UNIVERSITÀ
DEGLI STUDI DI TRIESTE

dipartimento
di ingegneria
e architettura

L’analisi di segnali audio

A.Carini – Elettronica per l’audio e l’acustica

Analisi audio

- La maggior parte delle tecniche di analisi opera su finestre (frame) di dati del segnale audio di ingresso.
- Molte richiedono che nella finestra di analisi il segnale possa essere considerato stazionario (nel senso che le statistiche del segnale e la distribuzione in frequenza non cambino apprezzabilmente durante la durata della finestra).
- La scelta della durata della finestra, nonché del metodo di analisi utilizzato dipendono fortemente dal tipo di segnale analizzato (voce, rumore, musica,…)
- Spesso useremo le stesse tecniche per segnali diversi ma con diversi periodi di analisi e diversi range per i risultati dell'analisi.
- Lavoreremo su frame di *N* campioni. In alcuni casi le misure sono calcolate nel dominio del tempo e non occorre moltiplicare i dati per una funzione finestra, in altri casi lavoreremo nel dominio della frequenza e applicheremo sempre finestratura e overlap.

Zero crossing-rate

- Metodo molto semplice (ma anche limitato) per estrarre l'informazione del pitch.
- Lavora bene in assenza di rumore.
- Conta il numero di attraversamenti dell'asse *y=0* in un certo intervallo di tempo:

$$\mathrm{ZCR}_i = \frac{1}{N} \sum_{n=0}^{N-1} \left| \mathrm{sign}\{x_i(n)\} - \mathrm{sign}\{x_i(n-1)\} \right| .$$

- Per una sinusoide il numero di attraversamenti è il doppio della frequenza.
- La tecnica è molto efficace in assenza di rumore ma produce risultati erratici in presenza di rumore.

Zero crossing-rate

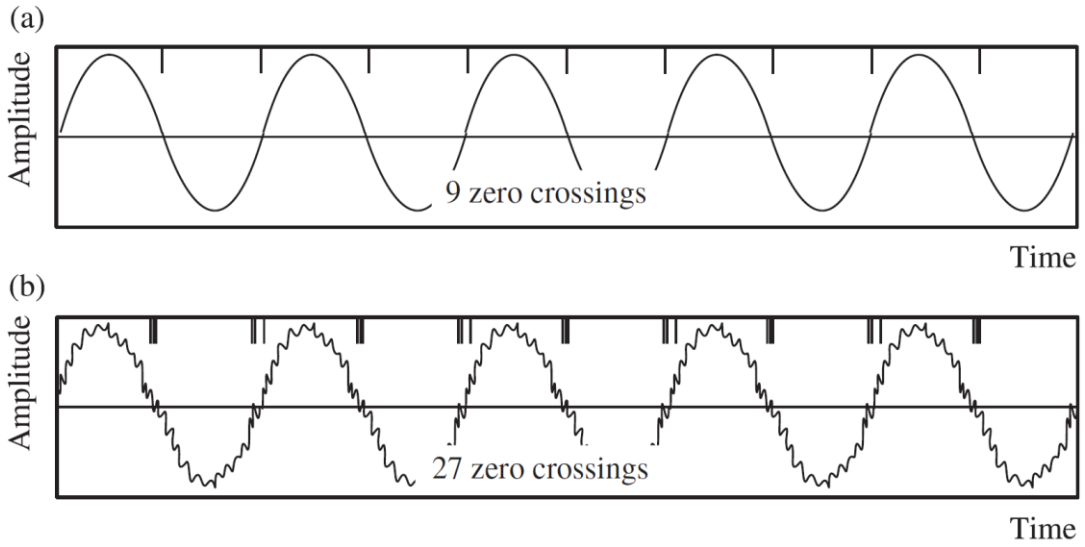


Figure 7.1 Zero-crossing rate calculation illustrated for a pure sinewave (a) and one corrupted with noise (b). Zero crossings are indicated by tick marks at the top of each plot. The noise-corrupted sinewave below exhibits many more spurious zero crossings due to additive noise taking the signal backwards and forwards over the trigger amplitude several times each period.

Zero crossing-rate

```
function [zcr]=zcr(segment)
zc=0;
for m=1:length(segment)-1
  if segment(m)*segment(m+1) > 0
    zc=zc+0;
  else
    zc=zc+1;
  end
end
zcr=zc/length(segment);
```

Zero crossing-rate

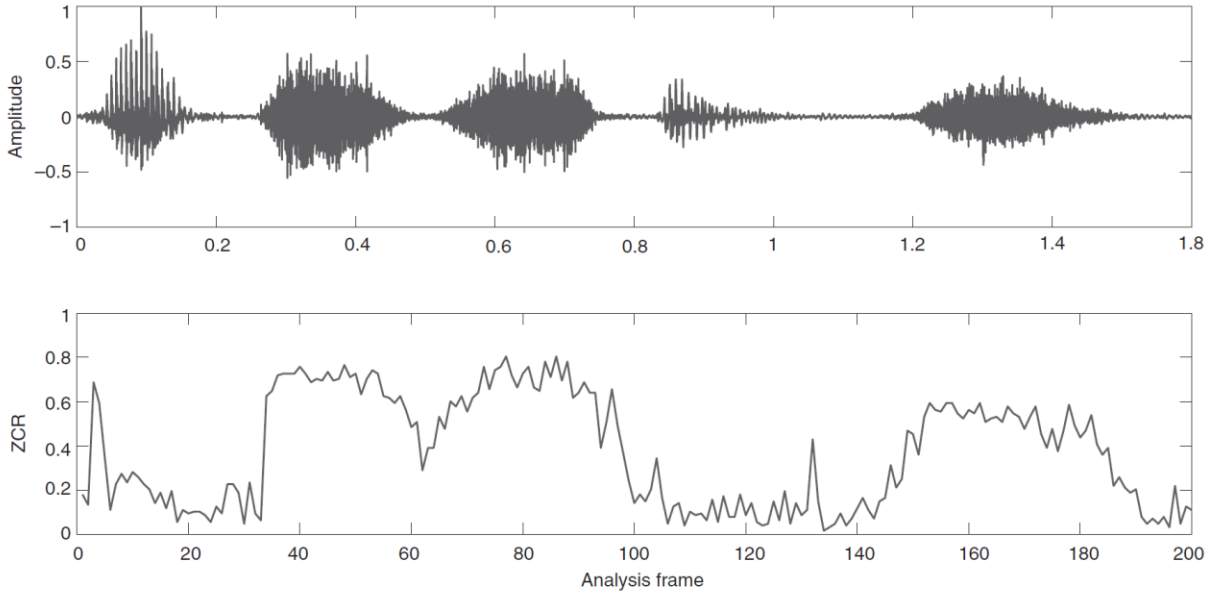


Figure 7.2 The upper graph shows a recording of a male speaker saying 'it's speech', which contains several stressed high-frequency sibilants and the lower graph is the frame-by-frame ZCR corresponding to this. The excursions of the ZCR plot closely match the high-frequency speech components, namely a breathy /i/ at the start, both /s/ sounds and the final /ch/.

Threshold crossing rate (TCR)

- Per contrastare l'effetto del rumore si considera spesso una regione di soglia attorno all'asse *y=0* contando gli attraversamenti pieni di questa zona di soglia.

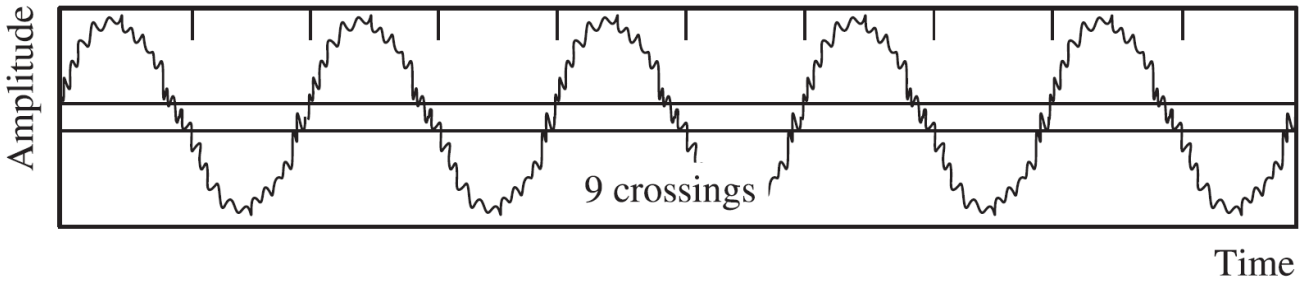


Figure 7.3 Threshold-crossing rate illustrated for a noisy sinewave, showing that the spurious zero crossings of Figure 7.1(b) are no longer present.

Zero crossing rate (ZCR)

- Risultati simili al threshold crossing rate possono anche essere ottenuti filtrando passabasso il segnale prima del calcolo dello zero crossing rate.
- La frequenza di taglio del filtro potrà essere determinata dal range atteso del pitch.
- Lo ZCR filtrato può essere usato per avere una indicazione approssimata del contenuto del segnale vocale, con il segnale non vocalizzato che fornisce valori di ZCR più elevati di quello vocalizzato.

Frame power

- E’ una misura dell’energia del segnale nel frame:

$$E_i = \frac{1}{N} \sum_{n=0}^{N-1} |x_i(n)|^2 .$$

- Fornisce una semplice rappresentazione della variazione del volume del segnale vocale.
- I suoni non vocalizzati sono emessi con minore potenza di quelli vocalizzati e per questo motivo può essere usata come indicatore di suoni vocalizzati o non.

```
[fpow]=fpow(segment)
  fpow=sum(segment.^2)/length(segment);
```

Average magnitude difference function (AMDF)

- Fornisce una indicazione simile alla potenza nel frame ma con minore complessità computazionale:

$$\mathrm{AMDF}_i = \frac{1}{N}\sum_{n=0}^{N-1} |x_i(n)| \,.$$

```
function [amdf]=amdf(segment)
  amdf=sum(abs(segment))/length(segment);
```

Average magnitude difference function (AMDF)

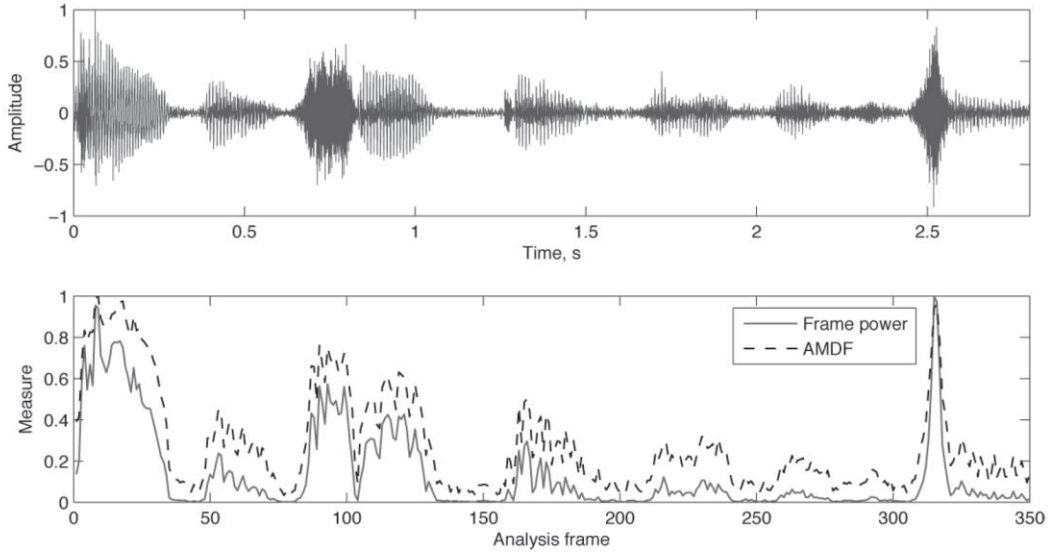


Figure 7.4 Average magnitude difference function (AMDF) and frame power plots (lower graph) for a recitation of the alphabet from A to G (plotted as a waveform on the upper graph). The duration of the letter C is enough for there to be almost two amplitude 'bumps' for the /c/ and the /ee/ sounds separated by a short gap, spanning the time period from approximately 0.7 to 1.2 seconds.

Spectral analysis

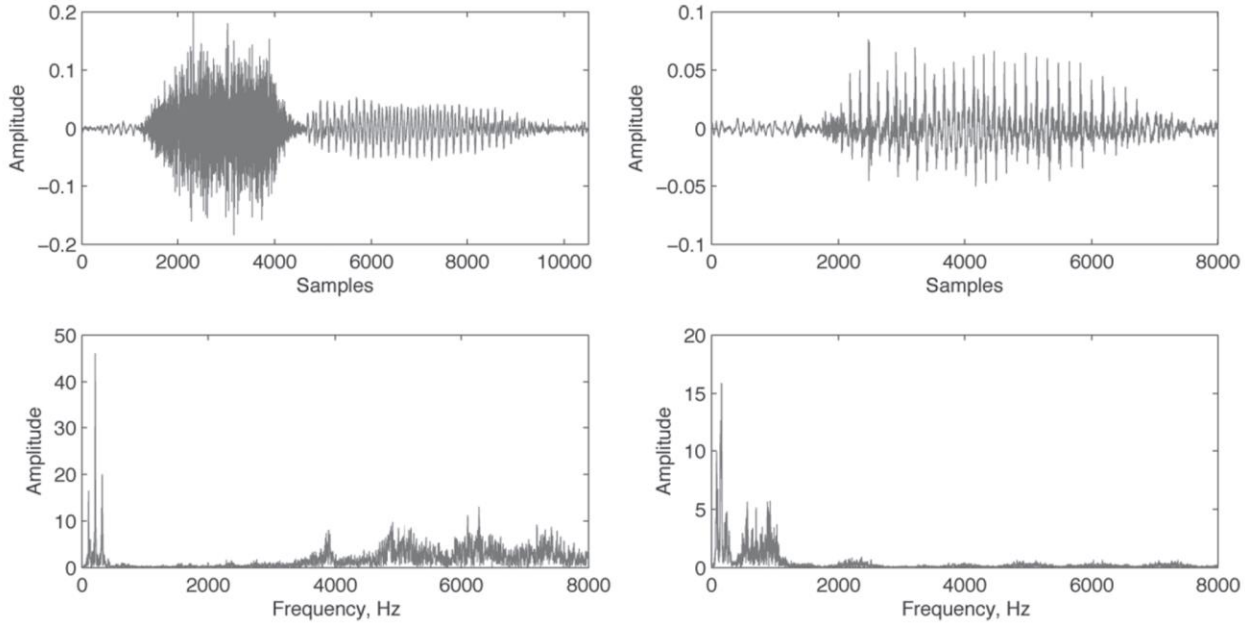


Figure 7.5 Plot of two 16 kHz sampled speech utterances of spoken letters C (left) and R (right), with time-domain waveform plots at the top and frequency-domain spectra plotted below them.

Cepstral analsysis

- L'utilità del cepstrum deriva dal fatto che è «la FFT inversa del logaritmo della FFT del segnale» e il logaritmo trasforma il prodotto nella somma dei logaritmi:

$$Y(e^{jw}) = H(e^{jw}) \cdot X(e^{jw})$$

$$\log[Y(e^{jw})] = \log[H(e^{jw})] + \log[X(e^{jw})]$$

- Supponiamo $X(e^{jw})$ sia la componente del pitch e $H(e^{jw})$ la componente legata al tratto vocale.
- Se nel dominio del tempo le due componenti sono legate da una convoluzione, nel dominio del cepstrum sono legate in modo additivo.
- In un plot del cepstrum la componente del pitch sarà separata da quella del tratto vocale.

Cepstral analsysis

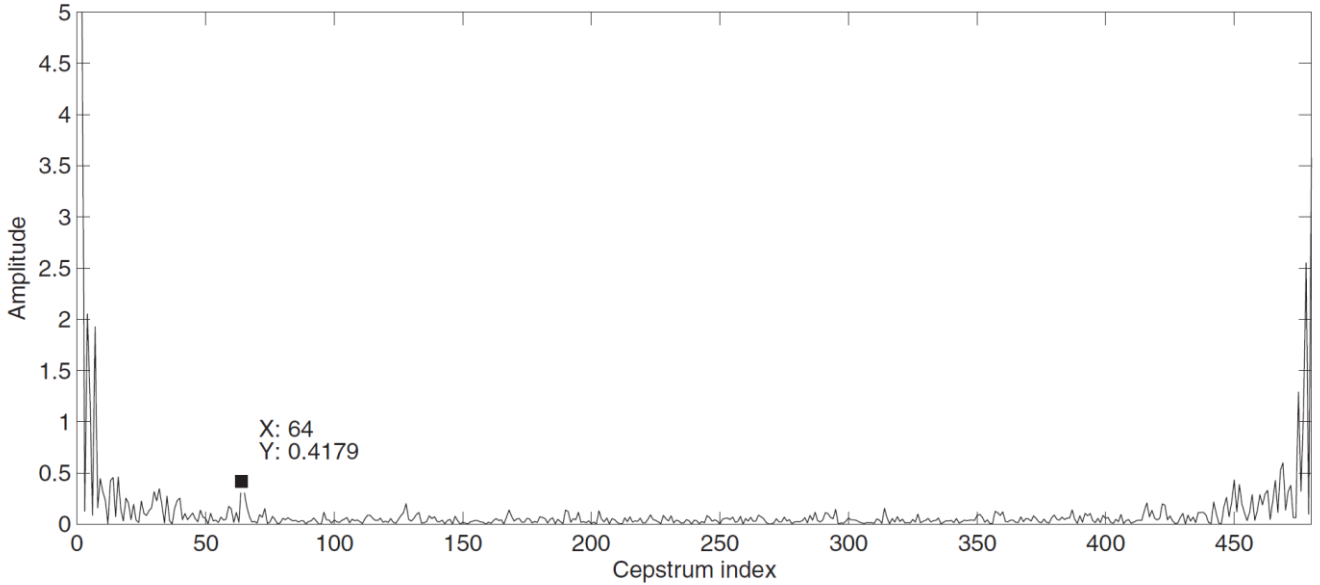


Figure 7.6 Cepstral plot of a segment of voiced speech, amplitude against cepstral index for a 480-sample analysis window. The likely pitch component has been selected, at index position 64.

Cepstral analsysis

```
len=length(segment);
%Take the cepstrum
ps=log(abs(fft(segment)));
cep=ifft(ps);
%Perform the filtering
cut=30;
cep2=zeros(1,len);
cep2(1:cut-1)=cep(1:cut-1)*2;
cep2(1)=cep(1);
cep2(cut)=cep(cut);
%Convert to frequency domain
env=real(fft(cep2));
act=real(fft(cep));
%Plot the result
pl1=20*log10(env(1:len/2));
pl2=20*log10(act(1:len/2));
span=[1:fs/len:fs/2];
plot(span,pl1,'k-.',span,pl2,'b');
xlabel('Frequency, Hz');
ylabel('Amplitude, dB');
```

Cepstral analsysis

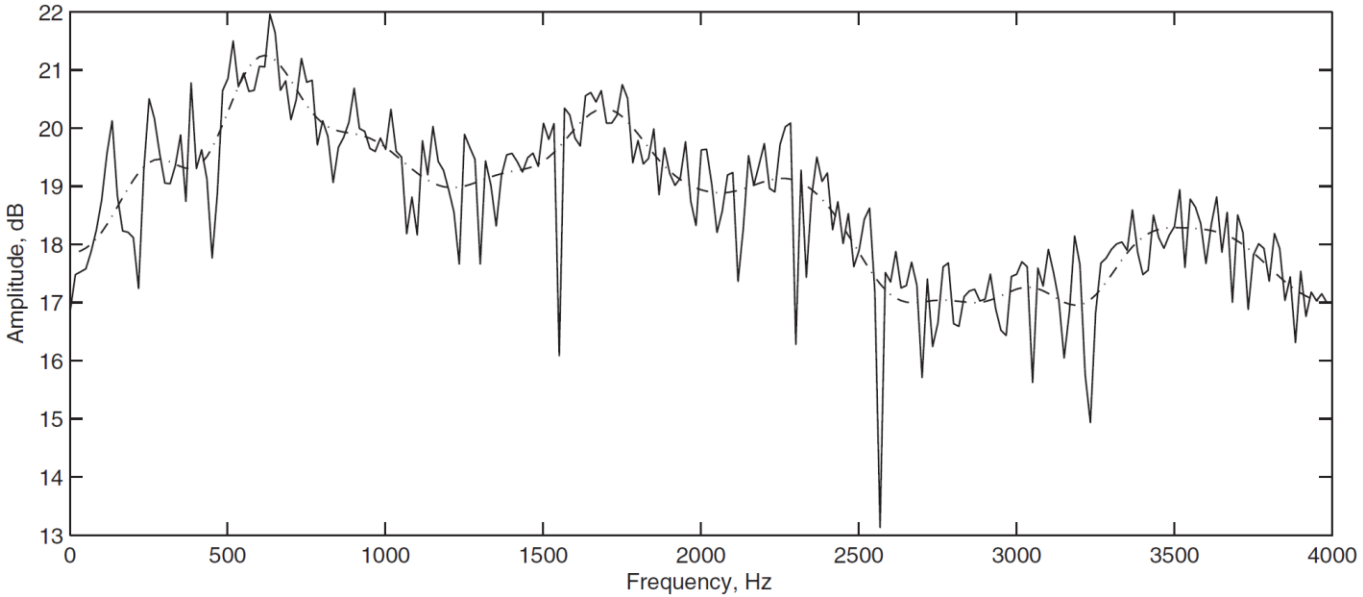


Figure 7.7 Frequency plot of a segment of voiced speech (solid line) overlaid with the frequency envelope obtained from the first few cepstral coefficients (dashed line).

LSP based measures

- L'analisi può essere anche effettuata usando misure basate su line spectral pairs.
- La tecnica viene normalmente applicata alla voce ma ha validità generale.
- Nel seguito assumeremo di aver calcolato le LSP ω_i per la finestra di analisi considerata.
- Lo *shift* indica il movimento spettrale dominante tra frame:

$$Shift[n] = \left\{ \sum_{i=1}^{p} \omega_i[n] \right\} - \left\{ \sum_{i=1}^{p} \omega_i[n+1] \right\}.$$

```
function [shift] = lsp_shift(w1,w2)
    shift=sum(w1) - sum(w2);
```

LSP based measures

- La media delle LSP è una utile misura del *Bias* in frequenza:

$$Bias[n] = \frac{1}{p}\sum_{i=1}^{p}\omega_i[n],$$

```
function [bias] = lsp_bias(w)
    bias=sum(w)/length(w);
```

LSP based measures

- E’ anche possibile specificare una posizione nominale per le LSP e calcolare la deviazione tra questo riferimento e le LSP nella finestra di analisi:

$$Dev[n] = \sum_{i=1}^{p} (\omega_i[n] - \bar{\omega}_i)^{\beta} .$$

```
function [dev] = lsp_dev(w,bar_w,b)
    dev=sum( (w-bar_w).^b);
```

- Spesso le $\bar{\omega}_i$ sono distribuite in modo uniforme nello spettro:

$$\bar{\omega}_i = i\pi/(p+1) \qquad \text{for} \quad i = 1, \ldots, p.$$

```
bar_w=[1:10]*pi/11
```

LSP based measures

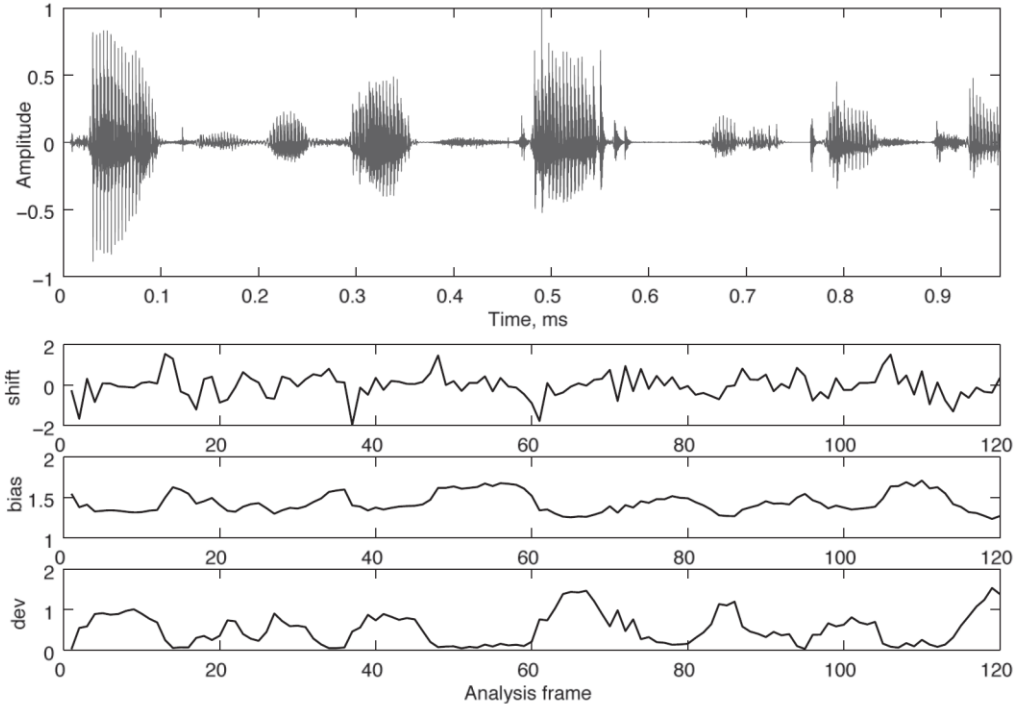


Figure 7.8 Deviation, bias and shift LSP analysis features collected for a 16 kHz sampled speech utterance, with waveform plotted on top, extracted from the TIMIT database.

LSP based measures

- L'evoluzione nel tempo delle LSP può essere usata per estrarre diverse informazioni.
- E' stata usata per la classificazione dei suoni (in 6 categorie: voiceless fricative e affricative, fricative, plosives, nasals, glides, silence).
- E' stata usata per il riconoscimento vocale
  - [Nei vecchi sistemi si usava template matching, statistical modeling, o multifeature vector quantization e il riconoscimento poteva essere effettuato con LSP. Oggi le soluzioni dominanti sono basate su approcci da big data o Hidden Markov Models-HMMs con Mel Freq. Cepstral C.- MFCC]
- L'analisi LSP è stata usata per il riconoscimento della lingua parlata (in quanto variano considerevolmente da lingua a lingua).

Speech analysis and classification

- L'analisi della voce è un importante requisito in diverse applicazioni e molte tecniche richiedono la *classificazione* della voce in diverse categorie.
- Le seguenti tecniche base forniscono un indicazione della varietà di applicazioni che richiedono l'analisi della voce e la sua classificazione:
    - detecting the presence of speech;
    - detecting voiced or unvoiced speech;
    - finding boundaries between phonemes or words;
    - classifying speech by phoneme type;
    - language detection;
    - speaker recognition;
    - speech recognition.

Speech analysis and classification

- La *classificazione* è un'area importante e in crescita della ricerca vocale. E' legata al *machine understanding,* alla comprensione da parte di una macchina.
- Per classificare in genere è necessario dapprima effettuare qualche forma di misura.
- Ad esempio, per rilevare se la voce è voiced o unvoiced potremmo misurare la potenza ed il pitch, e magari esaminare anche i dati LSP.
- Qualunque nuova tecnica di classificazione farà in genere uso di diverse tecniche base di misura/analisi.
- Esistono dei metodi di analisi che sono stati usati quasi esclusivamente per l'analisi della voce, in particolare i metodi di analisi del pitch.

Pitch analysis

- Abbiamo già visto il metodo basato sulla long-term prediction.
- Ci sono molti altri metodi alternativi. Quelli più accurati richiedono qualche forma di coinvolgimento dell'uomo, e.g., misure basare su accelerometri o elettroglottografia (EGG) su gola o glottide per estrarre direttamente il pitch.
- Dalla sola voce è impossibile identificare il pitch con un'accuratezza del 100%, ma molti algoritmi forniscono stime che sono comunque accurate.
- Le tecniche di maggiore successo impiegano:
  - L'analisi zero-crossing nel dominio del tempo.
  - L'autocorrelazione nel dominio del tempo.
  - L'analisi cepstrale nel dominio della frequenza.
  - Metodi basati sull'average magnitude difference function.
  - Metodi basati su analisi LPC o LSP.
  - Time-frequency domain analysis.

Audio analysis example

```
[speech,fs]=audioread('SA1_converted.wav');
```

```
spec=spectrogram(speech,1024,1024-128,1024);
```

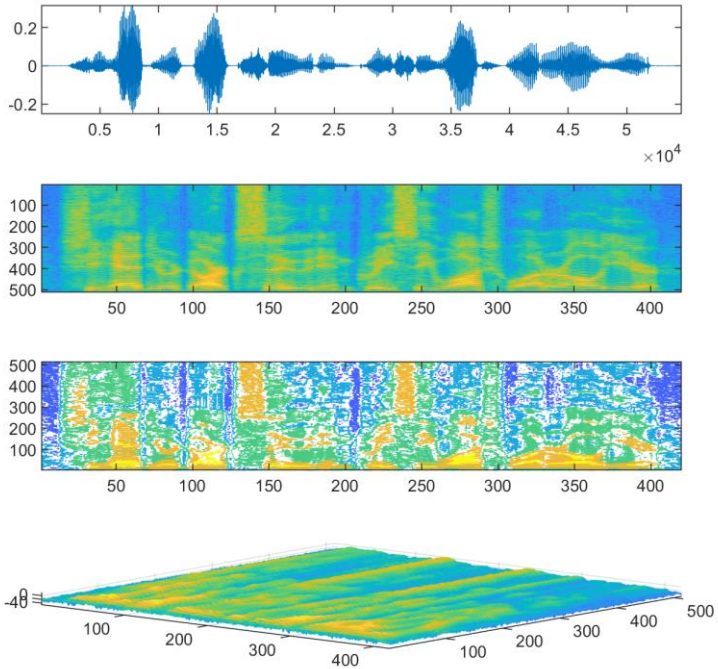
Analizza frame da 1024 campioni, con overlap di 1024-128, e fft su 1024 campioni

```
>> size(speech)
ans =
       49544           1
>> size(spec)
ans =
         513         380
>>
```

Audio analysis example

```
subplot(4,1,1)
plot(speech)
subplot(4,1,2)
imagesc(flipud(10*log10(abs(spec))))
subplot(4,1,3)
contour(10*log10(abs(spec)),6)
subplot(4,1,4)
mesh(10*log10(abs(spec)))
axis tight % make plot completely fill the X-Y axes
view(40,80) % change viewpoint of 3D plot
```

Audio analysis example



Statistics and classification

- La *statistica di un segnale* è semplicemente una raccolta di informazioni sugli eventi correnti e passati, più «qualche rudimentale analisi matematica».
- Viene spesso applicata ai dati di analisi della voce o dell'audio.
- Inizia confrontando lo *score* dell'analisi nella finestra corrente con quello medio (analisi del primo ordine); procede con il calcolo della deviazione standard dello score d'analisi (analisi del secondo ordine) e in molti casi l'analisi continua con il terzo ordine o ordini superiori.

Statistics and classification

- Nessun sistema di classificazione dà una accuratezza del 100%.
- Uno degli aspetti più importanti da considerare è cosa accade quando un classificatore sbaglia.
- In una classificazione o rilevazione binaria, si considerano quattro parametri relativi all'accuratezza:
  - *True-positive classification accuracy*
    - La proporzione di rilevazioni positive classificate correttamente.
  - *True-negative classification accuracy*
    - La proporzione di rilevazioni negative classificate correttamente.
  - *False-positive classification accuracy*
    - La prop. di rilevazioni negative classificate erroneamente come positive.
  - *False-negative classification accuracy*
    - La prop. di rilevazioni positive classificate erroneamente come negative.

Vedere:

- Ian Vince McLoughlin, “Speech and Audio Processing”- Cambridge Univesity Press (2016)
  - Cap. 7