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
Viale Marx, 15
00137 Roma (Italy)
www: http://www.istc.cnr.it

SEZIONE DI PADOVA "FONETICA E DIALETTOLOGIA"
Via G. Anghinoni, 10
35121 Padova (Italy)
e-mail: cosi@pd.istc.cnr.it
www: http://www.pd.istc.cnr.it

"ANALISI PROSODICA"
teorie, modelli e sistemi di annotazione
2° Convegno Nazionale AISV – 30/11- 2/12 2005
Università degli Studi di Salerno, Campus di Fisciano - - "Aula delle Lauree"

Obiettivo

Convertire un segnale vocale neutro (privo di caratterizzazioni emotive) in un segnale vocale "emotivo"

Metodo

Si utilizza una funzione di conversione dello spettro basata su un modello statistico a mistura di gaussiane (*GMM*)

Sintesi delle emozioni

- Il presente studio si colloca nell'ambito della sintesi vocale emotiva.
- Nei primi studi sulla sintesi vocale l'importante era ottenere l'**intellegibilità**, ora che questa è stata raggiunta, diventa oggetto di ricerca la **qualità della voce (Voice Quality)**.
- La sfida più importante è fornire **naturalezza** alla voce sintetizzata.
- La sintesi delle emozioni può essere effettuata con buoni risultati soprattutto con i sintetizzatori a **concatenazione di difoni** perché si può agire sulla forma dello spettro di ogni singolo fonema.

Schema del progetto

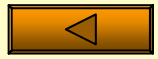
- Acquisizione di un *corpus* di analisi
- *Copy synthesis* tramite Mbrola
- Calcolo dei parametri della trasformazione dello spettro
- Creazione del modello statistico
- Trasformazione con il modello

Corpus **di analisi**

- Con il termine *corpus* si intende l'insieme di segnali audio da cui si è partiti per creare la funzione di conversione.
- È costituito da:
  - la voce di un parlatore che legge, in camera *anecoica,* il racconto "Il Colombre" di Dino Buzzati simulando l'emozione della **rabbia**.
  - un segnale sintetizzato per *copy synthesis*

Peculiarità:

la voce è la stessa che è stata utilizzata per creare il database di difoni del sintetizzatore vocale.



Copy Synthesis

Schema

Processo che permette la creazione di un segnale sintetizzato uguale ad un originale, per

- Testo pronunciato
- Durata dei fonemi
- Altezza del pitch

- Etichettatura dei fonemi tramite un riconoscitore vocale (HMM + ANN) ▷
- Estrazione del *pitch* con un analizzatore di segnali (PRAAT) ▷
- Creazione del file di istruzioni per il motore di sintesi (file ".pho") ▷
- Creazione della forma d'onda tramite il motore di sintesi Mbrola

◁

Coefficienti Mel-Cepstrum

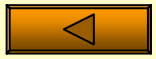
- Da entrambi i segnali del corpus vengono estratti i coefficienti cepstrali in scala Mel (MFCC)
- I coefficienti cepstrali si calcolano antitrasformando il logaritmo della trasformata di Fourier del segnale
- Scala Mel:

$$f_{mel} = 2595 \log_{10}\left(1 + \frac{f_{Hz}}{700}\right)$$

Caratteristiche

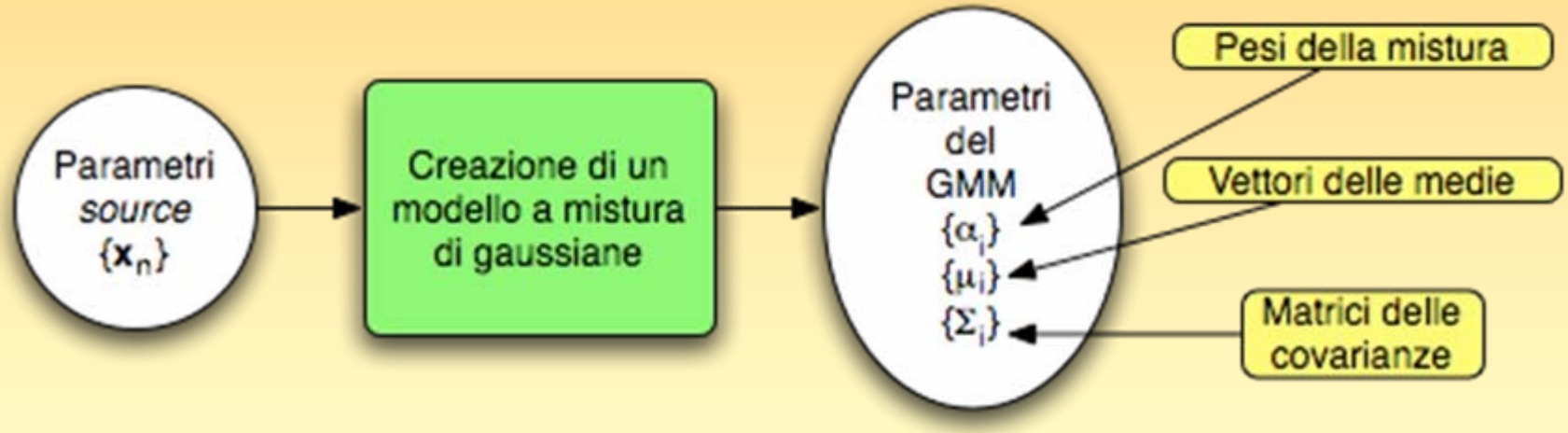
1. nel dominio cepstrale la convoluzione diventa una somma
2. enfatizzano le frequenze a cui l'orecchio umano è più sensibile
3. da essi si può ricavare l'inviluppo dello spettro del segnale

Specifiche



Funzione di conversione

- **Passo 1**: creazione di un modello statistico a mistura di gaussiane (*GMM, Gaussian Mixture Model*) dello spazio acustico dei vettori rappresentanti il segnale *synth* (1 mistura da 312 gaussiane).



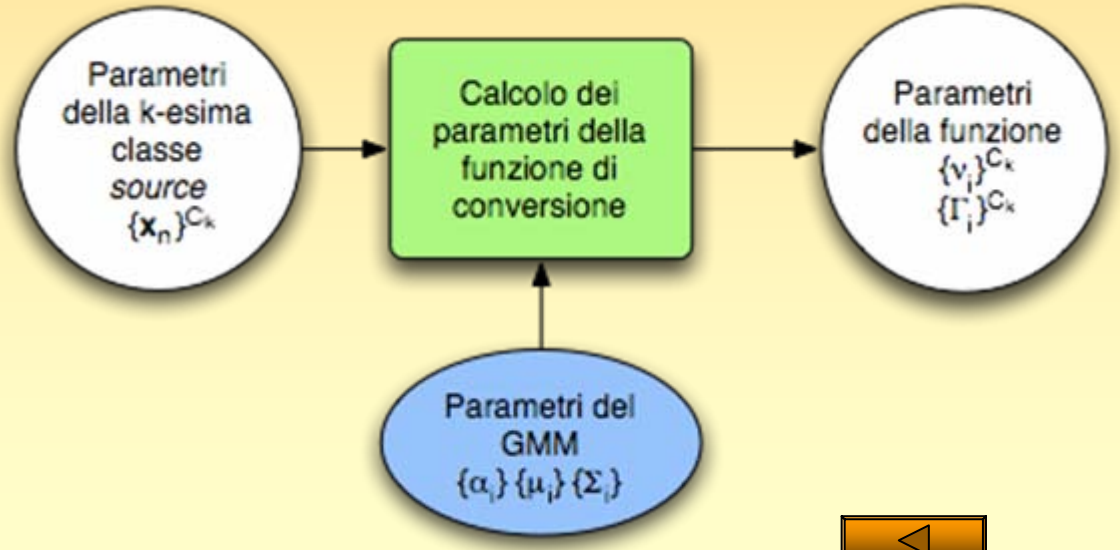
- **Passo 2**: divisione in classi C_k. Il numero delle classi è lo stesso delle gaussiane del GMM.

Funzione di conversione

- **Passo 3**: estrazione dall'insieme dei vettori *synth* e *target* di allenamento dei parametri della funzione di conversione:

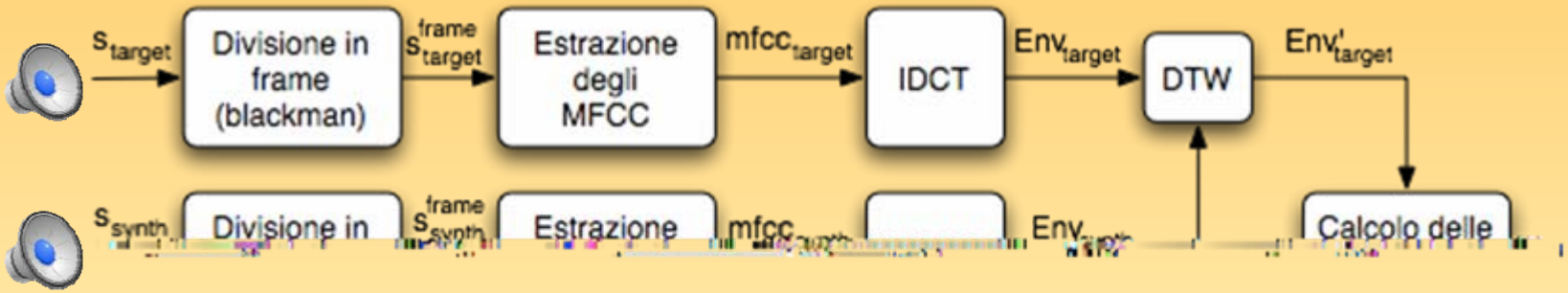
$$F(\mathbf{x}_n) = \sum_{k=1}^{M} P(C_k \mid \mathbf{x}_n)\left[\nu_k + \Gamma_k \Sigma_k^{-1}(\mathbf{x}_n - \mu_k)\right]$$

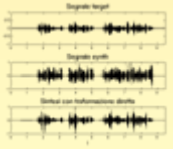
Per ogni classe identificata dal GMM, si applica la formula e si ottengono tante funzioni di conversione quante le gaussiane della mistura



Trasformazione spettrale

Metodo diretto

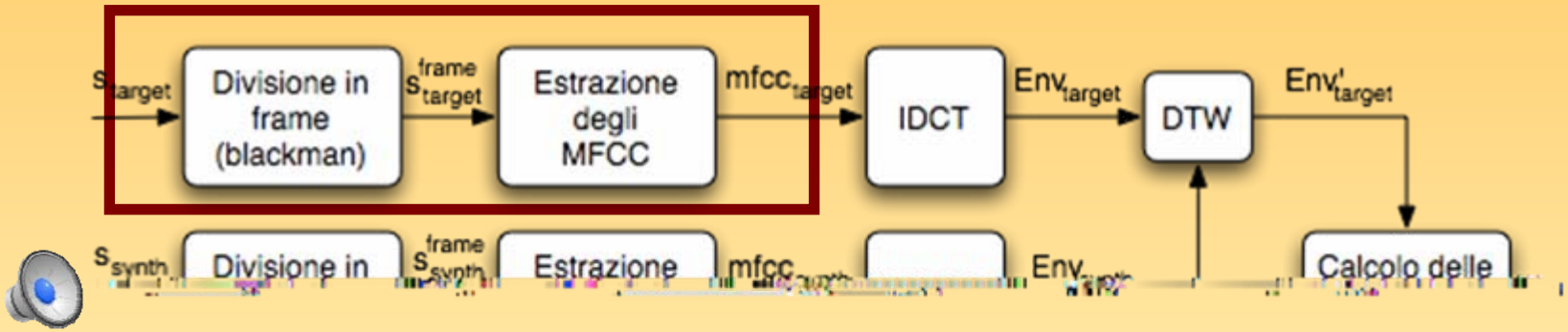




esempio

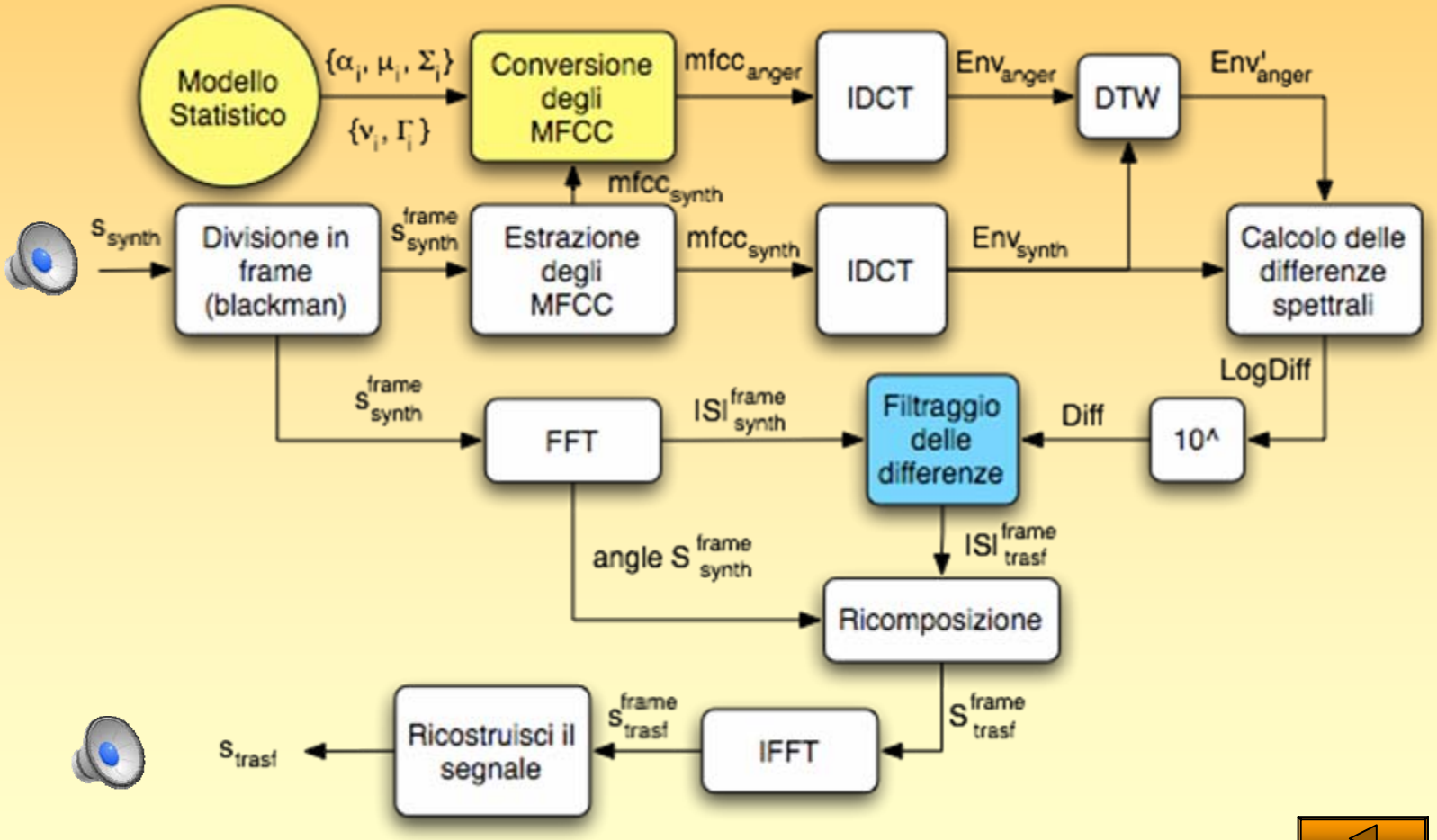
Trasformazione spettrale

Metodo diretto



Trasformazione spettrale

Modello statistico



Conclusioni

- La **trasformazione** dello spettro è un metodo molto efficace.
- Il modello statistico (GMM) ben rappresenta lo spazio acustico del segnale *synth*
- La **funzione** di conversione è efficace. Anche se si introducono dei disturbi sull'energia del segnale, la voce prodotta è riconoscibile come "arrabbiata".
- Questo metodo rappresenta quindi un miglioramento nello stato dell'arte della sintesi delle emozioni.

Sviluppi futuri

- Inserire la funzione di conversione come **modulo** del sintetizzatore vocale di **Mbrola**.
- Verificare i **parametri critici** del modello statistico come il numero di classi.
- Sviluppare ulteriori modelli per le **altre emozioni**.
- Sviluppare il metodo anche in un **sintetizzatore sinusoidale**.
- Contestualizzare l'analisi di un *frame* considerando le informazioni relative i precedenti e i successivi.

Copy Synthesis

- Etichettatura dei fonemi:
  - riconoscimento vocale con sistema sviluppato dall'Istituto di Fonetica e Dialettologia ISTC-CNR di Padova.
  - si basa su un modello ibrido di catene di Markov nascoste (HMM) e di rete neurale (ANN)
  - Creazione del file di testo ".plab"

```
...
3.53 26 #
3.57 26 v
3.68 26 0
13.73 26 L
...
```

Copy Synthesis

- Estrazione del **pitch**:
  - Analisi tramite il software per i segnali audio PRAAT
  - Estrazione del pitch medio su finestre di 20 ms
  - Creazione del file di testo "*.pitch*"

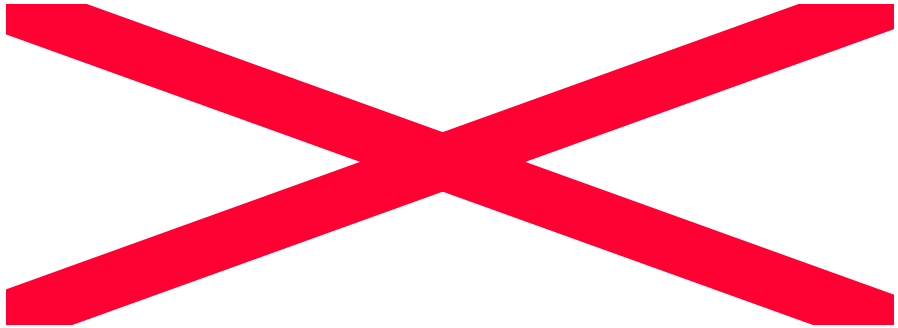
```
...
0
163.66711
174.42488
191.00139
...
```

Copy Synthesis

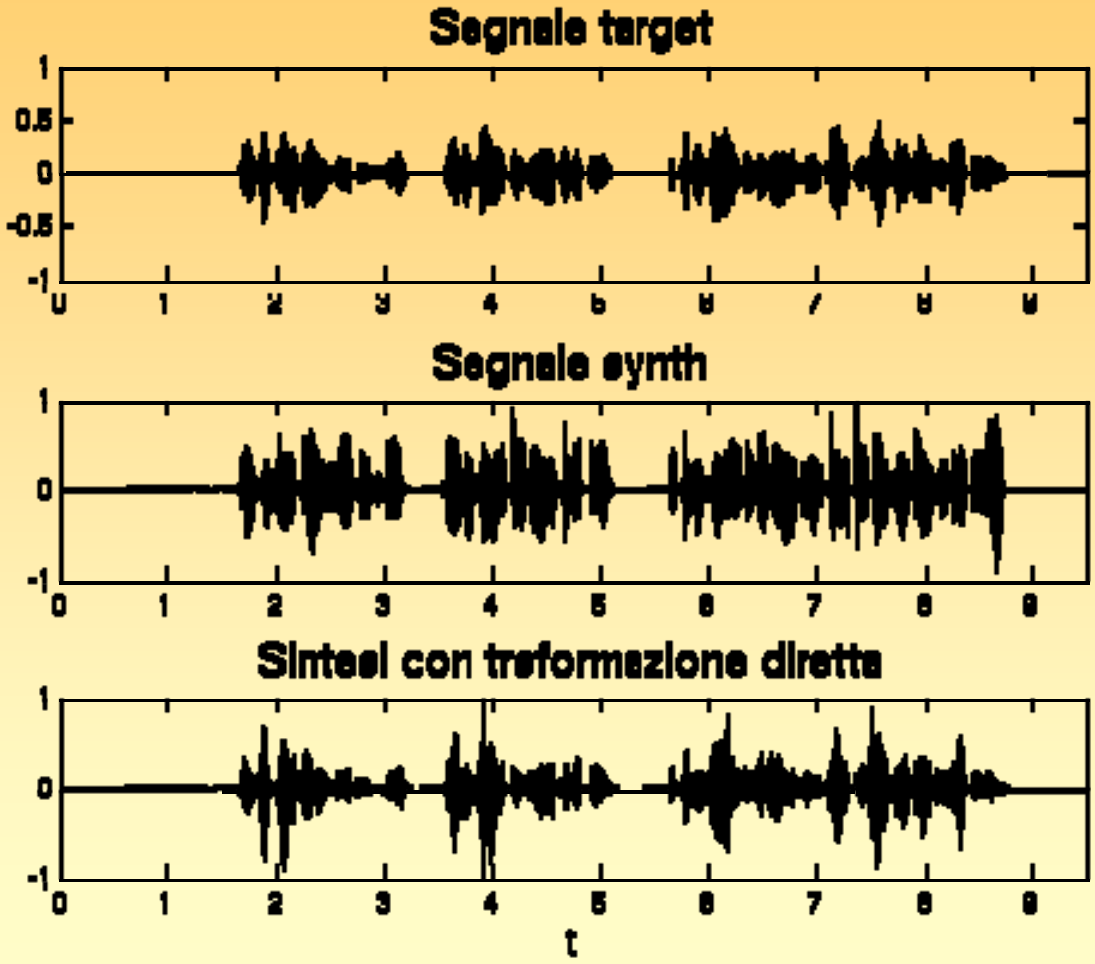
- Creazione della **forma d'onda**:
  - Creazione del file di testo "*.pho*" derivato dall'unione delle informazioni contenute nel file "*.plab*" e nel "*.pitch*"

```
...
v 40
O1 110 0 161.6486 20 172.2822 40 184.2199 60 195.3564 80 205.6733
L 50 0 214.4922 33 221.4594 67 227.1906
o 40 0 231.6248 50 232.4543
...
```

Schema del processo di Copy Synthesis

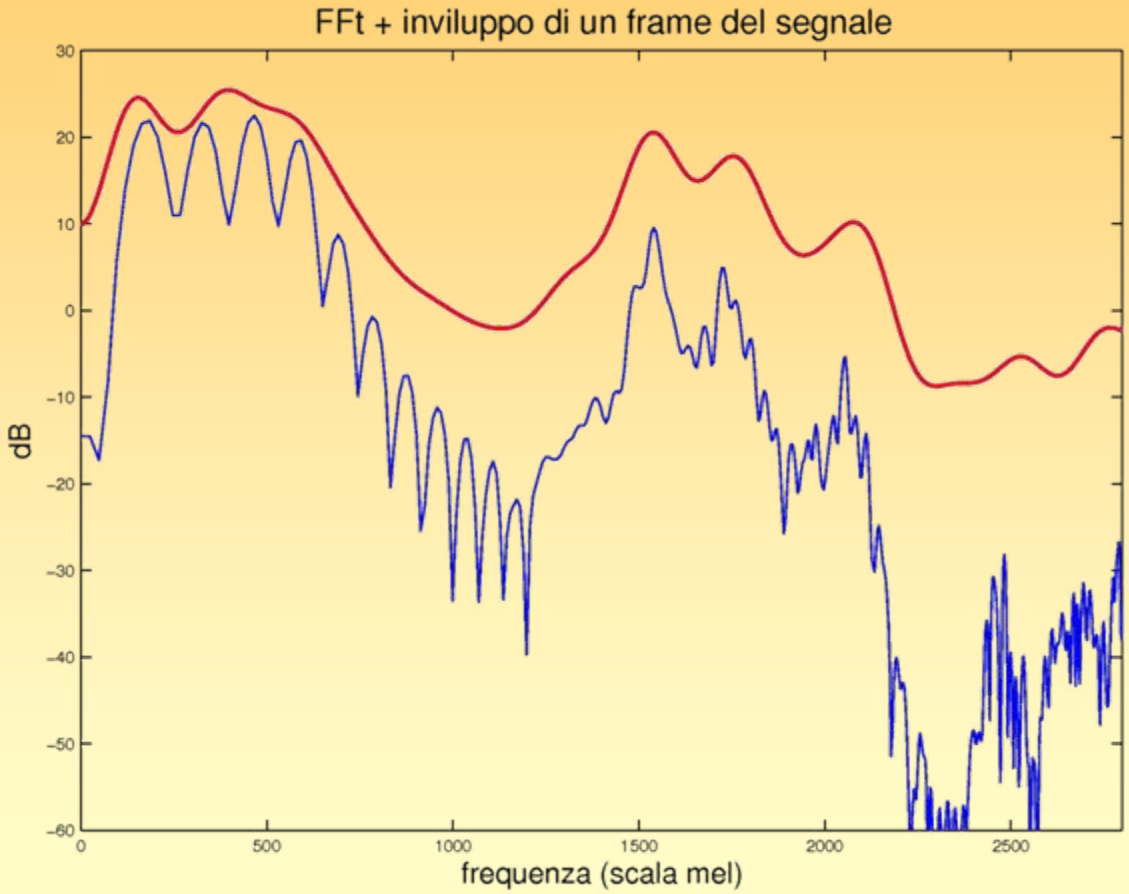


Analisi dei segnali nel tempo



Coefficienti MFCC

Coefficienti Mel-Cepstrum



Specifiche:

- Finestra di analisi: 32 ms (512 punti)
- Incremento delle finestre: 2 ms (32 punti)
- Punti della FFT: 1024
- Numero di filtri per il calcolo degli MFCC: 40
- Numero di MFCC: 26