Sergio Canazza
Lab. AVIRES – Università di Udine, http://avires.dimi.uniud.it

Metodologie di restauro audio

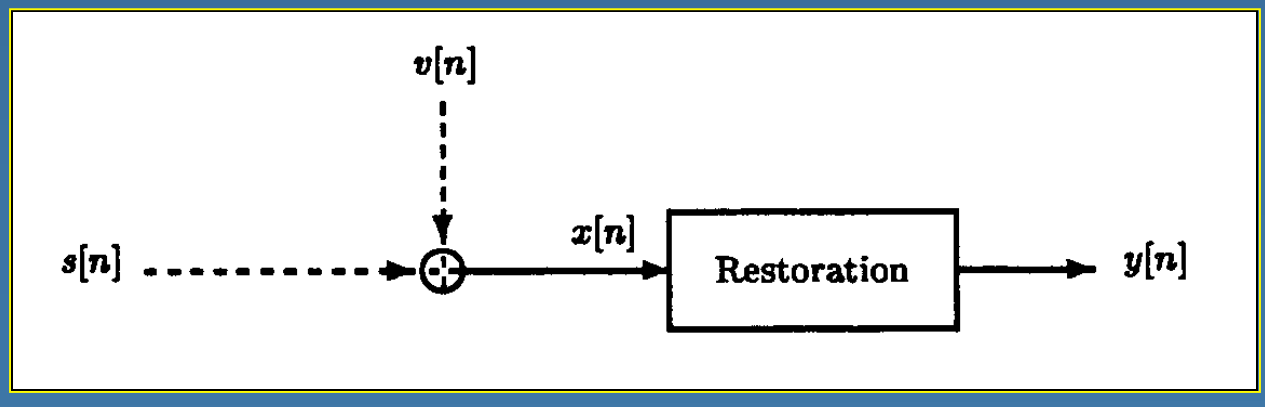
- *rumore*?
  - Diverse esigenze:
    - Tempo reale
    - Off-line
  - Diversi approcci

Restauro

- Diverse metodologie implicano risultati diversi.
- Necessità di operare scelte consapevoli:
  - Metodi in frequenza: (poca) informazione *a priori* (= impronta di rumore) + (molta) informazione *a posteriori*.
  - Restauro per modelli del segnale → necessità di informazione *a priori* per stimare la distribuzione di probabilità degli eventi → utilizzabile in segnali "semplici" (quasi-periodici).
    - Informazione *a priori* (segnale di eccitazione e coefficienti del filtro) + informazione *a posteriori* (*tracking* del segnale).
    - Modelli generalizzabili a diverse tipologie di segnali sono "non-informativi" (poca informazione *a priori*).
  - Restauro per modelli della sorgente: informazione *a priori*.

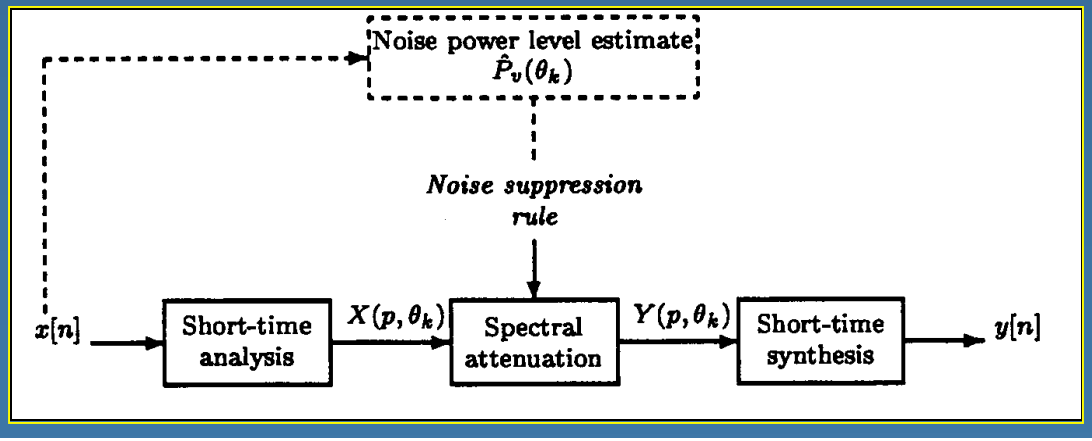
Sottrazione spettrale - ipotesi

- Attenuazione Spettrale a Breve Termine (STSA):
  - applica un'attenuazione tempo-variante allo spettro a breve termine del segnale deteriorato
  - non richiede la definizione di un modello del segnale
- Ipotesi:
  - rumore additivo e stazionario
  - incorrelato ad s
  - è stimabile la sua densità spettrale di potenza



Sottrazione spettrale - algoritmo

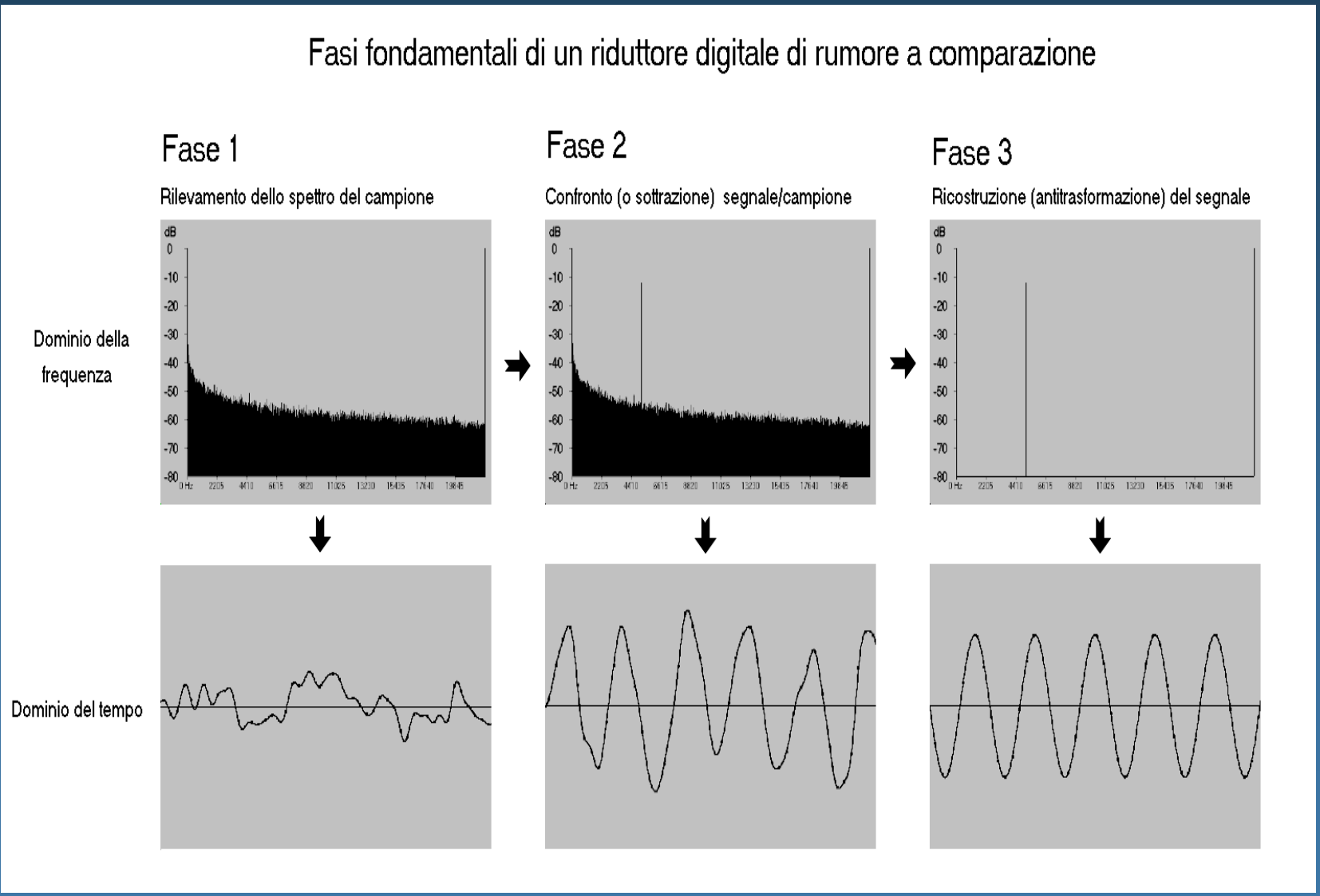
- STFT del segnale
- ogni frequenza viene attenuata con un guadagno positivo e minore di 1 (spectral attenuation)
- il guadagno tempo-variante applicato ad ogni canale viene determinato da una *noise suppression rule*
  - realizza una stima, per ogni frequenza, della potenza di rumore
- viene risintetizzato il segnale



Sottrazione spettrale - considerazioni

- Problema aperto: si processa solo il modulo
  - L'orecchio *risulterebbe* insensibile alla fase (è provato solo per i segnali stazionari)
- Nata negli anni 70 per la trasmissione del parlato
- Grande diffusione:
  - non si fanno ipotesi sul segnale (approccio non-parametrico)
  - intuitivo (banco di filtri, equalizzatore a bande)

Sottrazione spettrale



Sottrazione spettrale - regola di soppressione

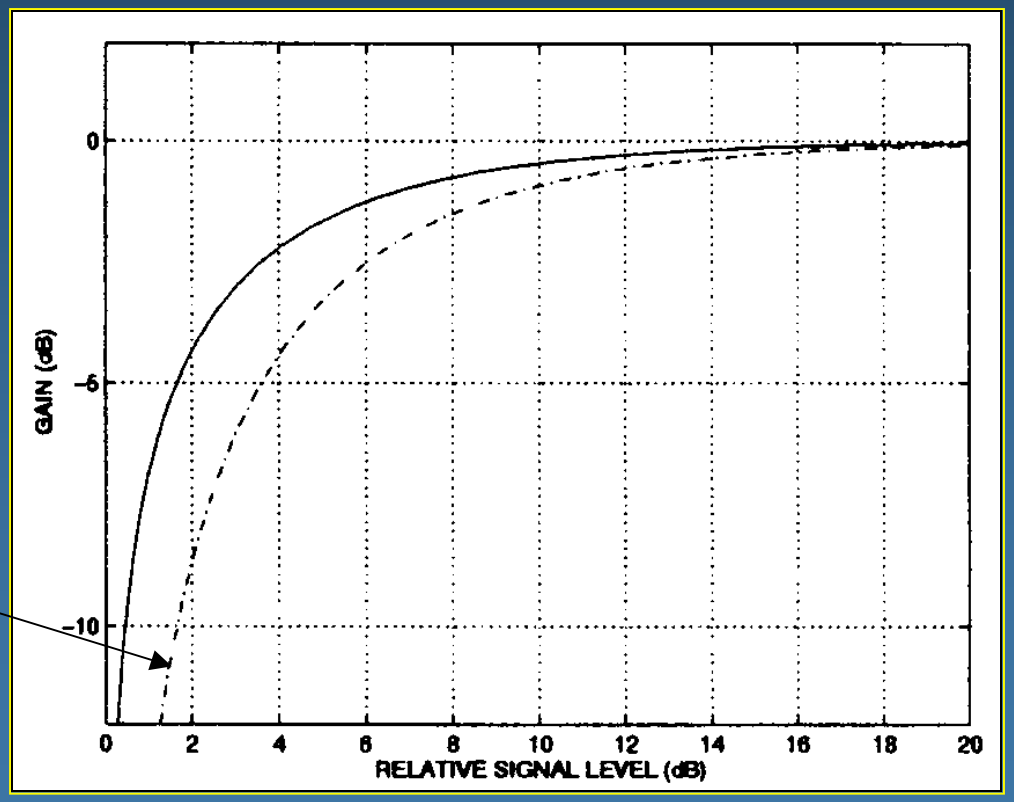
- $X(p,\theta_k)$ è la STFT del segnale rumoroso $x(n)$, dove p è l'indice temporale e θ_k l'indice frequenziale
- $G(p,\theta_k)$ è il guadagno
  - corrisponde ad una attenuazione del segnale (quindi limitato tra 0 e 1)
- dipende da $|X(p,\theta_k)|^2$
  - **misura** del livello della potenza del segnale rumoroso
  - **stima** della potenza del rumore alla frequenza θ_k, $\hat{P}_v(\theta_k) = E\{|V(p,\theta_k)|^2\}$
- si definisce il segnale relativo (>1):

$$Q(p,\theta_k) = \frac{|X(p,\theta_k)|^2}{\hat{P}_v(\theta_k)}$$

$$E\{Q(p,\theta_k)\} = 1 + \frac{E\{|S(p,\theta_k)|^2\}}{\hat{P}_v(\theta_k)}$$

Sottrazione spettrale - filtro di Wiener

$$G(p,\theta_k) = \frac{|X(p,\theta_k)|^2 - \hat{P}_v(\theta_k)}{|X(p,\theta_k)|^2}$$



Sottrazione spettrale - eliminazione di componenti utili del segnale

Nel caso di segnale sinusoidale:

$$E\{Q(p,\theta)\} = 1 + \frac{P_s}{V(\theta) W_\theta}$$

- P_s è la potenza della sinusoide (segnale non rumoroso)
- $V(\theta)$ la densità spettrale di potenza del rumore alla frequenza θ
- W_θ è la larghezza di banda della finestra utilizzata, centrata attorno alla frequenza θ
- il livello delle componenti di segnale che sono erroneamente cancellate dal processo di restauro aumentano in relazione alla larghezza di banda della finestra. Questa è inversamente proporzionale alla durata temporale della finestra
- Si dimostra che finestre inferiori ai 40 ms causano la soppressione di componenti udibili del segnale

Sottrazione spettrale - difetti

- Componenti di rumore non soppresse. Il segnale processato può presentare del rumore filtrato localizzato attorno alle componenti di s(n)
  - il rumore è fortemente correlato al segnale -> effetti di modulazione
  - considerando gli effetti di mascheramento si dimostra l'opportunità di avere delle finestre della STFT > 30÷40 ms
- Il rumore musicale. L'attenuazione è una quantità casuale (funzione del segnale relativo, correlato allo spettro del rumore da una varianza molto alta)
  - forte mancanza di correlazione tra i valori di frequenza corrispondenti tra successivi frame del segnale relativo
  - anche in presenza, al tempo p ed alla frequenza k, di valori non trascurabili della stima di Q (es. E[Q]=8dB) il valore reale potrebbe essere vicino allo 0 dB (ovvero presenza di solo rumore)
  - non è quindi possibile separare il rumore dalle componenti di segnale di ampiezza modesta.

Sottrazione spettrale - correzione del rumore musicale

- Sovrastima del livello di rumore
  - cancellazione di componenti di segnale utile
  - la sovrastima necessaria per ridurre le componenti del rumore musicale a livelli impercettibili (sotto lo 0.1%) può superare i 9 dB
- mascherare il rumore musicale tramite rumore a larga banda.
  - basta assume che il rumore da eliminare sia minore di quello effettivamente presente, attribuendo al guadagno valori superiori ad una soglia prefissata (noise floor)
- Regola di soppressione di Ephraim e Malah (EMSR). Il guadagno dipendente da due diverse stime di Q
  - SNR a posteriori, corrispondente alla consueta stima di Q. Usata quando Q >> 0 dB
  - SNR a priori, calcolata su più frame temporali (varianza minore). Usata quando Q ≈ 0 dB

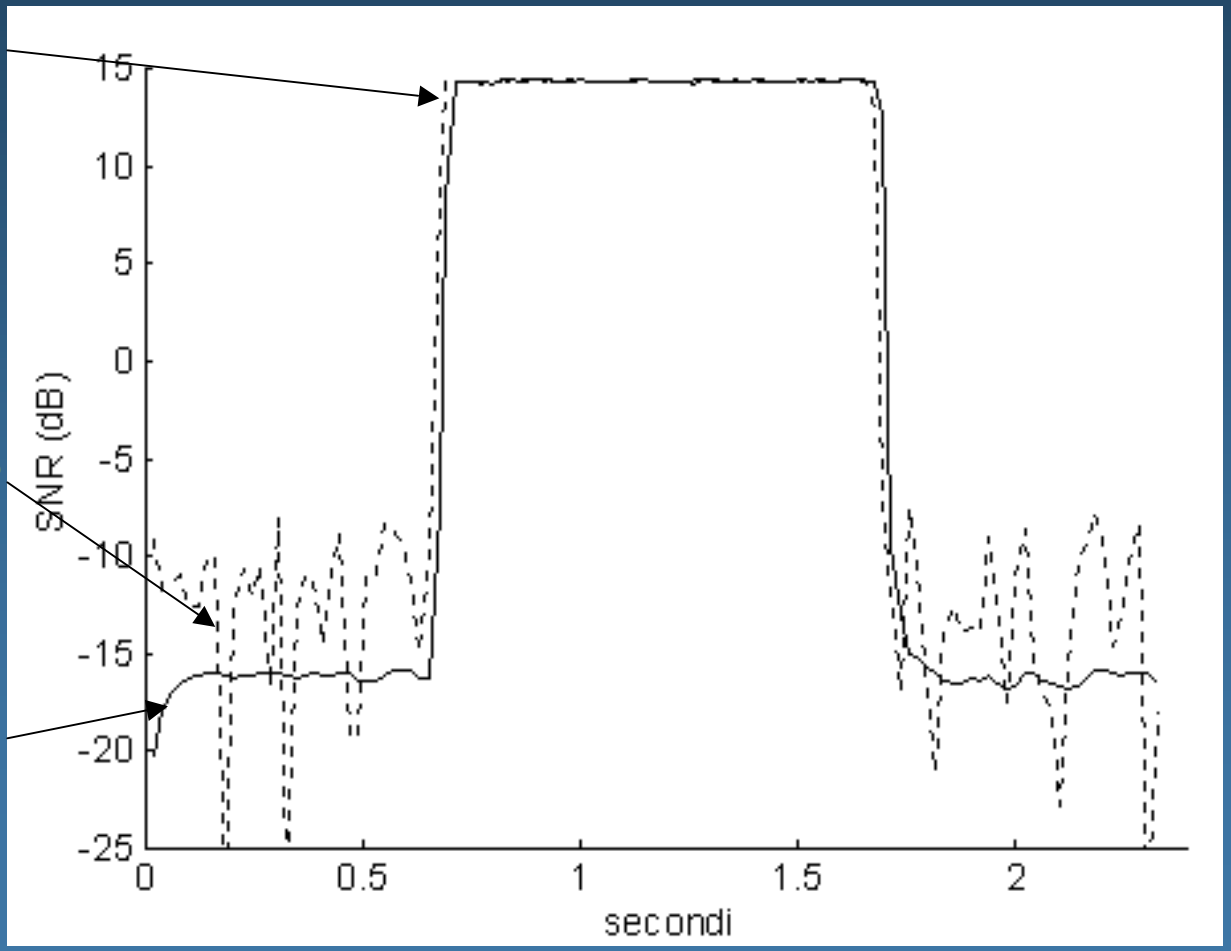
$$G = \frac{\sqrt{\pi}}{2}\sqrt{\left(\frac{1}{1+R_{post}}\right)\left(\frac{R_{prio}}{1+R_{prio}}\right)} * M\left[(1+R_{post})\left(\frac{R_{prio}}{1+R_{prio}}\right)\right]$$

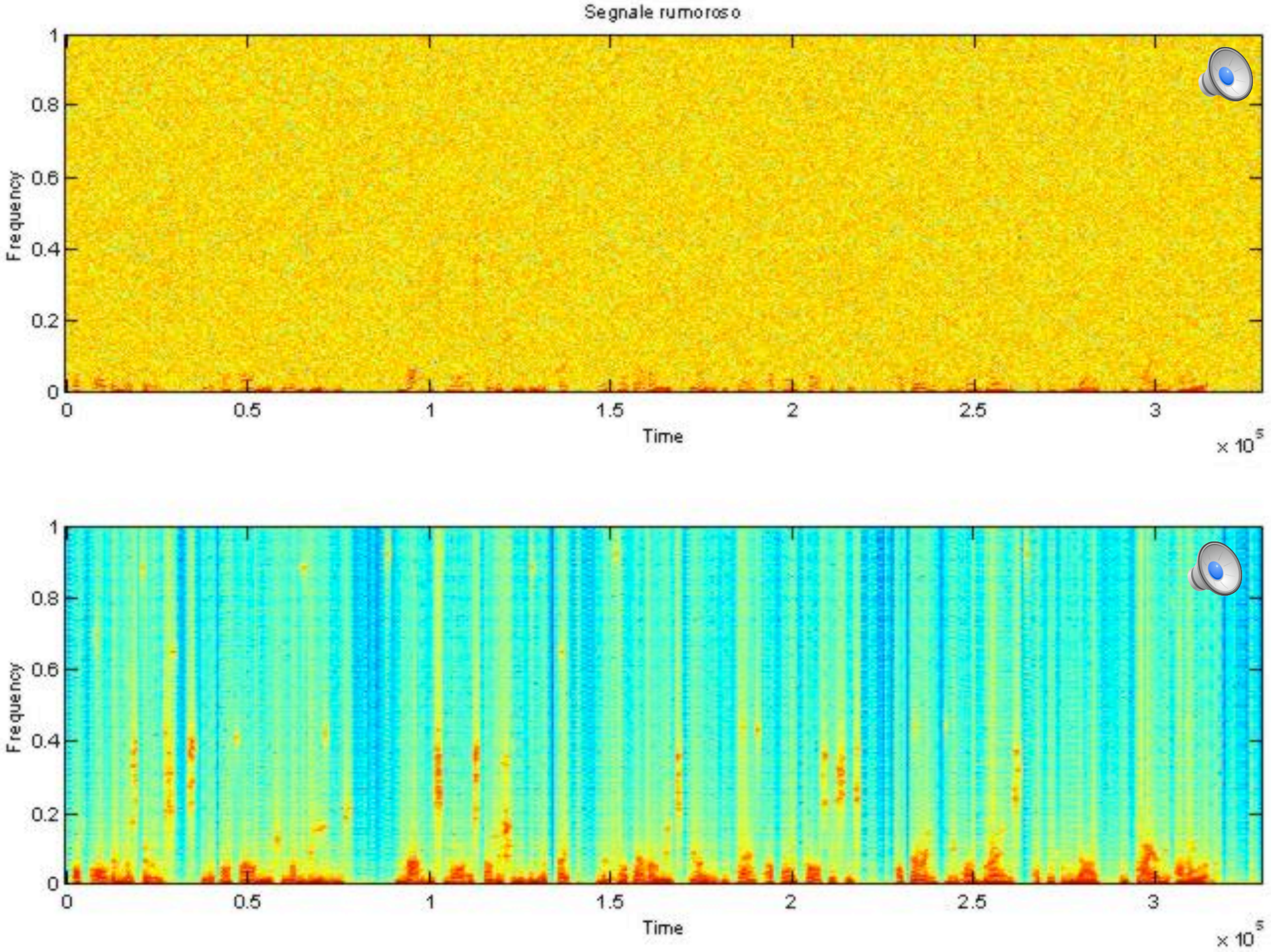
Il *rumore musicale*

comportamento ‘corretto’ (in presenza di segnale) grazie a Rpost

comportamento ‘a picchi’ di Wiener

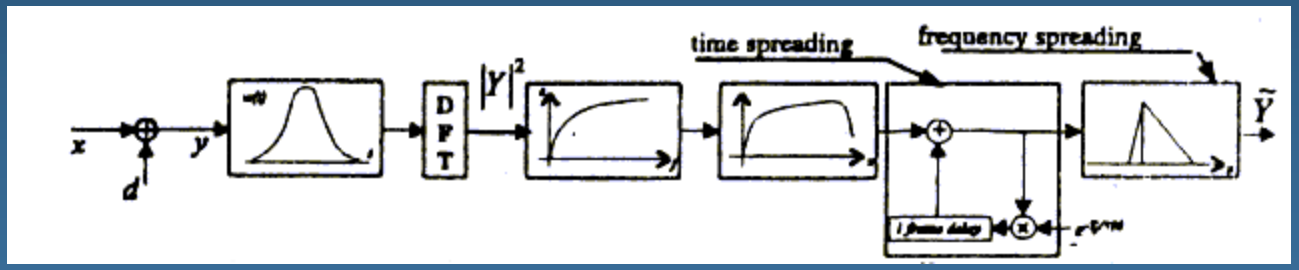
comportamento ‘smussato’ grazie ad Rprio



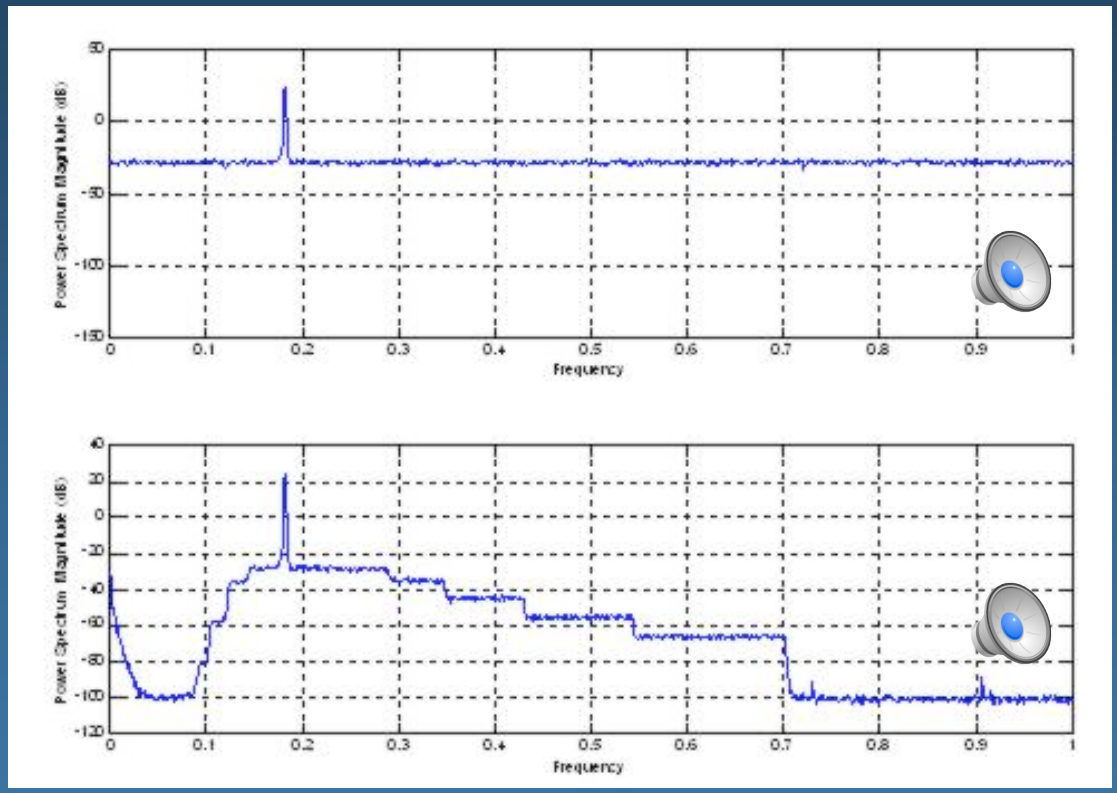

Segnale rumoroso
Frequency
Time
0
0.2
0.4
0.6
0.8
1
0.5
1.5
2
2.5
3
× 10^5

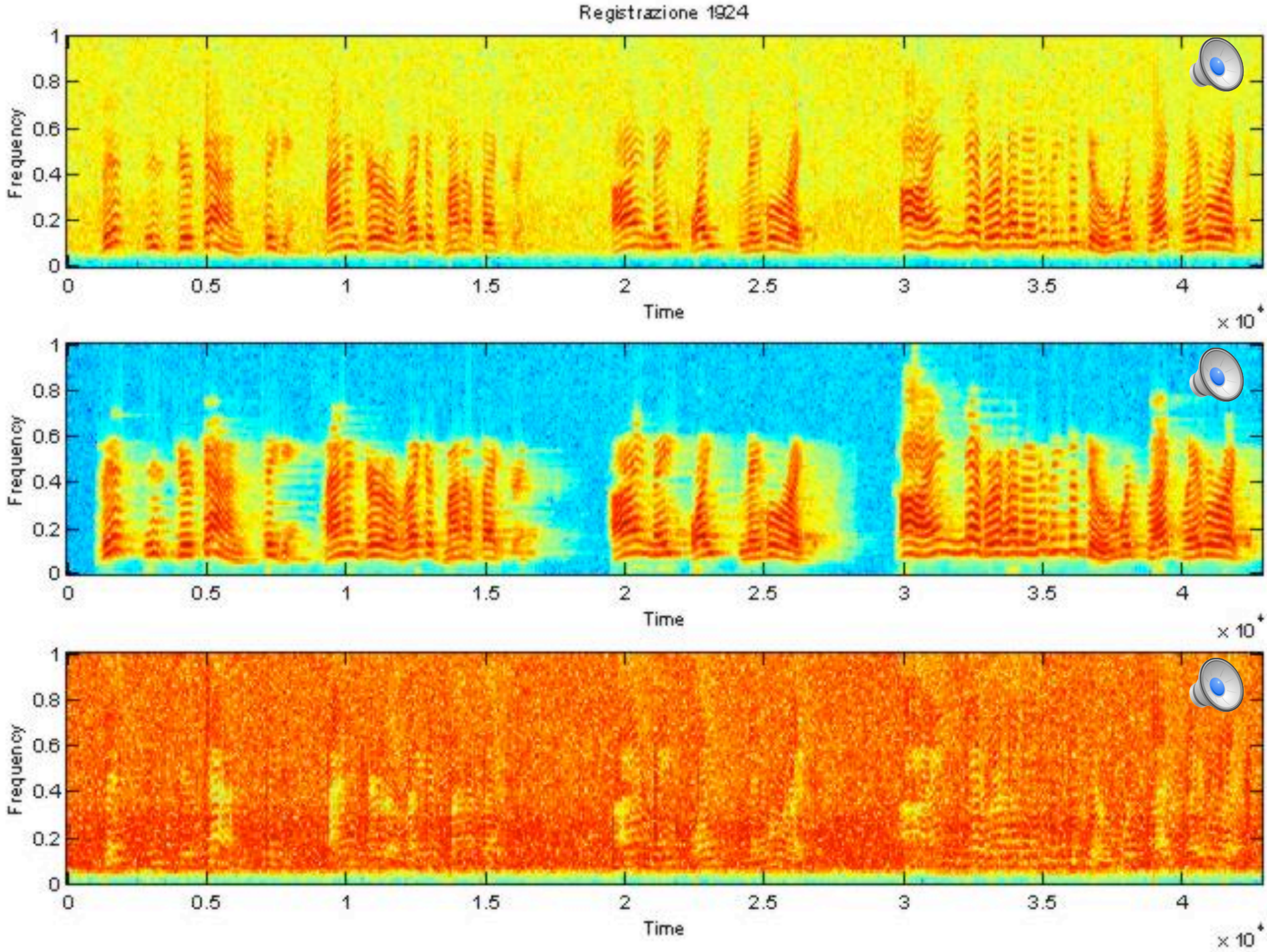
Filtro percettivo

- È basato sul fenomeno del mascheramento
- Alcune componenti rumorose percettivamente inudubili non vengono cancellate
- Poiché l'ammontare del rumore rimosso è minore, viene introdotto un numero minore di artefatti



- finestratura nel tempo e DFT con calcolo della potenza del segnale rumoroso;
- passaggio dalla scala degli Hertz a quella dei bark, tramite il calcolo dell'eccitazione del segnale relativo a ciascuna banda critica;
- *Outer to inner ear transformation*;
- mascheramento nel tempo (*time spreading*: operazione con memoria del frame precedente;
- mascheramento in frequenza (*frequency spreading*).


Power Spectrum Magnitude (dB)
Frequency
Power Spectrum Magnitude (dB)
Frequency


Registrazione 1924
Frequency
Time
0
0.2
0.4
0.6
0.8
1
0.5
1.5
2
2.5
3
3.5
4
× 10^4

Restauro per modelli (del segnale)

- Modello del segnale: $x(t) = f(\boldsymbol{a}, e(t))$

  il vettore $\boldsymbol{a}$ rappresenta i parameteri del modello, $e(t)$ è rumore, o 'eccitazione', termine che considera elementi inarmonici del segnale, e $f(\ )$ è una funzione che mappa i parametri e l'eccitazione nei valori del segnale audio.

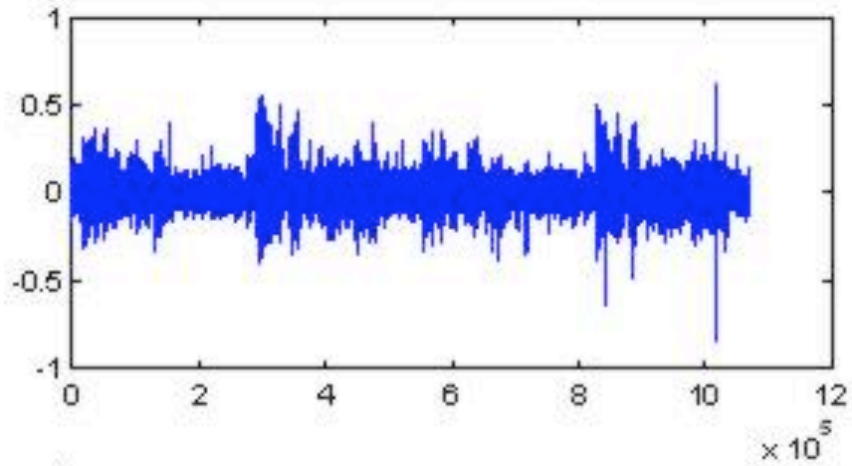
- Versione corrotta del segnale audio: $y(t) = g(x(t); v(t))$

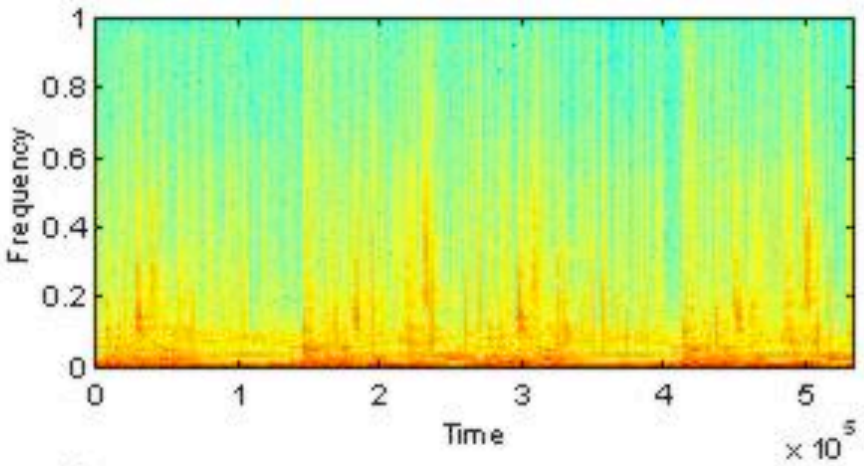
  $g(\ )$ è una funzione che descrive come i valori del segnale sono mappati nei valori corrotti e $v(t)$ è rumore che esprime qualsiasi randomness presente in questa mappa.

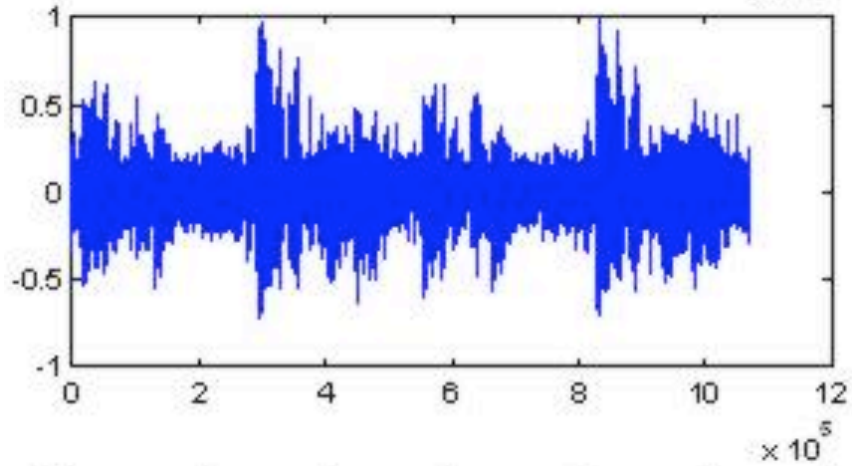
- Date le osservazioni corrotte $y(t)$, è possibile stimare i valori sconosciuti di $x(t)$ e $\boldsymbol{a}$?

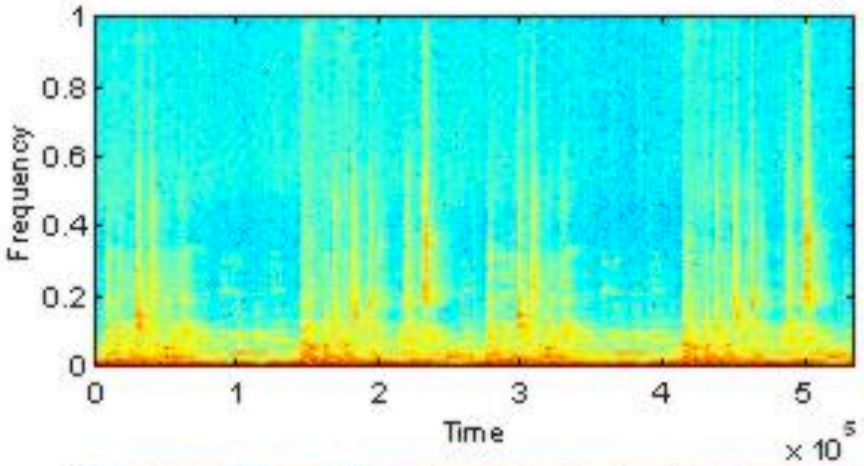
  Questo è un classico problema di stima, la cui soluzione adottata dipende dalla conoscenza a priori ipotizzata (modello del segnale)

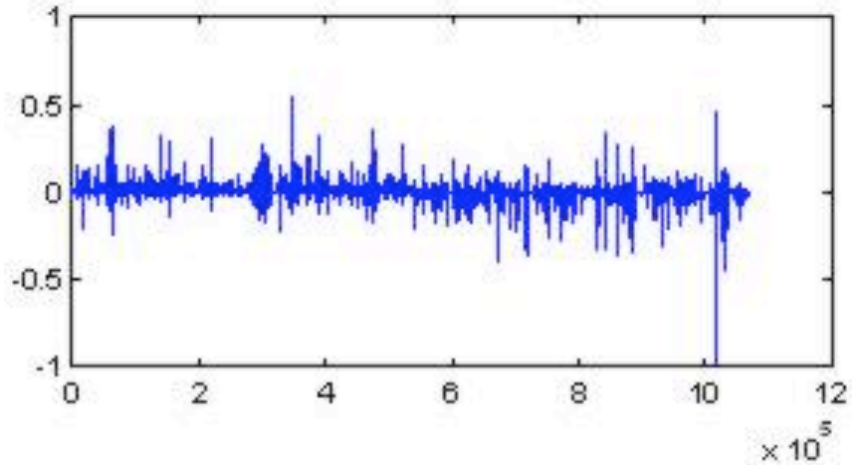
  - Extended Kalman Filter
  - Bayesian approach

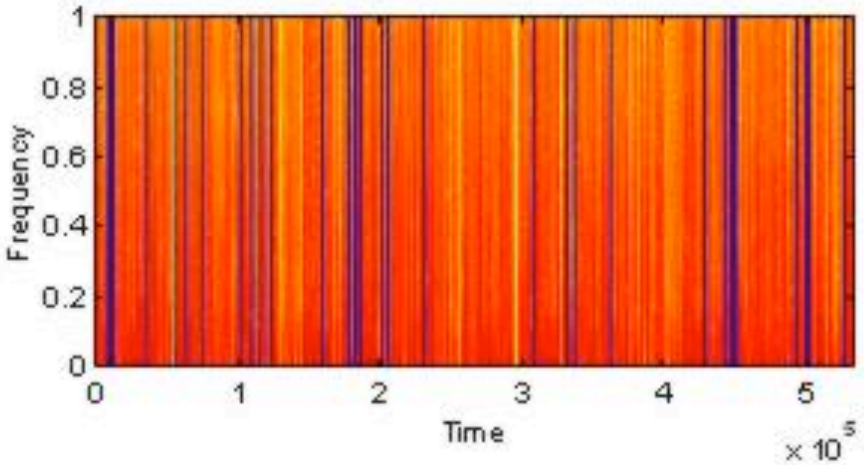

1
0.5
0
-0.5
-1
0
2
4
6
8
10
12
x 10^5


Frequency
Time
x 10^5


1
0.5
0
-0.5
-1
0
2
4
6
8
10
12
x 10^5


Frequency
Time
x 10^5

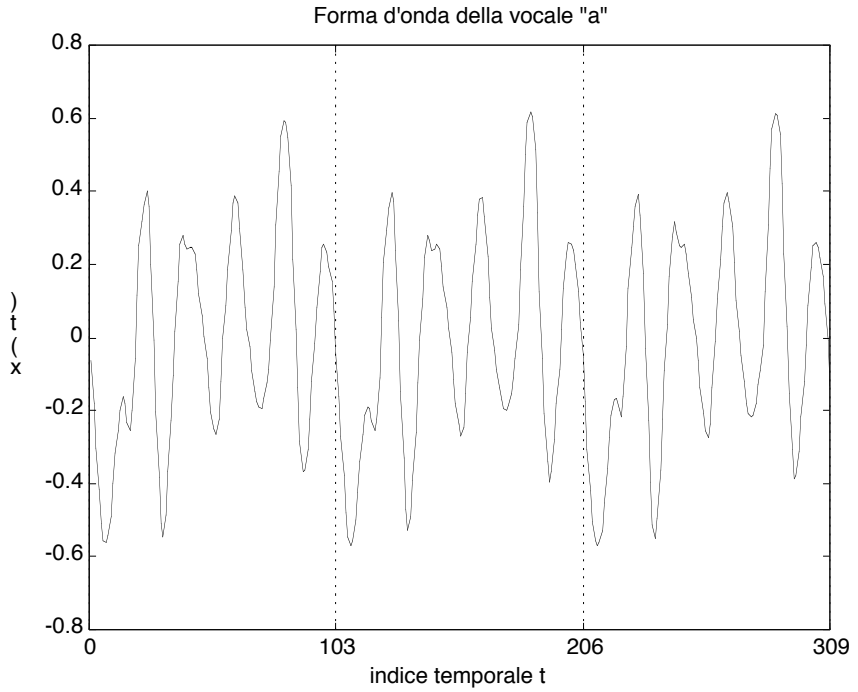

1
0.5
0
-0.5
-1
0
2
4
6
8
10
12
x 10^5

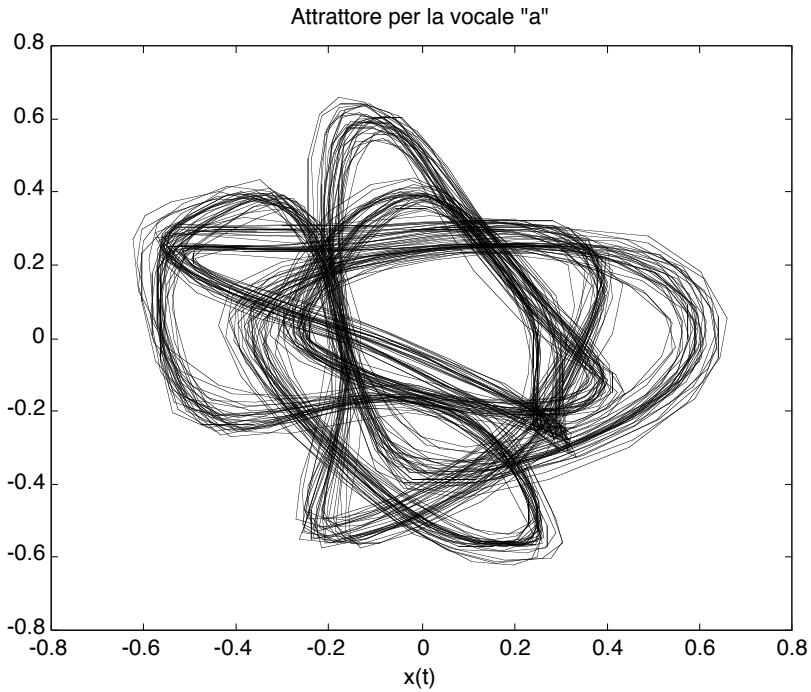

Frequency
Time
x 10^5

Restauro mediante proiezione locale

- Voce → fenomeno complesso e non stazionario
- Vocali → periodicità locale

Forma d'onda del segnale rappresentante la vocale "a" e attrattore per la vocale "a", frequenza di campionamento Fc=16kHz. L'intervallo di osservazione tra l'ascissa e l'ordinata è stato tenuto pari a 12 campioni, equivalente a un intervallo temporale di Δt=0.75ms.





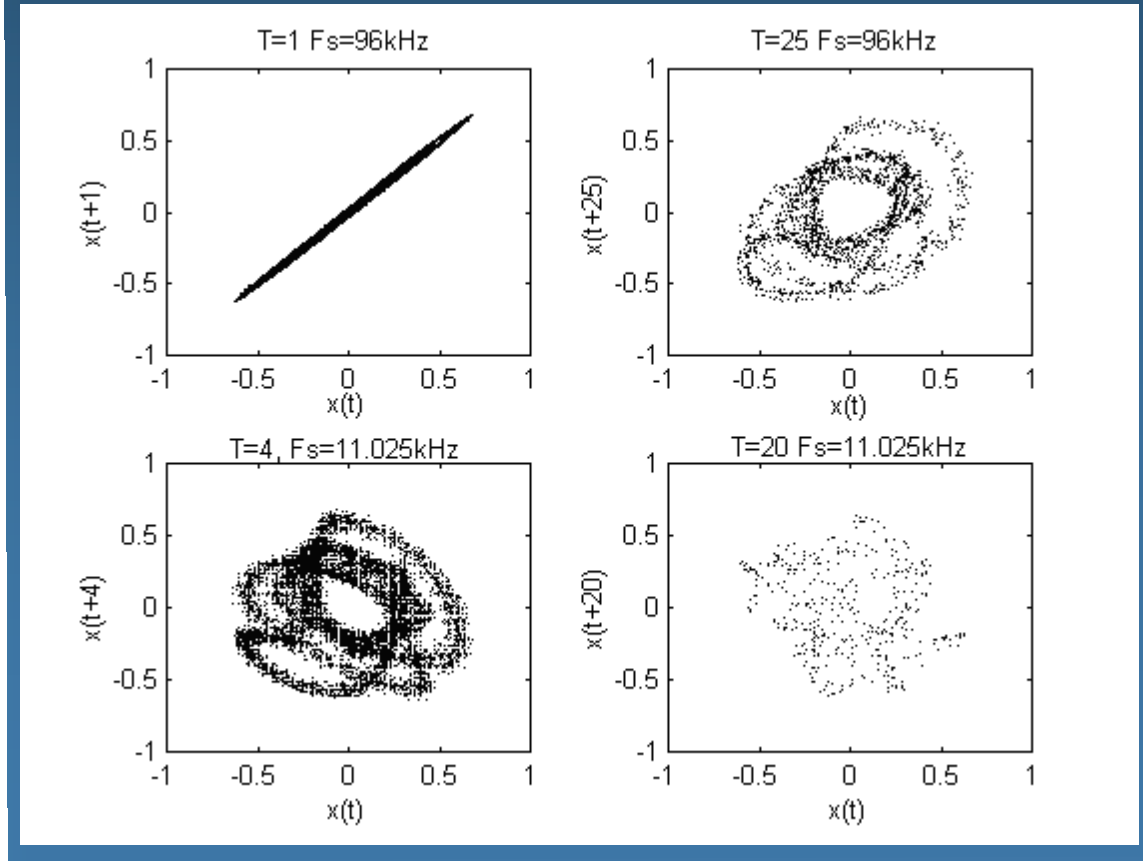
Restauro mediante proiezione locale

- Operando a livello locale sul segnale è possibile evitare il problema della non stazionarietà e utilizzare così strumenti di riduzione del rumore validi in contesti di caos deterministico, come ad esempio la tecnica di riduzione del rumore mediante proiezione locale.
- La ridondanza è data da forme d'onda simili all'interno del segnale stesso

- Dalla serie temporale del segnale rumoroso y_t, si formano i vettori di ritardo:
  - $s_t = \{y_t, y_{t-\tau}, \ldots, y_{t-(m-1)\tau}\}$
- Per ogni vettore s_t si considera l'insieme di vicinanza composto dai vettori simili
  - $U_t = \{s_k \text{ tali che } |s_t - s_k| < \varepsilon\}$
- Si calcola il vettore media tra i vettori simili e per ognuno di essi si calcola il vettore scarto dalla media z_k e la matrice di covarianza
  - $C_{u_{ij}} = \sum_{(n \in Ut)} (z_n)_i \cdot (z_n)_j$
- Si determinano gli autovalori e gli autovettori associati
- Si proietta il vettore s_t sugli autovettori relativi agli autovalori minori, ottenendo così una stima del rumore sovrapposto
- Si ricava una stima della componente utile (non rumorosa) del vettore effettuando una semplice sottrazione della componente rumorosa stimata
- Si effettua un'operazione di media tra le m diverse correzioni che si ricavano per ogni componente di s_t

Restauro mediante proiezione locale

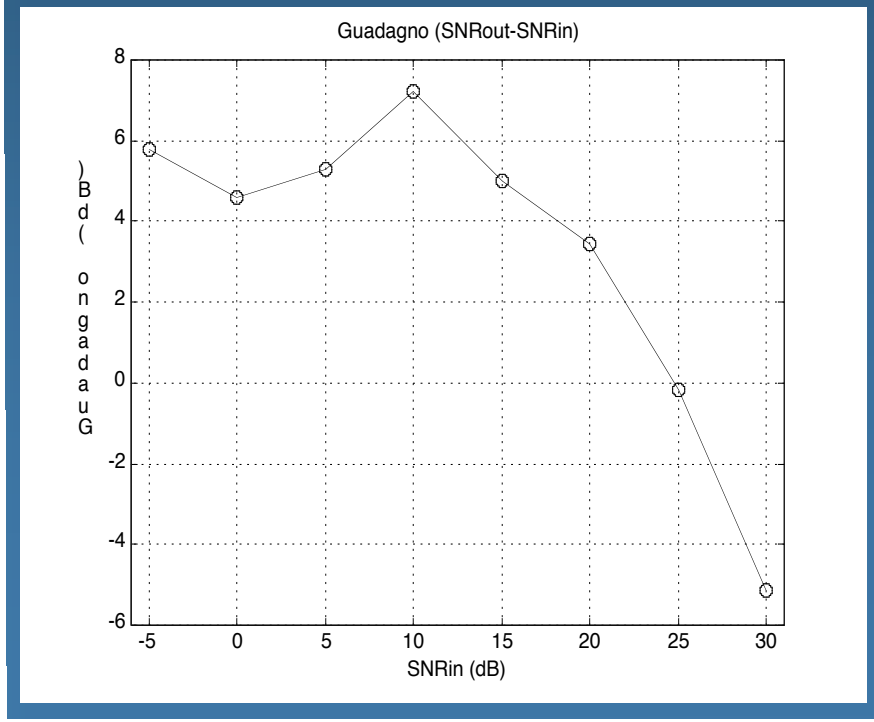
- Necessita di molte informazioni: raggio di vicinanza entro cui due forme d'onda vengono ritenute simili (ε), la dimensione m e il periodo di campionamento τ del vettore di ritardo, numero di valori singolari da considerare per la proiezione



Attrattori per la vocale "a" ricavati con diverse frequenze di campionamento e diversi valori del parametro τ. Se τ è troppo piccolo (nel riquadro in alto a sinistra $\Delta T \cong 10.4\mu s$) l'attrattore risulta schiacciato lungo la diagonale, mentre se è troppo grande (nel riquadro in basso a destra $\Delta T \cong 1.8ms$) si mettono in relazione dati incorrelati, creando strutture inconsistenti

Restauro mediante proiezione locale

- Problema: la periodicità è solo locale → scelta dei parametri strettamente dipendente dalle caratteristiche temporali del segnale e del rumore



Alice: originale - rumoroso - restaurato

Vega: rumoroso - restaurato

How..: originale - rumoroso - restaurato

Disturbi locali: Restauro per modelli (del segnale)

- Fase di rilevamento
  - Modello AR (anzichè passalto + rilevatore a soglia)
  - Si sottrae il modello dal segnale originale
  - Rilevatore a soglia
- Fase di rimozione
  - Least Squares AR-based (anziché interpolazione)

Disturbi locali - rilevazione basata su modelli AR

- Modello AR del segnale (P da 30 a 100)

$$s(n) = \sum_{i=1}^{P} s(n-i)a_i + e(n)$$

- Modello additivo (*i(n)*=0,1; *v(n)*=modello del click)

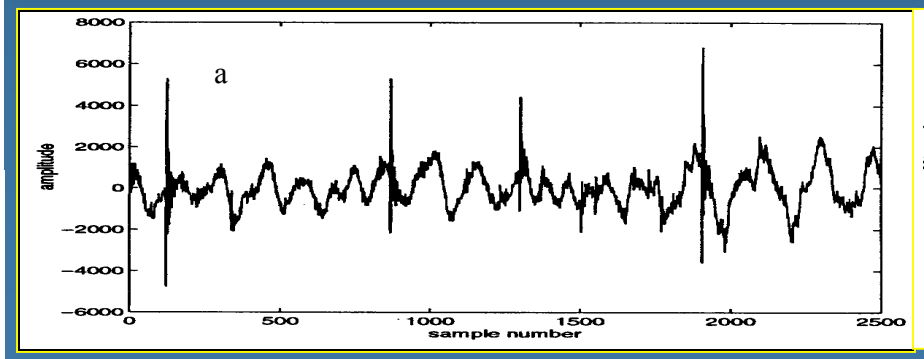
$$x(n) = s(n) + i(n)v(n)$$

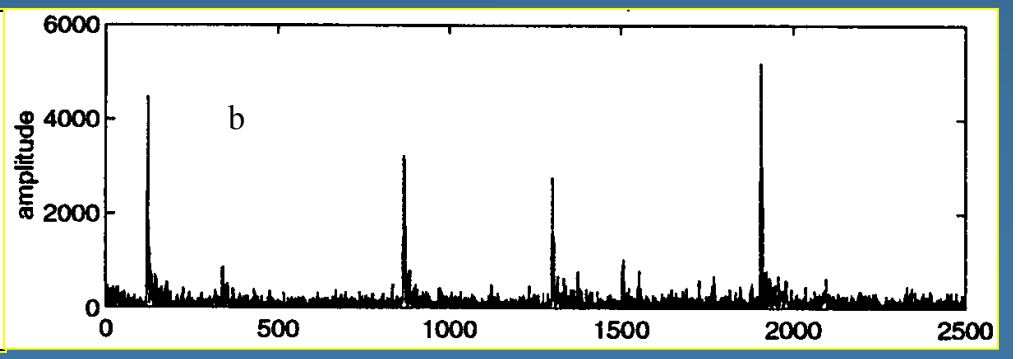
- Segnale di localizzazione $e_l(n) = x(n) - \sum_{i=1}^{P} x(n-i)a_i$

- e quindi $e_l(n) = e(n) + i(n)v(n) - \sum_{i=1}^{P} i(n-i)v(n-i)a_i$

Disturbi locali - individuazione

- Viene amplificato il rapporto disturbo/segnale_non_rumoroso
- Si perde precisione nella localizzazione (l'effetto del click influenza P+1 campioni)
- Si utilizza un rilevamento a soglia sul segnale $e_l(n)^2$





Rimozione dei disturbi locali

- Least Square AR-based (LSAR), con ipotesi sul segnale *mancante*
- Interpolazione *pura* (sino a poche centinaia di campioni) senza ipotesi sul segnale *mancante*:
  - sostituiscono i campioni mancanti con curve di grado *adeguato*
    - 'two sides' (pesa in modo uguale il segnale prima e dopo il click)
    - L $\rightarrow$ R (se il click è presente prima di un attacco rapido)
    - R $\rightarrow$ L (se il click è presente dopo un attacco rapido)

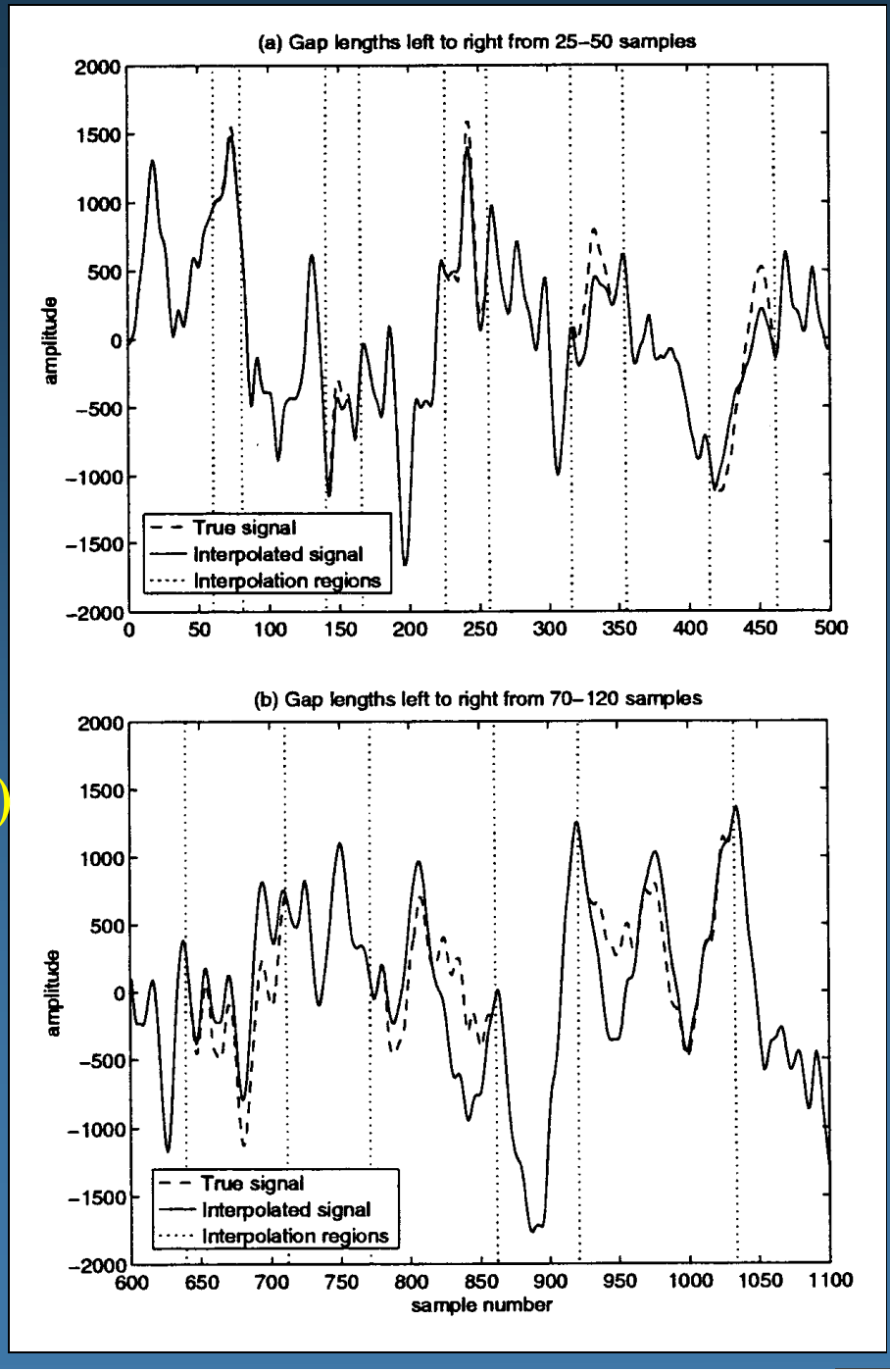
Rimozione dei disturbi locali (mediante LSAR)

- Si minimizza la somma dei quadrati E=**e'e** (esistono molti algoritmi per questo)
- Sono validi anche per click estesi
- Tendono a togliere la parte inarmonica del segnale

Interpolazione tramite LSAR

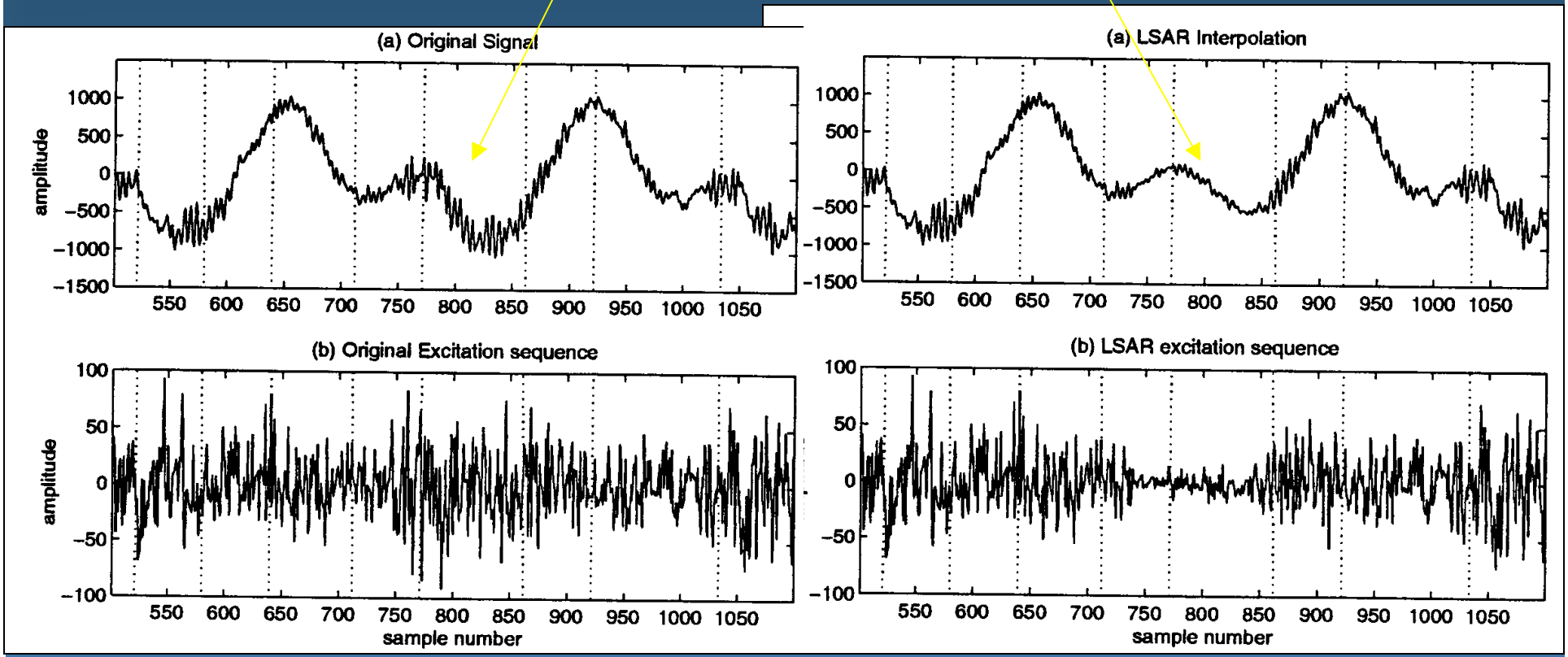
P=60, musica da camera.

(a) brevi click (<50 campioni) (b) click prolungati (70÷120 campioni)



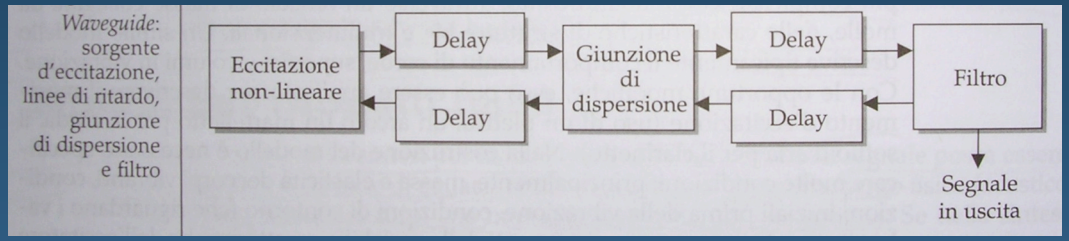
Rimozione dei disturbi locali (mediante LSAR)

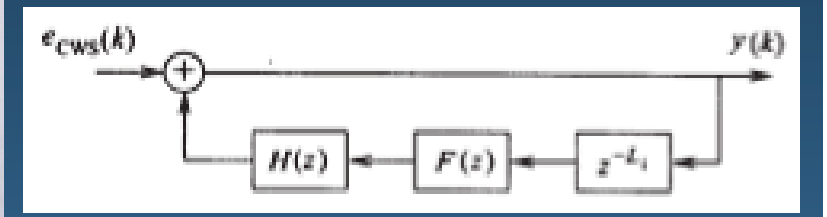
Parte inarmonica ridotta



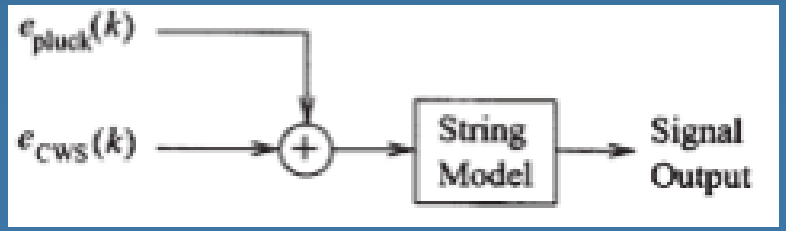
Restauro per modelli (della sorgente)

- Modello di sintesi waveguide (Esquef, JAES, 50(4))



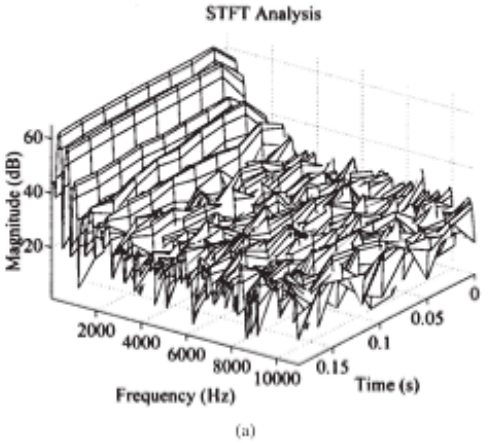


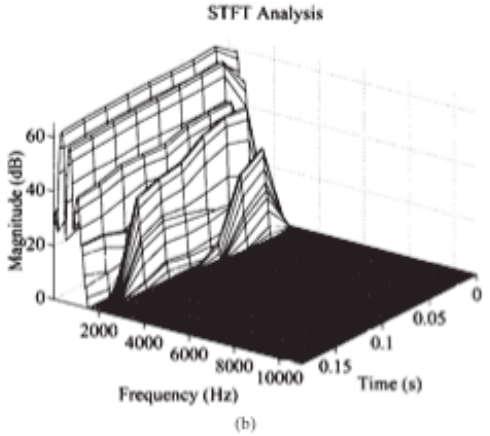
- Stima automatica dei parametri del modello a partire dalla registrazione corrotta (pitch-synchronized STFT analysis)
- Miglioramento della banda (per superare le limitazioni dovute agli strumenti di registrazione dell'epoca)
  - Aggiungere un rumore di eccitazione artificiale al modello della corda

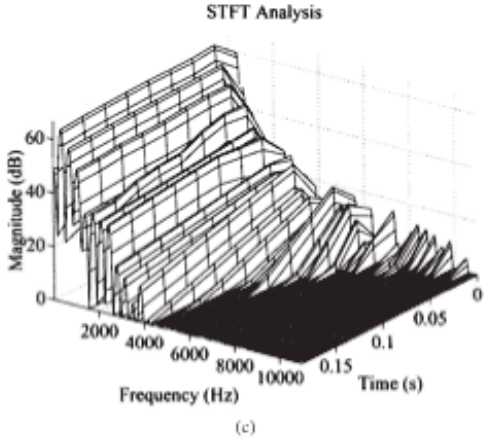


Restauro per modelli (della sorgente)

- Restauro con metodi in frequenza
- Miglioramento della banda







Misura *oggettiva* di un restauro

- Ipotizziamo di conoscere il segnale *pulito*
- Misure utilizzate (*distanze* dall'originale):
  - Scostamento medio e massimo (d=differenza rispetto al segnale pulito):

$$MD = 10 \cdot \log_{10}\left(\left|\bar{d}\right|^2\right)$$

$$MxD = 10 \cdot \log_{10}\left(\max\left|d\right|^2\right)$$

  - Distanza spettrale $\rightarrow$ $$SpD = \frac{10}{\log_e 10} \cdot \int_0^{2\pi} \left(\log_e \frac{S_i(\theta)}{S_u(\theta)}\right)^2 \frac{d\theta}{2\pi}$$

S_i e S_u $\rightarrow$ periodogrammi segnale ingresso e uscita

Conclusioni

- Metodi in frequenza:
  - Semplicità; generalità
  - Rumore musicale; rumore localizzato attorno alle componenti frequenziali del segnale
- Restauro mediante modelli del segnale
  - Rimozione disturbi locali e globali
  - Molti parametri da regolare; inefficace in caso di bassi SNR
- Restauro mediante modelli della sorgente
  - Efficace nei casi di bassissimo SNR (…segnale mancante?)
  - Limitato a casi semplici → separazione e riconoscimento e degli eventi musicali, modello fisico dell'evento. (…voce?)
- Restauro mediante proiezione locale
  - Efficace nei casi di basso SNR
    - Con SNRin < 10 dB, si arriva a SNRout ≅ 18 dB
  - Il rumore viene eliminato solo nelle vocali