MODELLIZZAZIONE DELLA PROSODIA E DEL TIMBRO PER LA SINTESI DEL PARLATO EMOTIVO

Mauro Nicolao, Carlo Drioli, Piero Cosi
Istituto di Scienze e Tecnologie della Cognizione - Sede di Padova "Fonetica e Dialettologia".
Consiglio Nazionale delle Ricerche
nicolao@pd.istc.cnr.it, cosi@pd.istc.cnr.it, drioli@pd.istc.cnr.it

1. SOMMARIO

Viene descritta una procedura per la creazione di una funzione di trasformazione di un segnale vocale neutro in uno caratterizzato emotivamente. Questa funzione è stata sviluppata sulla base di un modello statistico, a mistura di funzioni gaussiane, dello spettro del segnale vocale.

Sono utilizzati, come segnali di riferimento per l'addestramento del modello, due *database* di segnali vocali creati *ad hoc*: uno registrato da un parlatore, simulando l'emozione della collera, e uno neutro, con la stessa intonazione e durata dei fonemi, ottenuto con un sintetizzatore vocale per concatenazione di difoni, che utilizza la "voce" dello stesso parlatore.

Il modello a mistura di gaussiane, addestrato sui coefficienti *mel-cepstrali* estratti dal segnale neutro, è utilizzato per dividere questo spazio acustico in classi fonetiche equivalenti e per calcolare, per ogni classe identificata, i parametri delle funzioni di conversione.

Il metodo di trasformazione del segnale nel dominio delle frequenze ha fornito delle ottime prestazioni, come è stato dimostrato da un test percettivo in cui un segnale neutro convertito è stato riconosciuto come "*arrabbiato*".

2. INTRODUZIONE

2.1 Sintesi vocale

Con il termine *sintesi vocale* viene identificato l'insieme di sistemi tecnologici che permettono ai computer di parlare. Questo tipo di tecnologia si dimostra ogni giorno più utile in situazioni dove l'utente non può avere accesso alle informazioni in modo visivo. Ad esempio, quando la comunicazione avviene attraverso un apparecchio telefonico, quando la vista è impegnata in altri compiti (alla guida di un'automobile), quando si interagisce con strumenti senza interfaccia video oppure quando si possiedono degli handicap visivi. Lo scopo di queste interfacce uomo-macchina è di cercare di simulare la voce umana per ottenere un parlato sempre più naturale ed espressivo.

Si cerca di creare degli *agenti virtuali* che possano essere utilizzati in sistemi di apprendimento, tipo *e-learning*, in contesti commerciali come *front-end vocali* per fornire informazioni, o, più in generale, per interagire con una macchina senza l'uso di strumenti visivi.

2.2 Sintesi vocale emotiva

La lingua parlata è caratterizzata da numerosi importanti attributi, come il messaggio che si vuole esprimere, l'identità del parlatore, l'accento o le emozioni che si trasmettono.

Nei primi tentativi di sintesi della voce umana, ci si è concentrati principalmente sul primo aspetto: ottenere un segnale vocale intelligibile. Gli attuali sintetizzatori, ora,

riescono ad ottenere ottimamente questo obiettivo, e, quindi, si può focalizzare l'attenzione su strategie volte ad aumentare la sensazione di naturalezza del parlato.

Un importante elemento per caratterizzare e personalizzare una voce è l'inserimento delle emozioni, queste infatti possono influenzare direttamente il messaggio che si vuole trasmettere. Il tono allegro dato ad un messaggio a contenuto triste, ad esempio, può trasmettere ironia, oppure un tono arrabbiato dà maggiore incisività ad una frase normale. E' molto importante perciò poter caratterizzare emotivamente il parlato sintetizzato.

E' molto complesso identificare il concetto di *emozione*, esso infatti coinvolge nozioni di psicologia, soggettività e di senso comune. Comunque, in generale, si può dire che, con questo termine, si identifica la combinazione di tutte le caratteristiche del segnale vocale che forniscono a chi ascolta la percezione che chi l'ha prodotto si trova in uno stato emotivo non neutro.

3. ACQUISIZIONE E ANALISI DEI DATI SPERIMENTALI

Verranno illustrate, di seguito, le caratteristiche dell'insieme di dati che è stato utilizzato come modello per il sistema di conversione. Questi dati sono stati ricavati da un segnale audio registrato da un parlatore e, prima di poter essere utilizzati, sono stati adeguatamente preparati attraverso trasformazioni specifiche.

3.1 Acquisizione del corpus

In questo lavoro era necessario avere un segnale vocale di riferimento, di durata abbastanza lunga, che simulasse il parlato emotivo e in cui l'emozione risaltasse in modo evidente. Per questo è stato necessario registrare un *corpus* creato *ad hoc*.

Si è scelto di prendere come emozione di riferimento la *collera*, poiché è riconosciuta come la più facilmente caratterizzabile e riconoscibile.

Il segnale vocale è costituito dalla lettura enfatizzata del racconto *"Il colombre"* di Dino Buzzati ed è stato acquisito, in camera anecoica, in un'unica sessione, tramite un sistema di registrazione digitale e ed è stato memorizzato su un supporto magnetico digitale ad una frequenza di campionamento di 44 kHz.

Un'interessante peculiarità del segnale di riferimento è che è stato pronunciato dallo stesso parlatore che ha registrato i difoni del *database* di MBROLA[1] usato successivamente nella risintesi.

3.2 Copy synthesis del segnale registrato

Parte del lavoro di preparazione per la creazione del modello statistico è stata la produzione, tramite il motore di sintesi MBROLA, di una copia del segnale originale. Questo processo viene normalmente identificato con il termine *copy synthesis* e può essere schematizzato come in Figura 1.

Lo scopo di questa risintesi è creare un segnale, in tutto uguale all'originale (voce dello stesso parlatore, stesso enunciato, stessa durata dei difoni, stessa intonazione); la sola differenza sarà costituita dalle caratteristiche che caratterizzano l'emozione che si cerca di modellare.

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

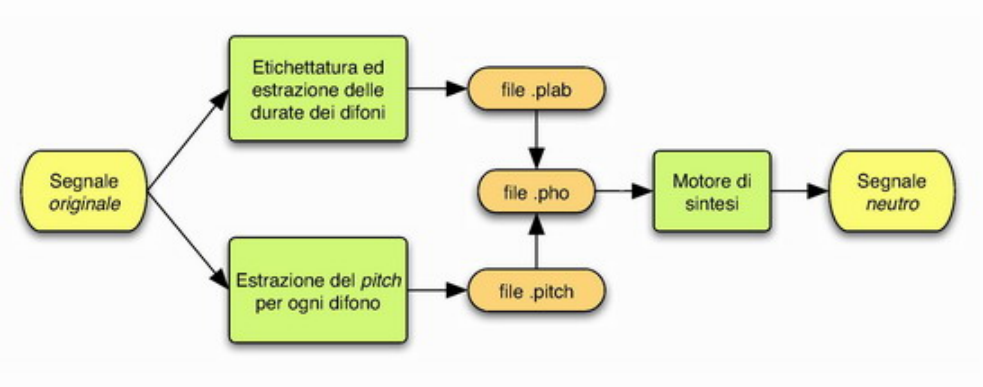


Figura 1: Schema del processo di *copy synthesis*.

3.2.1. Etichettatura dei fonemi

Una parte del processo di *copy synthesis* consiste nell'effettuare sul segnale originale un *riconoscimento vocale*. Questo serve per etichettare i fonemi pronunciati che dovranno essere risintetizzati. Per ottenere questa etichettatura è stato utilizzato il riconoscitore vocale per la lingua italiana sviluppato dalla Sezione di Padova "Fonetica e Dialettologia" dell'ISTC-CNR, descritto in (Cosi & Hosom, 2000). Questo metodo, che utilizza il pacchetto software *CSLU Speech Toolkit* (Sutton *et al*., 1996), si basa un modello ibrido a catene di Markov nascoste (*HMM, Hidden Markov Model*) e a rete neurale (*ANN, Artificial Neural Network*). Per una trattazione dettagliata del metodo si rimanda alla bibliografia.

Dall'analisi del segnale audio registrato sono stati ricavati dei *file* di testo (*file .plab*) che contengono informazioni sul tipo di fonema pronunciato, identificato secondo la notazione SAMPA, e sull'istante t di fine del fonema.

3.2.2. Estrazione del pitch

Ulteriori informazioni sono necessarie per produrre la *copy synthesis* del segnale vocale originale: le informazioni sul *pitch*[2] del segnale vocale.

L'estrazione di questo parametro è stata effettuata con una funzione del software PRAAT[3]. Sono stati così ottenuti i valori della frequenza fondamentale calcolati per ogni frammento ad intervalli costanti (nel nostro caso 20 ms).

Si sono unite le informazioni fonetiche e di durata con i valori del *pitch* in un unico *file*, denominato "*.pho*", che sarà usato come ingresso per il motore di sintesi.

3.2.3. Creazione della forma d'onda

Dopo aver estratto i parametri che identificano il contenuto del messaggio, le durate e l'intonazione del segnale originale, il passaggio successivo consiste nel sintetizzare la forma d'onda vocale. Per farlo si utilizza il *motore di sintesi* MBROLA. Questo costruisce un segnale audio seguendo le informazioni contenute in un *file* di testo e utilizzando un *database* di difoni precedentemente registrato.

I *file* audio così ottenuti avranno le seguenti caratteristiche:

- allineamento temporale con i *file* del segnale originale
- medesima durata di ogni singolo fonema
- stesso valore del *pitch* per ogni *frame* di analisi

[2] Con il termine *pitch* si identifica l'altezza di un suono vocalico.
[3] *Software* libero di elaborazione e analisi dei segnali audio (Boersma, 2001)

- stesso timbro di voce.

Il segnale vocale così elaborato (diviso in frammenti, campionato a 16 kHz) sarà identificato d'ora in avanti come segnale *target*.

3.3 Estrazione dei parametri spettrali

La trasformazione agisce nel dominio delle frequenze; è stato quindi necessario analizzare i segnali in quest'ambito. In particolare, la forma d'onda del segnale e le sue caratteristiche spettrali sono state convertite in alcuni tipi di coefficienti rappresentativi del loro andamento locale. Questi hanno il compito di far risaltare gli aspetti del segnale necessari all'analisi e all'elaborazione, escludendo le informazioni inutili o sovrabbondanti.

La caratterizzazione necessaria per la *voice quality* (*VQ*) è molto differente rispetto a quella utilizzata, ad esempio, per il riconoscimento vocale. Nella *VQ*, bisogna acquisire informazioni dettagliate sull'andamento del segnale. Oltre alle caratteristiche spettrali macroscopiche (f_0 e le *formanti*) è necessario ricavare anche informazioni sulle caratteristiche spettrali anche in alta frequenza.

3.3.1. Analisi spettrale

Lo strumento più importante per l'analisi spettrale è dato dalla trasformata di Fourier. In pratica, nelle simulazioni numeriche, risulta di molta utilità la sua versione discreta e *short-term* identificata come *stDFT*:

$$X(k, nT) = \sum_{m=nT-N+1}^{nT} x(m)h(nT - m)e^{-j\frac{2k\pi m}{N}} \tag{1}$$

dove è presente una funzione di finestratura, h(.), che, nel nostro caso, è una finestra di tipo *Blackman*, la cui espressione è:

$$h(n) = \begin{cases} 0.42 - 0.5\cos\left(\frac{2\pi n}{N-1}\right) + 0.08\cos\left(\frac{4\pi n}{N-1}\right), & 0 \le n \le N-1 \\ 0 & altrimenti \end{cases} \tag{2}$$

3.3.2. Rappresentazione percettiva

Per analizzare le peculiarità del segnale vocale si utilizza l'analisi *cepstrale* (Deller *et al*, 1993) trasformata attraverso alcune relazioni psicofisiche: il segnale, in questo modo, viene elaborato in maniera da "seguire" una caratteristica percettiva dell'orecchio umano.

Le *features* così calcolate sono perciò robuste a molte variazioni del parlato: se un cambio di forma d'onda non è percepito da un ascoltatore umano, i corrispondenti valori calcolati non devono cambiare.

In Figura 2 è descritto il metodo di calcolo di una tra le rappresentazioni percettive più usate: la codifica a coefficienti cepstrali in scala Mel (*Mel Frequency Cepstrum Coefficients*, *MFCC*). In questo tipo di coefficienti, una scala percettiva (scala Mel) viene applicata all'analisi cepstrale. Tale scala cerca di correlare la frequenza con la sensazione di altezza del suono.

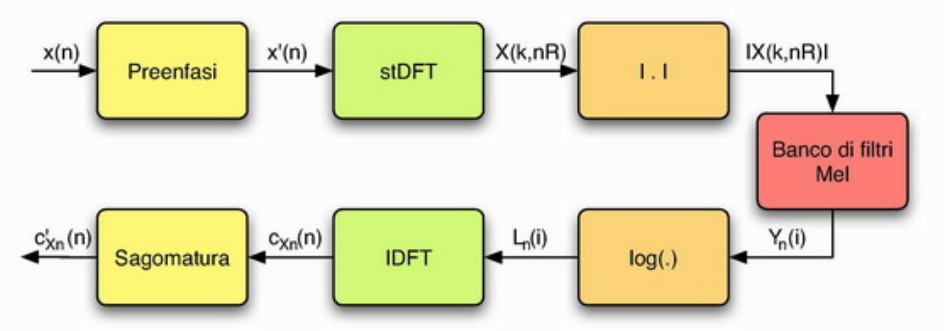


Figura 2: Schema a blocchi per il calcolo dei coefficienti MFCC.

La procedura usata per calcolare tali coefficienti può essere schematizzata nel seguente modo:

Filtro di preenfasi sul segnale: serve per enfatizzare le alte frequenze. Nel nostro caso, comunque, si è scelto di non modificare il segnale con preelaborazioni perché si perderebbe la corrispondenza diretta con lo spettro del segnale.

Calcolo del modulo dello spettro $|X(k;nR)|$, utilizzando la *stDFT* con una finestra *Hamming* di N campioni.

Calcolo dei coefficienti spettrali Mel: viene usato un banco di M filtri triangolari equispaziati secondo la scala Mel.

Trasformazione logaritmica:

$$L_n(i) = \log Y_n(i) \qquad 1 \le i \le M \tag{3}$$

Calcolo dei coefficienti cepstrali: dato che L(.) è pari si può utilizzare la trasformata coseno, al posto della IDFT, che si chiama *DCT* (*Discrete Cosine Transform*).

$$c_{x_n}(j) = \sqrt{\frac{2}{M}} \sum_{i=1}^{M} L_n(i) \cos\left(\frac{\pi i}{M} (j - 0.5) \right) \tag{4}$$

dove M è il numero di filtri del banco di analisi.

Sagomatura: solitamente i coefficienti $c_{x_n}(j)$ di ordine elevato vengono pesati. In questo caso, come per la preenfasi, si è scelto di non modificare in alcun modo i coefficienti.

Nel processo di estrazione degli *MFCC* sono stati individuati dei parametri critici che sono:

il *numero dei filtri* che compongono il banco,

il *numero di coefficienti* che verranno utilizzati.

Questi sono parametri importanti poiché determinano quanto l'inviluppo è aderente al profilo della trasformata di Fourier. Se si decide di modellare l'inviluppo su tutte le variazioni dello spettro (maggior numero di coefficienti), si perde però la generalità della trasformazione calcolata su di essi.

3.4 I correlati acustici spettrali

Effettuare un'analisi oggettiva di come un'emozione viene resa in un segnale vocale è estremamente complesso, principalmente perché l'emozione non è quantificabile. L'unico

metodo utilizzabile è quello di estrarre degli indicatori considerati significativi della forma dello spettro dei segnali prodotti e confrontarli poi con quelli estratti.

Gli indicatori che sono stati scelti per l'analisi sono quelli che si incontrano più comunemente, in letteratura scientifica, nell'analisi del parlato emotivo (Banse & Scherer, 1996; Alter *et al.*, 2003; Drioli *et al*., 2003) e sono:

Shimmer: con questo indicatore si misura la rapida variazione, tra un periodo e l'altro, dell'ampiezza del segnale.

Jitter: misura la variazione della durata del periodo fondamentale del segnale in tratto di segnale periodico.

Harmonic to Noise Ratio (HNR): è definito come il rapporto tra l'energia della parte armonica del segnale e il resto del segnale (parte rumorosa).

Glottal to Noise Excitation ratio Index (GNE): è il rapporto tra l'energia del segnale glottale e la parte rumorosa.

Hammarberg Index (HammI): indica la differenza tra la massima energia nella banda di frequenze tra 0 e 2000 Hz e quella della banda tra 2000 e 5000 Hz.

Do1000: indica la caduta di energia spettrale sopra i 1000 Hz e viene calcolata come il gradiente dell'approssimazione quadratica minima dell'inviluppo spettrale sopra i 1000 Hz.

Pe1000: è il rapporto tra l'energia totale in alta frequenza (oltre i 1000 Hz) e quella in bassa frequenza (da 0 a 1000 Hz).

Spectral Flatness Measure (SFM): è la misura della piattezza dello spettro che è misurata come il rapporto tra la media geometrica e la media aritmetica della distribuzione dell'energia spettrale.

Questi indicatori hanno senso solo se calcolati su parti armoniche (*voiced*) del segnale poiché si basano sul confronto tra parti simili del segnale.

Questi comunque non hanno valenza assoluta, ma sono ugualmente importanti perché, se estratti dallo stesso fonema proveniente da segnali vocali diversi, danno una misura della maggiore o minore somiglianza tra essi.

Qui di seguito un esempio di valori calcolati per lo stesso *frame* di 2 diversi tipi di segnale:

Indici	Segnale copy synthesis	Segnale originale
Jitter	0,7	1,145
Shimmer	4,971	4,05
HNR (dB)	18,117	11,766
Do1000	-4,843	-5,1664
GNE	2,2841	1,0364
SFM	0,29808	1,1672
Pe1000	0,33882	0,22404
HammI	9,2966	1,7232

In questa tabella si possono vedere come variano i parametri nei segnali di *copy synthesis* e *originale*. Questi parametri sono estratti dallo stesso *frame* dello stesso fonema[4] e si può notare facilmente come essi siano molto differenti, nonostante molta della parte prosodica sia esattamente la stessa.

Si nota, ad esempio, un maggiore valore dell'*HNR* del segnale originale rispetto a quello degli altri due; questo è dovuto alle componenti rumorose di *harsh* e di *breathy* (Laver,

[4] Il fonema considerato è la vocale "*a*" nella parola "*mare*".

1980) presenti in questo segnale. Questi sono tra i parametri di *VQ* che più caratterizzano la qualità della voce rabbiosa e ovviamente non sono presenti nel segnale sintetizzato "neutro".

4. CREAZIONE DELLA FUNZIONE DI CONVERSIONE

Nell'analisi del paragrafo precedente si è visto come la sola procedura di *copy synthesis* non sia sufficiente. Nel segnale sintetizzato, seppur allineato col segnale *target* nella durata dei fonemi con lo stesso livello di *pitch* e seppur con fonemi pronunciati dallo stesso parlatore, non si riconosce una grande affinità emotiva con l'originale.

Il segnale di *copy synthesis* costituisce l'indispensabile punto di partenza privilegiato per estrarre le differenze prettamente spettrali rispetto al segnale obiettivo. Si vuole infatti ottenere un metodo che possa essere utilizzato come *post-elaborazione* del segnale, indipendente dal *pitch* e dalla durata dei fonemi.

4.1 Voice Conversion

Questo tipo di approccio si inserisce nel contesto più generale che, in letteratura scientifica, viene denominato *voice conversion*. Quest'ambito della sintesi vocale si occupa di sviluppare dei metodi per convertire una voce in un'altra. Queste possono differire per l'identità del parlatore che le ha generate o per il contesto, emotivo o ambientale, in cui sono state prodotte.

Su questo argomento si trovano numerosi articoli che illustrano metodi più o meno differenti, ma che si basano, per lo più, sulla conversione dell'inviluppo spettrale del segnale (Abe *et al.*, 1988; Baudoin & Stylianou, 1996; Stylianou *et al.*, 1998; Kain & Macon, 1998).

Stylianou *et al.*, in particolare, dividono lo spazio acustico del segnale sorgente usando un modello basato su una mistura di gaussiane (*GMM, Gaussian Mixture Model*). Questi propongono poi una funzione di conversione statistica, basata appunto sul *GMM* creato, per trasformare gli inviluppi spettrali delle parti armoniche del segnale.

Il presente modello si sviluppa seguendo le linee guida introdotte da questo metodo.

4.2 Filtraggio in frequenza

Lo scopo è creare un filtro che permetta, con una semplice operazione di filtraggio nel domino delle frequenze, di trasformare lo spettro e quindi il segnale nel tempo.

4.2.1. Metodo diretto

Per testare le modalità con cui il metodo statistico deve essere applicato al segnale, si è scelto di provarle attraverso una *conversione diretta* dello spettro del segnale sintetizzato da MBROLA, *synth*.

Con il termine conversione diretta si indica l'applicazione ad ogni *frame* di segnale di una trasformazione spettrale "*ad hoc*" calcolata sulle differenze tra questo e il corrispondente *frame* del segnale *target* che, si ricorda, è allineato temporalmente.

Questo metodo, ovviamente, prevede la conoscenza a priori del segnale *target*, quindi non può essere generalizzato. Il segnale così trasformato, però, costituisce un ottimo punto di riferimento per il segnale che verrà generato con il modello. Sarà infatti il massimo risultato ottenibile con questo tipo di trasformazione in frequenza e quindi un'implicita misura di qualità.

Il metodo, illustrato in Figura 3, sarà ora descritto in dettaglio.

Dati i due segnali $\mathbf{s}_{target}$ e $\mathbf{s}_{synth}$ si è operata la preelaborazione illustrata nel paragrafo precedente:

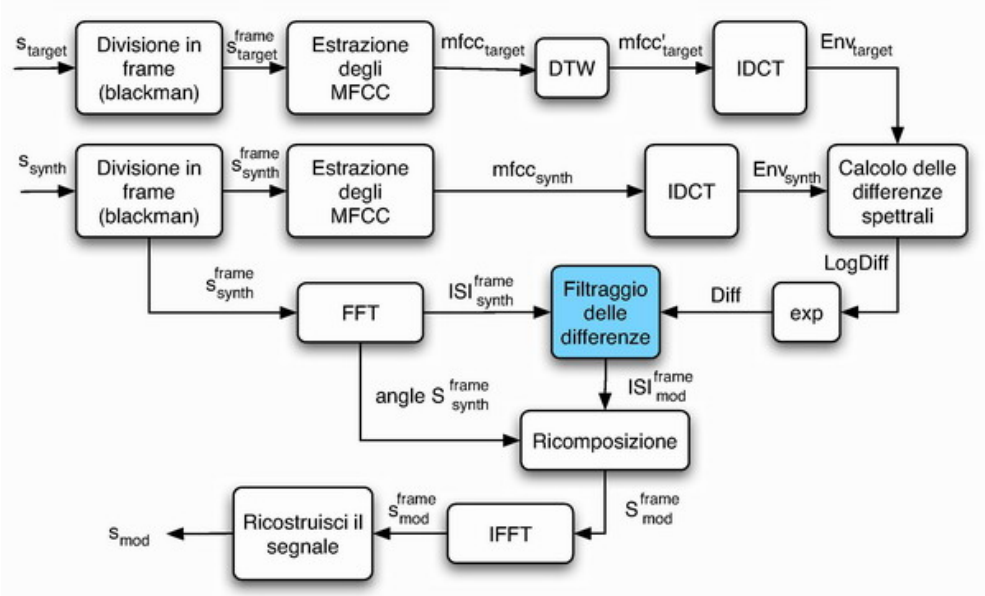


Figura 3: Schema del metodo di trasformazione spettrale utilizzato. E' colorato il filtro che opera la trasformazione.

- divisione in *frame*, tramite una finestra di *Blackman*, con un numero di N_{window} campioni, che corrispondono a N_{window} / Fs secondi, dove Fs è la frequenza di campionamento, che, nel nostro caso, sarà sempre Fs = 16 kHz. Si ottengono i segnali $\mathbf{s}_{target}^{frame}$ e $\mathbf{s}_{synth}^{frame}$;
- trasformazione del segnale nel dominio delle frequenze tramite una *stFFT* su N_{fft} campioni ($N_{fft} \geq N_{window}$). Si ottengono i vettori contenenti i campioni del segnale in frequenza come *modulo*, $|\mathbf{S}_{target}^{frame}|$e $|\mathbf{S}_{synth}^{frame}|$, e come *fase*, $\arg \mathbf{S}_{target}^{frame}$ e $\arg \mathbf{S}_{synth}^{frame}$;
- estrazione dei coefficienti cepstrali in scala Mel (*MFCC*). Si ottengono i vettori $\mathbf{mfcc}_{target}$ e $\mathbf{mfcc}_{synth}$ che costituiscono il punto di partenza per il calcolo degli inviluppi spettrali e quindi saranno i dati su cui verrà poi creato il modello.

Come già detto dai coefficienti *mel-cepstrali* è possibile ricavare la forma dell'inviluppo spettrale del segnale. A tal fine è stata utilizzata la trasformazione inversa rispetto a quella usata per calcolarli, cioè la *IDCT:*

$$\mathbf{Env}(n) = w(k) \cdot \sum_{j=1}^{N} \mathbf{x}(j) \cos\left(\frac{\pi(2k-1)(j-1)}{2N_{fft}} \right) \qquad k = 1, \ldots, N_{fft} \tag{5}$$

con

$$w(n) = \begin{cases} \dfrac{1}{2\sqrt{N_{fft}}} & k = 1 \\ \dfrac{2}{2\sqrt{N_{fft}}} & 2 \leq k \leq N \end{cases} \tag{6}$$

Il risultato di questa ricostruzione è la versione campionata della trasformata logaritmica del segnale vocale senza la parte derivante dall'eccitazione delle corde vocali. In altre parole, si ottiene un inviluppo spettrale logaritmico, tanto più smussato quanto più si elimina la parte glottale.

Si ottengono così delle curve che seguono abbastanza fedelmente il profilo delle relative *stFFT*, vedi la Figura 4. I vettori che contengono i valori di queste curve si chiamano $\mathbf{Env}_{synth}$ e $\mathbf{Env}_{target}$ e hanno la stessa dimensione $\mathbf{N}_{fft}$ dei vettori che contengono i valori le trasformate di Fourier.

L'idea che sta alla base del metodo di conversione è che, se si è in grado di sapere quanto varia l'inviluppo della *stFFT* di un segnale quando si introduce un'emozione, le stesse differenze si possono applicare alla *stFFT* stessa e, quindi, ottenere un segnale modificato emotivamente.

I parametri critici che indicano quanto bene un inviluppo segua la *stFFT*, sono il numero di filtri del banco e il numero di coefficienti *MFCC* utilizzati per calcolare l'inviluppo con la *IDCT*.

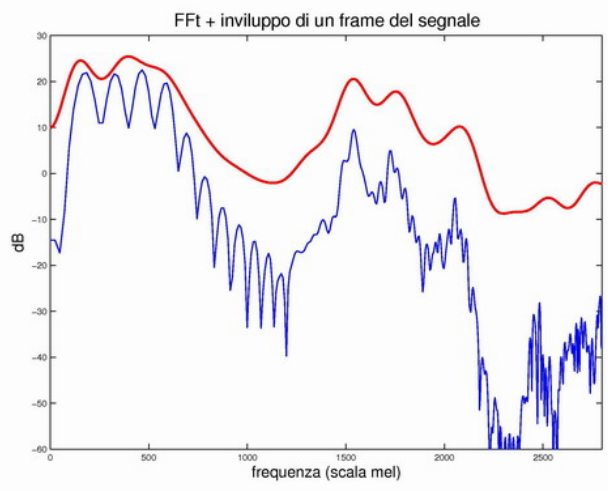


Figura 4: Inviluppo spettrale di una porzione di segnale.

Si calcola quindi la differenza tra gli inviluppi spettrali su ogni porzione di segnale finestrato:

$$\mathbf{LogDiff}(n) = \mathbf{Env}_{target}(n) - \mathbf{Env}_{synth}(n) \tag{7}$$

dove n è il numero del *frame* su cui si sta operando.

Queste quantità vengono poi aggiunte allo spettro del segnale sintetizzato *synth*. Il vettore **LogDiff**(n) è un vettore di differenze logaritmiche, quindi è necessario tornare in un dominio lineare prima di operare il vero filtraggio. Si ha quindi,

$$\mathbf{Diff}(n) = e^{\mathbf{LogDiff}(n)}$$

E' importante sottolineare che il filtraggio coinvolge solo il modulo $|\mathbf{S}_{synth}^{frame}|$, mentre per quanto riguarda la fase, si ipotizza non subisca variazioni dato che il segnale *synth* e il *target* sono allineati nel tempo.

$$\left| \mathbf{S}_{\text{mod}}^{\text{frame}}(n,k) \right| = \left| \mathbf{S}_{\text{synth}}^{\text{frame}}(n,k) \right| \cdot \mathbf{Diff}(n,k) \qquad k = 1,\ldots,P \tag{8}$$

Si ricompone quindi il segnale unendo il modulo modificato, $|\mathbf{S}_{\text{mod}}^{\text{frame}}|$, con la fase lasciata invariata, $\arg \mathbf{S}_{\text{mod}}^{\text{frame}} = \arg \mathbf{S}_{\text{synth}}^{\text{frame}}$.

Si ottiene $\mathbf{S}_{\text{mod}}^{\text{frame}}$ che dovrà essere antitrasformato, tramite *stIFFT*, tenendo conto del valore di N_{fft}.

A questo punto si ottiene il segnale nel tempo modificato $\mathbf{s}_{\text{synth}}^{\text{frame}}$, ma ancora finestrato. Si ricostruisce con una procedura detta *overlap and add*, che tiene conto della funzione di finestratura utilizzata e dell'energia di questa.

Il segnale risultante che aveva già la stessa prosodia del segnale *target* per merito del processo di *copy synthesis*, avrà ora anche delle caratteristiche spettrali molto simili, che serviranno a rendere l'emozione desiderata.

4.3 Metodo statistico

Lo stesso metodo verrà utilizzato anche per applicare la trasformazione risultante dal modello statistico. Questo modello si inserisce nello schema in Figura 3 e lo trasforma come descritto in Figura 5.

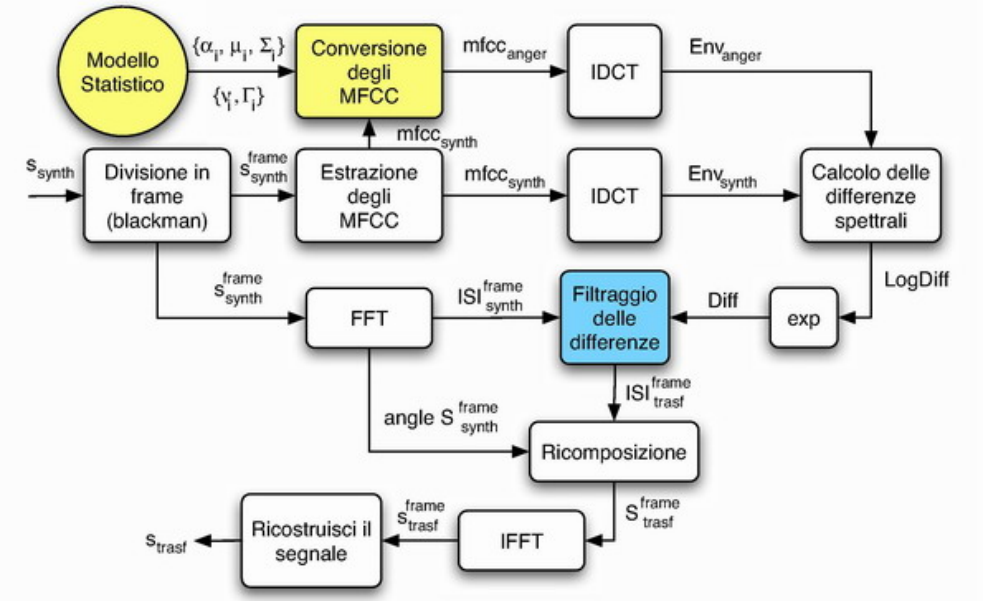


Figura 5: Schema del metodo di trasformazione spettrale basata sul metodo statistico.

Il modello fornisce un vettore di *MFCC* trasformato per ogni vettore di *MFCC* del segnale neutro che viene dato in ingresso. Quindi, come riferimento per il calcolo delle differenze, al posto degli inviluppi del segnale *target* si usano quelli calcolati con la funzione di conversione statistica.

4.3.1. Rappresentazione dei dati

La creazione del modello prevede l'utilizzo di due insiemi di dati paralleli: i coefficienti *MFCC* del segnale neutro (*synth*) e quelli del segnale che si vuole imitare (*target*) caratterizzato dall'emozione della *collera*.

I dati disponibili sono raggruppati in due insiemi di vettori: **x**(n) e **y**(n). Ognuno di questi è un vettore P-dimensionale di *MFCC* che identifica univocamente l'inviluppo spettrale dell'intervallo di segnale relativo.

I due insiemi $\{\mathbf{x}(n), n=1,\ldots,N_T\}$ e $\{\mathbf{y}(n), n=1,\ldots,N_T\}$, dove n rappresenta l'istante temporale, hanno la stessa lunghezza, N_T, e si suppone che siano allineati nel tempo. Questo è assicurato dal metodo di costruzione del segnale neutro (*copy synthesis*), ciononostante, per evitare possibili dilatazioni temporali introdotte dal motore di sintesi, è stata applicata un'ulteriore procedura di allineamento temporale (*DTW, Dynamic Time Warping*)[5].

4.4 Il modello a mistura di gaussiane

Lo scopo del modello è creare una funzione $\Im(.)$ tale che la trasformazione $\mathsf{F}(\mathbf{x}(n))$ permetta di ottenere un vettore che si avvicini in modo ottimo al *target*, **y**(n), per ogni coppia di vettori dell'insieme di dati per addestramento (n=1,..., N_T). Questo verrà fatto tramite l'uso di un modello statistico.

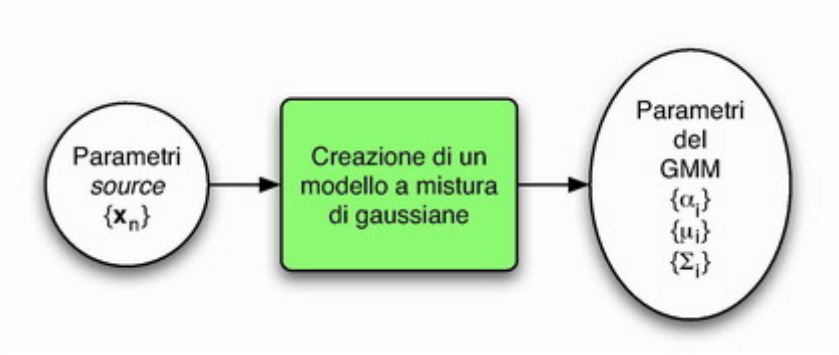


Figura 6: Schema della creazione di un modello a mistura di gaussiane.

Il modello a mistura di gaussiane o *GMM* (*Gaussian Mixture Model*) è un modello parametrico largamente utilizzato in molti sistemi di riconoscimento vocale la cui efficienza è ormai dimostrata e consolidata.

Il *GMM* si basa sull'assunto che la distribuzione di probabilità dei parametri osservati, **x**(n), abbia la seguente forma:

$$p(\mathbf{x}) = \sum_{i=1}^{M} \alpha_i N\left(\mathbf{x}, \mu_i, \Sigma_i\right) \tag{9}$$

dove $N(\mathbf{x},\boldsymbol{\mu}_i,\boldsymbol{\Sigma}_i)$ indica la distribuzione di probabilità di un vettore aleatorio gaussiano P-dimensionale con media $\boldsymbol{\mu}$ e matrice di covarianza $\boldsymbol{\Sigma}$.

Nella (9) il termine α_i è un coefficiente positivo che rappresenta il peso con cui deve essere considerata la gaussiana presente nella i-esima mistura e deve soddisfare le seguenti condizioni:

[5] È stata usata la funzione DTW del toolbox MATLAB denominato Auditory Toolbox.

$$\sum_{i=1}^{M} \alpha_i = 1 \qquad e \qquad \alpha_i \geq 0$$

Un'ipotesi fondamentale che ha permesso l'utilizzo del *GMM* è che i vettori dell'osservazione {**x**(n)} fossero indipendenti tra loro. Questa semplificazione, infatti, permette di poter considerare irrilevante nel modello la dipendenza dal tempo n.

In questo modo non si è reso necessario l'utilizzo di un modello più complesso come quello a catene di *Markov* nascoste (*HMM*, *Hidden Markov Model*), in cui il modello *GMM* che si utilizza dipende dallo stato in cui ci si trova.

Nel nostro caso, l'indipendenza dal tempo è giustificata dal fatto che cerchiamo una funzione di conversione su segmenti molto piccoli (alcuni millisecondi) e si possono quindi trascurare le informazioni sulle informazioni linguistiche e lessicali sui difoni e sulle durate.

Un ulteriore motivo per cui è stato scelto uno strumento come il *GMM* è la sua capacità di operare un blanda classificazione degli innumerevoli tipi di inviluppi del segnale in classi corrispondenti alle componenti della mistura. Con il termine componente si identifica la distribuzione gaussiana unimodale, $N(\mathbf{x}, \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)$. Quando questo modello viene utilizzato nell'ambito dei segnali vocali ed ogni classe così identificata rappresenta un evento fonetico diverso, come, ad esempio, una peculiare accezione di un fonema.

Ogni classe acustica poi è univocamente determinata da due elementi: da un valore centrale (il vettore delle medie $\boldsymbol{\mu}_i$) e da una dispersione caratteristica intorno a questo valore (la matrice della covarianza $\boldsymbol{\Sigma}_i$). I pesi della mistura, α_i, rappresentano la frequenza statistica con cui si presenta un vettore appartenente ad una classe all'interno del fenomeno osservato.

La probabilità condizionata che un dato vettore **x** appartenga ad una determinata classe acustica C_i del *GMM* è ricavata facilmente dall'applicazione della regola di *Bayes* e si ha:

$$P(C_i \mid \mathbf{x}) = \frac{\alpha_j N(\mathbf{x}, \boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)}{\sum_{j=1}^{M} \alpha_j N(\mathbf{x}, \boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)} \tag{10}$$

I parametri del *GMM* sono stimati utilizzando un caso particolare dell'algoritmo conosciuto col nome di *Baum-Welch* (chiamato anche *Forward-Backward*) (cfr. Deller *et al*., 1993). Questo usualmente viene utilizzato per stimare i parametri di un *GMM* congiuntamente alla probabilità di transizione tra gli stati di un *HMM*[6], ma si ipotizza che la catena abbia un solo stato contenente una mistura di M gaussiane, ci si riconduce ad un semplice *GMM*.

Un fattore critico dell'algoritmo è l'inizializzazione dei parametri, per questo sono stati utilizzati alcuni accorgimenti. Prima di tutto, è stata utilizzata una procedura di quantizzazione vettoriale (Vector Quantization) che ha creato una sommaria divisione in classi.

[6] Questo è l'algoritmo di stima per la mistura di gaussiane del pacchetto software, *HTK*, *HMM ToolKit* (Young *et al*., 2002), che è stato utilizzato per il calcolo del GMM.

Successivamente è stato utilizzato l'algoritmo di *Viterbi* (cfr. Deller *et al*., 1993) che ha fornito la prima stima di $\boldsymbol{\mu}$ e di $\boldsymbol{\Sigma}$ che ha permesso al metodo *Baum-Welch* di convergere verso dei valori ottimi per i pesi, le medie e le covarianze di ogni gaussiana del *GMM*[7].

Un fattore critico per il nostro modello si è dimostrato il problema delle componenti gaussiane con bassi valori di covarianza. E' stato verificato che il metodo di stima utilizzato non converge quando la norma di almeno una delle matrici tende a zero. Il metodo, per ovviare a questo inconveniente, è imporre una soglia minima oltre alla quale un elemento della diagonale di una matrice di covarianza non può scendere (Reynolds *et al*., 1995). Questo valore di soglia inoltre è importante perché influenza direttamente la capacità di riconoscimento del *GMM*. Infatti più la covarianza, che indica la dispersione dei valori rispetto al valor medio, è piccola, più la probabilità del *GMM* di riconoscere valori diversi da quelli dell'insieme dei vettori di allenamento sarà bassa.

Dopo aver allenato un *GMM* di M misture di gaussiane su un insieme di N_T vettori di coefficienti mel-cepstrali (*MFCC*) del segnale *synth*, otteniamo i valori ottimi dei parametri α_i, $\boldsymbol{\mu}_i$, $\boldsymbol{\Sigma}_i$ relativi ai vettori $\mathbf{x}(n)$, n=1,..., N_T.

4.5 La funzione di conversione

Dopo aver ottenuto un buon modello di riconoscimento del segnale *synth*, ora ci focalizziamo sul problema di trovare la funzione di conversione che trasformi i vettori di $\{\mathbf{x}(n)\}$ nei corrispettivi *target* $\{\mathbf{y}(n)\}$ per ogni n=1,…, N_T.

Si assume che la funzione di conversione abbia la seguente forma (Stylianou *et al*., 1998):

$$\mathfrak{F}(\mathbf{x}(n)) = \sum_{i=1}^{M} P(\mathfrak{C}_i \mid \mathbf{x}(n)) \left[\mathbf{v}_i + \boldsymbol{\Gamma}_i \boldsymbol{\Sigma}_i^{-1} (\mathbf{x}(n) - \boldsymbol{\mu}_i) \right] \qquad (11)$$

I parametri che definiscono questa funzione sono il vettore P-dimensionale $\mathbf{v}_i$ e la matrice $\boldsymbol{\Gamma}_i$ di dimensione P x P, con i=1,...,M (M, il numero di componenti della mistura).

Questa forma per la funzione è stata scelta per analogia con il caso limite di una singola gaussiana (M=1). Si è deciso di estendere questo risultato al caso di una mistura con M>1, operando una somma pesata di termini analoghi, ognuno per ogni gaussiana. I pesi della somma sono le probabilità condizionate che il vettore $\mathbf{x}(n)$ appartenga alle differenti classi $\mathfrak{C}_i$.

Anche se la funzione di conversione (11) non è supportata da un adeguato modello statistico teorico, può essere utile continuare ad interpretare i parametri $\mathbf{v}$ e $\boldsymbol{\Gamma}$, analogamente al caso della singola gaussiana, come vettore delle medie e matrice della covarianza di un modello a mistura di gaussiane dello spazio acustico *target*.

4.6 Ottimizzazione della funzione di conversione

Partendo dalla funzione (11) si possono distinguere tre tipi di conversione.

Conversione a parametri completi: corrisponde al caso generale in cui le matrici che compaiono nella funzione sopra citata, vengono considerate nella loro forma completa e la funzione applicata senza semplificazioni.

Conversione a parametri diagonalizzati: L'uso del modello a mistura di gaussiane viene spesso utilizzato supponendo le matrici delle covarianze, $\boldsymbol{\Sigma}_i$, in forma diagonale. Questa semplificazione è giustificata teoricamente perché, nel caso di coefficienti cepstrali, la

[7] Per un'accurata descrizione dell'algoritmo di Viterbi e di Baum-Welch utilizzato si veda Young *et al*., (2002).

correlazione tra distinti vettori è molto bassa. Questo sistema permette di ridurre sensibilmente i tempi di calcolo del modello.

La stessa semplificazione può essere adottata anche nel calcolo della funzione di conversione. In questo caso, consideriamo diagonale oltre alla matrice $\mathbf{\Sigma}_i$ anche la matrice di conversione $\mathbf{\Gamma}_i$. Diminuisce così la mole di calcoli necessaria a ricavare i parametri di conversione perché il problema viene ridotto a P (numero di coefficienti *MFCC*) sottoproblemi indipendenti e scalari.

Conversione a quantizzazione vettoriale: Se omettiamo il termine di correzione, che dipende dalla differenza tra il vettore *synth* **x**(n) e la media delle componenti del *GMM*, $\boldsymbol{\mu}_i$, nella funzione di conversione (11), questa si riduce a:

$$\mathfrak{F}(\mathbf{x}(n)) = \sum_{i=1}^{M} P(\mathfrak{C}_i \mid \mathbf{x}(n))\mathbf{v}_i \tag{12}$$

La (12) è la somma dei vari contributi dei vettori di conversione $\mathbf{v}_i$ (che ricordiamo possono essere considerati i valori medi dei vettori in cui è stato partizionato lo spazio acustico *target*), pesati dalla funzione di probabilità condizionata. In questo modo si ottiene una forma di interpolazione dell'inviluppo spettrale trasformato.

Nel presente lavoro, abbiamo utilizzato il metodo a parametri diagonalizzati, tralasciando invece il primo poiché implica calcoli molto onerosi e il terzo perché poco performante.

Da qui in avanti si indicherà con $p_n(i)$ la probabilità condizionata $P(C_i \mid \mathbf{x}(n))$.

4.7 Conversione a parametri diagonalizzati

Grazie alla natura lineare della funzione di conversione (11) l'ottimizzazione della stima dei parametri di conversione può essere vista come equivalente alla soluzione del seguente sistema di equazioni lineari:

$$\mathbf{y}(n) = \sum_{i=1}^{M} p_n(i)\left[\ \mathbf{v}_i + \mathbf{\Gamma}_i\mathbf{\Sigma}_i^{-1}(\mathbf{x}(n) - \boldsymbol{\mu}_i)\ \right] \tag{13}$$

Quando le matrici delle covarianze del *GMM*, $\mathbf{\Sigma}_i$, e le matrici di conversione, $\mathbf{\Gamma}_i$, sono entrambi diagonali, è possibile dividere il problema in P sottoproblemi scalari indipendenti tra loro, considerando la stima di ogni singolo coefficiente mel-cepstrale separatamente. Il k-esimo elemento della (13) può essere riscritto come:

$$y_n^{(k)} = \sum_{i=1}^{M} p_n(i)\left[\ \gamma_i^{(k)}\frac{(x^{(k)}(n) - \mu_i)}{\sigma_i^{(k)}} + v_i^{(k)}\ \right] \tag{14}$$

dove l'apice (k) indica la k-esima coordinata del vettore, mentre nel caso di $\sigma_i^{(k)}$ e di $\gamma_i^{(k)}$ indica il k-esimo elemento della diagonale delle matrici $\mathbf{\Sigma}_i$, e di $\mathbf{\Gamma}_i$.

Sviluppando il calcolo come nel caso generale si ottiene una formula matriciale semplificata per il calcolo dei parametri della conversione.

$$\begin{bmatrix} \mathbf{P}^T\mathbf{P} & \vdots & \mathbf{P}^T\boldsymbol{\Delta}^{(k)} \\ \cdots & \vdots & \cdots \\ \boldsymbol{\Delta}^{(k)^T}\mathbf{P}^T & \vdots & \boldsymbol{\Delta}^{(k)^T}\boldsymbol{\Delta}^{(k)} \end{bmatrix} \cdot \begin{bmatrix} \mathbf{v}^{(k)} \\ \cdots \\ \boldsymbol{\gamma}^{(k)} \end{bmatrix} = \begin{bmatrix} \mathbf{P}^T\mathbf{y}^{(k)} \\ \cdots \\ \boldsymbol{\Delta}^{(k)^T}\mathbf{y}^{(k)} \end{bmatrix} \tag{15}$$

dove $\boldsymbol{\Delta}^{(k)}$ è definita come illustrato di seguito:

$$\boldsymbol{\Delta}^{(k)} = \begin{bmatrix} p_1(1)\dfrac{(x_1^{(k)} - \mu_1^{(k)})}{\sigma_1^{(k)}} & \cdots & p_1(M)\dfrac{(x_1^{(k)} - \mu_M^{(k)})}{\sigma_M^{(k)}} \\ p_2(1)\dfrac{(x_2^{(k)} - \mu_1^{(k)})}{\sigma_1^{(k)}} & \cdots & p_2(M)\dfrac{(x_2^{(k)} - \mu_M^{(k)})}{\sigma_M^{(k)}} \\ \vdots & \ddots & \vdots \\ p_{N_T}(1)\dfrac{(x_{N_T}^{(k)} - \mu_1^{(k)})}{\sigma_1^{(k)}} & \cdots & p_{N_T}(M)\dfrac{(x_{N_T}^{(k)} - \mu_M^{(k)})}{\sigma_M^{(k)}} \end{bmatrix}_{(N_T \times M)} \tag{16}$$

$\mathbf{y}^{(k)}$ indica il vettore

$$\mathbf{y}^{(k)} = \left[y_1^{(k)}, \ldots, y_{N_T}^{(k)} \right]^T_{(N_T \times 1)}$$

e la matrice $\mathbf{P}$ è definita come

$$\mathbf{P} = \begin{bmatrix} p_1(1) & p_1(2) & \cdots & p_1(M) \\ p_2(1) & p_1(2) & \cdots & p_1(M) \\ \vdots & \vdots & \ddots & \vdots \\ p_{N_T}(1) & p_{N_T}(2) & \cdots & p_{N_T}(M) \end{bmatrix}_{(N_T \times M)} \tag{17}$$

Inoltre, poiché consideriamo solo una coordinata alla volta, i parametri da calcolare sono ridotti ai due vettori,

$$\mathbf{v}^{(k)} = \left[v_1^{(k)}, \ldots, v_M^{(k)} \right]^T$$

e

$$\boldsymbol{\gamma}^{(k)} = \left[\gamma_1^{(k)}, \ldots, \gamma_M^{(k)} \right]^T$$

cioè i valori del vettore $\mathbf{v}_i$ e della diagonale della matrice $\boldsymbol{\Gamma}_i$ relativi ad ogni coordinata dei vettori di allenamento $\{\mathbf{x}(n)\}$ e $\{\mathbf{y}(n)\}$. L'equazione (15) deve essere quindi applicata per ogni coefficiente, cioè per ogni k = 1,…, P dove P è la dimensione dei vettori che descrivono lo spazio acustico.

Per risolvere l'equazione di matrici (15) ci sono vari metodi. Si possono adottare strategie di soluzione di sistemi lineari oppure, come nel nostro caso, invertendo la matrice a blocchi più a sinistra. La formula risolutiva del problema sarà quindi la seguente:

$$\begin{bmatrix} \mathbf{v}^{(k)} \\ \cdots \\ \boldsymbol{\gamma}^{(k)} \end{bmatrix} = \begin{bmatrix} \mathbf{P}^T\mathbf{P} & \vdots & \mathbf{P}^T\boldsymbol{\Delta}^{(k)} \\ \cdots & \vdots & \cdots \\ \boldsymbol{\Delta}^{(k)^T}\mathbf{P}^T & \vdots & \boldsymbol{\Delta}^{(k)^T}\boldsymbol{\Delta}^{(k)} \end{bmatrix}^{-1} \cdot \begin{bmatrix} \mathbf{P}^T\mathbf{y}^{(k)} \\ \cdots \\ \boldsymbol{\Delta}^{(k)^T}\mathbf{y}^{(k)} \end{bmatrix} \quad \text{per k} = 1, \ldots, \text{P} \qquad (18)$$

Il maggior costo computazionale è dovuto al calcolo e all'inversione della matrice a blocchi a sinistra nella (15). Un problema inerente a questa matrice è la sua quasi singolarità, sulla sua diagonale infatti compaiano elementi molto vicini al valore 0. Questo inconveniente è stato ovviato aggiungendo una piccola perturbazione sulla diagonale. In particolare al posto di ogni 0 che si presenta viene sostituito un valore molto piccolo (dell'ordine di 10^{-30}) ma diverso da 0.

4.8 Miglioramento del modello

Il modello di calcolo di riferimento (Stylianou *et al.*, 1998) per questo metodo di conversione applica la trasformazione non a tutto lo spettro ma solo alla parte armonica di esso. Il presente lavoro invece non fa distinzione tra parte armonica e rumorosa del segnale vocale e mira ad inserirsi come processo di post-elaborazione del segnale in cascata al motore di sintesi vocale. Questo introduce una serie di difficoltà nella creazione del modello.

Il numero di parametri caratterizzanti gli spazi acustici *synth* e *target* è troppo alto e i loro valori sono troppo variabili per poter essere ricavata un'unica funzione di conversione che valga per ogni classe di vettori del modello *GMM*.

Si è deciso quindi di specializzare le funzioni di conversione, una per ogni classe C_k. Questo è stato fatto in due passaggi.

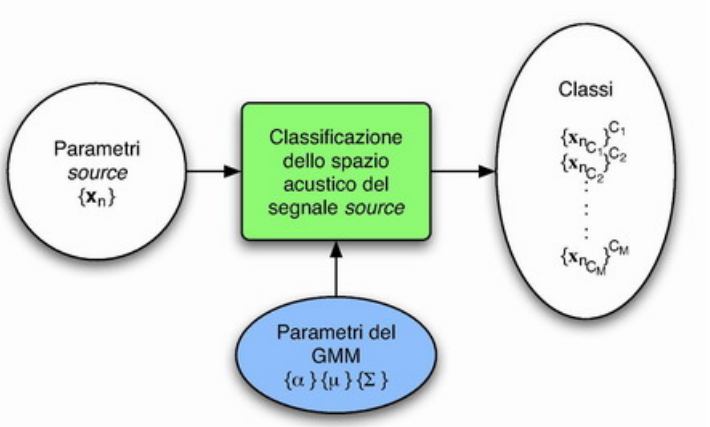


Figura 7: Schema della suddivisione in classi C_k

Il primo, illustrato in Figura 7, in cui si è ricavato all'interno dell'insieme dei vettori di allenamento $\{\mathbf{x}(n), n=1,\ldots,N_T\}$ l'elenco dei vettori appartenenti ad ogni classe, $\{\mathbf{x}(n_{C_k})\}^{C_k}$ con $n_{C_k}=1,\ldots,N_{C_k}$ e k=1,…,M. Per fare questo si sono utilizzati i dati relativi al *GMM* come base per un semplice riconoscitore che calcoli la probabilità di un vettore di appartenere ad una classe C_k e lo assegni a quella con probabilità maggiore.

Il secondo passaggio, Figura 8, consiste nel calcolare la funzione di conversione relativa ad ogni classe. Questo viene fatto usando solo gli elementi, $\mathbf{x}(n_{C_k})$ appartenenti ad una determinata classe e i corrispondenti vettori *target* come insieme di allenamento.

Sono state così create M diverse funzioni di conversione, una per ogni classe che potranno essere applicate solo ai vettori di quella determinata classe.

Se si ipotizza che ogni classe corrisponda ad un fonema in un particolare stato (dipendente dal fonema precedente e da quello successivo) allora quello che è stato creato è un sistema di conversione fonema-dipendente. Ad ogni fonema dello spazio acustico del segnale di partenza, riconosciuto con un opportuno *GMM*, verrà applicata una funzione specifica che lo convertirà nel suo corrispettivo dello spazio acustico *target*.

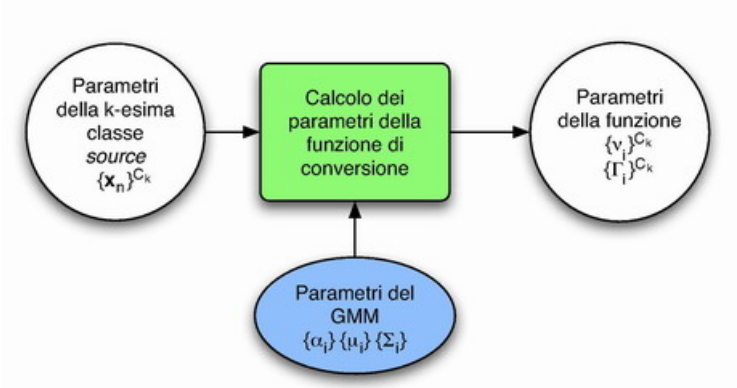


Figura 8: Schema del calcolo dei parametri della funzione di conversione, diversi per ogni classe individuata.

5. RISULTATI SPERIMENTALI

In questo paragrafo saranno analizzati i segnali prodotti con i metodi descritti nei paragrafi precedenti, sottolineando i principali risultati ottenuti.

Questi sono:

il segnale originale o target: è un segnale audio, registrato a 44 kHz, ricampionato successivamente a 16 kHz, diviso infine in 47 frammenti con una durata media di 10 s;

il segnale synth, sintetizzato tramite *copy synthesis*: è un segnale audio a 16 kHz, prodotto dal sintetizzatore vocale *MBROLA*, allineato con il segnale *target* per tipo e durata dei fonemi e per altezza del *pitch*.

il segnale mod modificato con la trasformazione diretta dell'inviluppo: segnale audio derivante dal segnale *synth* modificando ogni *frame* di quest'ultimo nel corrispondente del segnale *target*.

il segnale trasf trasformato con funzione di conversione basata sul modello statistico: segnale audio derivante dalla trasformazione del segnale *synth* tramite i parametri della trasformazione appresi con il modello statistico.

L'analisi dei segnali sarà effettuata tramite metodi oggettivi e soggettivi.

5.1 Parametri dell'esperimento

Le caratteristiche dei segnali sopra citati, imposte dalle specifiche della voce originale o frutto di scelte progettuali, talvolta molto delicate, sono le seguenti:

Frequenza di campionamento: Fs = 16 kHz.

Finestra di analisi: sono state utilizzate una finestra di *hamming*, per l'estrazione dei parametri mel-cepstrali, e una finestra di *blackman*, più performante nel processo di ricostruzione del segnale, per il filtraggio in frequenza.

Larghezza della finestra, N_{window}: la finestra ha una durata di 512 campioni che equivalgono a 32 ms.

Sovrapposizione delle finestre: le finestre di analisi si sovrappongo. In particolare è stato utilizzato un incremento tra una finestra e l'altra di 32 campioni, equivalenti a 2 ms.

Numero di campioni per la FFT N_{fft}: si è deciso di utilizzare 1024 campioni per calcolare la FFT, per cui è stata necessaria un'operazione di aggiunta di zeri alla fine del segnale finestrato (*zero padding*).

Numero di filtri: è uno dei parametri critici per il calcolo della trasformazione, sarà oggetto una maggiore analisi in seguito. Comunque il valore che si è trovato che dà le migliori prestazioni complessive è di 40 banchi.

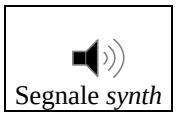
Numero di coefficienti MFCC: anche il numero di coefficienti *mel-cepstrali* si è rivelato molto critico e sarà oggetto di discussione. Comunque il valore scelto è stato di 26 coefficienti.

Ci sono inoltre dei parametri che caratterizzano il modello statistico utilizzato per la creazione del segnale *trasf* e sono:

Numero di gaussiane della mistura: più questo parametro è elevato, meglio il modello riuscirà a definire le varie istanze di vettori *MFCC* che si presentano, lo svantaggio però è che il peso computazionale diventa troppo elevato e quindi ingestibile. Il valore scelto è perciò di 240 gaussiane.

Numero di classi C_k: è il numero di classi in cui si è scelto di dividere il segnale sarà pari al numero delle gaussiane cioè 240.

5.2 I segnali target e synth


Segnale *synth*


Segnale *target*

Di seguito verranno illustrate le caratteristiche spettrali dei segnali che sono stati utilizzati come base per il calcolo delle trasformazioni. Come si vede dalla Figura 9 i due segnali, pur avendo la stessa durata e gli stessi fonemi, non hanno una forma d'onda molto simile.

La differenza più evidente è sicuramente l'altezza del segnale. Infatti l'intensità del segnale *synth* è più elevata del segnale *target*.

Nella Figura 10 si può notare come il *pitch* e le prime formanti siano molto simili, mentre, in alta frequenza, gli spettri presentano caratteristiche molto diverse.

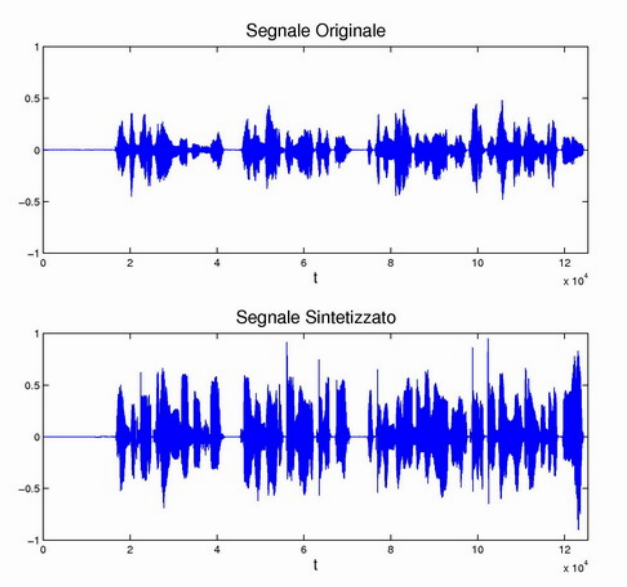
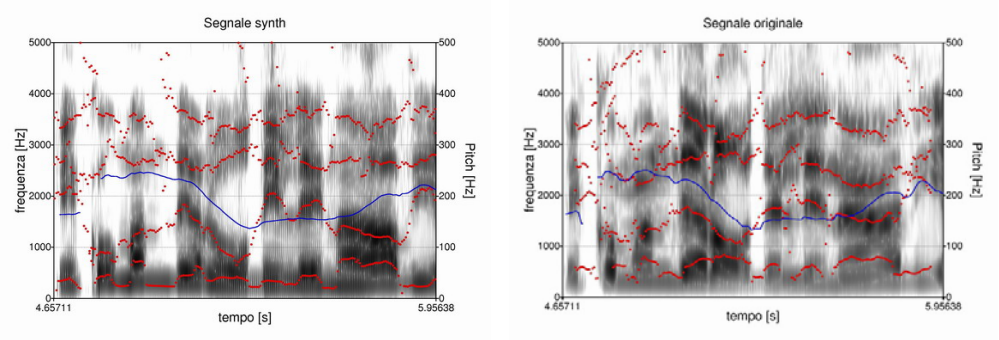


Figura 9: Forma d'onda dei due segnali vocali. Sono estratte, come tutti gli esempi seguenti, dalla terza parte del segnale originale e la sua copy synthesis.

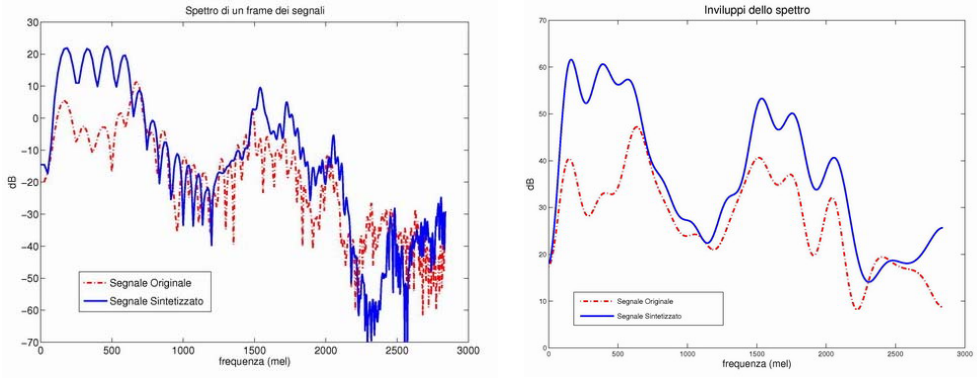


(a) (b)

Figura 10: Spettrogramma del segnale sintetizzato (a) e del segnale originale (b). Sono evidenziate le formanti e il tracciato del *pitch* (in scala enfatizzata).

Questo dimostra che gli strumenti della copy synthesis non bastano per riprodurre le caratteristiche di voice quality che forniscono alla voce un'emozione.

Si possono vedere bene queste differenze, evidenziando lo spettro dello stesso *frame* dei due segnali e il relativo inviluppo (Figura 11). Queste curve sono molto simili, per esempio nella posizione dei principali picchi, ma presentano differenze tra le ampiezze, diverse per ogni picco.



(a) (b)

Figura 11: Spettro di un *frame* del segnale originale e della sua copy synthesis (a). I relativi inviluppi (b). Il frammento è relativo alla vocale "e".

5.3 Trasformazione diretta

In questo paragrafo verranno illustrati gli effetti della trasformazione dello spettro del segnale derivante dalla copy synthesis tramite il metodo diretto, cioè la modifica di ogni *frame* del *synth* nello corrispettivo del *target*.

Segnale *mod*

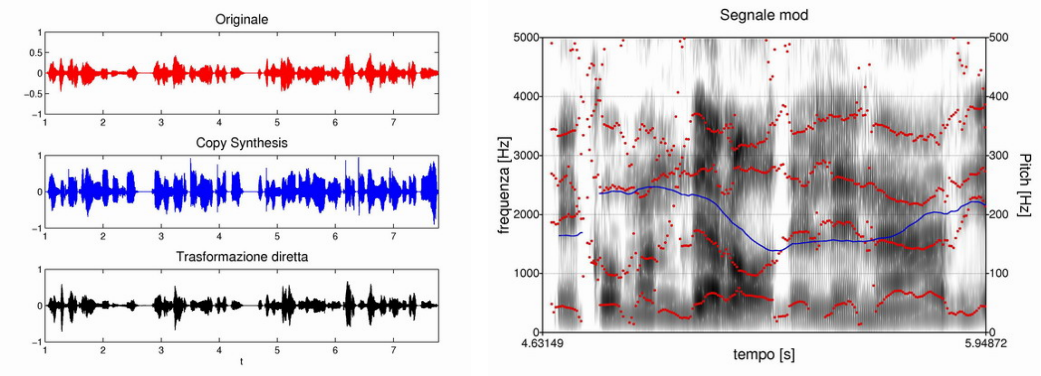
Nelle Figure 12 e 13 sono riportati alcuni esempi di *frame* di segnale confrontati l'obiettivo da imitare.

Si noti che in questo caso non si è ritenuto necessario operare il riscalamento delle energie. Quindi il segnale ottenuto avrà la stessa ampiezza del segnale *target*, inferiore a quella del segnale *synth*.

Il risultato ottenuto è abbastanza soddisfacente. E' necessario sottolineare come il metodo funzioni meglio alle basse frequenze (0-1000 Hz) mentre a quelle alte la conversione dello spettro non è sempre precisa. Questa è una conseguenza diretta dell'utilizzo della scala percettiva Mel che calcola gli inviluppi e le relative differenze più accuratamente in bassa frequenza. Comunque questa imprecisione non viene percepita all'ascolto poiché la banda penalizzata dalla scala Mel è anche quella a cui l'orecchio umano è meno sensibile.

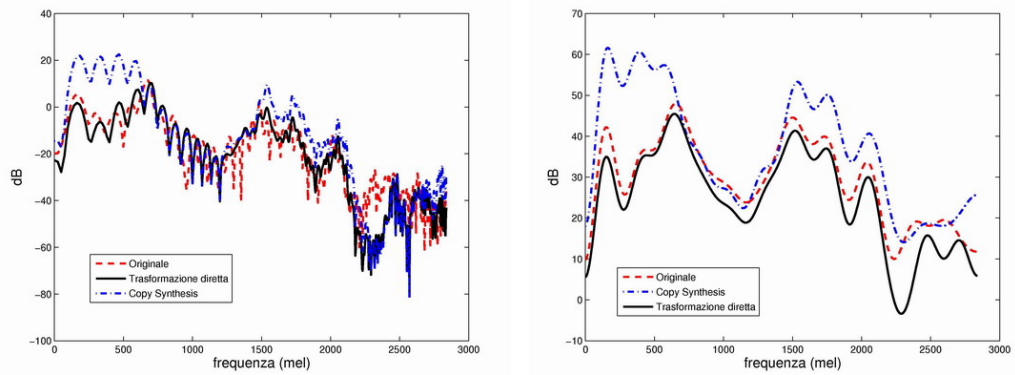
La trasformazione diretta ha dimostrato l'efficacia delle modifiche applicate allo spettro di un segnale, calcolate sulle differenze tra il suo inviluppo e quello del segnale *target*. Fornisce ad esso proprio le caratteristiche che mancano per rendere credibile l'emozione che si vuole esprimere.

Rappresenta quindi un limite superiore per la trasformazione con il modello statistico. Il modello infatti non riuscirà mai a predire il segnale *target* meglio di quanto faccia il metodo diretto. Si può vedere, nelle figure qui riportate, come gli spettri e le forme d'onda degli esempi di segnale trasformato (*mod*) siano abbastanza aderenti a quelli del segnale obiettivo. Il metodo quindi opera un'ottima conversione dello spettro e fornisce un limite molto elevato dando ampio margine di miglioramenti per il modello statistico.



(a) (b)

Figura 12: (a) Forma d'onda del segnale modificato tramite trasformazione diretta (in basso). In alto il segnale *target*, al centro il segnale *synth*. (b) Spettrogramma di una parte del segnale modificato con la trasformazione spettrale diretta. Sono evidenziate le formanti e il tracciato del *pitch* (in scala enfatizzata).



(a) (b)

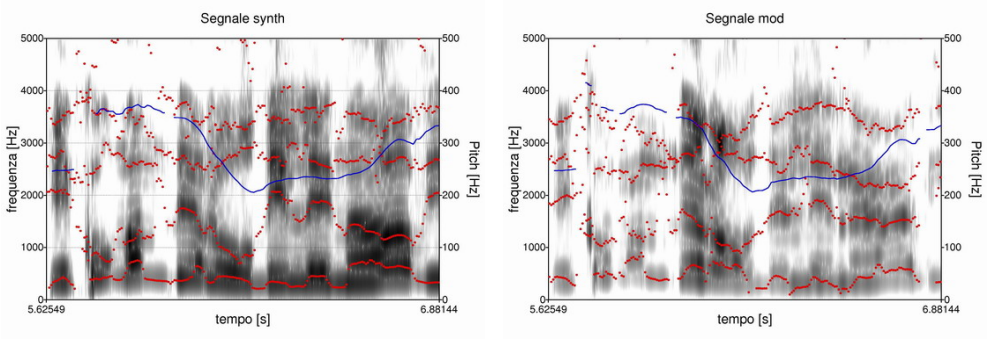
Figura 13: (a) Spettro di un *frame* del segnale modificato con la trasformazione spettrale diretta. Sono disegnati anche gli spettri del *target* e del *synth*. Il *frame* è relativo ad una vocale "e". (b) Inviluppo dello stesso *frame* del segnale. Sono disegnati anche gli inviluppi del *target* e del *synth*. Il *frame* è relativo ad una vocale "e".

5.3.1. Indipendenza dal pitch

Una caratteristica importante che deve avere una trasformazione di questo tipo è l'indipendenza dal *pitch* del segnale che si vuole modificare. Se così non fosse, infatti, il cambiamento dello spettro potrebbe modificare l'intonazione della frase e peggiorare la qualità della voce. Inoltre una trasformazione di questo tipo non sarebbe applicabile su segnali diversi da quelli preallineati nel *pitch.*

Per verificare che questa condizione sia soddisfatta è stata calcolata una matrice di differenze[8] calcolate tra un segnale *synth* e un segnale *target* su ogni *frame* ed è stata riapplicata allo stesso segnale di partenza, ma con *pitch* alzato o abbassato di un'ottava.

[8] La matrice *Diff* del metodo Diretto.



(a) (b)

Figura 14: Spettrogramma del segnale sintetizzato (a) e del segnale modificato con la trasformazione spettrale diretta (b). Il *pitch* del segnale di partenza è stato alzato di 1 ottava.

Dalla Figura 14 si vede che il tracciato del *pitch* (la linea continua) non varia tra segnale *synth* e *mod*; si modificano invece le formanti secondarie e si percepisce una trasformazione emotiva.

Questo dimostra che la trasformazione ha un valore che prescinde dal valore di *pitch* del segnale di partenza e può essere, quindi, applicata efficacemente anche se i segnali non sono allineati come intonazione.

5.3.2. Correlati acustici spettrali

Per decidere i parametri da utilizzare nella sintesi sono state effettuate numerose prove con molti valori. I segnali così ricavati sono stati poi analizzati per valutare quello con le migliori prestazioni. A questo scopo sono stati utilizzati gli *indicatori spettrali* introdotti nel paragrafo precedente.

Come ovvi valori di riferimento per queste verifiche sono stati presi quelli del segnale *synth* e quelli del segnale *target*.

Il campione analizzato è costituito da un *frame* di 1024 campioni all'interno della vocale "*a*" pronunciata nel terzo frammento del corpus. Fa parte della frase "…voglio andare per m-**a**-re come te…". Si è scelto un segnale audio vocalizzato perché gli indicatori hanno senso solo in questo contesto.

Sono state eseguite varie trasformazioni con diversi valori del numero di banchi di filtri e dei coefficienti *MFCC*. All'aumentare del valore di entrambi i parametri, i correlati acustici del segnale trasformato si avvicinano mediamente a quelli del segnale originale.

Gli indicatori relativi alla misura della differenza tra l'energia del segnale in alta e bassa frequenza invece tendono a convergere verso i valori del segnale *target*.

Non si nota un eccessivo miglioramento, all'aumentare dei filtri nel banco di analisi. Si nota invece l'esigenza di incrementare il numero di *MFCC* proporzionalmente al numero di filtri.

Per questi motivi, si è scelto di utilizzare come parametri della trasformazione spettrale, un banco da 40 filtri e un numero di coefficienti di 26.

Se si riassumono in un unico grafico i parametri calcolati per i vari segnali: originale, *copy synthesis*, con modello, si può verificare come variano i valori per questi tre segnali.

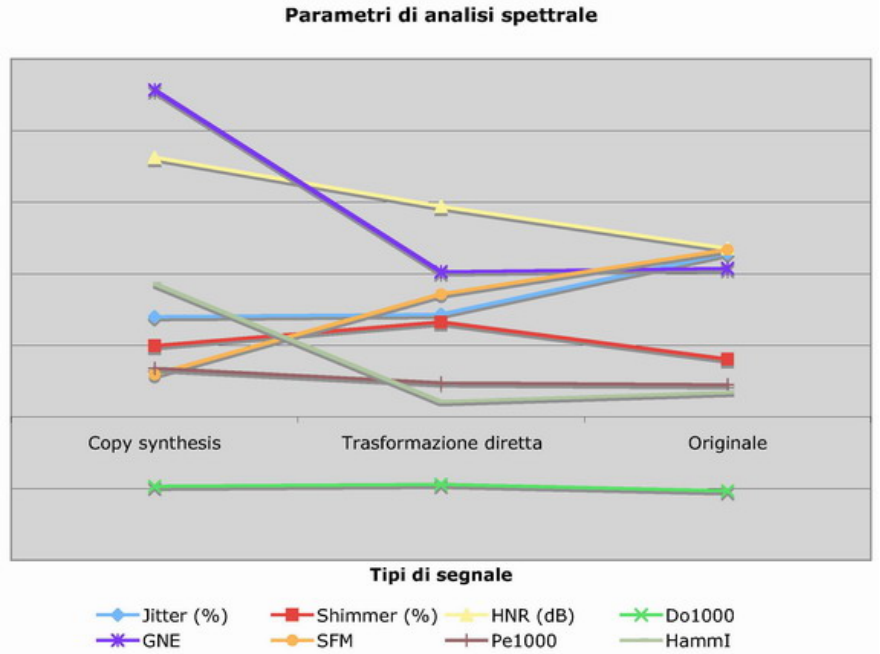


Figura 15: Confronto, a parità di numero di filtri nel banco e di coefficienti mel-cepstrali, degli indicatori spettrali dei segnali *synth*, modificato e *target*.

In Figura 15 si può vedere come gli indicatori relativi al segnale trasformato direttamente si collochino in una zona intermedia tra quelli del segnale derivante dalla *copy synthesis* e quelli relativi al segnale originale.

5.4 Trasformazione con il modello

Una volta decisi i parametri della trasformazione dello spettro, è stato inserito in essa il modello statistico precedentemente calcolato. Al posto dei coefficienti mel-cepstrali estratti dal segnale *target*, sono stati utilizzati i coefficienti forniti dalla funzione di conversione (11).

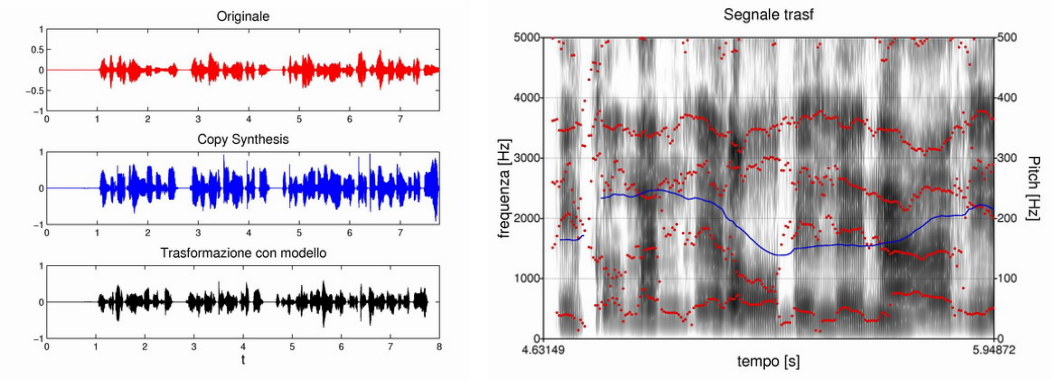
Nell'applicazione di questo metodo si presentano due possibilità: il segnale neutro di partenza appartiene all'insieme di quelli usati per l'addestramento della funzione, il segnale neutro è preso dello stesso *corpus* ma è esterno all'insieme di addestramento.

5.4.1. Trasformazione di un segnale dell'insieme di addestramento

Segnale *trasf*

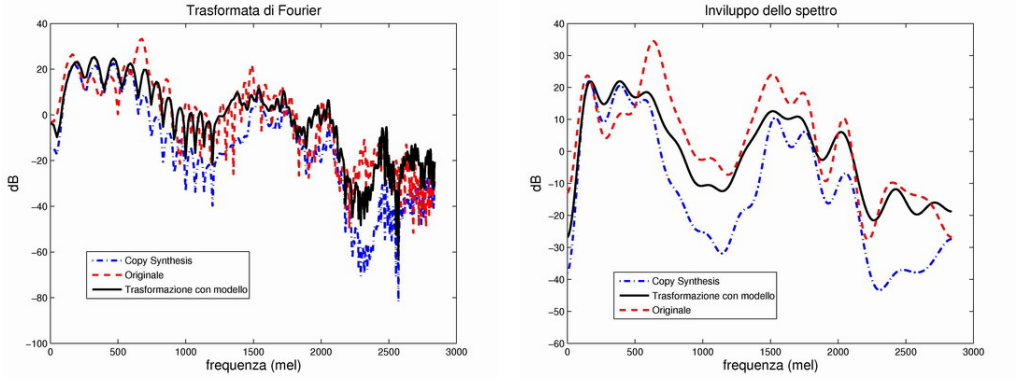
Dall'esempio di segnale qui rappresentato, si vede che la trasformazione con il modello applicata ad un segnale dell'insieme di addestramento agisce sullo spettro del segnale sintetizzato, ma non produce risultati ottimi come il metodo diretto. Introduce infatti

disturbi dovuti ad un'inadeguata identificazione del inviluppo verso cui il segnale deve essere convertito.



(a) (b)

Figura 16: (a) Esempio di forme d'onda del segnale modificato tramite trasformazione con modello (in basso). In alto il segnale *target*, al centro il segnale *synth*. (b) Spettrogramma di una parte del segnale modificato attraverso la trasformazione con modello. Sono evidenziate le formanti e il tracciato del *pitch* (in scala enfatizzata).



(a) (b)

Figura 17: (a) Spettro di un *frame* del segnale modificato attraverso la trasformazione spettrale con modello. Sono disegnati anche gli spettri del *target* e del *synth*. Il *frame* è relativo ad una vocale "e". (b) Inviluppo dello stesso *frame* del segnale trasformato con il modello. Sono disegnati anche gli inviluppi del *target*} e del *synth*. Il *frame* è relativo ad una vocale "e".

Le problematiche sono numerose. Il principale problema deriva da una difficoltà nel modello a mistura di gaussiane (*GMM*) ad identificare tutti i tipi di vettori *MFCC* presenti nello spazio acustico da imitare e nel creare una quindi una funzione di conversione che li crei.

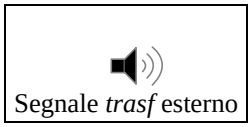
Il modello infatti individua con precisione la classe di appartenenza di ciascun vettore, ma all'interno di essa c'è ancora una grossa variabilità che si ripercuote sui risultati della trasformazione. Questo potrebbe essere ovviato aumentando il numero delle gaussiane del

modello, il che, però, aumenterebbe troppo il tempo di calcolo e renderebbe il metodo non efficiente.

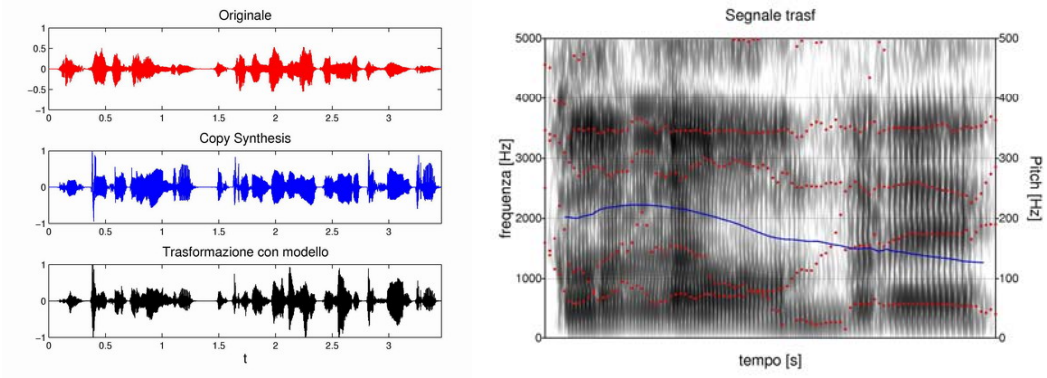
Un'ulteriore problema è il valore utilizzato come limite inferiore per gli elementi della diagonale della matrice di covarianza del *GMM*. Inoltre un difetto del modello *GMM*, noto anche in letteratura, è l'appiattimento dello spettro e una perdita di informazioni in alta frequenza.

5.4.2. Trasformazione di un segnale esterno all'insieme di allenamento

Per quanto riguarda, infine, il caso più generale della trasformazione di un segnale del *corpus*, ma non appartenente all'insieme di addestramento, le prestazioni del metodo sono inferiori alla situazione precedente.



In questo caso, infatti, c'è una probabilità maggiore che il modello debba trasformare un vettore che non riconosce. Può capitare che per un vettore sorgente, **x**(n), nessuna classe, C_k , dia un valore di $P(C_k \mid \mathbf{x}(n))$ maggiore di 0.5. Questo implica un'incertezza che si traduce in un errore nella trasformazione.

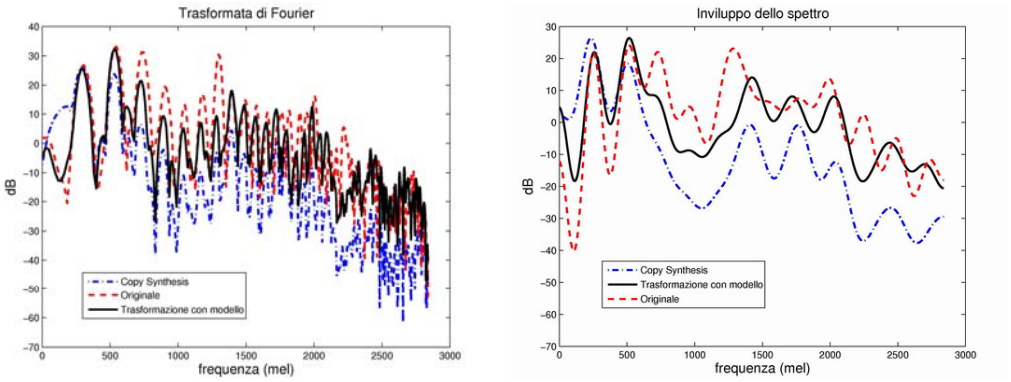


(a) (b)

Figura 18: (a) Esempio di forme d'onda di un segnale, esterno all'insieme di allenamento, modificato tramite trasformazione con modello (in basso). In alto, il segnale *target*, al centro il segnale *synth*. (b) Spettrogramma di una parte di un segnale, esterno all'insieme di allenamento, modificato attraverso la trasformazione con modello. Sono evidenziate le formanti e il tracciato del *pitch* (in scala enfatizzata).

Il modello prova a ricostruire il vettore tramite una somma pesata di vettori *target* di classi simili, ma, per come è definito il metodo, la matrici della trasformazione sono addestrate solo sui vettori di una classe, quindi, quando il vettore non è riconosciuto viene trasformato con una funzione diversa da quella ottima.

Nonostante questo problema che introduce dei disturbi e corrompe leggermente il segnale, l'intelligibilità del messaggio e la percezione di una emozione diversa da quella neutra è sempre riscontrata.



(a) (b)

Figura 19: (a) Spettro di un *frame* di un segnale, esterno all'insieme di allenamento, modificato attraverso la trasformazione spettrale con modello. Sono disegnati anche gli spettri del *target* e del *synth*. (b) Inviluppo dello stesso *frame* di un segnale, esterno all'insieme di allenamento, trasformato con il modello. Sono disegnati anche gli inviluppi del *target* e del *synth*.

5.4.3. Correlati acustici spettrali

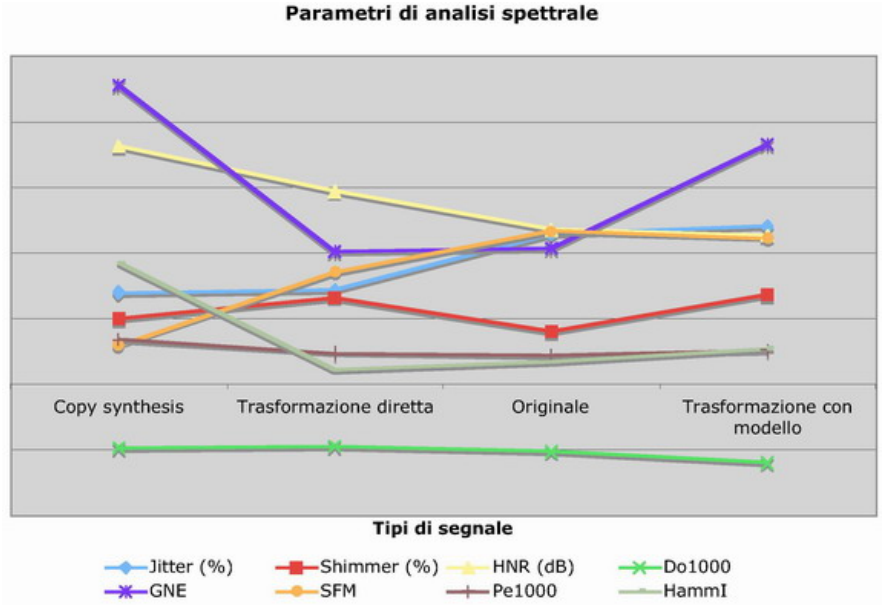


Figura 20: Confronto tra i correlati acustici spettrali dei segnali *synth*, modificato con la *trasformazione diretta*, *target* e modificato con il *modello*.

Se si estraggono i correlati acustici spettrali della voce trasformata con il modello e si confrontano con quelli ricavati dagli altri segnali, si nota che c'è una tendenza a convergere verso i valori del segnale *target* (Figura 20)

È sicuramente una convergenza meno marcata rispetto al caso diretto, che rappresenta, come detto in precedenza, il limite superiore per la qualità della trasformazione, però i valori sono mediamente sempre migliori rispetto a quelli del segnale sintetizzato da MBROLA.

6. CONCLUSIONI

In questo lavoro è stato ricavato un metodo di trasformazione dello spettro di un segnale locale basato su un modello statistico a mistura di gaussiane (*GMM*) che riconosce il tipo di *frame* di segnale che si vuole convertire e in base ad esso applica una funzione di conversione calcolata *ad hoc*.

Il modello è stato addestrato su un segnale vocale registrato da un parlatore e sull'uscita, allineata con il primo per durata dei fonemi e intonazione, di un sintetizzatore vocale.

Questo sistema funziona come modulo di post-elaborazione del segnale.

È stato dimostrato che la funzione di conversione non necessita del processo di allineamento dell'intonazione per operare la conversione della voce, perché indipendente dal *pitch*.

Il risultato della trasformazione sarà inferiore a quello che idealmente si potrebbe ottenere tramite una conversione spettrale diretta, perché i vettori mel-cepstrali del segnale potranno assumere valori non modellati, però, nella maggior parte dei *frame*, produrrà un risultato accettabile e riconoscibile come voce emotiva.

6.1 Valutazione del metodo

La trasformazione spettrale cambia in modo ottimo lo spettro di ogni *frame* di segnale e trasmette ad esso quasi tutte le caratteristiche spettrali che rappresentano l'emozione della "collera". La conferma di ciò si ha perché il segnale, prodotto con il metodo diretto, è quasi indistinguibile, all'ascolto, dall'originale registrato.

Il sistema di riconoscimento, basato sul *GMM*, è in grado di riconoscere tutti i vettori che sono stati dati come ingresso.

La funzione calcolata tramite i parametri del *GMM* ha dei problemi dovuti all'estrema variabilità del segnale e alla difficoltà di costruire una funzione di conversione per ogni classe in cui è stato suddiviso lo spazio acustico del segnale sintetizzato.

Il segnale che si è riusciti a ricavare applicando il modello statistico di conversione si attesta, come qualità, tra il segnale trasformato con il metodo diretto e il segnale originale.

6.2 Prospettive future di sviluppo

In questo lavoro si è dimostrato che il metodo di conversione statistico è buono, ma necessita di ulteriori perfezionamenti e sviluppi.

Prima di tutto devono essere fatti ancora numerosi test per valutare l'influenza del numero di classi in cui si divide lo spazio acustico. Probabilmente il valore ottimo è intermedio tra una classe unica e tante quante il numero delle gaussiane.

Si potrebbe applicare ad un motore di sintesi sinusoidale, come quello si sta sviluppando presso l'Istituto di Scienze e Tecnologie della Cognizione, Sezione di Padova "Fonetica e Dialettologia" del CNR, che permetterebbe di applicare la trasformazione esclusivamente sulla parte armonica del segnale. Questo comporterebbe una semplificazione del modello statistico e un miglioramento della qualità audio. La parte non armonica del segnale sarebbe trasformata invece attraverso un altro tipo di filtraggio in frequenza.

Ulteriori sviluppi si potrebbero avere cercando di contestualizzare meglio il *frame* che si vuole convertire. Se si tenesse conto dei fonemi che precedono e seguono quello che

vogliamo modificare, il modello riuscirebbe a predire meglio la trasformazione. Questo si può fare in molti modi: introducendo i coefficienti δ e δ^2 nel calcolo dei coefficienti mel oppure attraverso alberi di decisione CART, oppure attraverso delle catene di Markov nascoste (HMM). Per effettuare questo sarebbe necessario un corpus di riferimento più ampio.

Un'altra caratteristica di questo sistema di conversione è che, allenando nuovamente la funzione su un adeguato insieme di riferimento, è possibile ottenere una trasformazione per ogni emozione desiderata (gioia, paura, disgusto, ecc…). Il sistema è esattamente la stesso, è necessario solamente ricalcolare le matrici della trasformazione relative ad ogni emozione.

Questo modello infine si presta alla creazione di un modulo indipendente da inserire come post-elaborazione nel motore di sintesi vocale. Senza modificare in alcun modo l'architettura già esistente, questo modulo si inserirebbe per modificarne l'uscita in base alle indicazioni dettate dall'utente. Oltre all'opzione neutro o emotivo, potrebbe essere implementata anche la possibilità di avere diversi gradi di emozione. Basta semplicemente pesare la matrice delle differenze con cui viene modificato lo spettro.

7. BIBLIOGRAFIA

Abe, M.; Nakamura, S.; Shikano, K.; Kuwabara, H., 1988. Voice conversion through vector quantization. *Proceedings of International Conference on Acoustics, Speech, and Signal Processing*, 655-658.

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*, 40 (2-3), 61-70.

Banse, R.; Scherer, K. R., 1996. Acoustic profiles in vocal emotion expression. *Journal of Personality and Social Psychology*, 70 (3), 614-636.

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

Boersma, P., 2001. PRAAT, a system for doing phonetics by computer. *Glot International*, 5, (9/10), 341-345. http://www.fon.hum.uva.nl/praat.

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, 2, 527-530.

Deller, J. R.; Proakis, J. G.; Hansen, J. H., 1993. *Discrete Time Processing of Speech*. Lebanon, Indiana, U.S.A: Prentice Hall PR.

Dempster, A. P.; Laird, N. M.; Rubin, D. B., 1977. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society B*, 39, 1-38.

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

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

Laver, J., 1980. *The phonetic description of Voice Quality*. Cambridge: Cambridge University Press.

Reynolds, D. A.; Rose, R. C., 1995. Robust text-indipendent speaker identification using Gaussian mixture speaker models. *Institute of Electrical and Electronics Engineers Trans. Speech Audio Processing*, 3, January, 72-83.

Stylianou, Y.; Cappè, O.; Moulines, E., 1998. Continuos probabilistic transform for voice conversion. *Institute of Electrical and Electronics Engineers Transactions on Speech and Audio Processing*, March, vol. 6 (2), 131-142.

Sutton, S.; Novick, D.G.; Cole, R. A.; Fanty, M., 1996. Building 10,000 spoken-dialogue systems. In *Proceedings of the International Conference on Spoken Language Processing '96*, Philadephia, P.A., October, 2, 709-712.

Young, S.; Evermann, G.; Hain, T.; Kershaw, D.; Moore, G.; Odell, J.; Ollason, D.; Povey, D.; Valtchev, V.; Woodland, P., 2002. *The HTK Book* (for HTK Version 3.2.1). Cambridge University Engineering Department, http://htk.eng.cam.ac.uk.