca, dopo l'estrazione degli MFCC del segnale *neutral* e può essere rappresentata dal seguente schema.

Figura 8: Blocco di calcolo da inserire nello schema di Figura 7 per il caso di modello HMM.

Anche i parametri, che vengono calcolati, sono delle triplette di vettori, che andranno poi utilizzate in modo opportuno nell'effettivo sistema di sintesi vocale, inserendo il blocco di Figura 8 anche nello schema di Figura 6.

5. SISTEMA COMPLETO DI EMOTIVE TTS

Le funzioni con i parametri così allenati sono pronte ad essere inserite in un sistema di sintesi vocale completo.

5.1 Parametri sperimentali

Per eseguire il processo di trasformazione dei file audio *neutral* in file audio emotivi, sono utilizzati i seguenti parametri.

- i vettori sono composti da 36 coefficienti MFCC (incluso c0, velocità e accelerazioni) calcolati attraverso un banco di 100 filtri triangolari spaziati secondo la scala percettiva Mel.
- le finestre di osservazione con cui è stato diviso il segnale sono di 32 ms e sono separate da un passo incrementale di 8 ms.
- il segnale prima di essere convertito è normalizzato rispetto all'intensità.
- le classi fonetiche considerate sono 34.
- nel modello GMM, le funzioni gaussiane sono 32.
- per il modello HMM, sono stati creati modelli a 3 stati con 8 funzioni gaussiane per ogni stato.

Sono presi in esame ora i due metodi attraverso cui è possibile applicare la trasformazione.

5.2 Dopo la generazione del segnale audio

Il primo sistema di Emotive TTS, che è stato costruito, consiste in un sistema "classico" di sintesi seguito dall'applicazione della funzione di conversione (Figura 9).

Si parte da un testo scritto e dal database di difoni, registrato in precedenza, e si genera un file audio, che sarà di tipo *neutral*. Il sistema TTS utilizzato è il sistema di sintesi Festival per l'italiano sviluppato all'ISTC-SPFD-CNR (Cosi et alii, 2001) con il generatore di forma d'onda a scelta tra uno dei due visti in precedenza (MBROLA o SMS).

Generato il file audio, si applica la trasformazione spettrale, comandata dalle funzioni di conversione allenate, in base alle informazioni sui fonemi fornite dal sistema TTS.

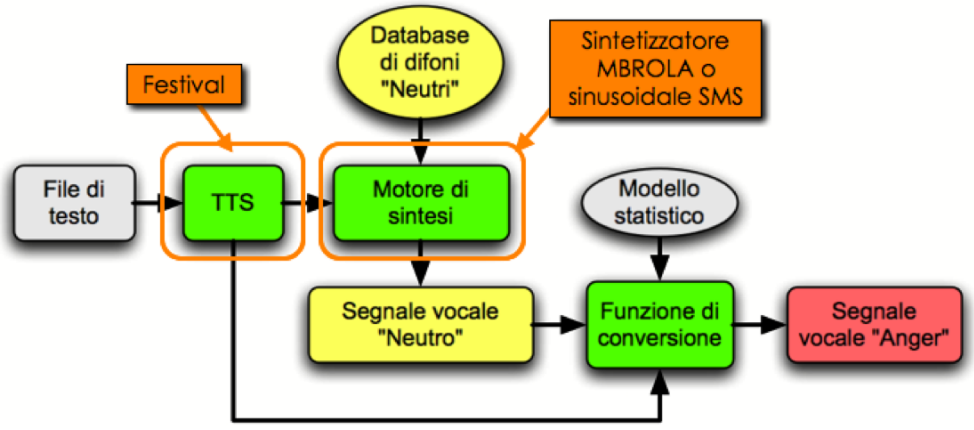

Figura 9: Schema completo di TTS "emotiva" con l'applicazione della funzione di conversione *dopo* la generazione del segnale audio.

Si tratta, quindi, di un metodo di *post processing* di un segnale audio precedentemente generato e presenta alcuni pregi e alcuni difetti.

La principale caratteristica positiva è che può essere utilizzato a prescindere dal sistema di sintesi, usando addirittura come punto di partenza un file audio non sintetizzato, basta solo avere la segmentazione del file audio con i fonemi pronunciati.

Le caratteristiche negative invece sono l'introduzione di artefatti non desiderati a causa di possibili salti tra i modelli, che non vengono gestite se non attraverso il modello a stati, che introducono discontinuità nella funzione globale di conversione. Inoltre il carico computazionale dell'intero processo è abbastanza alto e mediamente raddoppia i tempi di sintesi effettivi.

5.3 Dopo la generazione del segnale audio

Il secondo metodo, invece, si basa sull'inserimento della funzione di conversione, prima della creazione del file audio. Si utilizza per traformare il database di difoni registrato da *neutral* ad *anger* (Figura 10).

Ai file audio di difoni *neutral*, che costituiscono il database del sintetizzatore vocale, sono applicate le funzioni di conversione sulla base dei fonemi che contengono. Questo processo è molto delicato soprattutto per tutti i problemi di concatenazione tra i fonemi. Il modello che risolve meglio questo problema è sicuramente l'HMM, perché tiene conto negli stati delle zone di transizione tra fonemi.

Questa operazione ha la peculiarità di dover essere eseguita una sola volta per ogni emozione che di desidera modellare. Infatti, dopo che il database emotivo è stato creato, non sarà più necessario applicare la funzione di conversione. Così, pur moltiplicando la mole di dati da gestire, si velocizzano molto le operazioni di sintesi. A questo punto, i tempi di esecuzione dipenderanno solo dai tempi del motore di sintesi.

Questo metodo necessita, per essere applicato, di un sistema di sintesi in cui il database di difoni possa essere manipolato. Per questo, è stato fondamentale il lavoro sul sistema di sintesi SMS (Spectral Model Synthesis) svolto all'ISTC-SPFD-CNR (Sommavilla et alii, 2006), mentre non si è potuto utilizzare MBROLA.

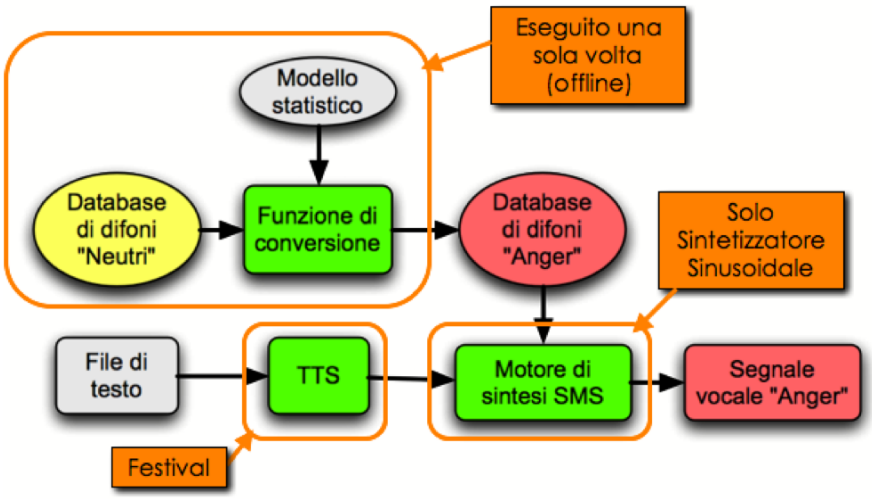

Figura 10: Schema completo di TTS "emotiva" con l'applicazione della funzione di conversione al *database* di difoni.

6. SEGNALI OTTENUTI

Come risultato delle varie prove effettuate, abbiamo ottenuto numerosi segnali vocali, che sono stati usati per eseguire i test, sia oggettivi che percettivi.

Si è partiti da 2 segnali audio *neutral*, uno appartenente al *training set* e uno proveniente direttamente da sintesi da testo scritto. Sono riportati i due segnali di partenza e il segnale *target* relativo al segnale del *training set.*

File audio *neutral* del training set	
File audio *target* del training set	
File audio da testo scritto	

I segnali ottenuti sono stati raggruppati per il metodo di conversione che è stato applicato. Un esempio dei risultati è contenuto nella Tabella seguente.

	Applicato al segnale				**Applicato al DB**			
	GMM		**HMM**		**GMM**		**HMM**	
Full Conversion								
Vector Quantization								
Smoothed GMM and MAP Adaptation								

6.1 Test oggettivi

Dare una misura oggettiva dell'emozione espressa da un segnale è molto complesso. Una prima osservazione può essere fatta, comunque, già con un confronto tra le forme d'onda e tra le gli inviluppi spettrali del segnale.

Nella Figura 11, sono disegnate le forme d'onda di un segnale *target,* del relativo segnale *neutral* e di una sua trasformazione con una funzione di conversione.

Per capire come la trasformazione agisca, si può notare la Figura 12 che contiene l'inviluppo spettrale dello stesso istante dei tre segnali sopra citati.

In entrambe le Figure precedenti, si vede come ci sia una tendenza del segnale *neutral* a trasformarsi ed a convergere verso quello *target.*

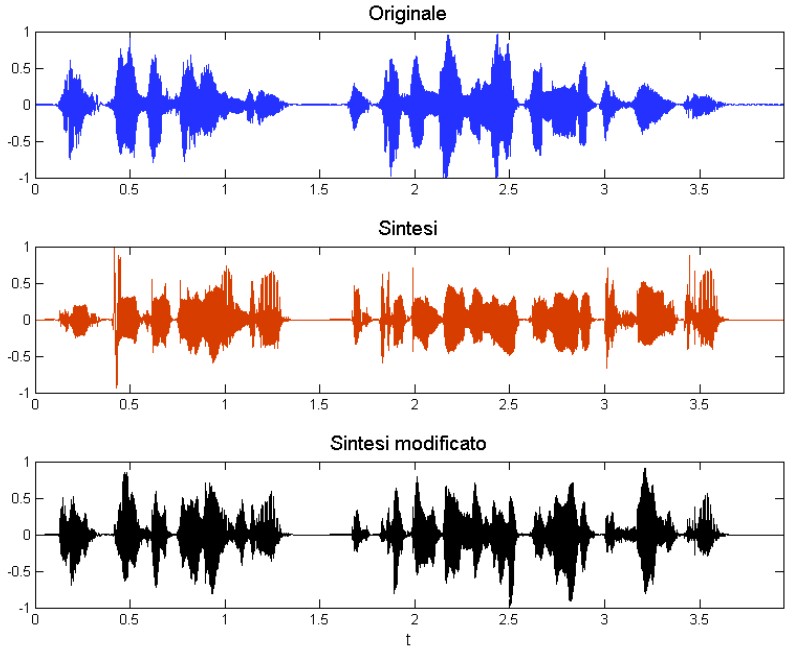

Figura 11: Confronto tra le forme d'onda del segnale originale (*target*), il segnale sintetizzato (*neutral*) e un segnale sintetizzato modificato (*anger*).

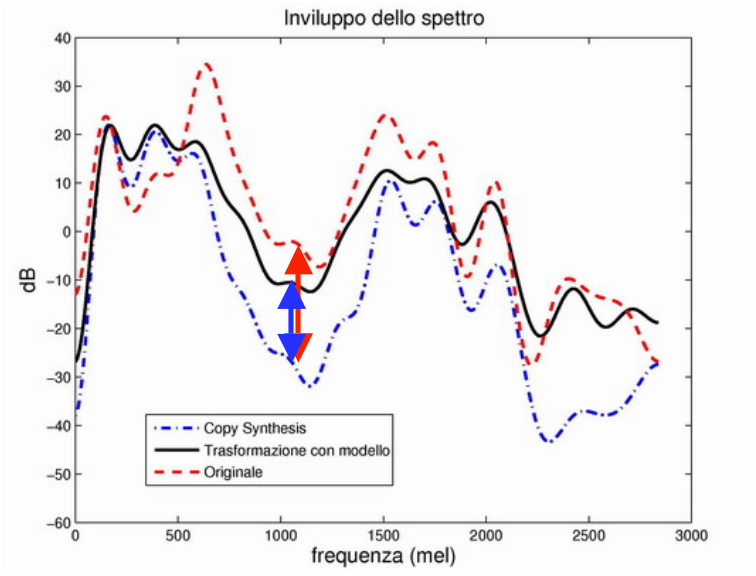

Figura 12: Esempio di trasformazione dell'inviluppo dello spettro di una finestra di segnale di *copy synthesis* (*neutral*) rispetto alla finestra di segnale originale (*target*) corrispondente.

Oltre al confronto visivo, si è cercato di associare anche una quantità numerica al problema. Si è deciso di valutare la qualità della conversione attraverso la misura delle differenze tra il segnale *trasformato* ottenuto e il segnale *target* che si sarebbe voluto imitare. Uno strumento che, in letteratura, è utilizzato spesso per questi scopi, è la distanza di *Itakura-Saito*. Essa misura, appunto, le variazioni dello spettro di un segnale rispetto ad un segnale di riferimento.

I risultati medi ottenuti sono rappresentati in Figura 13.

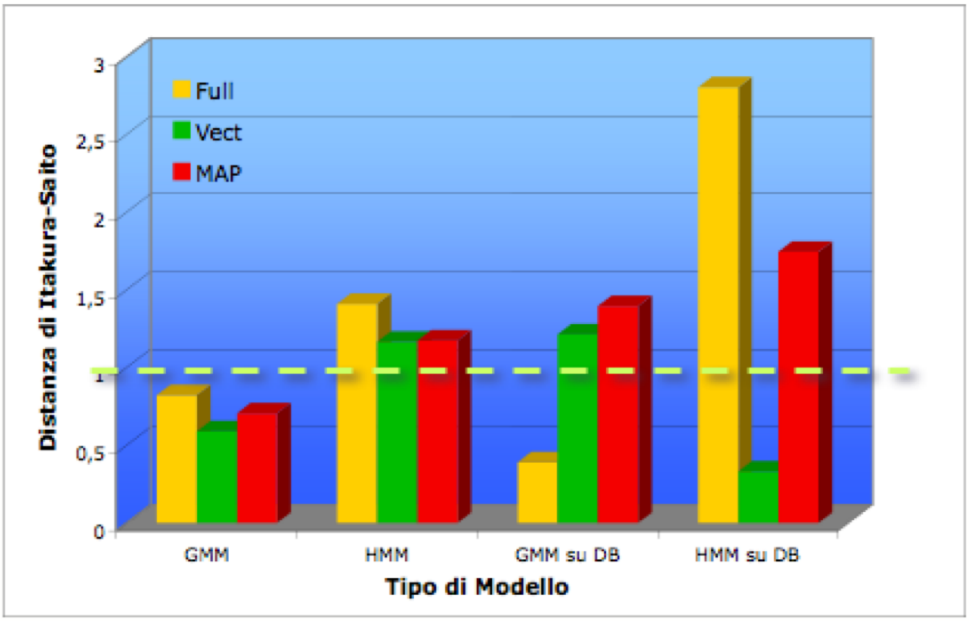

Figura 13: Differenze tra segnale *trasformato* e segnale *target*, normalizzate rispetto alla differenza tra segnale *neutral* e segnale *target* (linea tratteggiata).

In questa Figura sono rappresentate le distanze medie dei segnali ottenuti rispetto al segnale di riferimento (freccia blu la Figura 12). I risultati sono stati normalizzati rispetto alla distanza media tra il segnale di partenza e il segnale di riferimento (freccia rossa la Figura 12). In questo modo, se la distanza è minore di 1 (valore indicato dalla linea tratteggiata di Figura 13) vuol dire che il segnale trasformato s'avvicina, come inviluppo spettrale, al segnale target desiderato. Se è maggiore, invece, vuole dire che s'allontana.

E' stato possibile effettuare questo tipo di misure esclusivamente sui segnali sintetizzati attraverso il processo di *copy synthesis*, infatti solo per essi è possibile identificare un preciso segnale di riferimento (*target*).

Si vede che talvolta gli inviluppi dei segnali convertiti si discostano dagli inviluppi dei segnali *target* più di quanto si discostino quelli del segnale *neutral*. Questo non significa però che la trasformazione non stia sortendo gli effetti desiderati; è dovuto alla introduzione di disturbi che purtroppo alterano localmente il segnale ottenuto.

Si arriva a scoprire che sebbene la distanza di Itakura-Saito per un certo segnale sia maggiore di 1, il segnale risulta però più gradevole e percettivamente accettato come emotivo.

Inoltre, deve essere tenuto presente che i modelli che abbiamo generato sono stati allenati con relativamente poche ore di parlato. Questo implica che, per i fonemi meno frequenti e per i modelli più complessi, come gli HMM, si verificano più errori. Questo spiega perché in Figura 13, mediamente i risultati dei modelli HMM sono peggiori dei risultati dei modelli GMM.

6.2 Test percettivi

I test percettivi rimangono comunque il vero metro di giudizio per la qualità della sintesi emotiva. Servono a capire quanto un'emozione venga effettivamente percepita in un segnale audio. Come abbiamo visto in precedenza, alcuni metodi di conversione analizzati aumentano l'emozione riconosciuta nel segnale, ma, contemporaneamente, introducono disturbi, rumore e salti di fase che degradano il segnale. Si vede, però, che questi disturbi sono ben tollerati dall'ascoltatore che riconosce lo stesso l'emozione presente.

La misura soggettiva effettuata è costituita da un test percettivo informale in cui ai soggetti è stato chiesto di rispondere alla seguente domanda:

"La voce che ascoltate è arrabbiata?"

e di dare un punteggio (da 1 a 5) all'intensità della collera riconosciuta.

I risultati sono rappresentati in Figura 14.

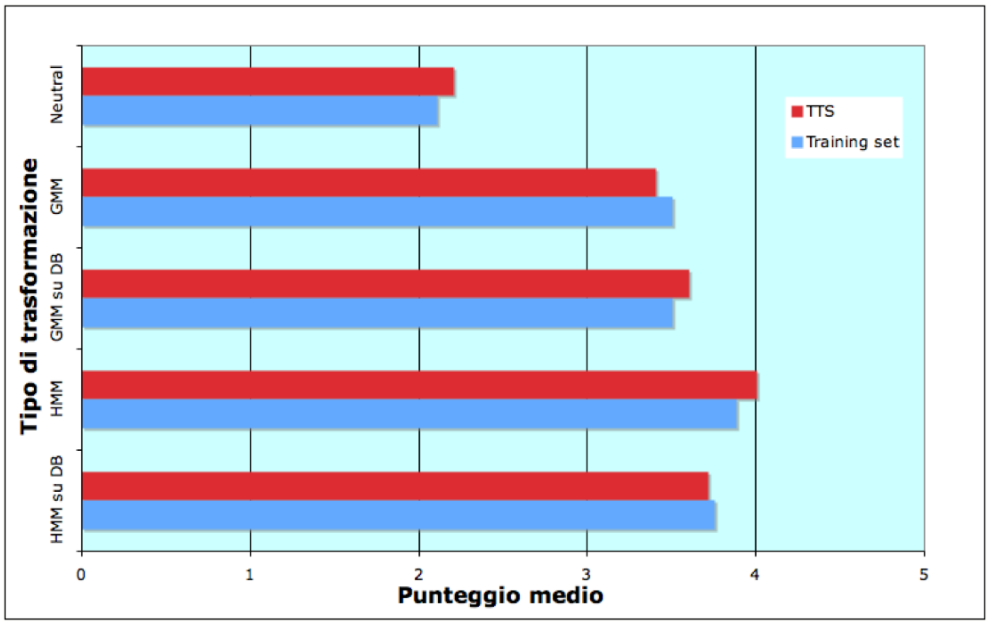

Figura 14: Risultati del test percettivo.

I soggetti interrogati, in generale, concordano nel riconoscere un miglioramento nella resa emotiva nel segnale *trasformato*, rispetto al corrispondente segnale *neutral*, ma non è stata trovata una preferenza tra i vari metodi di conversione, se non una leggera preferenza per il sistema a modello HMM applicato come *post processing*.

7. CONCLUSIONI

E' stato presentato un confronto tra vari metodi di conversione dello spettro di un segnale audio sintetizzato in modo da ottenere un segnale caratterizzato dall'emozione della *collera*.

Ricordando che, ogni volta che si fanno delle elaborazioni di un segnale, si introducono disturbi, possiamo concludere che:

- con ogni trasformazione si aumenta l'intensità emotiva del segnale
- si migliora in parte la naturalezza del parlato
- si introducono però disturbi dovuti ad errori della conversione
- vi sono degli errori che dipendono anche dal motore di sintesi

Come detto in precedenza, una delle cause principali dei problemi, che affliggono i sistemi presentati, sta nel fatto che i modelli utilizzati finora sono sviluppati usando un corpus di allenamento troppo ridotto, quindi, anche le funzioni di conversione non possono essere così precise e introdurre degli errori derivanti dalla errata identificazione dei vettore da convertire.

Andando nel dettaglio dei vari metodi di applicazione della conversione, se la trasformazione è applicata al segnale vocale:

- ha dei risultati migliori perché si può basare su un motore di sintesi migliore.
- non richiede più memoria di un normale motore di sintesi (un solo database di difoni)
- costituito da un "semplice" filtro di elaborazione del segnale
- moltiplica il tempo di produzione del segnale sintetizzato (di norma raddoppia)

Invece, se la trasformazione viene applicata al database di difoni:

- velocizza la generazione del segnale vocale (impiega lo stesso tempo della normale sintesi)
- richiede una quantità maggiore di memoria (è necessario un database per ogni emozione)
- la qualità dipende dalla bontà del database di difoni su cui si agisce e dalla capacità del motore di sintesi di gestire la concatenazione dei difoni modificati.

7.1 Sviluppi futuri

Questo progetto vuole essere solo un confronto tra varie tecniche possibili, sicuramente i risultati ottenuti sono ancora molto migliorabili.

Le principali strade da percorrere che, a nostro parere, porteranno considerevoli sviluppi sono:

- migliorare la qualità del motore di sintesi SMS, sia dal punto di vista del database utilizzato e sia per quanto riguarda l'algoritmo che è ancora in fase di sviluppo
- modellare e agire direttamente sui parametri sinusoidali e residuali del motore SMS
- consolidare i modelli statistici che stanno alla base delle funzioni di conversione
- aumentare le ore di parlato "emotivo" che costituiscono il corpus di allenamento per ottenere dei modelli statistici più solidi e significativi

Una volta perfezionato il sistema, sarà possibile estendere l'analisi e il metodo di conversione alle altre emozioni. Inoltre sarà necessario eseguire un maggior numero di test percettivi per rendere le valutazioni più accurate.

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 (2-3), April, 61-70.

Baudoin, G. & Stylianou, Y. (1996), On the trasformation of the speech spectrum for voice conversion, *International Conference on Spoken Language Processing*, 1405-1408.

Chen, Y., Chu, M., Chang, E., Liu, J. & Liu, R. (2003), Voice conversion with smoothed GMM and MAP adaptation, in *Proceedings of Eurospeech*, Geneva, Switzerland, pp. 2413–2416.

Cosi, P. & Hosom, J.P. (2000), High performance general purpose phonetic recognition for Italian, in *Proceedings of International Conference on Spoken Language Processing*, Beijing, Cina, October, vol. 2, 527-530.

Cosi, P., Tesser, F., Gretter, R., & Avesani, C. (2001), Festival speaks Italian!, *Proceedings of Eurospeech*, Aalborg, Denmark, Sept, pp. 509–512.

Drioli, C., Tisato, G., Cosi, P. & Tesser, F. (2003), Emotions and voice quality: experiments with sinusoidal modeling, *Proceedings of VOQUAL workshop*, Geneva, Switzerland, 27-29 August, 127-132.

Dutoit, T. & Leich, H. (1993), MBR-PSOLA : Text-To-Speech synthesis based on an MBE re-synthesis of the segments database, *Speech Communication*, vol. 13, no. 3-4, pp. 167–184, November.

Kain, A. & Macon, M.W. (1998) Spectral Voice Conversion for Text-to-Speech Synthesis, *Proceedings of International Conference on Acoustics, Speech, and Signal Processing*, vol. 1, 285-288.

Nicolao, M., Drioli, C. & Cosi, P. (2005), Modellazione della prosodia e del timbro nel parlato emotivo, *Atti del 2° congresso AISV 2005*, dicembre 2005, Salerno.

Nicolao, M., Drioli, C. & Cosi, P. (2006), Voice GMM modelling for FESTIVAL / MBROLA emotive TTS synthesis, *Proceedings of INTERSPEECH 2006*, September 2006, Pittsbourgh (PA).

Reynolds, D. A. & Rose, R. C. (1995) Robust text-indipendent speaker identification using Gaussian mixture speaker models, *IEEE Trans. Speech Audio Processing*, vol. 3, January, 72-83.

Serra, X. & Smith, J. (1990), Spectral Modeling Synthesis: A Sound Analysis/Synthesis Based on a Deterministic plus Stochastic Decomposition, *Computer Music Journal*, vol. 14, no. 4, pp. 12-24.

Sommavilla, G., Drioli, C. & Cosi, P. (2006), SMS-FESTIVAL: Un nuovo ambiente di lavoro per la sintesi vocale da testo scritto, *3° congresso AISV 2006*, dicembre, Povo (TN).

Stylianou, Y., Cappè, O. & Moulines, E. (1998), Continuos probabilistic transform for voice conversion, *IEEE Transactions on Speech and Audio Processing*, March, vol. 6 (2), 131-142.