Enhancement of Noisy Speech

State-of-the-Art and Perspectives

Rainer Martin

Institute of Communications Technology (IFN)

Technical University of Braunschweig

July, 2003

Applications of Noise Reduction

- Hands-free telephony.
- Robust speech recognition.
- Robust speech coding (ETSI/3GPP AMR, MELPe, ITU-T 4 kbit/s codecs).
- Hearing aids and cochlear implants.
- Restoration of historic recordings.
- Forensic applications.

Ingredients

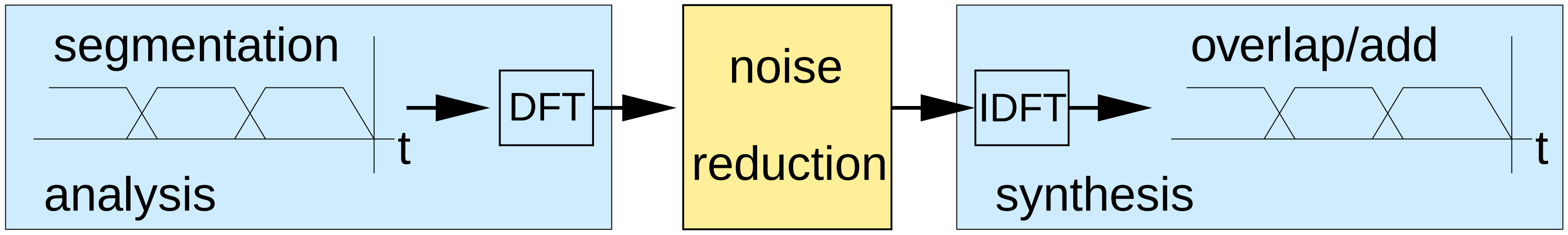
- Models of speech production
- Signal theory
- Room acoustics
- Psychoacoustics
- Models of speech perception

Objective: Improve quality and intelligibility!

Combine signal theoretic and perceptive approaches!

Noise Reduction in the Spectral Domain

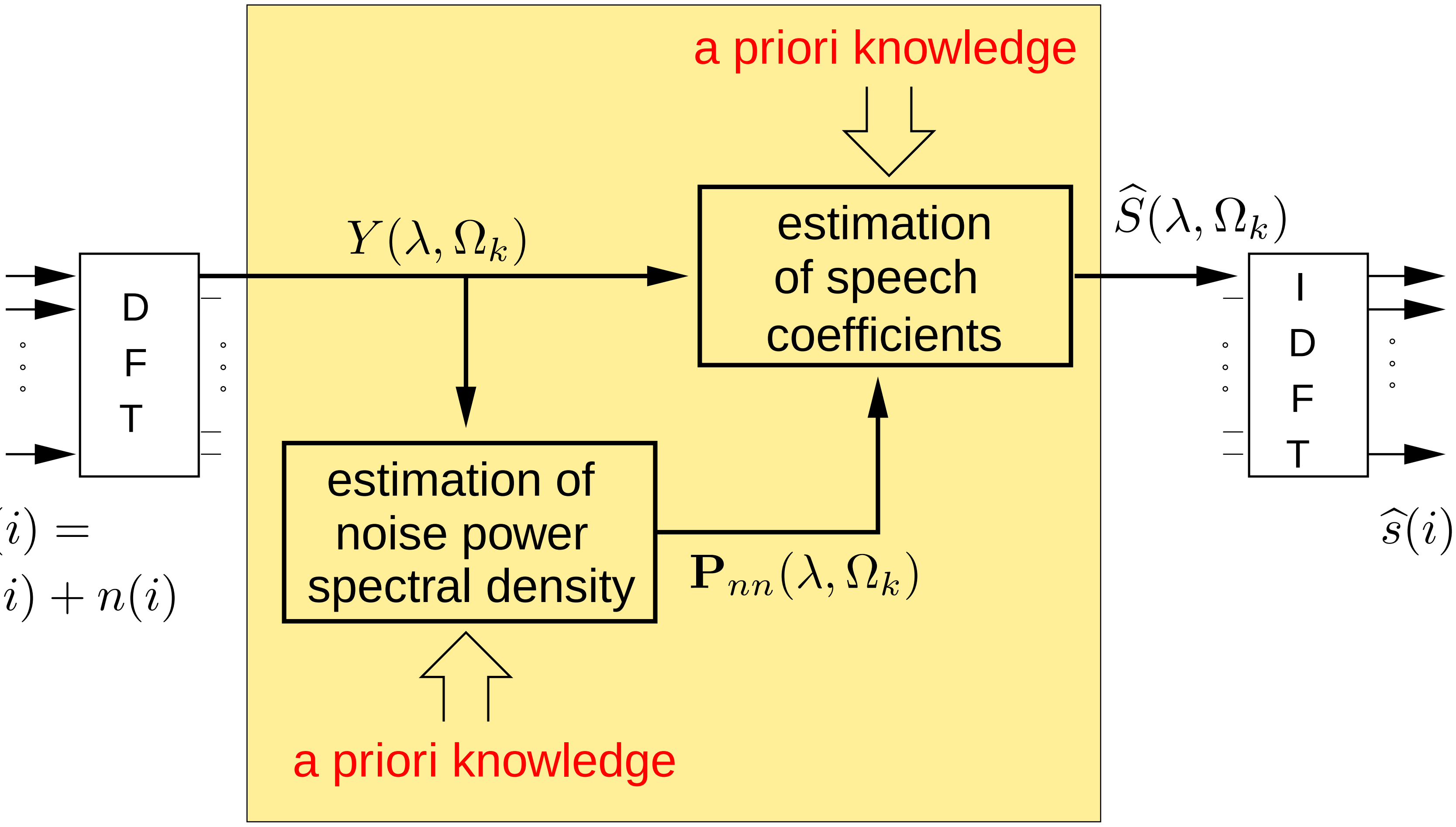
- **Spectral analysis – noise reduction – synthesis:**



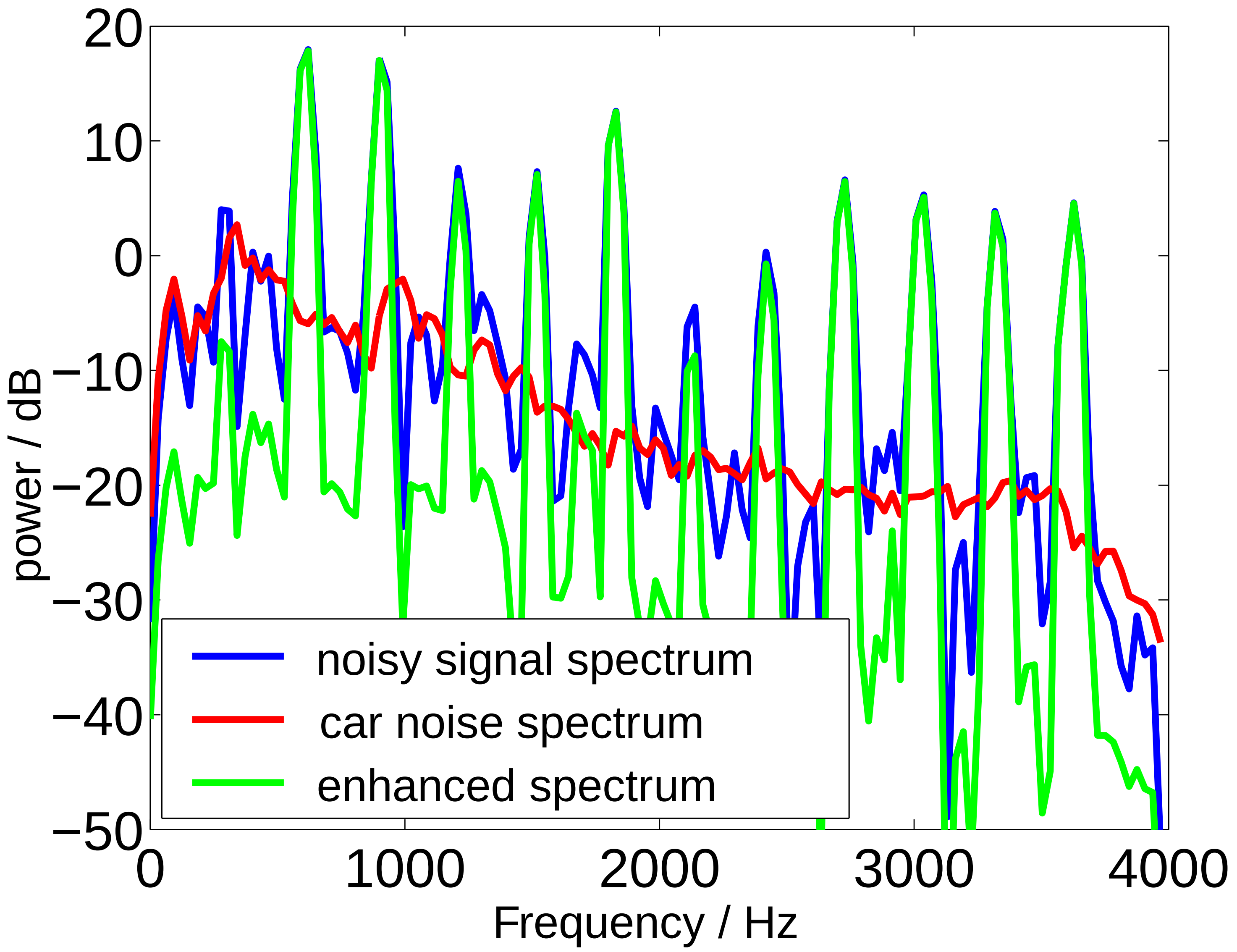
- **Advantages of spectral processing:**
  - good separation of speech and noise
  - decorrelation of spectral components
  - integration of psychoacoustic models

Principles of Noise Reduction

$\lambda \longrightarrow$ frame index $\qquad k \longrightarrow$ frequency bin index


a priori knowledge
estimation of speech coefficients
$\widehat{S}(\lambda, \Omega_k)$
$Y(\lambda, \Omega_k)$
D F T
I D F T
$y(i) = s(i) + n(i)$
$\widehat{s}(i)$
estimation of noise power spectral density
$\mathbf{P}_{nn}(\lambda, \Omega_k)$
a priori knowledge

Principles of Noise Reduction



Estimation of Speech Coefficients

- **Linear estimators**
  - e.g. Wiener Filter
- **Non-linear estimators**
  - MMSE Short Time Spectral Amplitude estimator

    [Ephraim & Malah, 1984, 1985]
  - Psychoacoustic methods [Gustafsson et al. 1998]
  - MMSE estimation based on supergaussian priors

    [Martin 2002]

MMSE Estimation

- **Optimal estimate for independent real and imaginary parts:**

$$E\{S \mid Y\} = E\{S_R \mid Y_R\} + jE\{S_I \mid Y_I\}$$

- **Estimation of either the real or the imaginary part:**

$$E\{S_\diamondsuit \mid Y_\diamondsuit\} = \int_{-\infty}^{\infty} S_\diamondsuit p(S_\diamondsuit \mid Y_\diamondsuit) dS_\diamondsuit$$

- **Application of Bayes theorem:**

$$E\{S_\diamondsuit \mid Y_\diamondsuit\} = \frac{1}{p(Y_\diamondsuit)} \int_{-\infty}^{\infty} S_\diamondsuit {\color{blue}p(Y_\diamondsuit \mid S_\diamondsuit)} {\color{red}p(S_\diamondsuit)} dS_\diamondsuit$$

- **What is the appropriate prior density ${\color{red}p(S_\diamondsuit)}$?**

Some Answers and Some Questions

- DFT coefficients are asymptotically complex Gaussian distributed ! [Brillinger, 1981]
- Typical frame size in mobile communications: 10-30 ms $<$ span of correlation of (voiced) speech !
- Do the asymptotic assumptions hold for speech signals ???
- No! See, e.g., [Porter and Boll, 1984].

Prior Densities for Real and Imaginary Part

- **Gaussian pdf:**

$$p(S_\Diamond) = \frac{1}{\sqrt{\pi}\sigma_s} \exp\left(-\frac{S_\Diamond^2}{\sigma_s^2}\right)$$

$\rightarrow$ Wiener filter

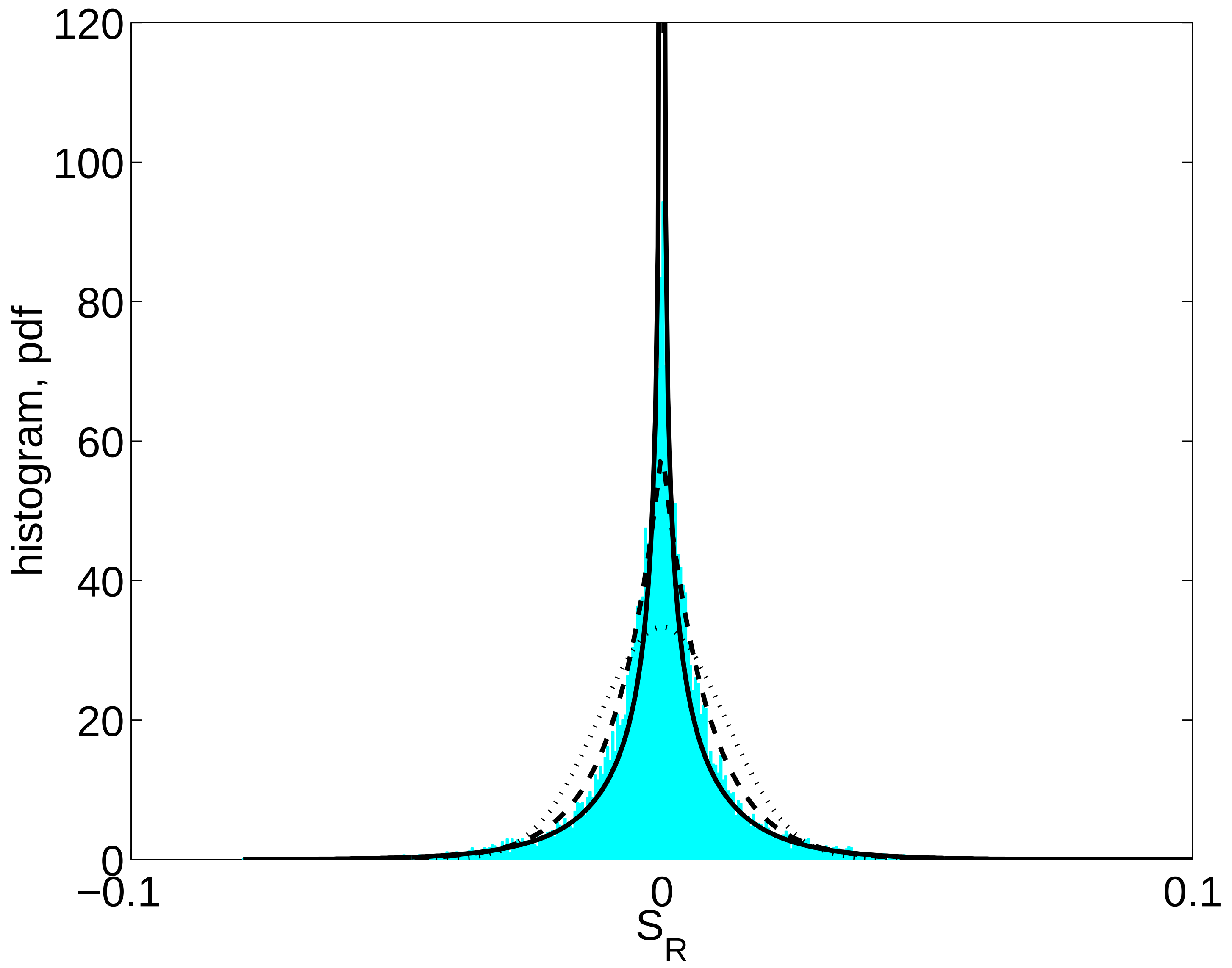
- **Laplacian pdf:**

$$p(S_\Diamond) = \frac{1}{\sigma_s} \exp\left(-\frac{2|S_\Diamond|}{\sigma_s}\right)$$

- **Gamma pdf:**

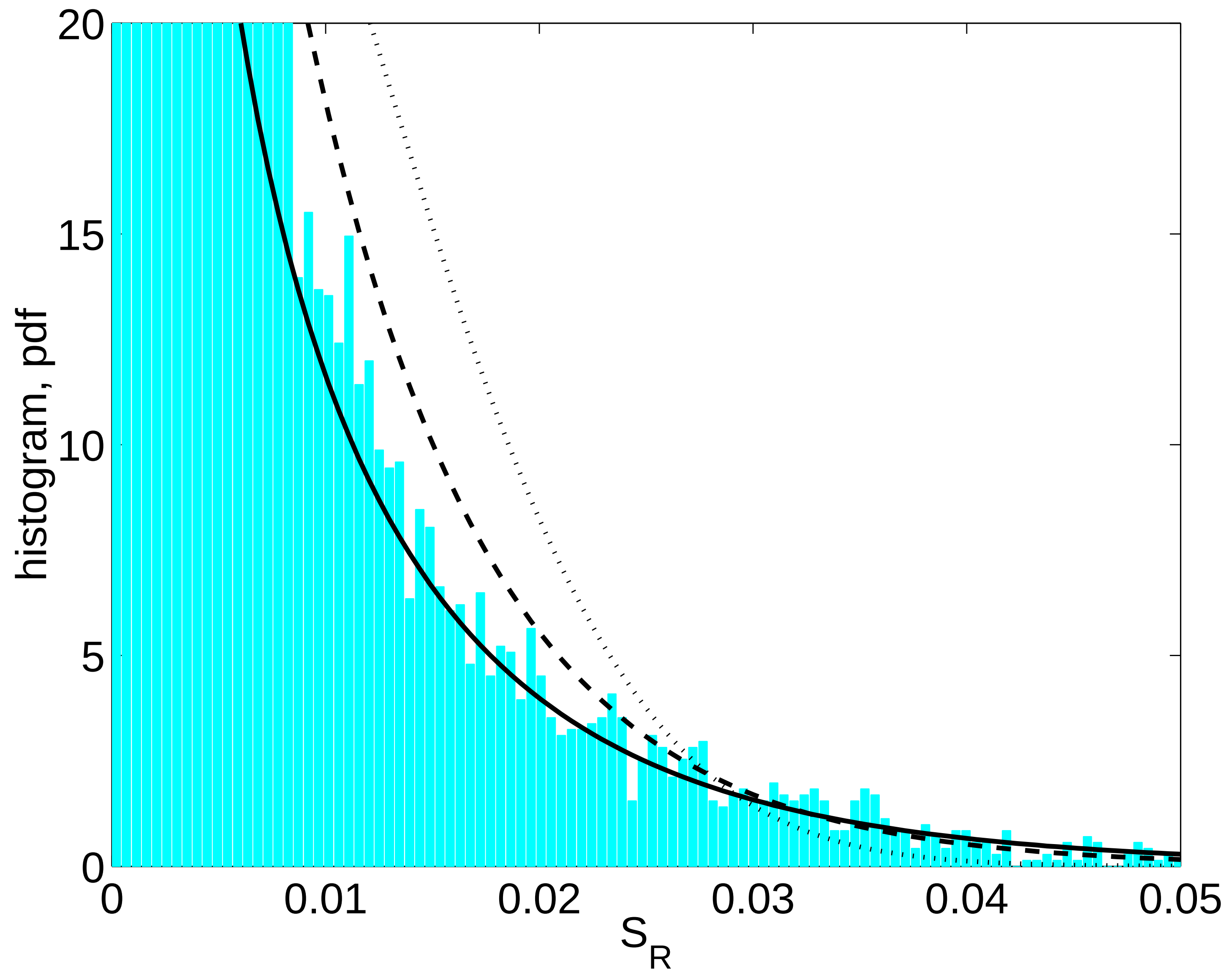
$$p(S_\Diamond) = \frac{\sqrt[4]{3}}{2\sqrt{\pi\sigma_s}\sqrt[4]{2}} |S_\Diamond|^{-\frac{1}{2}} \exp\left(-\frac{\sqrt{3}|S_\Diamond|}{\sqrt{2}\sigma_s}\right)$$

Histogram of DFT Coefficients for Speech



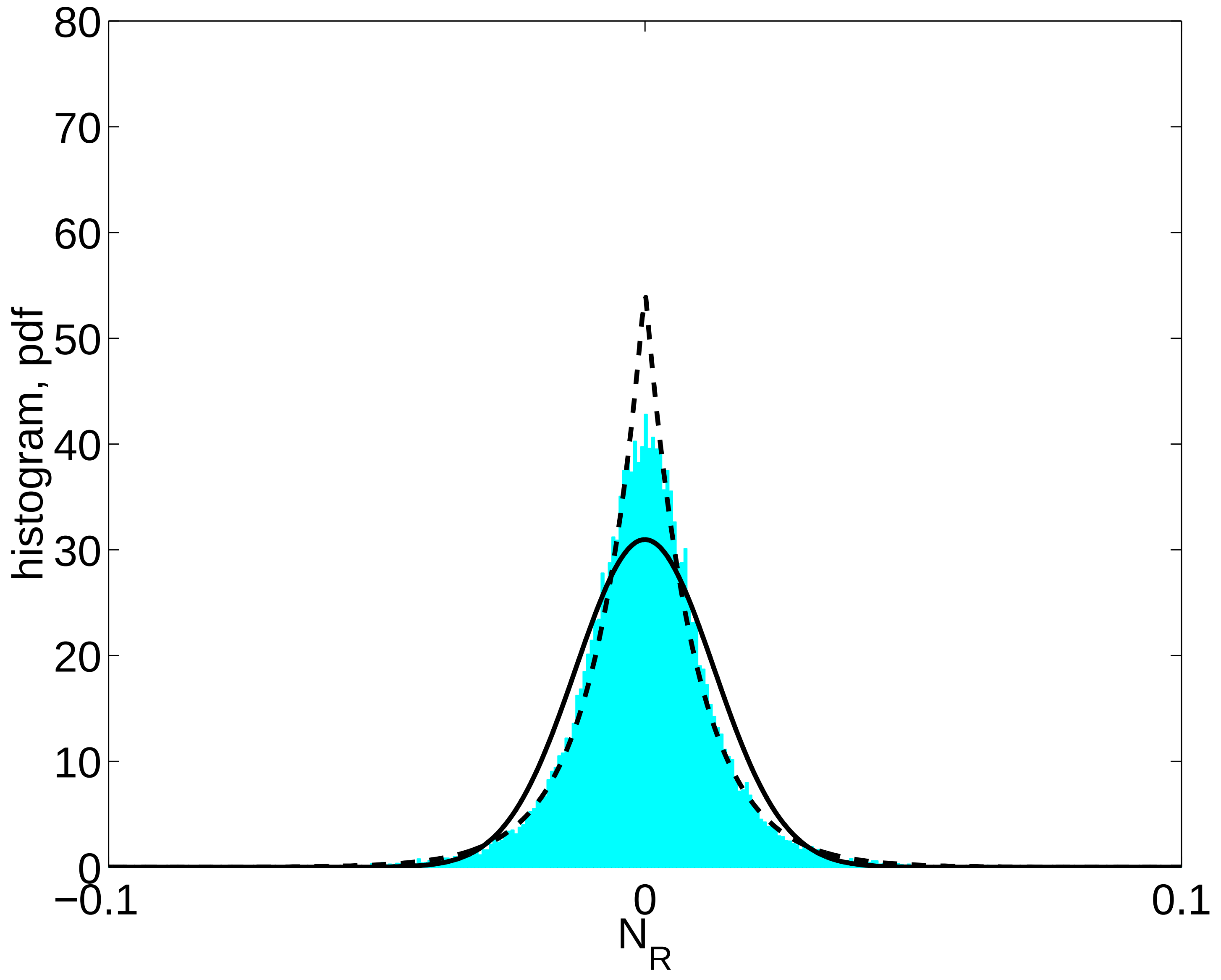
dotted: Gaussian pdf　　dashed: Laplacian pdf　　solid: Gamma pdf

Histogram of Speech Coefficients (enlarged)



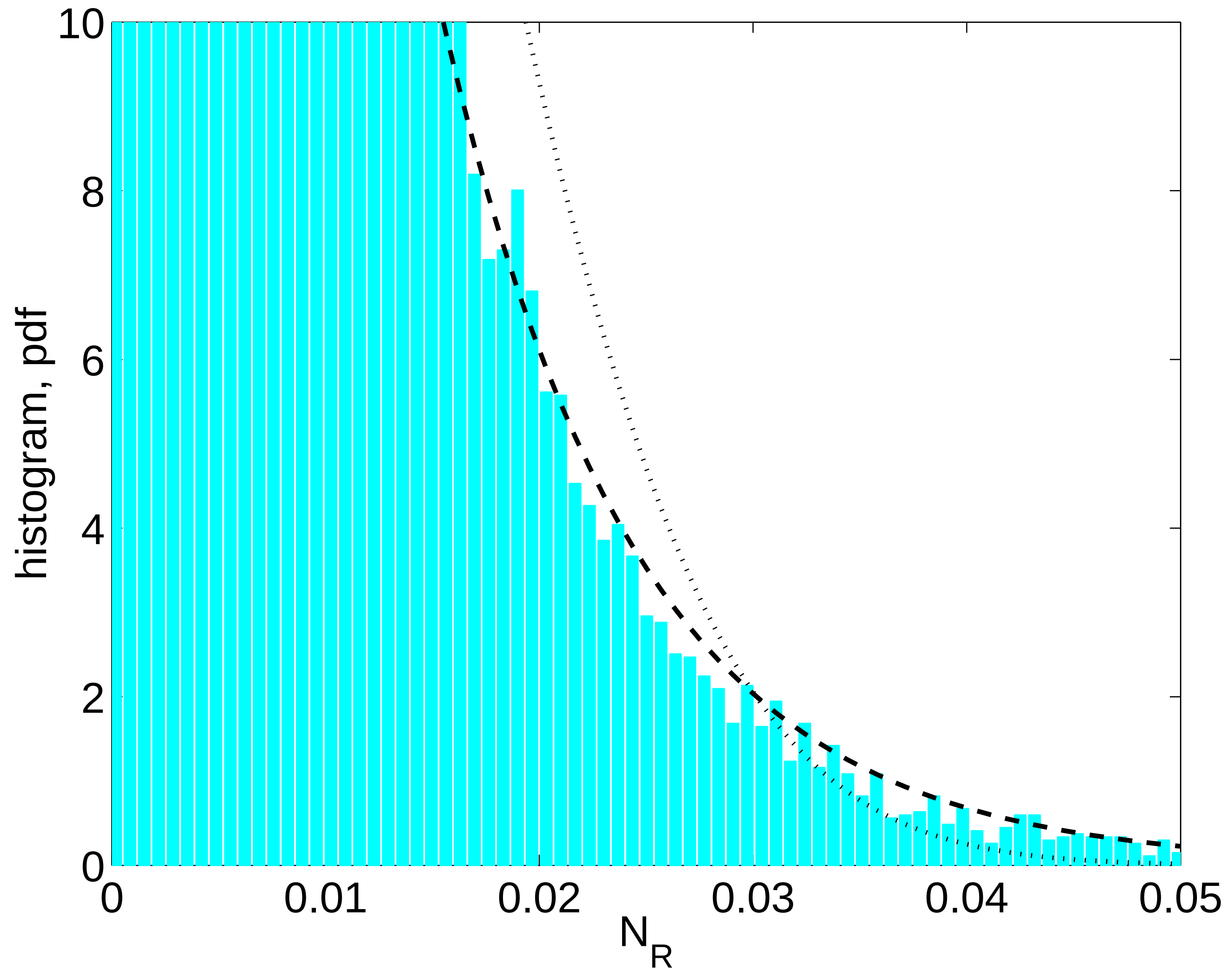
dotted: Gaussian pdf   dashed: Laplacian pdf   solid: Gamma pdf

Histogram of DFT Coefficients for Car Noise



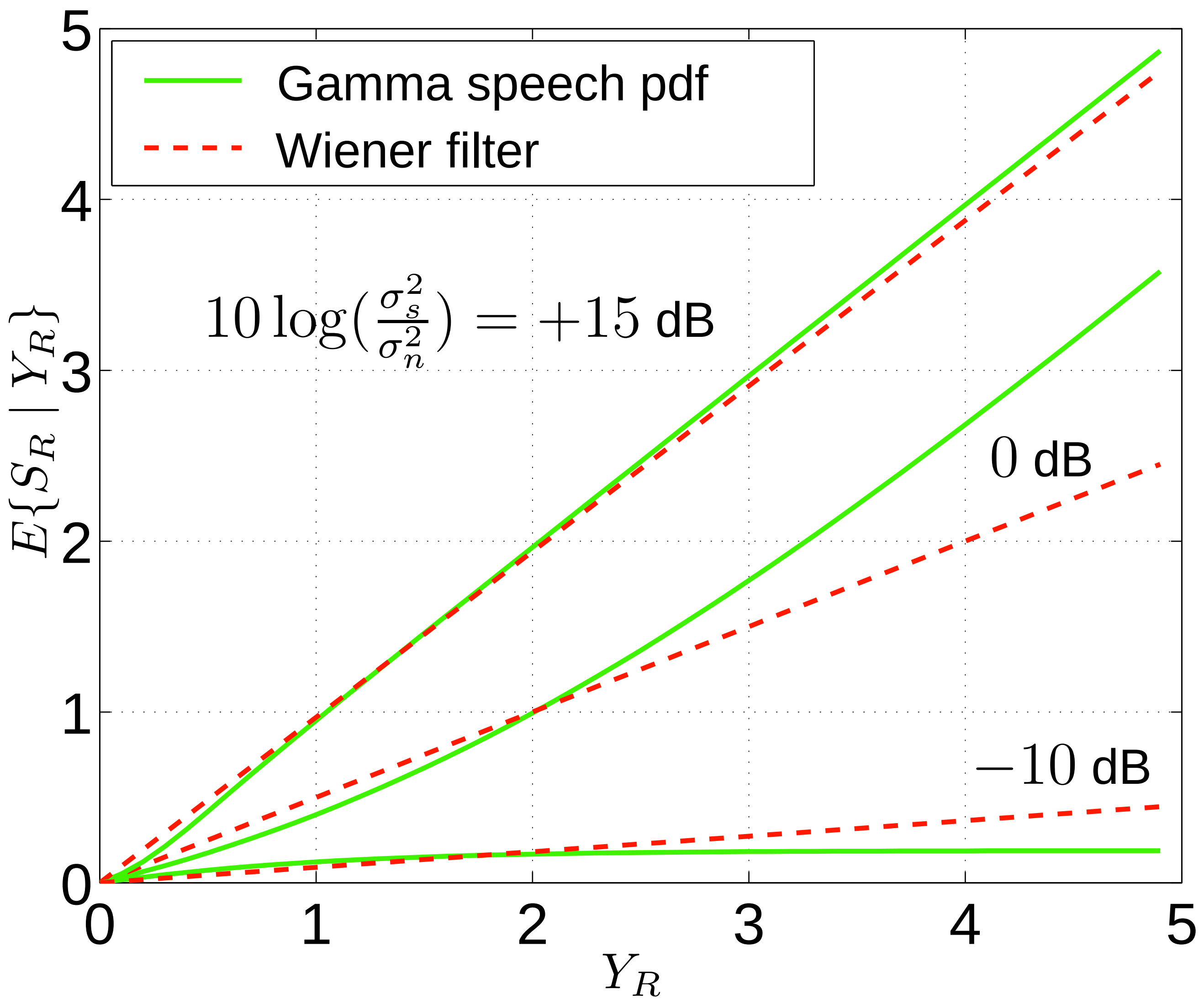
dotted: Gaussian pdf   dashed: Laplacian pdf

Histogram of Car Coefficients (enlarged)



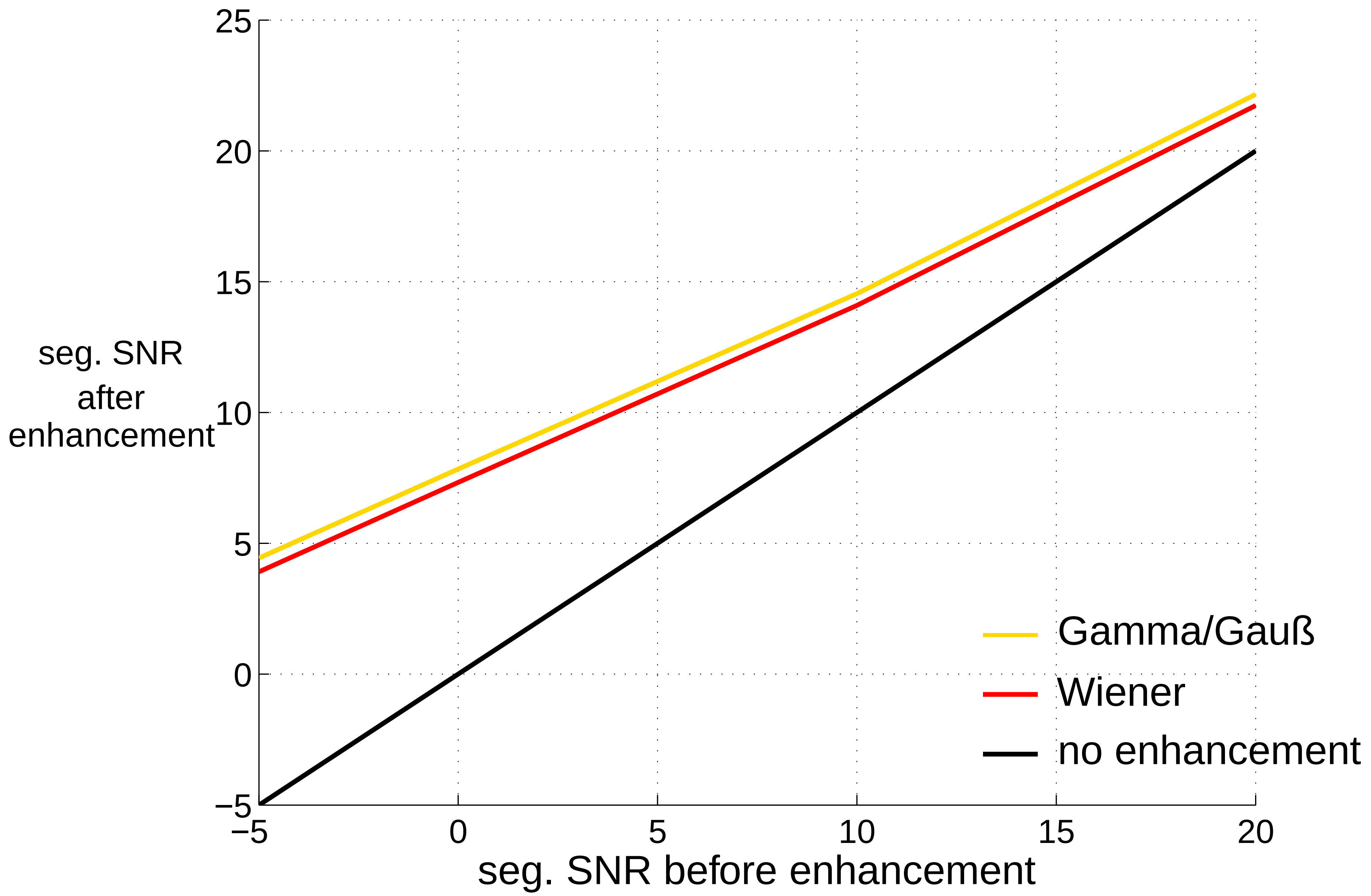
dotted: Gaussian pdf dashed: Laplacian pdf

Non-linear MMSE Estimator

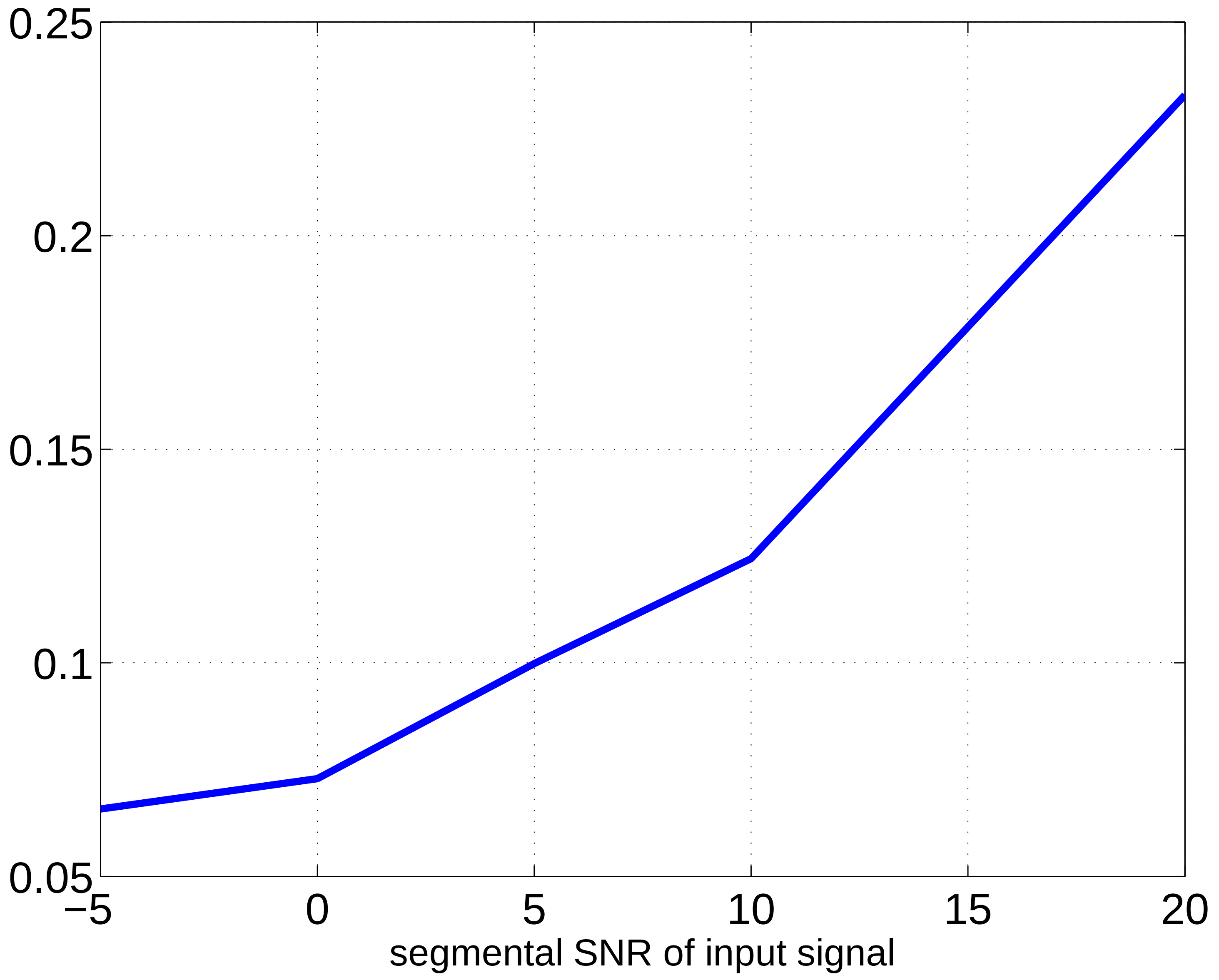


Laplacian Noise and Gamma Speech Prior $\sigma_s^2 + \sigma_n^2 = 2$

Segmental SNR Improvement (White Noise)



Relative Improvement w.r.t. Wiener Filter

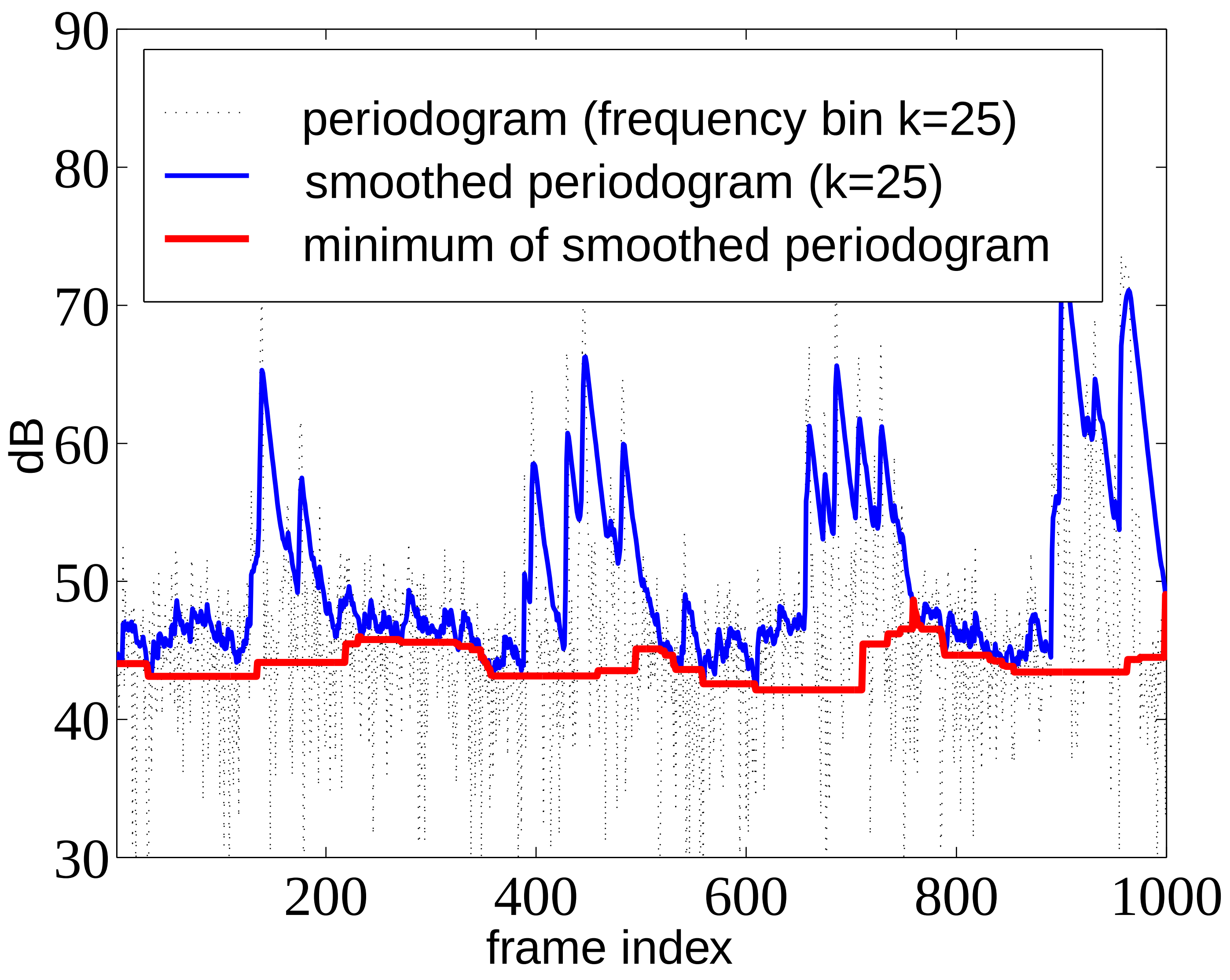


Background Noise PSD Estimation

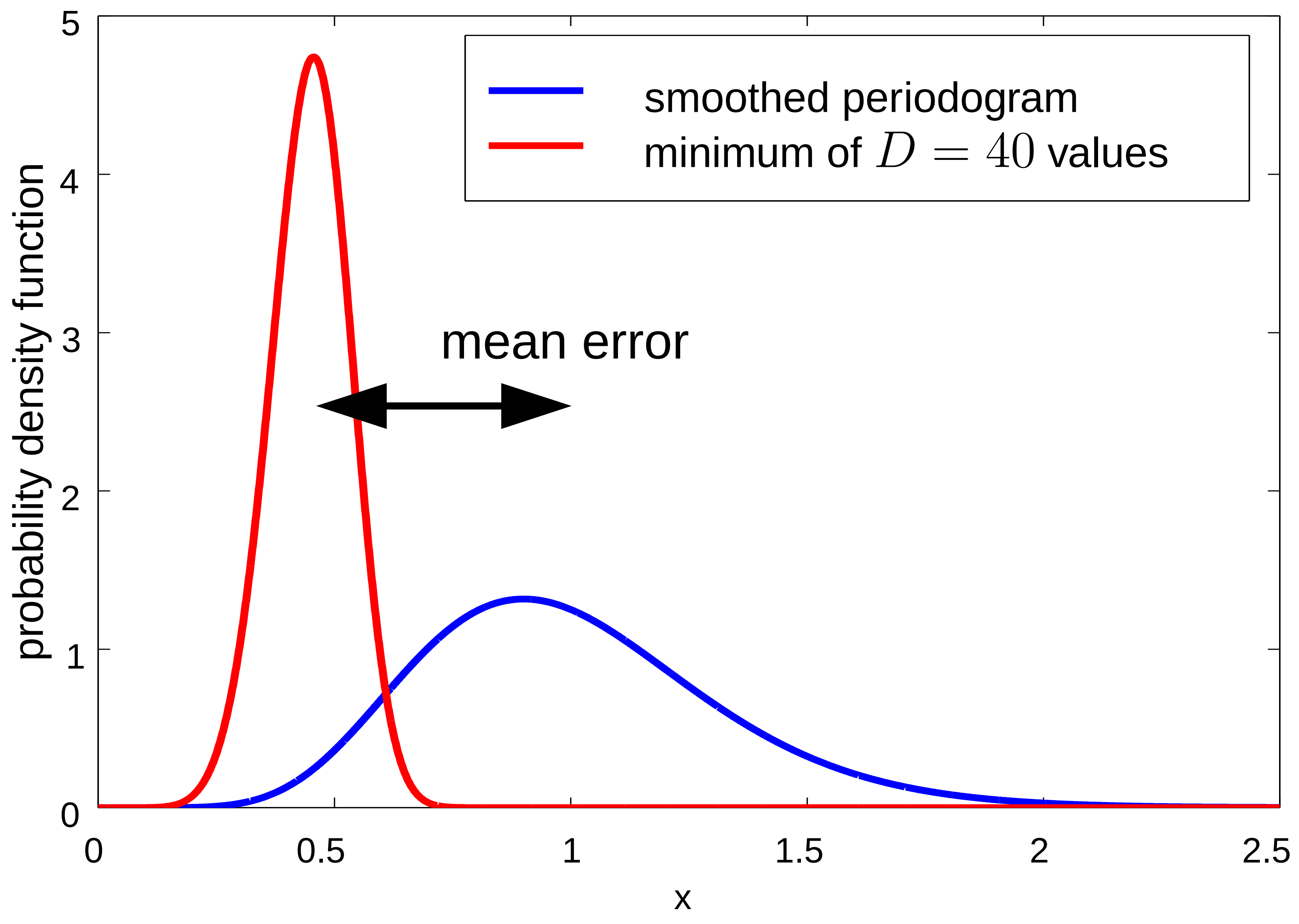
- **Methods:**
  - Voice activity detection;
  - Soft-decision methods;
  - Biased compensated tracking of spectral minima

    [Martin 1994, 2001]
- **Assumptions:**
  - Speech and noise are statistically independent;
  - Speech is not always present;
  - Noise is more stationary than speech.

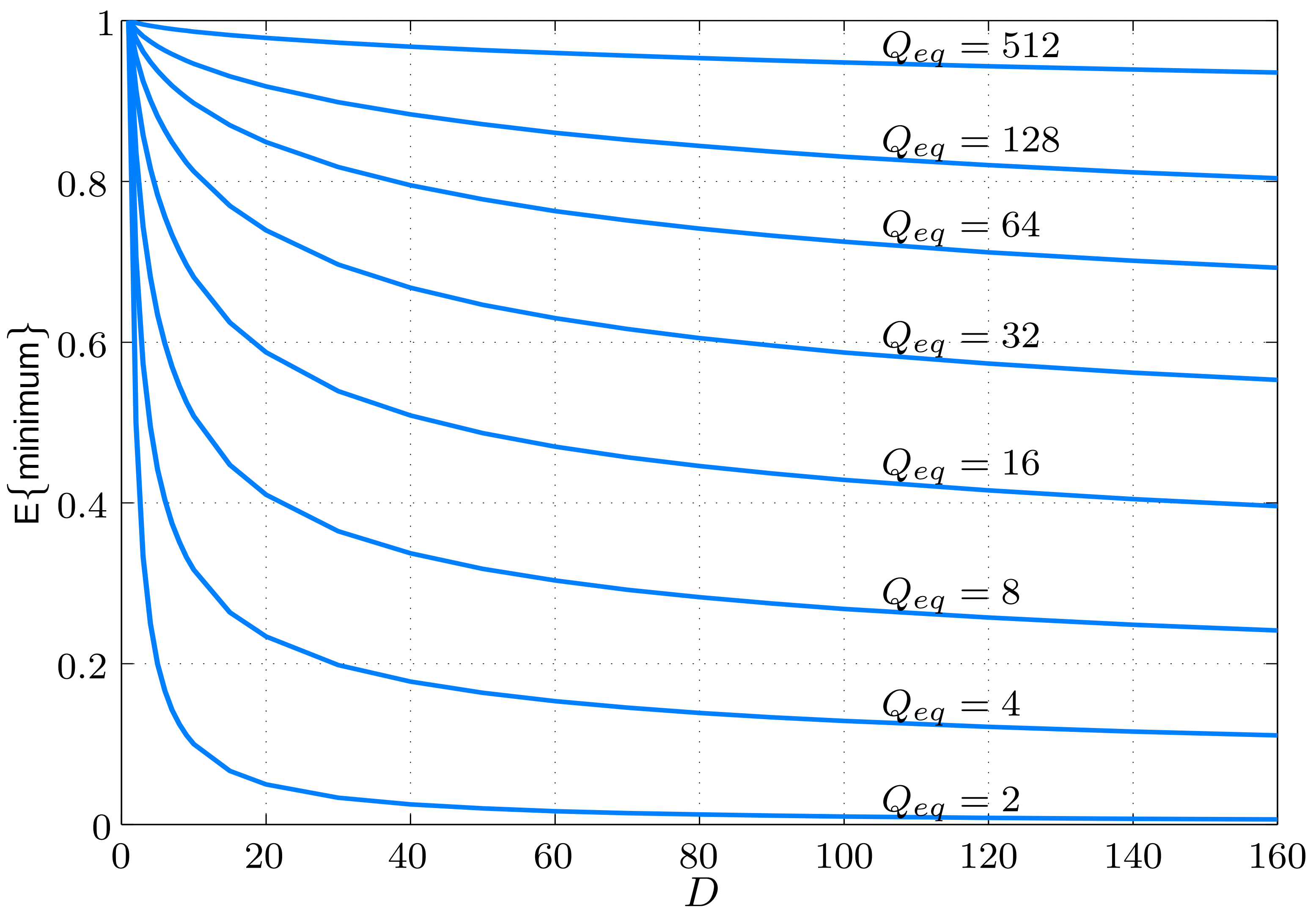
Minimum Statistics: Basic Principle



Minimum Statistics: Bias



Mean of Minimum

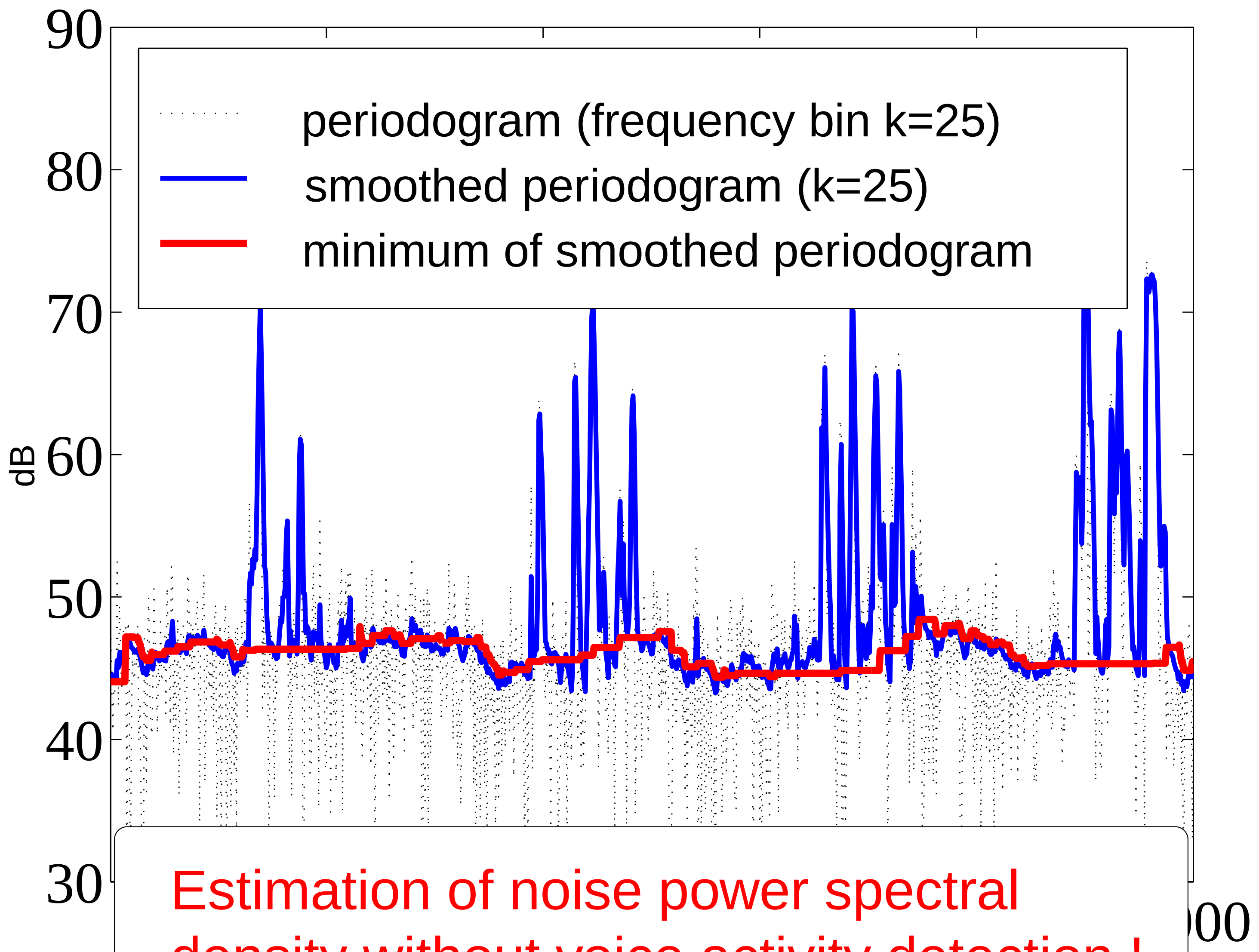


D: length of minimum search window $Q_{eq} = 1/var\{P(\lambda, \Omega_k)\}_{norm}$

Minimum Statistics: What's New ?

- **Minimum Statistic, version 1994**
  - fixed smoothing parameter α
  - fixed bias compensation
- **Minimum Statistic, version 2001**
  - signal dependent optimal smoothing
  - signal dependent bias compensation
  - fast minimum update

Minimum Statistics (version 2001)



Estimation of noise power spectral density without voice activity detection !

Relative Estimation Error

- **Speech pause:**

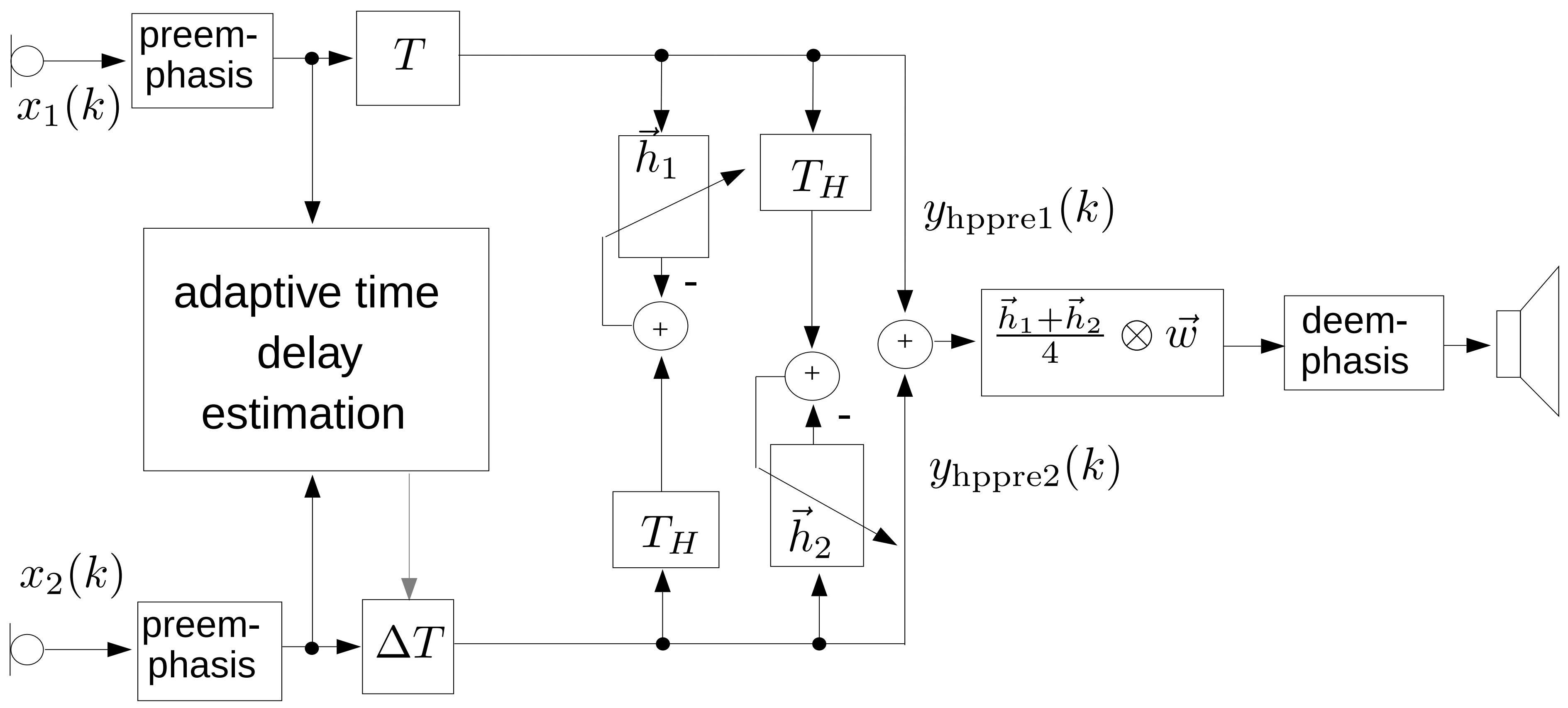
Algorithms	white noise	vehicular noise	street noise
MinStat 1994 ($\alpha = 0.6$)	0.059 (0.11)	0.062 (0.13)	-0.15 (0.21)
MinStat 2001	-0.006 (0.041)	-0.016 (0.041)	-0.27 (0.13)

(in parentheses: variance of estimation error)

- **Speech activity (3 min without speech pauses):**

Algorithms	white noise	vehicular noise	street noise
MinStat 1994 ($\alpha = 0.6$)	0.64 (0.77)	0.77 (1.04)	0.59 (1.9)
MinStat 2001	-0.04 (0.14)	0.02 (0.17)	-0.20 (0.28)

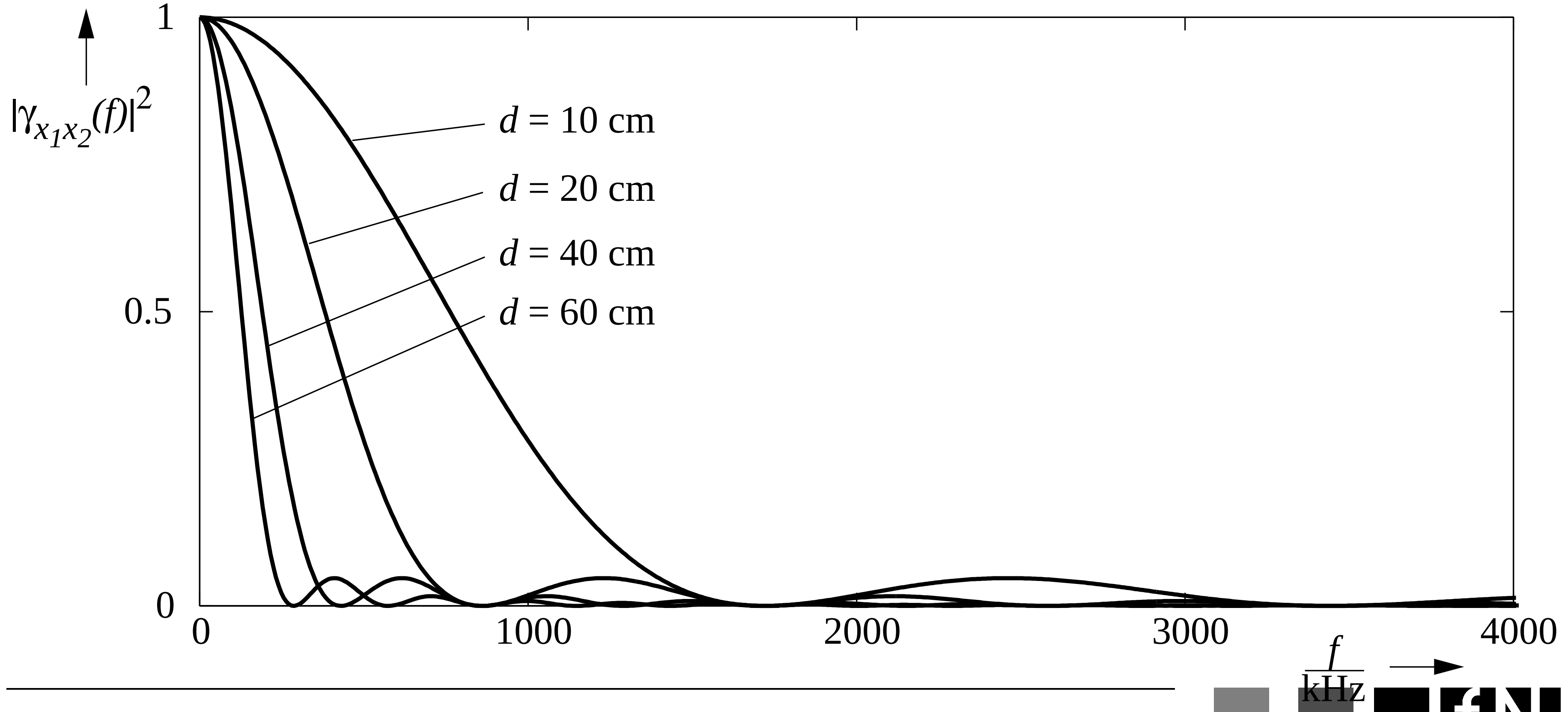
Two Channel Noise Reduction



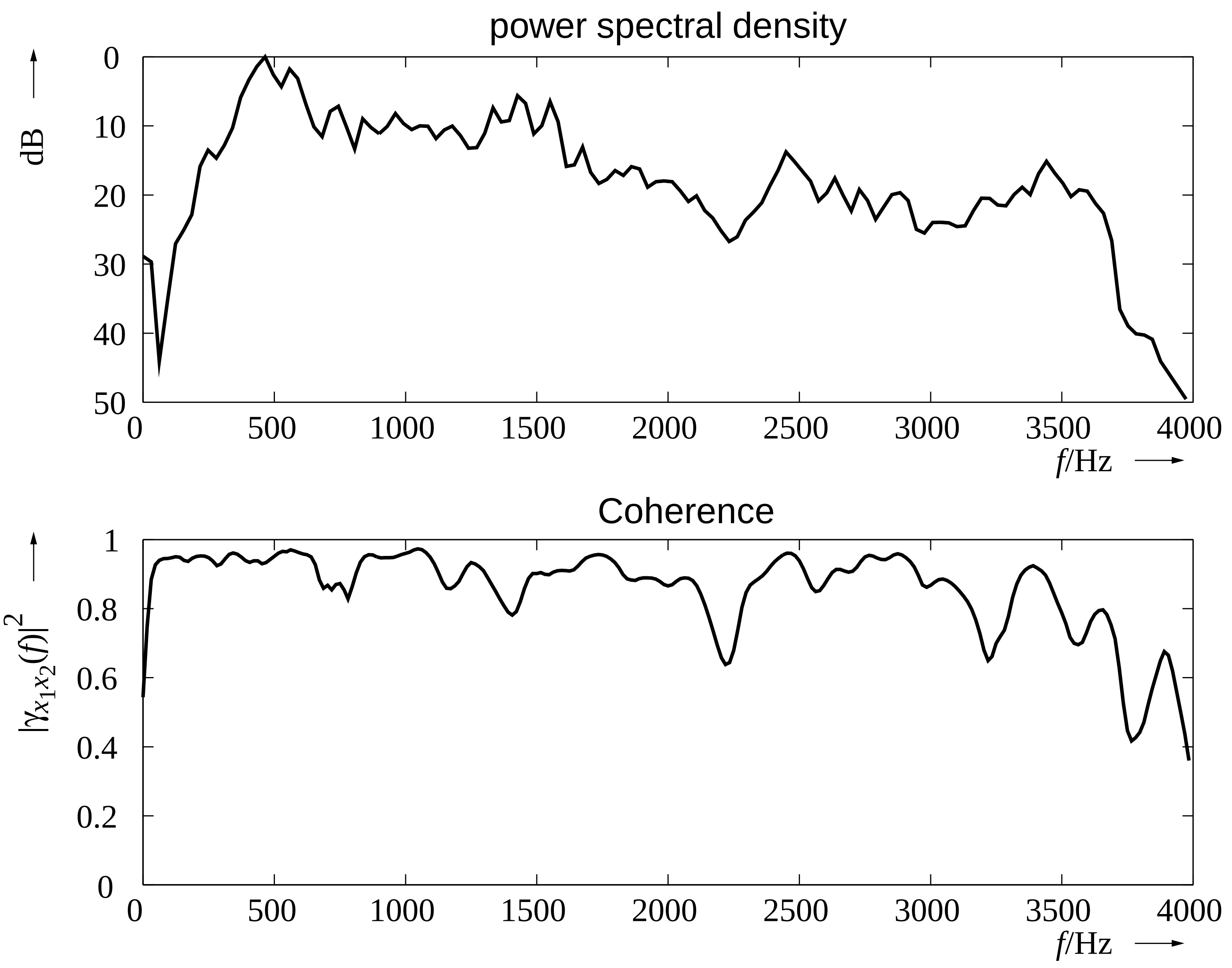
Coherence of Noise (Diffuse Sound Field)

The complex coherence $\gamma_{x_1x_2}(\Omega)$ of two signals $x_1(k)$ and $x_2(k)$ is defined as

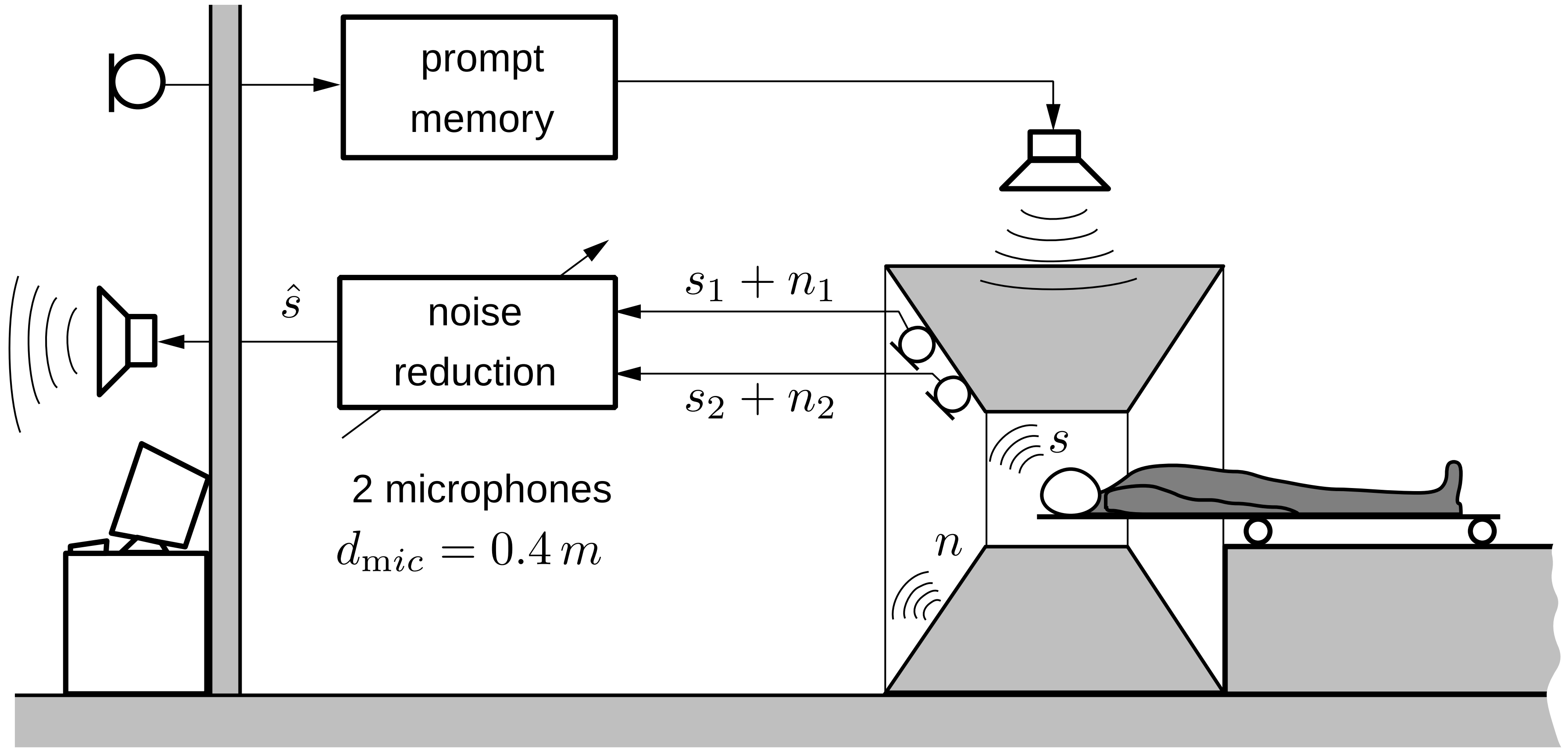
$$\gamma_{x_1x_2}(\Omega) = \frac{\Phi_{x_1x_2}(e^{j\Omega})}{\sqrt{\Phi_{x_1x_1}(e^{j\Omega})\ \Phi_{x_2x_2}(e^{j\Omega})}}\ .$$



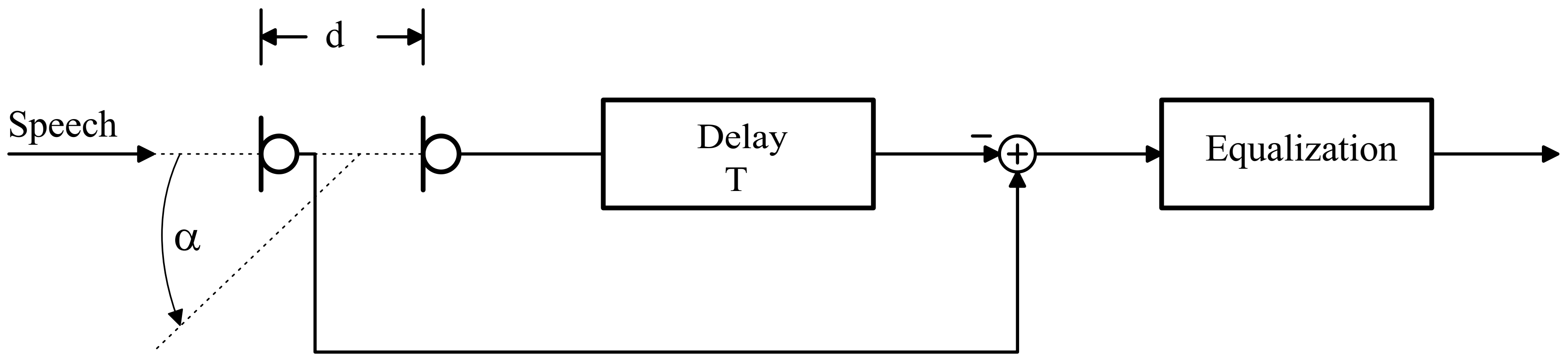
Coherence of Speech in a Car



Two Channel Noise Reduction



First-Order Differential Microfone

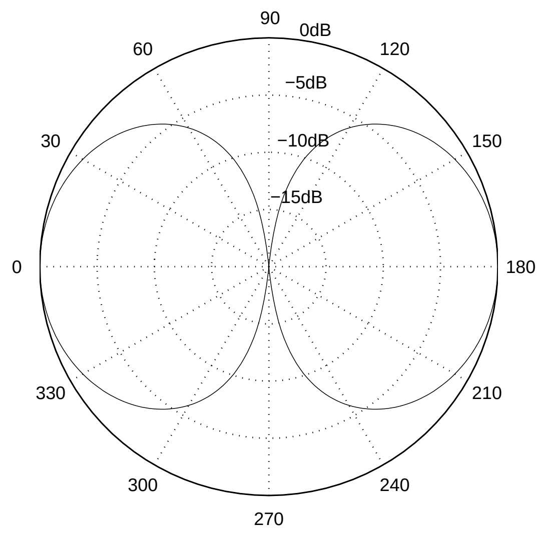


$$Y(j\omega) = S(j\omega)e^{j\omega\left(\frac{d}{2c}\cos(\alpha)\right)}\left[1 - e^{-j\omega\frac{d}{c}\left(\cos(\alpha)+\frac{cT}{d}\right)}\right]$$

$$\left|\frac{Y(j\omega)}{S(j\omega)}\right| = 2\left|\sin\left(\frac{\omega d}{2c}\left(\cos(\alpha) + \frac{cT}{d}\right)\right)\right|$$

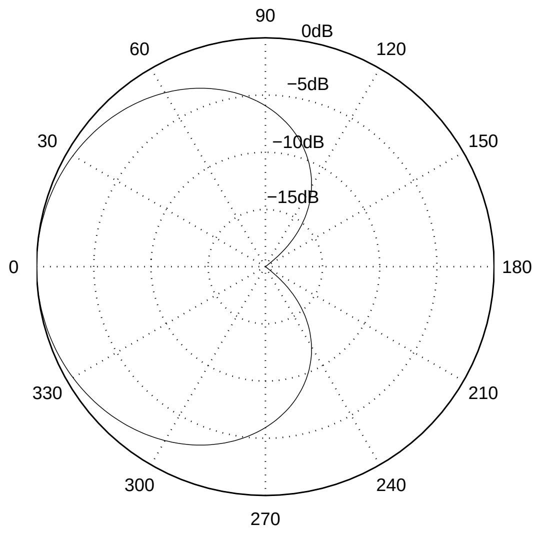
Directivity Patterns ($d = 0.015$ m, $f = 1kHz$)

Dipole (Tc/d = 0), f = 1000 Hz



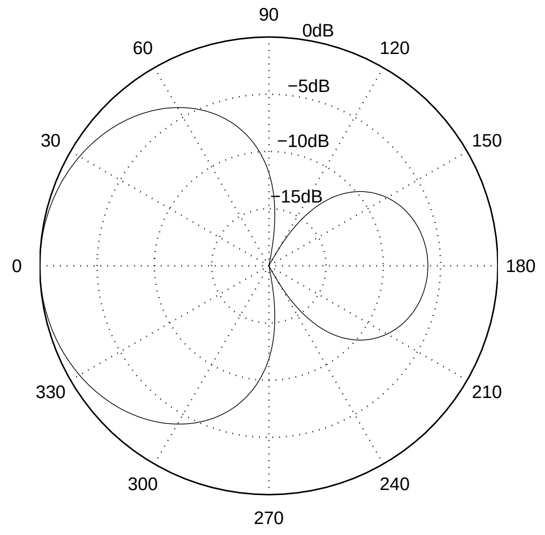
Azimuth angle in degrees

Cardioid (Tc/d = 1), f = 1000 Hz



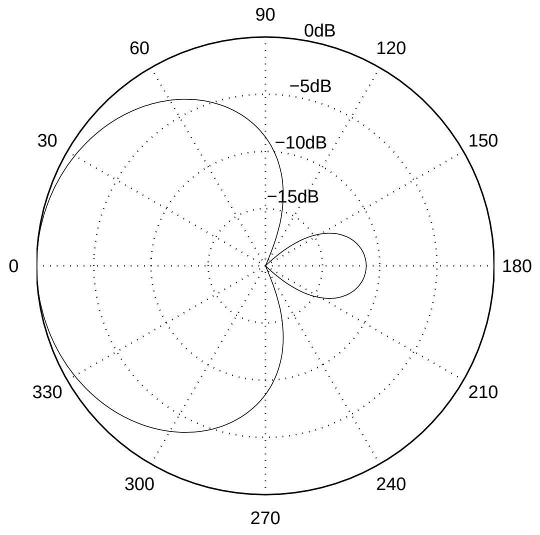
Azimuth angle in degrees

Hyper Cardioid (Tc/d = 0.34), f = 1000 Hz



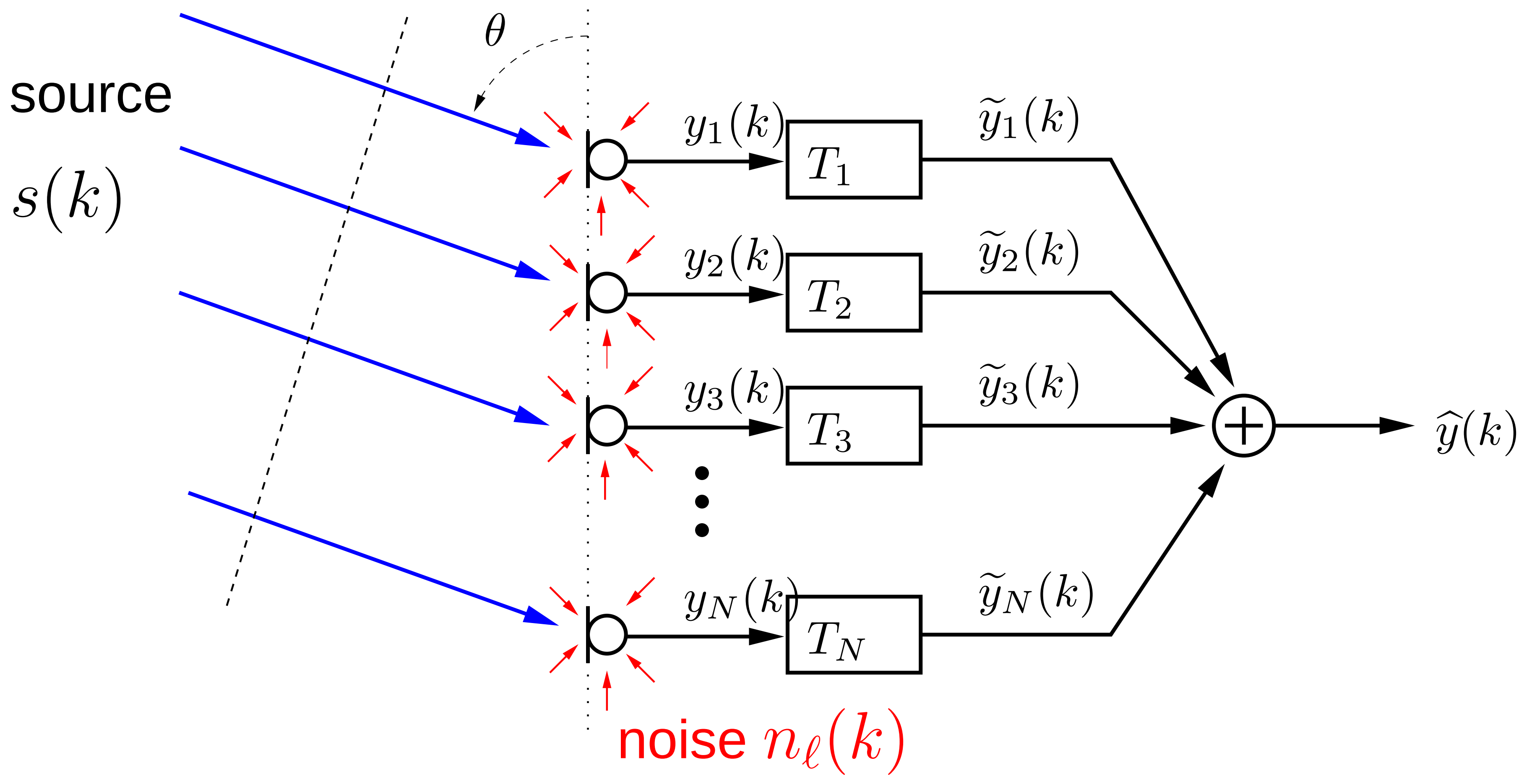
Azimuth angle in degrees

Super Cardioid (Tc/d = 0.57), f = 1000 Hz



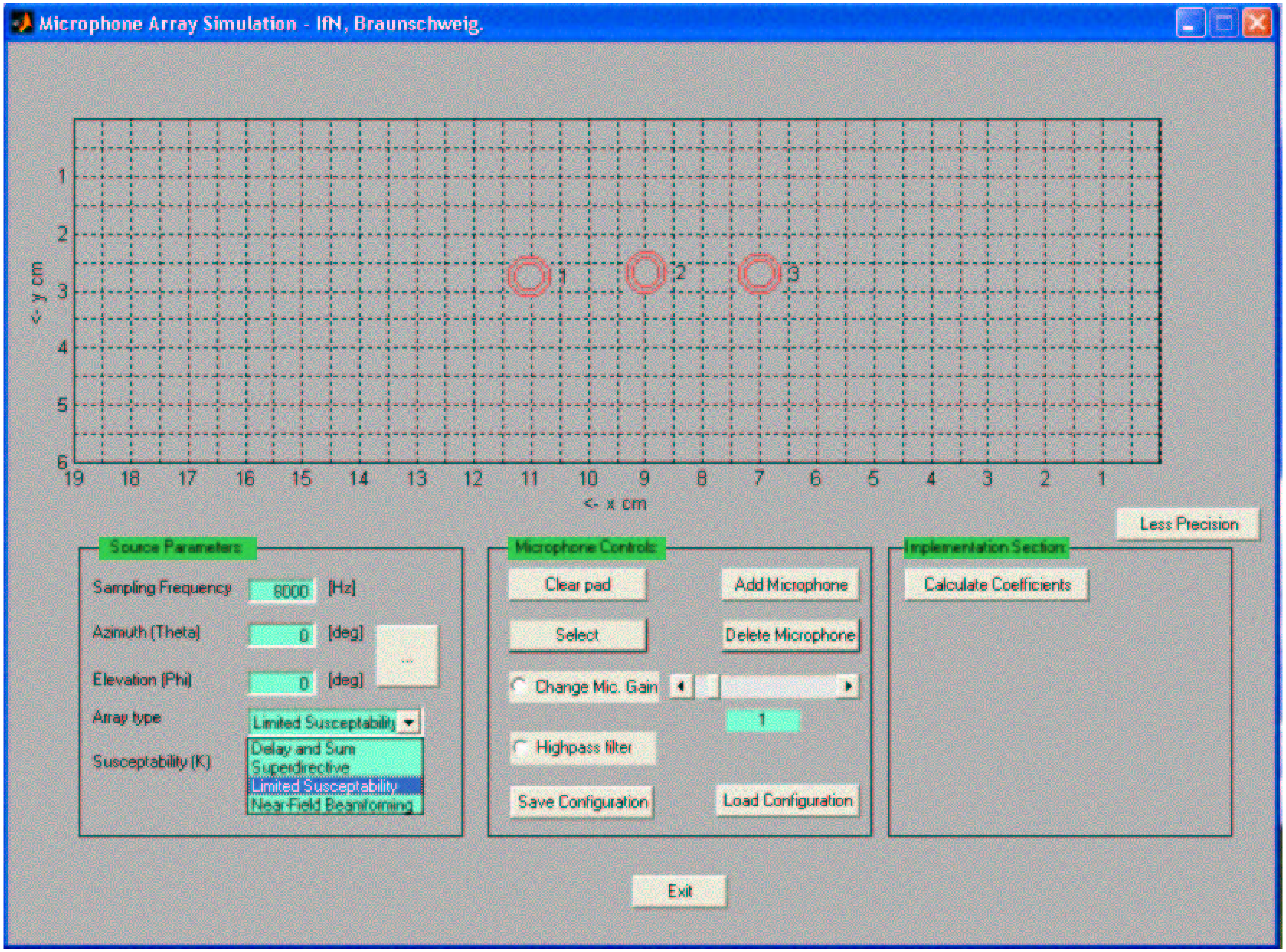
Azimuth angle in degrees

Delay-and-Sum Beamformer



i.i.d. noise: Gain $G = 10 \log(N)$

Design of Fixed Beamformers with MATLAB



Sampling Frequency 8000 [Hz]

Azimuth (Theta) 0 [deg]

Elevation (Phi) 0 [deg]

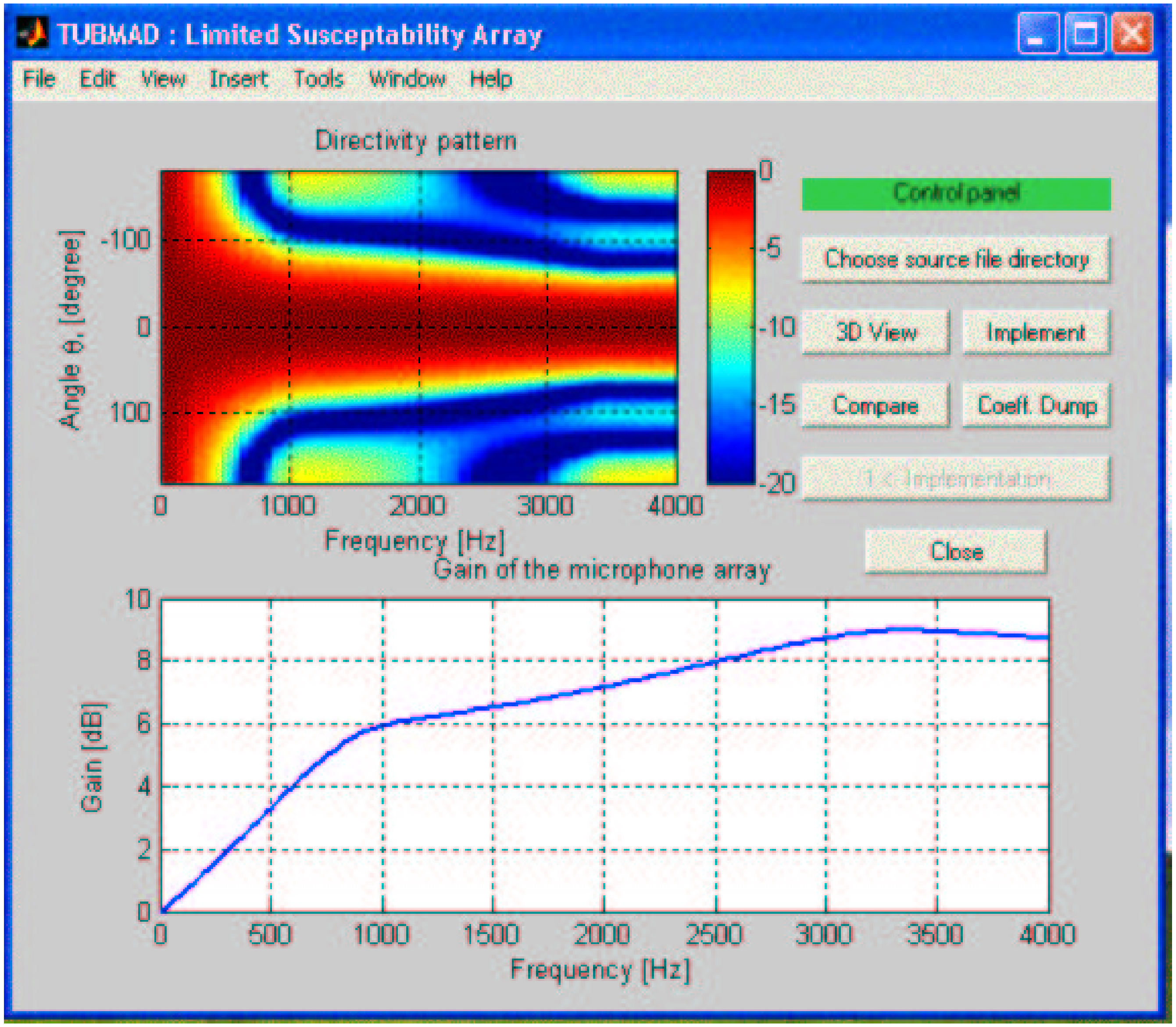
Array type: Limited Susceptibility

- Delay and Sum
- Superdirective
- Limited Susceptibility
- Near-Field Beamforming

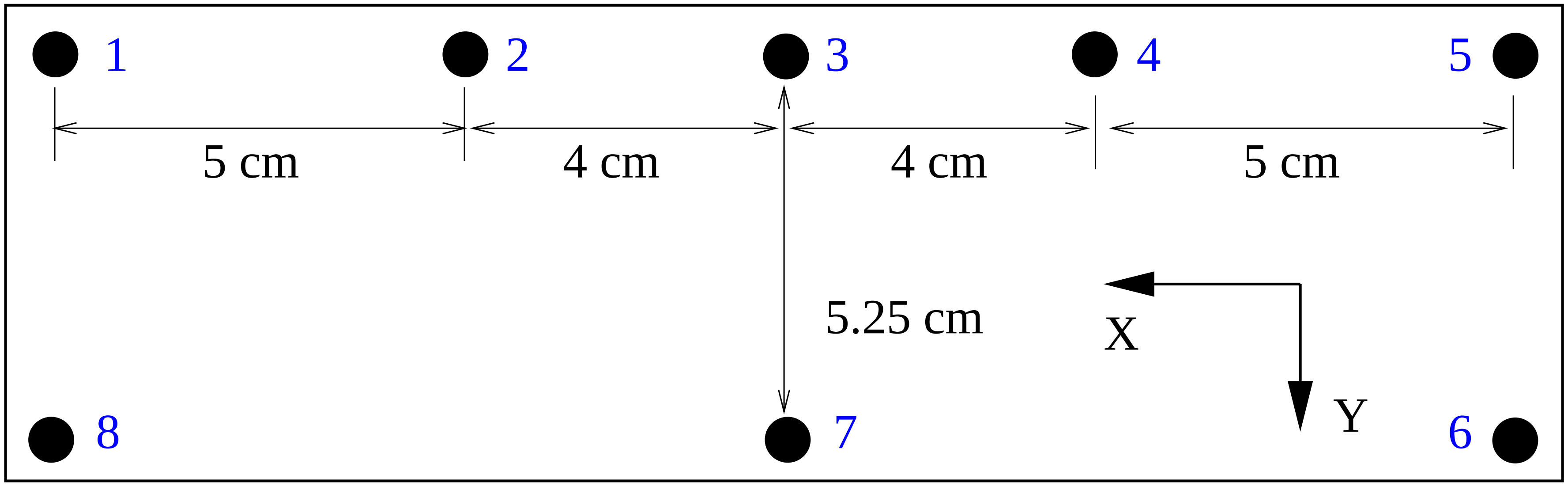
Susceptibility (K)

Change Mic. Gain: 1

Directivity Pattern



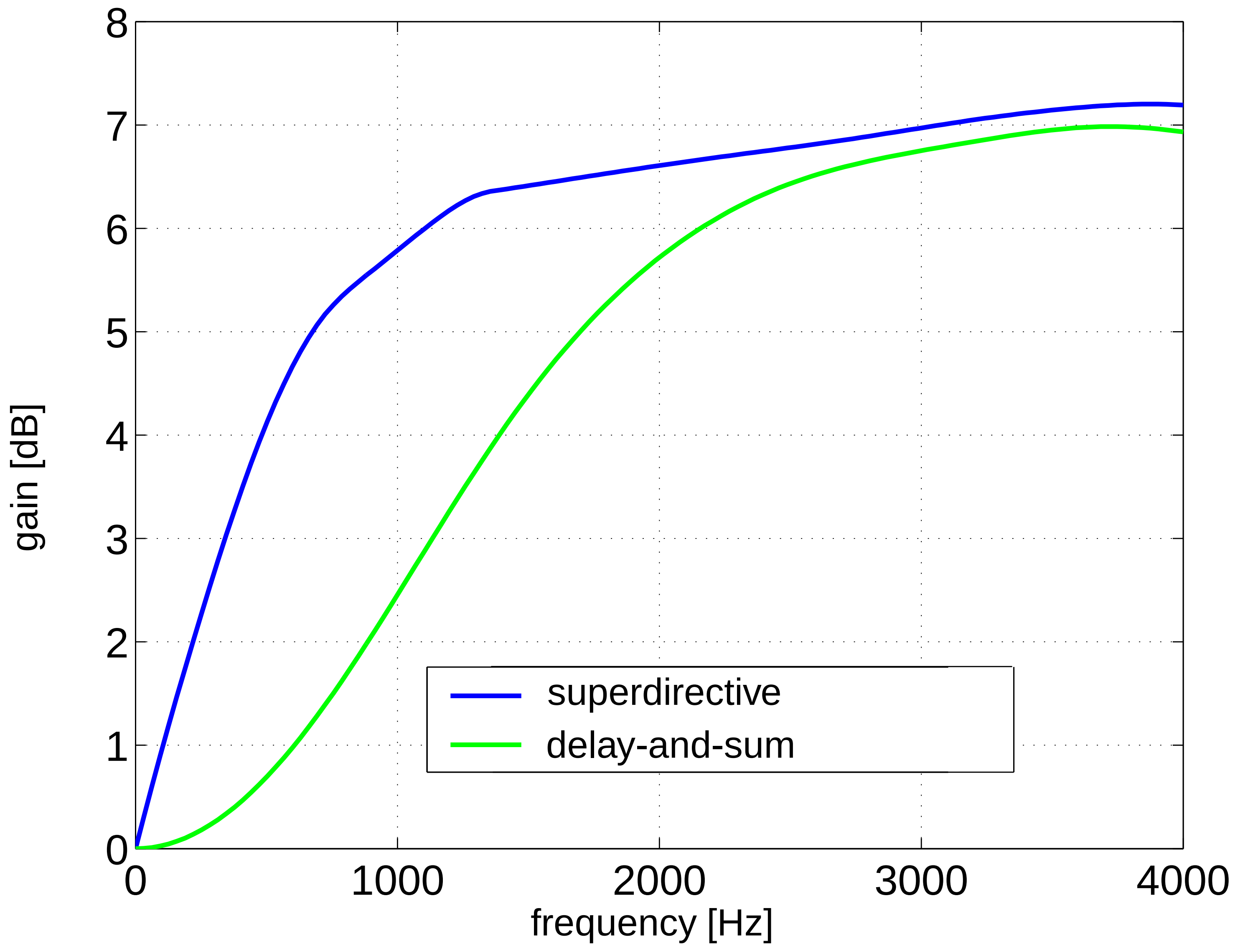
Arrays for Speech Acquisition in Cars



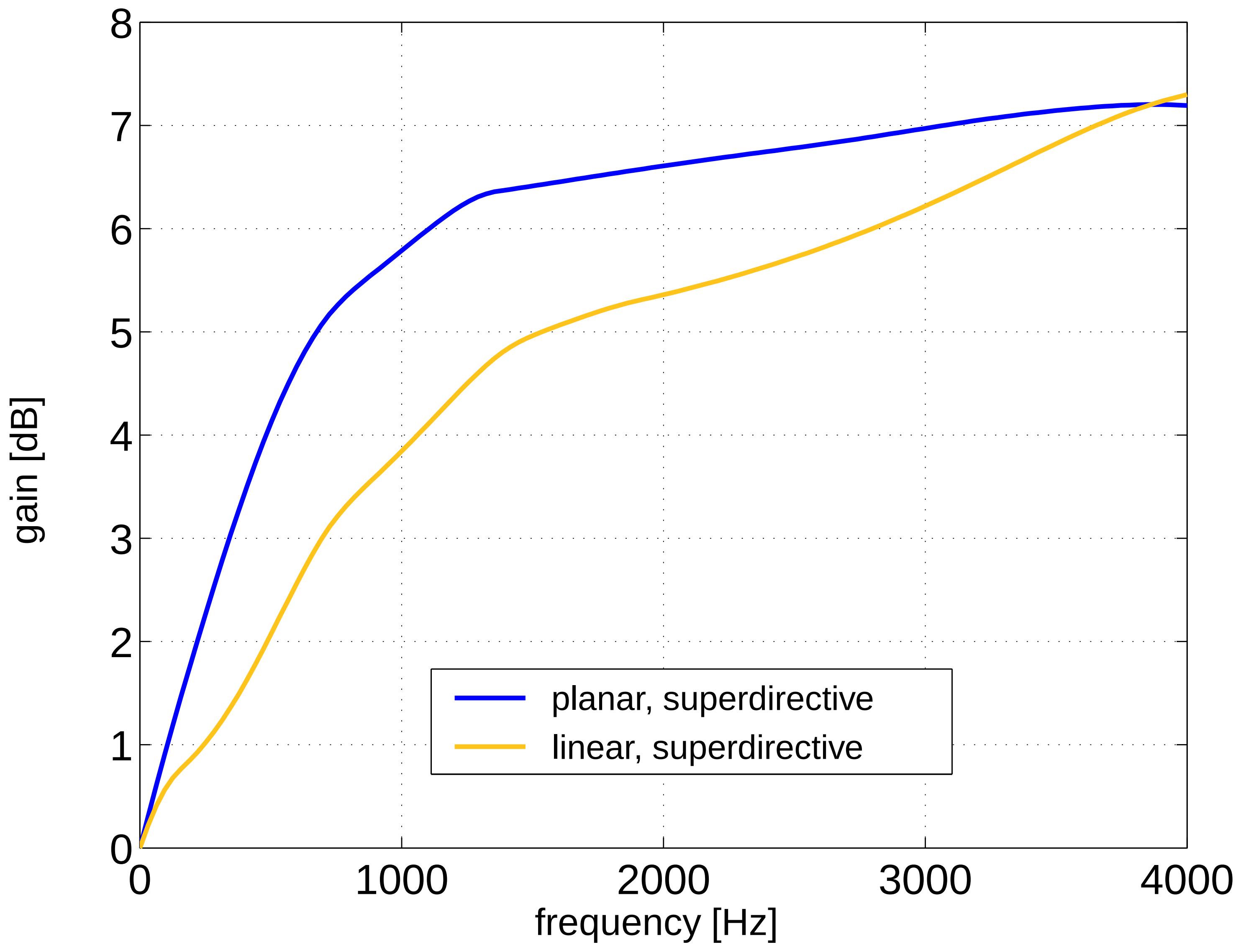
- microphones 1, 2, 3, 4, 5 $\longrightarrow$ linear array
- microphones 1, 2, 7, 4, 5 $\longrightarrow$ planar array

[Martin et al. 2001]

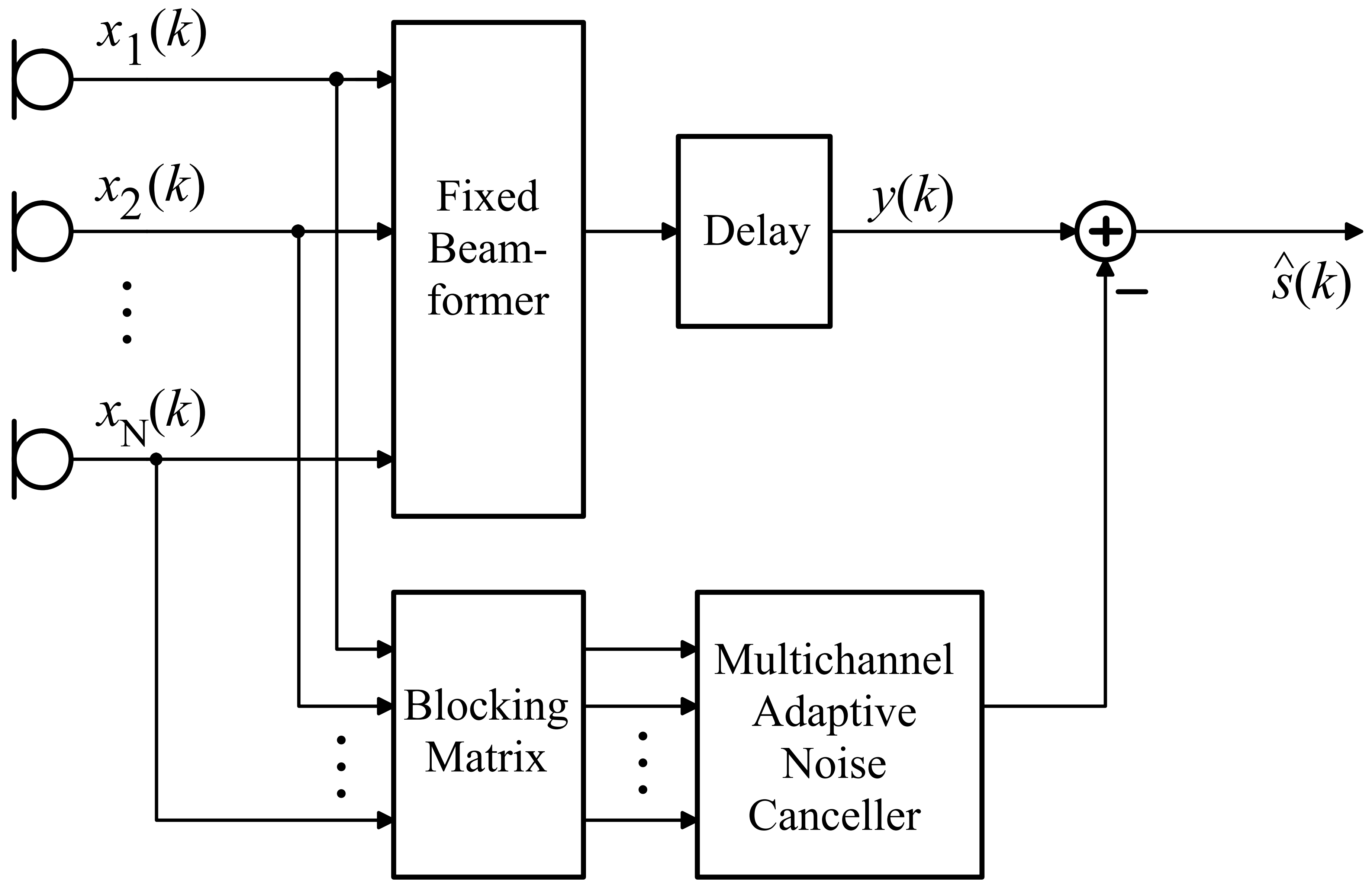
Delay-and-sum vs. Superdirective Arrays



Linear and Planar Microphone Arrays



Adaptive Beamformer (GSC)



Conclusions

- **Find better ways to exploit statistics of signals!**
  - Incorporate models of speech production
  - Develop better background noise estimation methods
  - Design algorithms for high quality and intelligibility
  - Exploit spatial selectivity using multiple microphones
- **Understand processing in the auditory system:**
  - Enhance perceptionally important features
  - Use perceptive models to reduce complexity of algorithms

Selected References