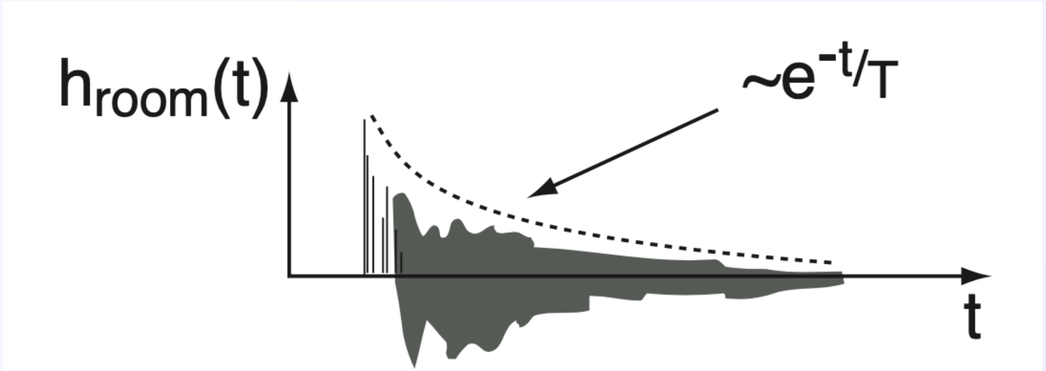
L3
Noise and reverberation - Enhancing speech

COM4511/COM6511 - Speech Technology



Thomas Hain
t.hain@sheffield.ac.uk
Spring Semester

Outline

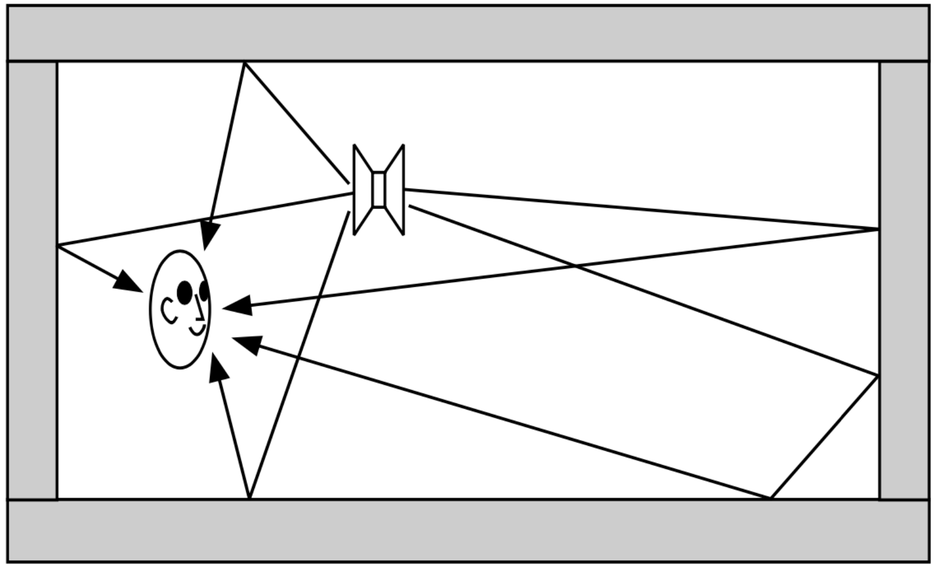
- Foundations
- Spectral substraction
- Wiener filtering
- Model based processing
- Beamforming

Foundations

- When speech is recorded it is often distorted by noise
  - acoustic channel
  - recording equipment
  - other people speaking in the background
  - other acoustic events in the background

Reverberation

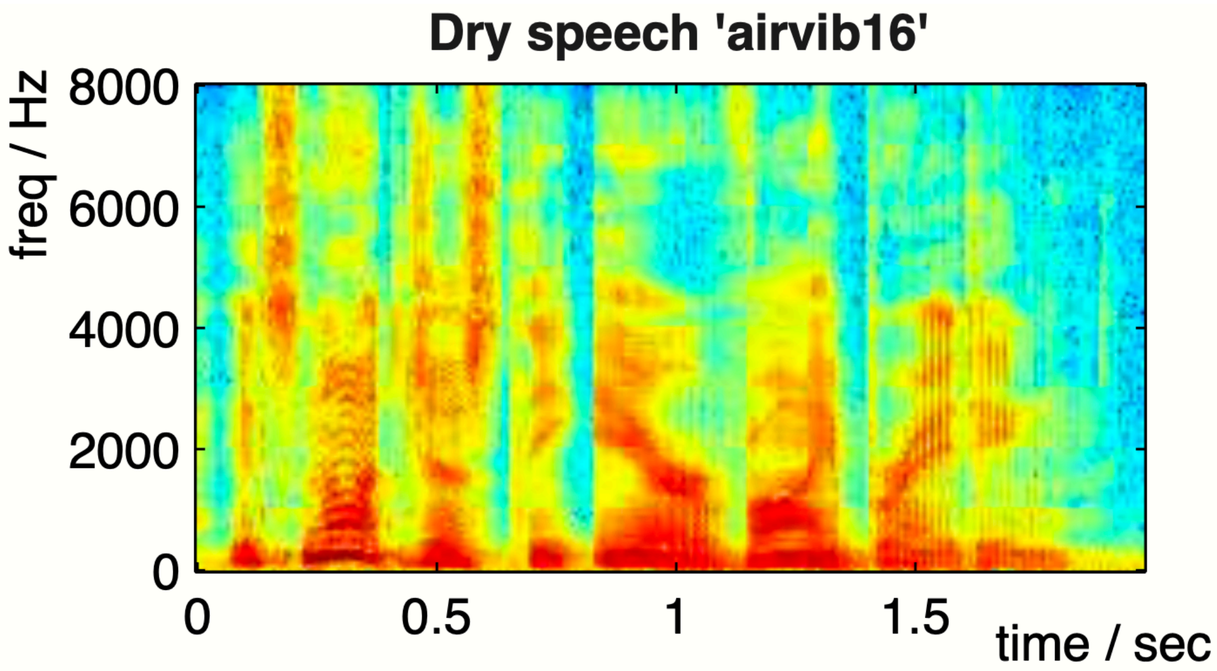
- Sound waves reflect off walls in an enclosed space
  - Different paths creates a series of echoes
- Sound scattered by different objects in the room
- Energy absorbed at walls
  - Depends on wall materials
  - Frequency dependent effects

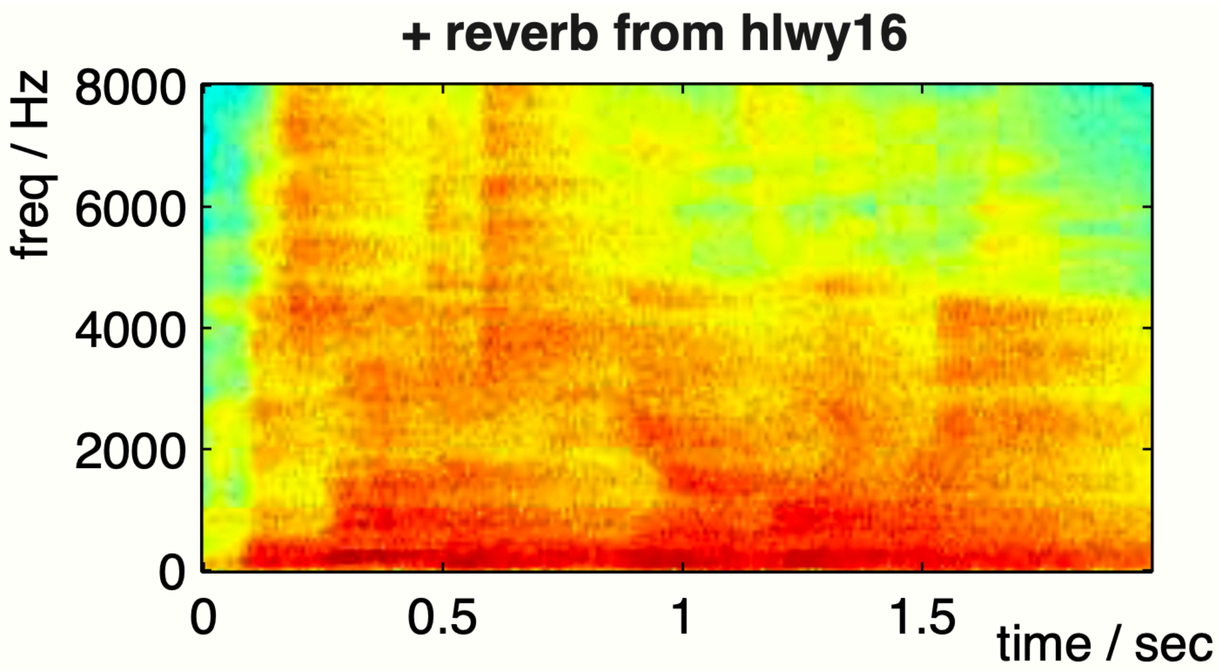


Reverberation (2)

$$\sum_i x[k - \tau_i] * g_i[k]$$

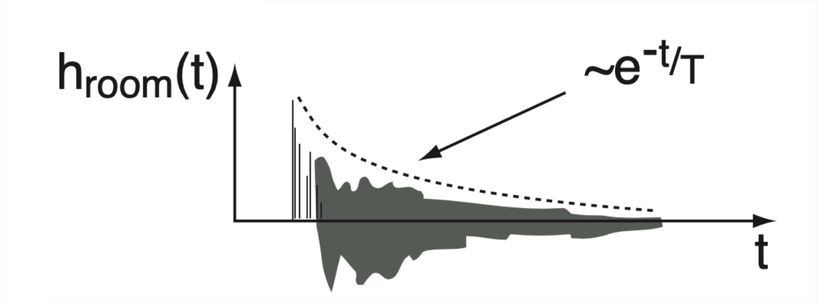
- (infinite) Sum of delayed and filtered signals
- Causes “smearing” of signal energy

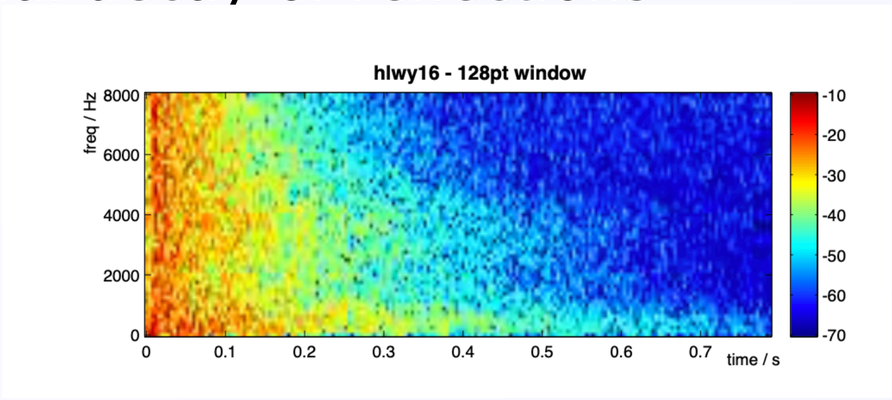




Reverberation time

- The room can be viewed as an LTI system
- The effect of the room is an exponential decay of reflections



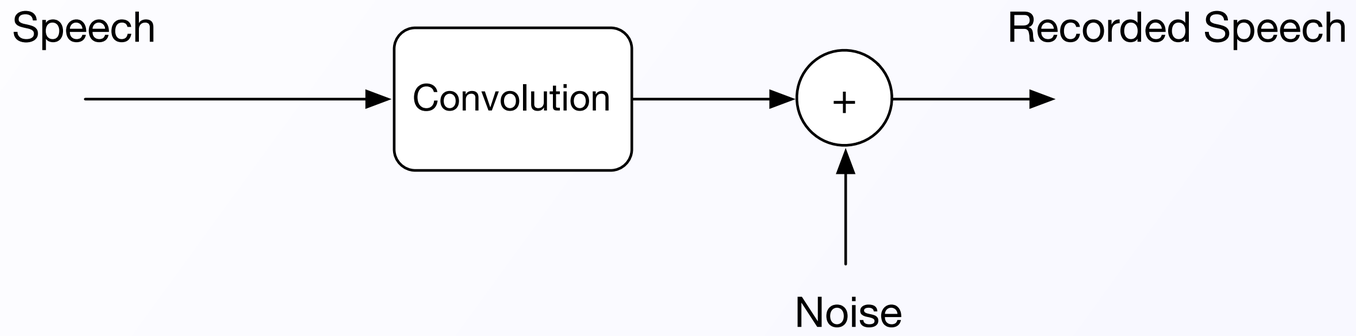


- The impulse responses are frequency dependent
  - greater absorption at higher freqeuncies means faster decay
- Length
  - Large rooms meean slow decay
- Measurement: Reverb time

$$R_{T60} = \frac{0.049V}{R\alpha}$$

Volume V, Surface area S, reflection coefficient α

Noise generally



- Additive

$$\mathrm{x[k]} = \mathrm{s[k]} + \mathrm{n[k]}$$

- Convolutional

$$x[k] = s[k] * h_n[k]$$

Additive noise

- time domain (discrete)

$$x[k] = s[k] * h_n[k]$$

- spectrum

$$X(\omega) = S(\omega) + N(\omega)$$

- in practice the signal is windowed, window p

$$X(p, \omega) = S(p, \omega) + N(p, \omega)$$

- target (no need to restore original phase)

$$\hat{X}(p, \omega) = |X(p, \omega)| e^{\angle Y(p,\omega)}$$

Convolutional noise

- Time domain
  $h[k]$ is linear time invariant filter

$$x[k] = s[k] * h[k]$$

- Frequency domain is a multiplication

$$X(p, \omega) = S(p, \omega)H(\omega)$$

Note: $H(\omega)$ is assumed not time dependent.

Spectral subtraction and additive noise

- Let us assume that the noise and the speech are uncorrelated.

$$|\hat{X}(p,\omega)|^2 = |Y(p,\omega)|^2 - S_N(\omega)$$

$S_N(\omega)$ is the noise power spectrum that is assumed not to change over time.

We need an estimate for the noise !

- To obtain the Short term Fourier transform of a the signal after substraction

$$\hat{X}(p,\omega) = |\hat{X}(p,\omega)|e^{\angle Y(p,\omega)}$$

Spectral subtraction as Filtering

- The subtraction operation can be rewritten:

$$\begin{aligned} |\hat{X}(p,\omega)|^2 &= |Y(p,\omega)|^2 - S_N(\omega) \\ &\approx |Y(p,\omega)|^2 \left[1 + \frac{1}{R(p,\omega)}\right]^- 1 \end{aligned}$$

with the signal to noise ratio (estimate)

$$R(p,\omega) = \frac{|X(p,\omega)|^2}{\hat{S}_N(\omega)}$$

- The suppression filter is

$$H_s(p,\omega) = \left[1 + \frac{1}{R(p,\omega)}\right]^{-\frac{1}{2}}$$

Wiener filtering

- Objective: compute a filter such that (additive noise)

$$\hat{x}[k] = y[k] * h_W[k]$$

- Find by minimising the error

$$e = \mathcal{E}\{(\hat{x}[k] - x[k])^2\}$$

  Several possible solutions.

- Frequency domain solution

$$H_W(\omega) = \frac{S_X(\omega)}{S_X(\omega) + S_N(\omega)}$$

Time-varying Wiener filter

- Again we need to consider windowed signals !

$$H_W(p,\omega) = \frac{\hat{S}_X(p,\omega)}{\hat{S}_X(p,\omega) + S_N(\omega)}$$

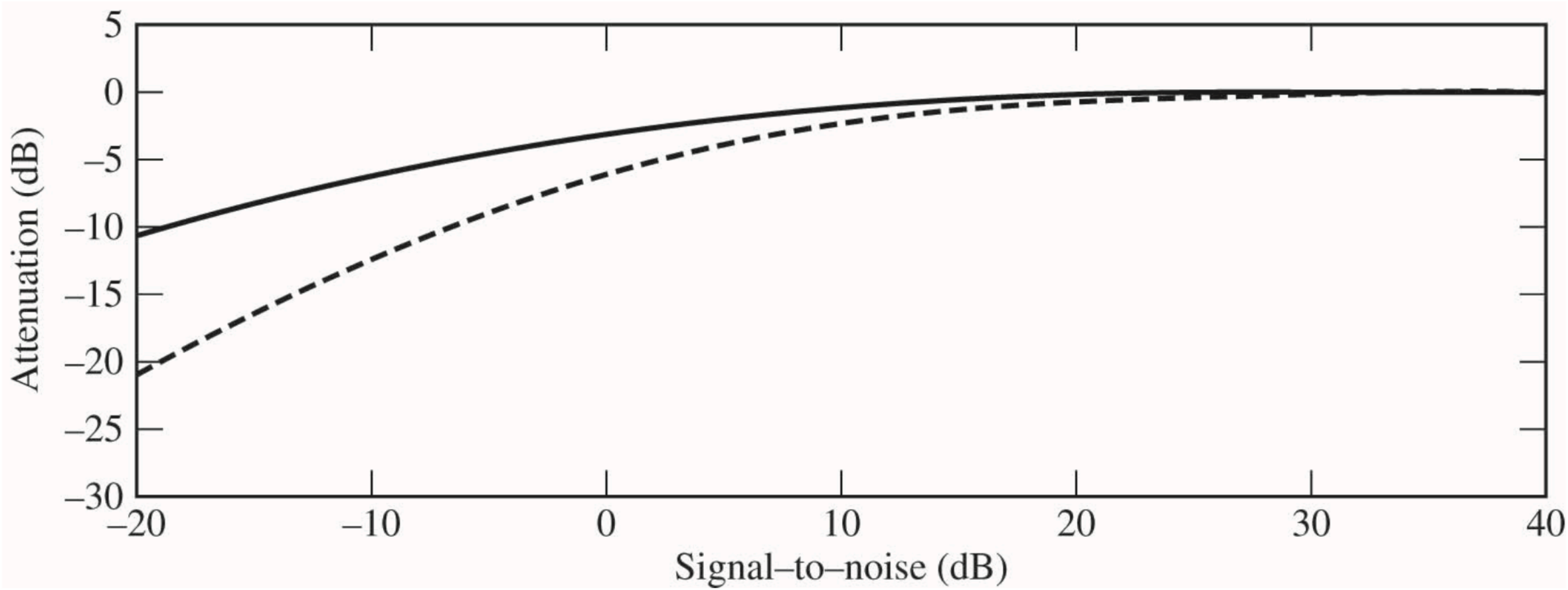
- which again can be written as

$$H_W(p,\omega) = \left[1 + \frac{1}{R(p,\omega)}\right]^{-1}$$

with

$$R(p,\omega) = \frac{\hat{S}(p,\omega)|^2}{\hat{S}_N(\omega)}$$

Comparison of filters - Suppresion



Solid line: Spectral Subtraction. Dashed-line: Wiener filter

A basic approach to estimation

- We assume that the Wiener filter for the previous frame, $p-1$ is known.

$$\hat{X}(p,\omega) = Y(p,\omega)H_W(p-1,\omega)$$

- We recursively update the Wiener filter

$$H_W(p,\omega) = \frac{|\hat{X}(p,\omega)|^2}{|\hat{X}(p,\omega)|^2 + \hat{S}_N(\omega)}$$

- We compute a smoothed version of the power spectrum

$$\hat{S}_X(p,\omega) = \tau\hat{S}_X(p-1,\omega) + (1-\tau)\hat{S}_X(p,\omega)$$

with

$$\hat{S}_X(p,\omega) = |\hat{X}(p,\omega)|^2$$

Initialise with spectral subtraction !

Compuitng in practice

- We assume that the Wiener filter of $p-1$ frame is known, then:

$$\hat{X}(pL,\omega) = Y(pL,\omega)H_w((p-1)L,\omega)$$

- Updating the Wiener filter:

$$H_w(pL,\omega) = \frac{|\hat{X}(pL,\omega)|^2}{|\hat{X}(pL,\omega)|^2 + \hat{S}_b(\omega)}$$

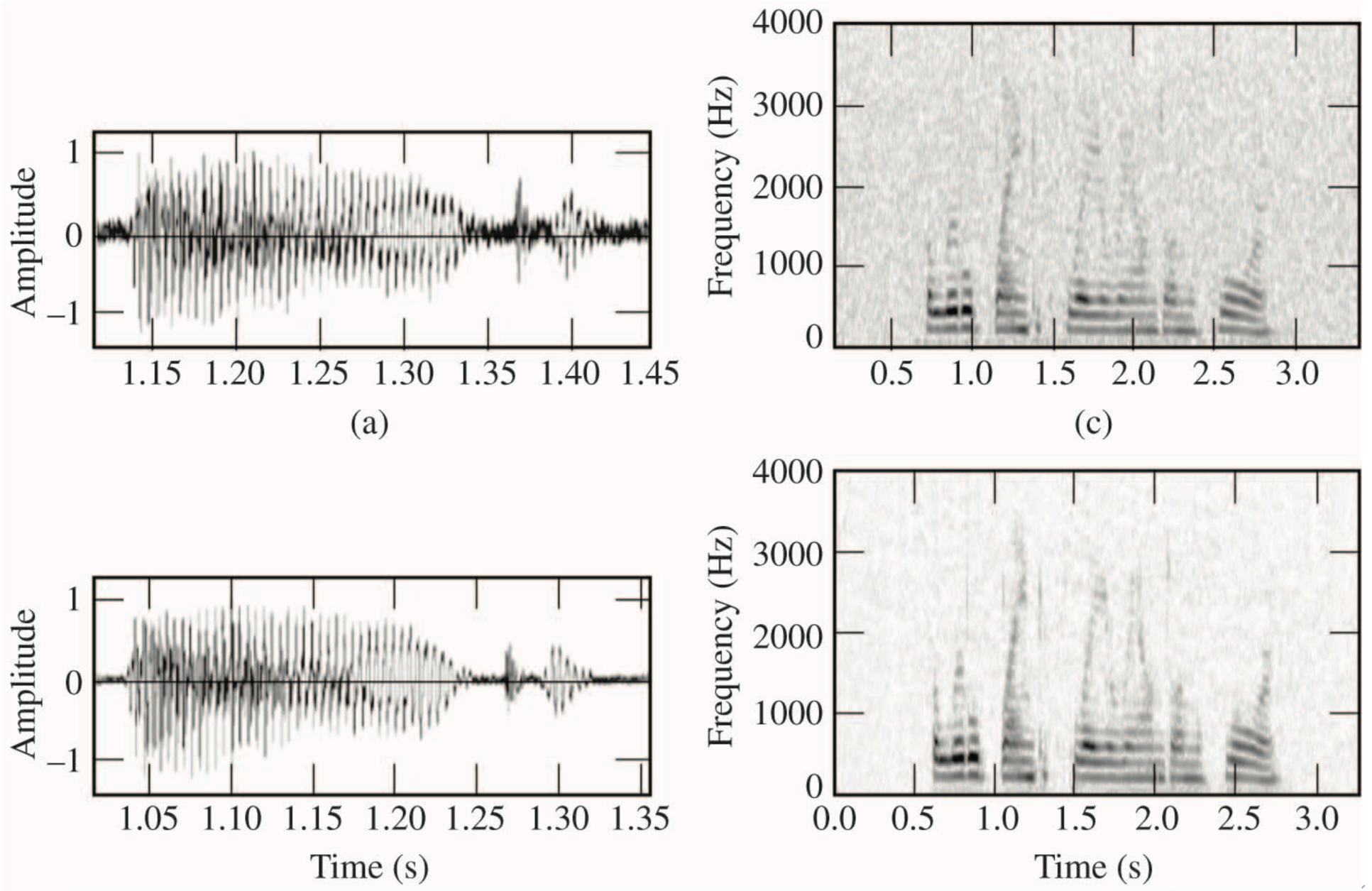
- Smooth power spectrum:

$$\tilde{S}_x(pL,\omega) = \tau\tilde{S}_x((p-1)L,\omega) + (1-\tau)\hat{S}_x(pL,\omega)$$

  where $\hat{S}_x(pL,\omega) = |\hat{X}(pL,\omega)|^2$

- Initialization: spectral subtraction

Enhancement examples

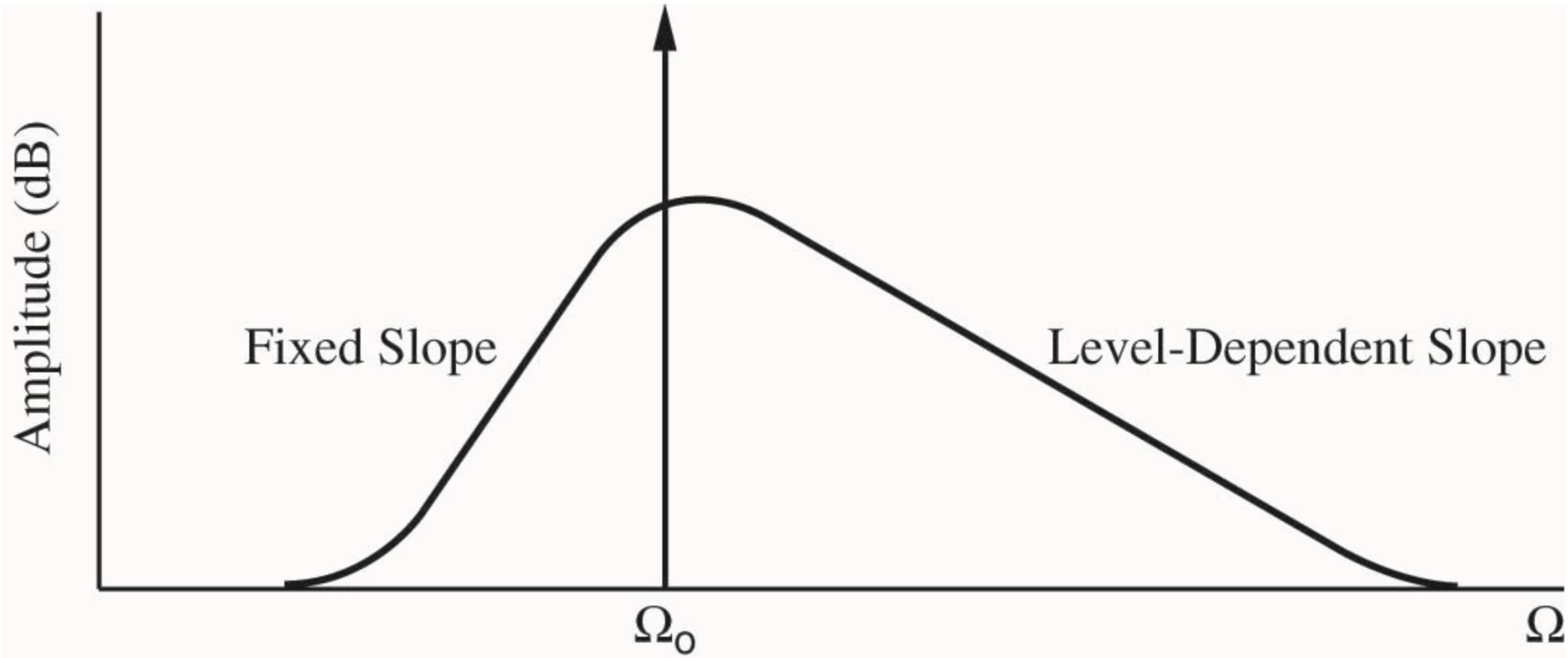


Auditory masking

Auditory masking: one sound component is concealed by the presence of another sound component.

- Frequency masking
- Temporal masking
- Critical band
- Masking threshold
- Maskee
- Masker

Masking threshold curve



Frequency-Domain Masking Principles

- Physiologically-based/Psychoacoustically-based filters
- Critical Bands: Bandwidth of Psychoacoustically-based filters
- Quantized critical bands (Bark Scale):
  - z = 13 arctan (0.76f) + 3.5 arctan (f /7500)
- Quantized critical bands (Mel Scale):
  - m = 2595 log1 0(1 + f /700)

Masking Threshold Calculation

1. Compute energy E_k in each k-th bark filter in the estimated speech spectrum (after spectral subtraction)

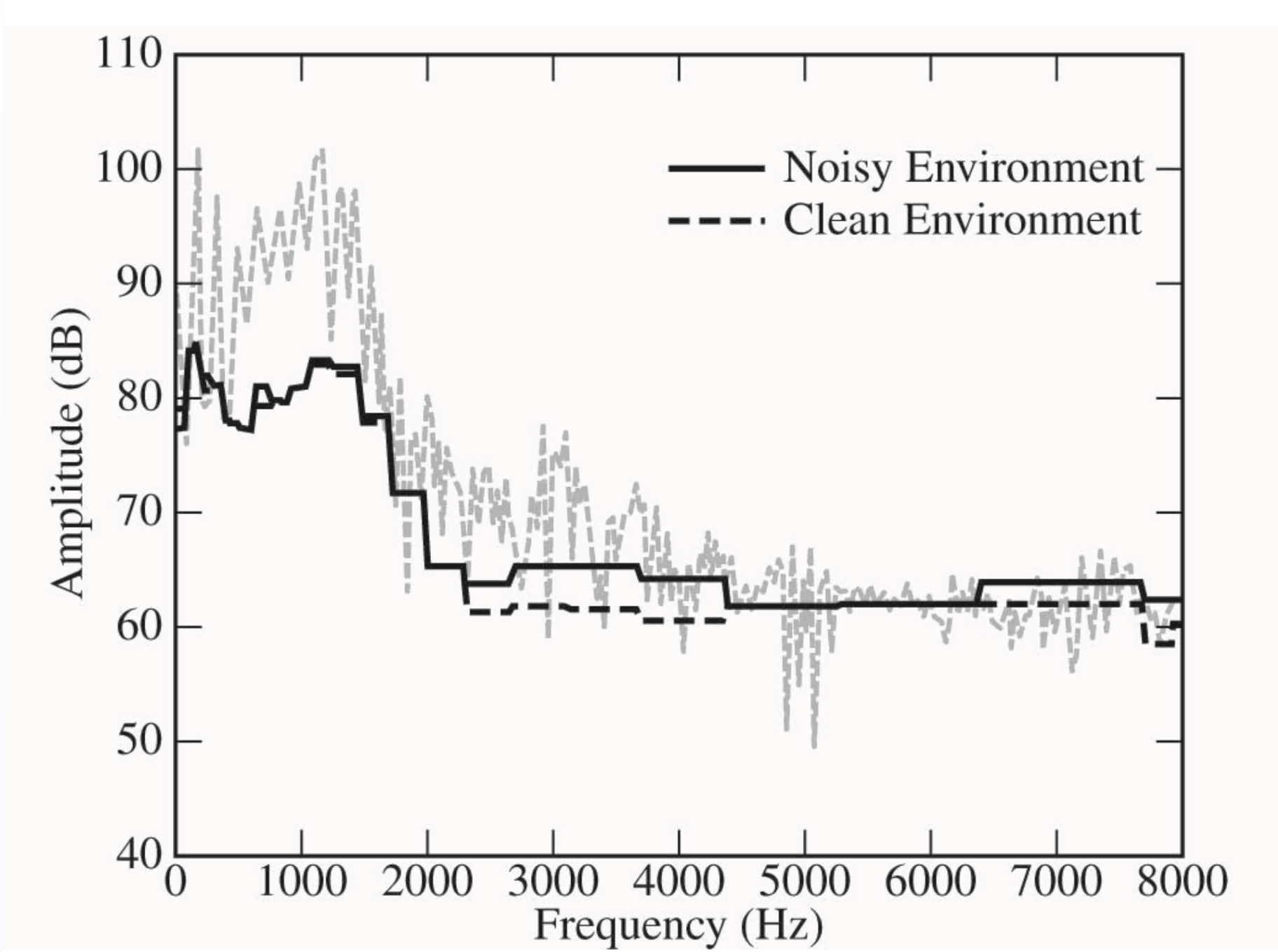
2. Convolve each Ek with a "spreading function" hk :

$$Tk = Ek * hk$$

3. Subtract a threshold offset depending if the masker is noise-like or tone-like.

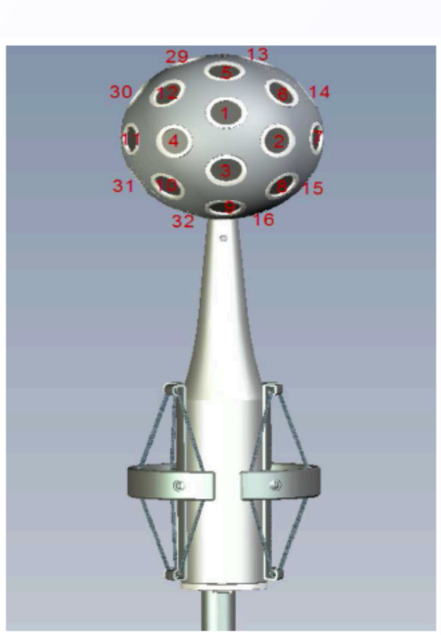
4. Map Tk to linear frequency scale to obtain $T(p, \omega)$

Auditory Masking Threshold Curves

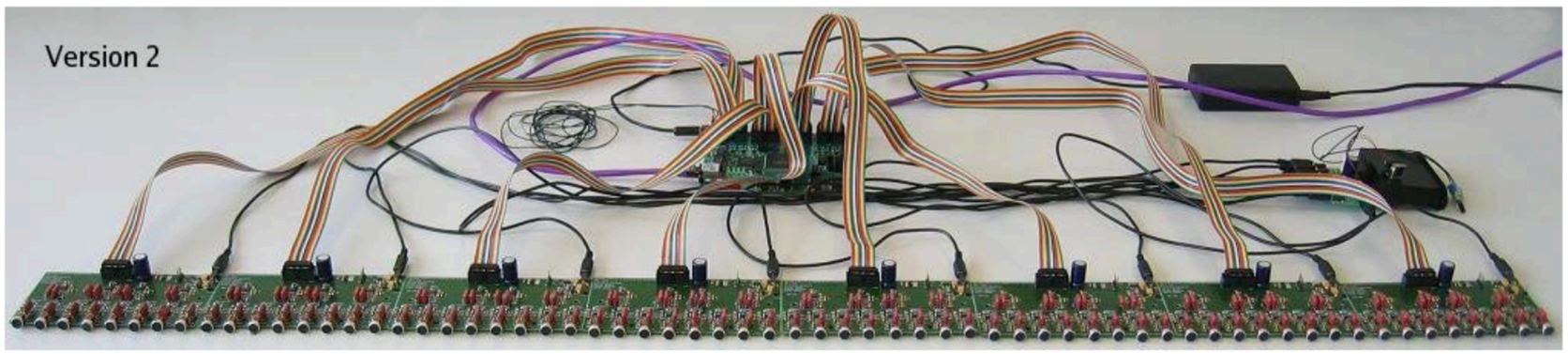


Beam-forming

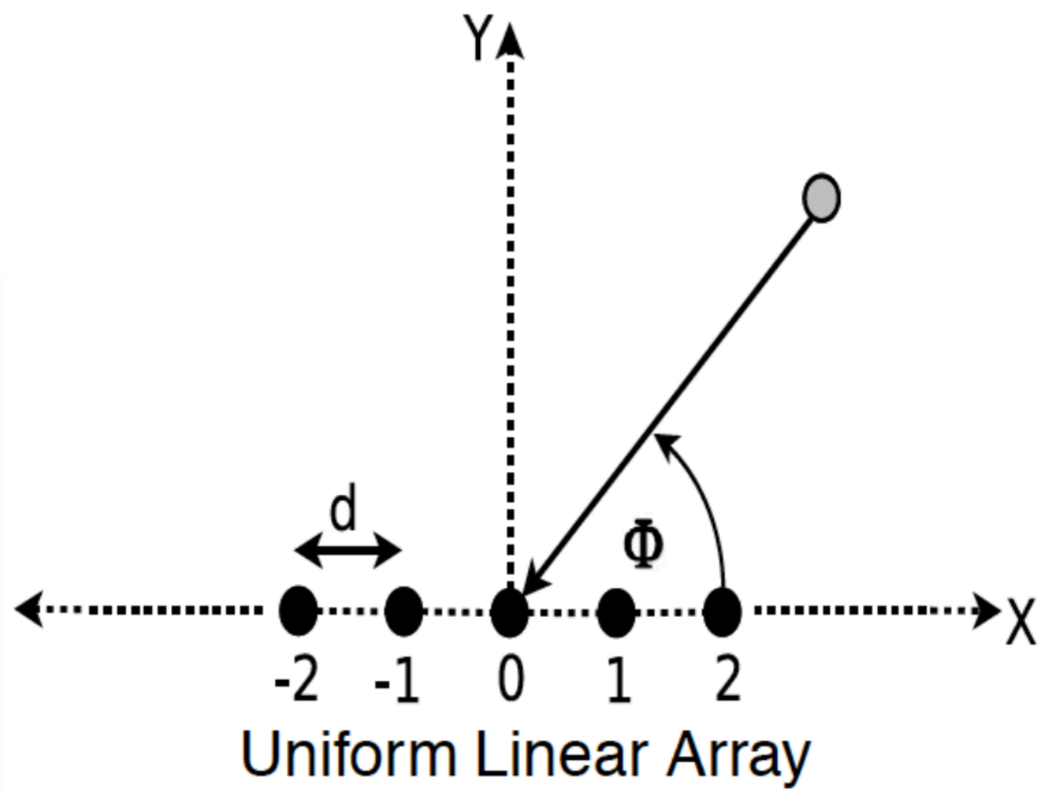
- Alternative - use more microphones
  - Listen into the direction of the sound source
  - Spatial filtering

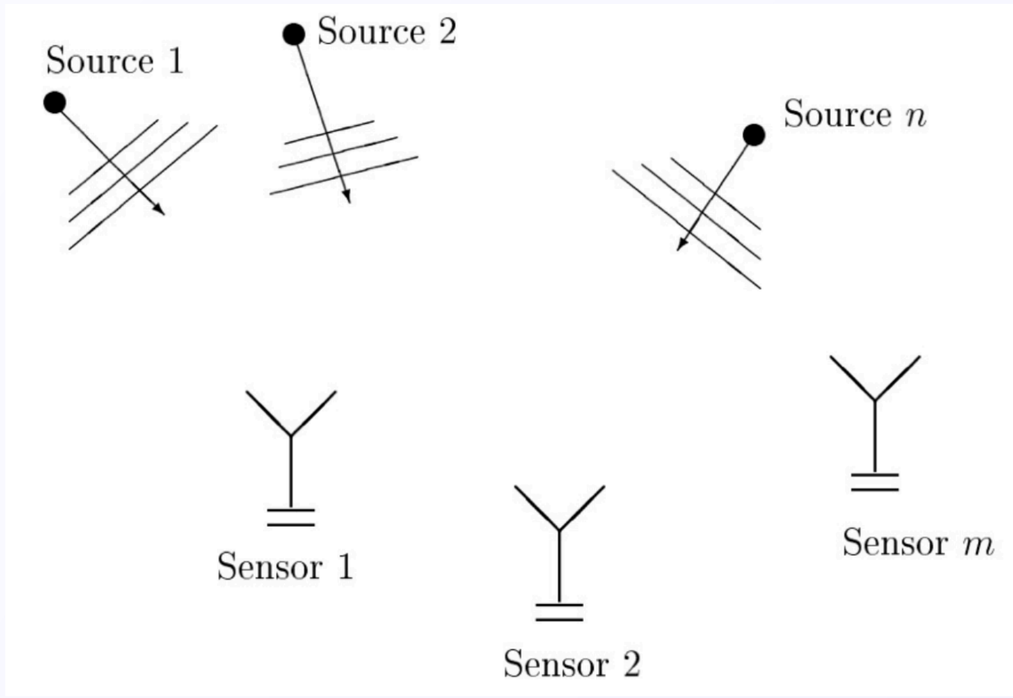


Uniform Circular Array



Microphone array configurations





$$y_i[k] = h_i[k] * x[k - \tau_i] + n_i[k]$$

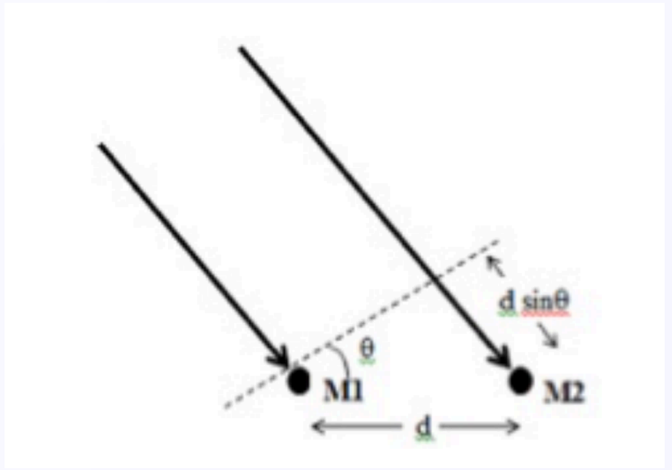
- Each sensor receives a time delayed and distrorted version of the signal

Time delay estimation - the trivial way !

- Start with simplified assumptions

$$y_i[k] = x[k - \tau_i] + n_i[k]$$

Only delay is assumed to exist.



- Autocorrelation peaks allow to find the delay
  Noise and signal are assumed to be uncorrelated.

$$r_{y_i}[k] = \sum_j y_I[i] y_i[j - k]$$

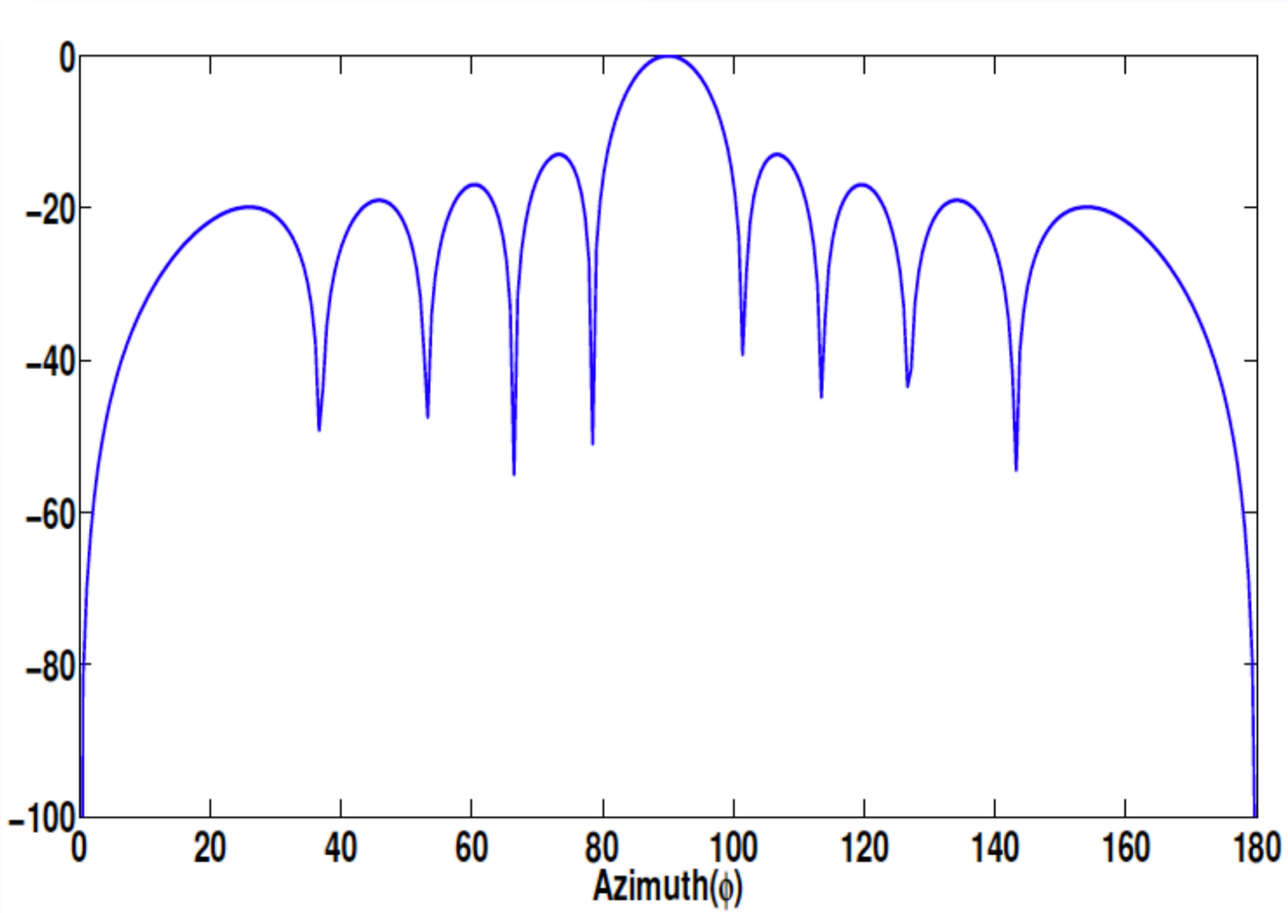
Therefore

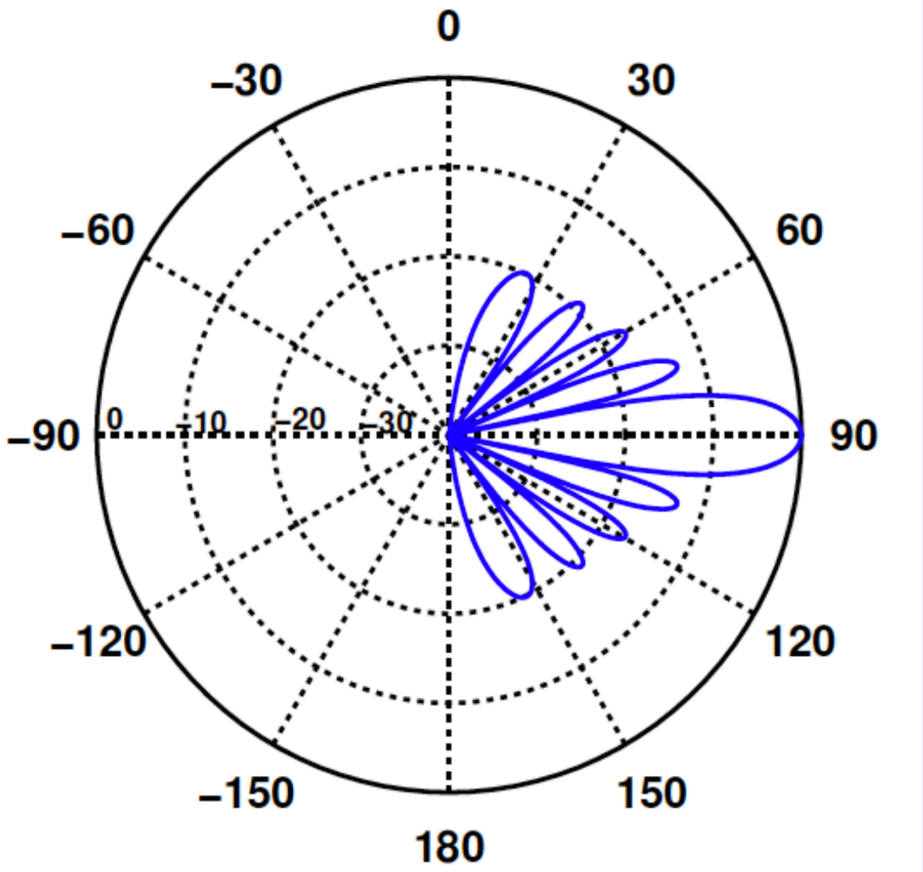
$$\hat{\tau}_i = \arg\max_k r_{y_i}[k]$$

- The enhanced signal will be

$$\hat{y}_i[k] = \sum_i y_i[k - \hat{\tau}_i]$$

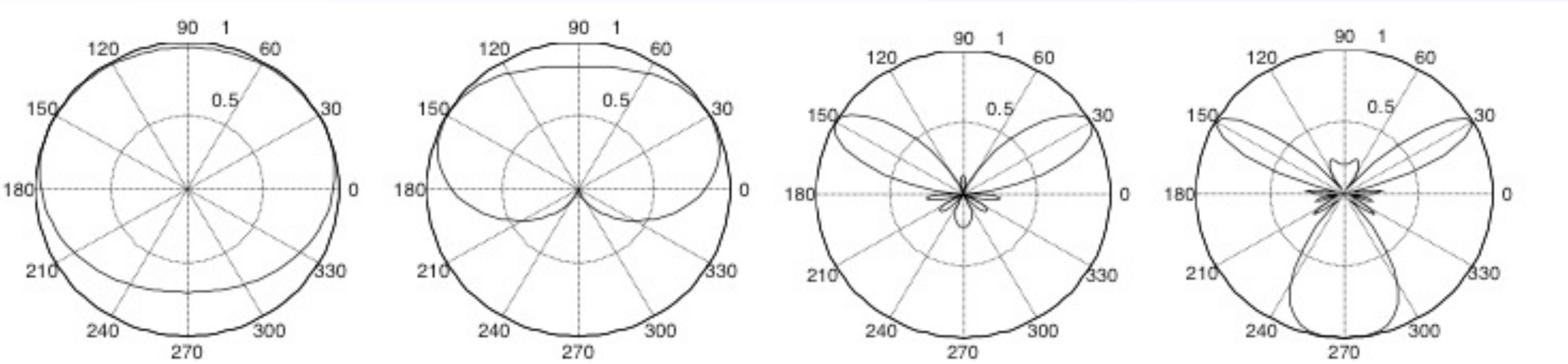
Directivity





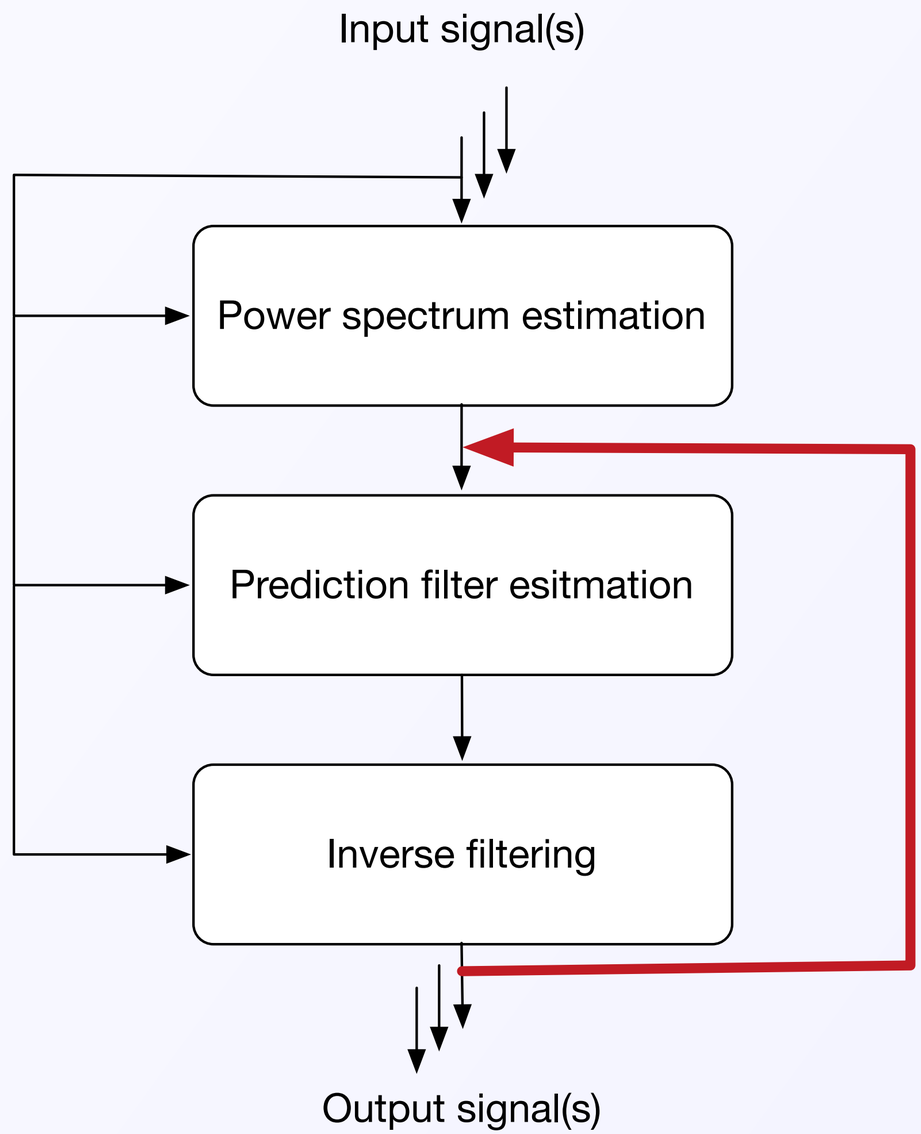
Directivity and frequency !

- Unfortunately the listening directivity is frequency dependent



Modern dereberation - WPE

- Reverberation implies weighted sums of identical signals
- Dereverberation through inverse filtering of original signal
  - maximum likelihood
- Power spectral density estimation of the target signal
  - iterative
- Prediction filter estimate



Nakatani et al. Speech Dereverberation Based on Variance-Normalized Delayed Linear Prediction (2010, TASLP