Lecture

Speech and Audio Signal Processing

Lecture 1: Introduction

TECHNISCHE UNIVERSITÄT DARMSTADT

Administrative issues

- **Contents:** In this course topics of audio signal processing are treated.
- **Lecturer:** Prof. Dr.-Ing. Henning Puder, TU Darmstadt,
  Honorary Professor at SPG and
  WSA / Sivantos GmbH (former Siemens Hearing Aids), Erlangen
- **Exercises:** Four exercise tutorials during the lecture: practical Matlab examples.

  Problem formulations will be distributed 1 week before, results will be presented and analyzed in the plenum
- **Time:** Mondays, 8:00 h – 10:30 h (3 x 45 min + break) including the exercises
- **Language:** German or English, Lecture Notes in English

Content of the lecture

- Introduction to the properties of speech and audio signals
- Methods of vector quantization and codebook processing
- **Audio quality measures**, basic methods of audio signal processing
- **Audio coding**:
  - Predictive coding (speech coding for mobile transmission, e.g., GSM, CELP coder)
  - Subband / Frequency domain coders (e.g., MP3, AAC)
- **Noise reduction:** classic and DNN (deep neural network) based
- **Beamforming** with multi-microphone setups
- **Cepstral processing & pitch estimation**; “Mel frequency cepstral coefficients” (MFCC)
- “Hidden Markov Models” (HMM)
- **Acoustic classification methods**: Bayes, Gaussian mixture (GMM) and DNN methods
- Applications of MFCC, GMM & HMM for **speech and speaker recognition**
- **Music signal processing**, e.g., beat detection and music retrieval (Shazam)
- Loudspeaker sound reproduction systems: Wave-field synthesis (WFS), Higher order ambisonics (HOA)

Administrative issues

- **Credit points:**
  - 6 credit points for Master students

- **Desirable pre-requisites:**
  - Digital signal processing & basics in “Adaptive Filters”

- **Exam:**
  - Oral, ca. 20 min. per student
  - **Two dates planned: Feb. 2025 after lecture period and April 2025 before SoSe**

- **Seminar presentation :**
  - Mandatory, based on a scientific paper study, ca. 15 min. per student, or couple of 2 students (25 min. together).
  - **Two dates fixed: Dec. 2, 2024 or Jan. 27, 2025**
  - Can improve the mark of oral exam (no degradation possible)
  - Topic selection during the semester, now already possible

Literature (I)

TECHNISCHE UNIVERSITÄT DARMSTADT

- Statistische Signaltheorie:
  - E. Hänsler: *Statistische Signale: Grundlagen und Anwendungen*, Springer, 2001
  - A. Papoulis: *Probability, Random Variables, and Stochastic Processes*, McGraw-Hill, 1965
- Noise reduction, beamforming, adaptive Filter:
  - E. Hänsler, G. Schmidt: *Acoustic Echo and Noise Control*, Wiley, 2004
  - S. Haykin: *Adaptive Filter Theory*, Prentice Hall, 2002
  - A. Sayed: *Fundamentals of Adaptive Filtering*, Wiley, 2004

- Examples for speech signal processing:
  - E. Hänsler, G. Schmidt: *Topics in Acoustic Echo and Noise Control*, Springer, 2006
  - B. Iser, et al.: *Bandwidth Extension of Speech Signals*, Springer, 2008
  - E. Hänsler, G. Schmidt: *Speech and Audio Processing in Adverse Environments*, Springer, 2008
  - J. Benesty, et al.: *Speech Enhancement*, Springer, 2005

Literature (II)

- Speech signal processing:
  - L. R. Rabiner, R. W. Schafer: Digital Processing of Speech Signals, Prentice Hall, 1978
  - P. Vary, U. Heute, W. Hess: Digitale Sprachsignalverarbeitung, Teubner, 1998
  - P. Vary, R. Martin: Digital Speech Transmission, Wiley, 2006
  - L. R. Rabiner, R. W. Schafer: Introduction to Digital Speech Processing, Now, 2008
  - B. Pfister, T. Kaufman: Sprachverarbeitung, Springer, 2008

- Audio signal processing:
  - U. Zölzer: DAFX – Digital Audio Effects, Wiley, 2002
  - E. Larsen, R. M. Aarts: Audio Bandwidth Extension, Wiley, 2004
  - M. Talbot-Smith: Audio Engineer‘s Reference Book, Focal Press, 1998

Content of the first lecture

- Notations

- Speech signal analysis
  - Human speech generation
  - Acoustic signal propagation
  - Acoustic signal perception => The human ear

- Sample based vs. block-based processing

- Basic processing schemes
  - Power estimation
  - Non-linear smoothing
  - Minimum power / noise power estimation
  - Speech activity detection
  - Short-Term Fourier Transform (STFT)
  - Power Spectral Density (PSD) estimation

Notation: Part 1

TECHNISCHE UNIVERSITÄT DARMSTADT

- Scalar notation:
  - Signals: $x(n)$ — *Discrete time index*
  - Impulse responses (time varying): $h_i(n)$ — *Coefficient index*
  - Example for a (real-value) convolution: $$y(n) = \sum_{i=0}^{N-1} x(n-i)\, h_i(n)$$
- Vector notation:
  - Signal vectors: $\boldsymbol{x}(n) = \Big[x(n),\, x(n-1),\, \ldots,\, x(n-N+1)\Big]^{\mathrm{T}}$ — *Bold, lower case*
  - Vectors of impulse responses (time varying): $\boldsymbol{h}(n) = \Big[h_0(n),\, h_1(n),\, \ldots,\, h_{N-1}(n)\Big]^{\mathrm{T}}$
  - Example for a (real-value) convolution : $$y(n) = \boldsymbol{x}^{\mathrm{T}}(n)\, \boldsymbol{h}(n) = \boldsymbol{h}^{\mathrm{T}}(n)\, \boldsymbol{x}(n)$$
- Matrices:

$$\boldsymbol{A}(n) = \begin{bmatrix} a_{00}(n) & a_{01}(n) & \ldots & a_{0N}(n) \\ a_{10}(n) & a_{11}(n) & \ldots & a_{1N}(n) \\ \vdots & \vdots & & \vdots \\ a_{M0}(n) & a_{M1}(n) & \ldots & a_{MN}(n) \end{bmatrix}$$

Bold, upper case

Notation: Part 2

- Random processes („Ensemble of signals“):

  No differentiation in notation of deterministic signals and random processes – other notations: $\boldsymbol{x}(\eta, n),\ x(\omega, n),\ X(n)$

  - Notation: $x(n),\ x_1(n),\ x_2(n)$
  - Probability density function: $f_x(x, n),\ f_{x_1 x_2}(x_1, x_2, n_1, n_2)$
  - Stationary random processes:

$$
\begin{aligned}
f_x(x,\, n) &= f_x(x,\, n + n_0) = f_x(x) \\
f_{x_1 x_2}(x_1,\, x_2,\, n_1,\, n_2) &= f_{x_1 x_2}(x_2,\, x_2,\, n_1 + n_0,\, n_2 + n_0) = f_{x_1 x_2}(x_1,\, x_2,\, n_2 - n_1)
\end{aligned}
$$

  - Expectation values for stationary random processes:

$$
\begin{aligned}
m_x^{(1)}(n) = \mathrm{E}\Big\{x(n)\Big\} &= \int\limits_{x=-\infty}^{\infty} x\, f_x(x, n)\, dx \\
m_x^{(2)}(n) = \mathrm{E}\Big\{x^2(n)\Big\} &= \int\limits_{x=-\infty}^{\infty} x^2\, f_x(x, n)\, dx \\
\mathrm{E}\Big\{g\big(x(n)\big)\Big\} &= \int\limits_{x=-\infty}^{\infty} g(x)\, f_x(x, n)\, dx
\end{aligned}
$$

Notation: Part 3

- Auto- und cross-correlation for real-value, stationary random processes:
  - Auto-correlation function:

$$\mathrm{E}\Big\{x(n)\,x(n+l)\Big\} \;=\; r_{xx}(l)$$

  - Cross-correlation function:

$$\mathrm{E}\Big\{x(n)\,y(n+l)\Big\} \;=\; r_{xy}(l)$$

  - Auto power spectral density:

$$S_{xx}(\Omega) \;=\; \sum_{l=-\infty}^{\infty} \mathrm{E}\Big\{x(n)\,x(n+l)\Big\}\,e^{-j\Omega\,l} \;=\; \sum_{l=-\infty}^{\infty} r_{xx}(l)\,e^{-j\Omega\,l}$$

  - Cross power spectral density:

$$S_{xy}(\Omega) \;=\; \sum_{l=-\infty}^{\infty} \mathrm{E}\Big\{x(n)\,y(n+l)\Big\}\,e^{-j\Omega\,l} \;=\; \sum_{l=-\infty}^{\infty} r_{xy}(l)\,e^{-j\Omega\,l}$$

Notation: Part 4

- Stationary, white noise:
  - Auto-correlation function:

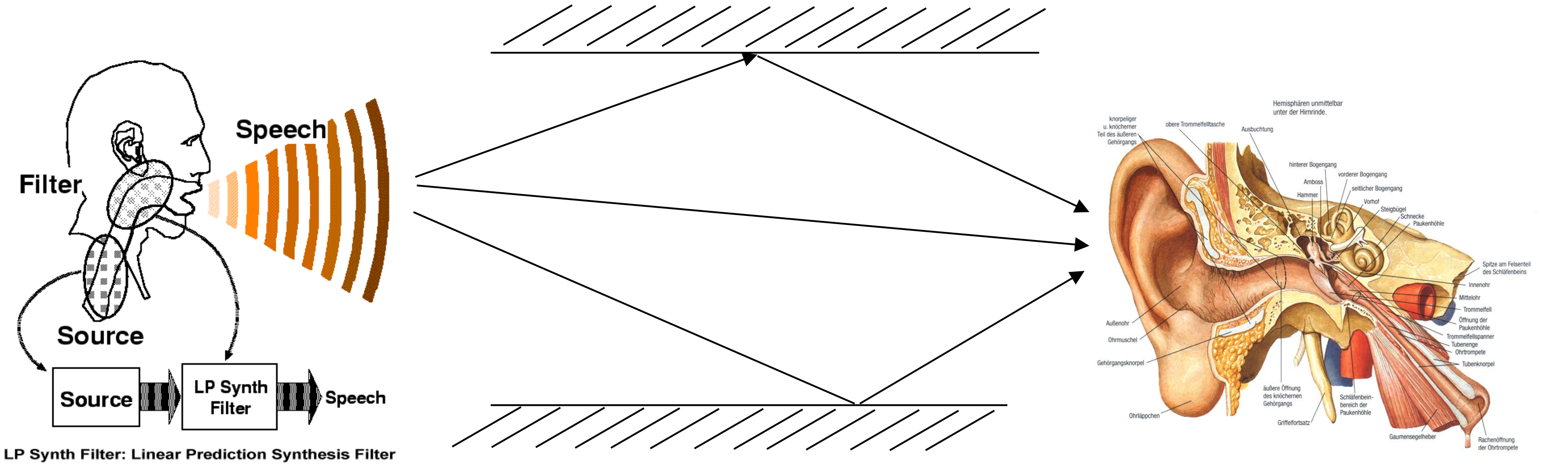
$$r_{xx}(l)\Big|_{\text{white noise}} = \sigma_x^2\,\delta_K(l) = \begin{cases} \sigma_x^2, & \text{if } l = 0, \\ 0, & \text{else} \end{cases}$$

  - Auto power spectral density:

$$S_{xx}(\Omega)\Big|_{\text{white noise}} = \sigma_x^2$$

- Literature:
  - E. Hänsler: *Statistische Signale: Grundlagen und Anwendungen, Kapitel 3 – Zufallsprozesse,* Springer, 2001
  - A. Zoubir: *Digital Signal Processing, Chapter 7 – Random Variables and Stochastic Processes*, Vorlesungsskript, Darmstadt, 2005

Speech signal analysis

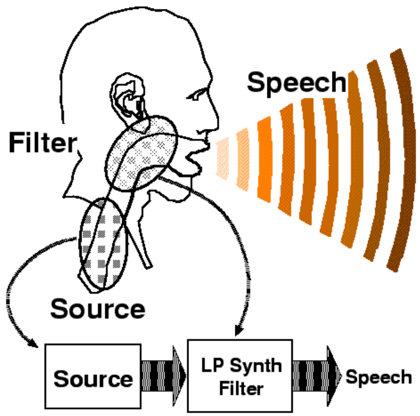


- **Speech generation**

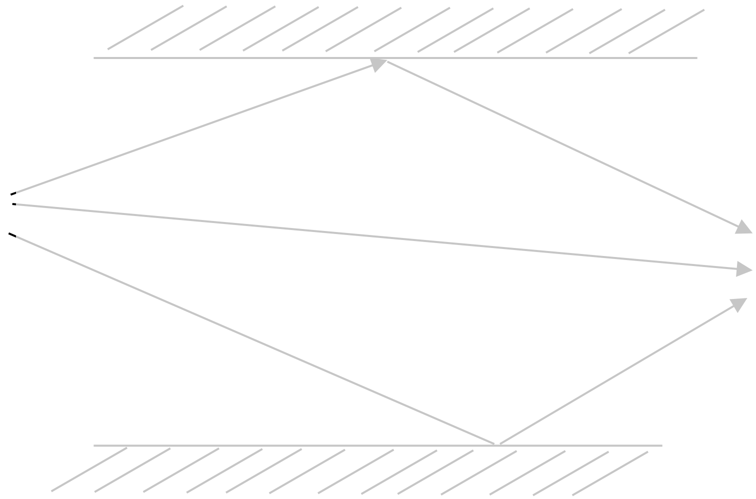
- **Speech signal propagation:**
  - direct path and reflections in the acoustic environment
  - head shading and different time of arrival at the human ears

- **Speech perception:**
  - outer, middle and inner ear effects

Speech signal analysis



LP Synth Filter: Linear Prediction Synthesis Filter



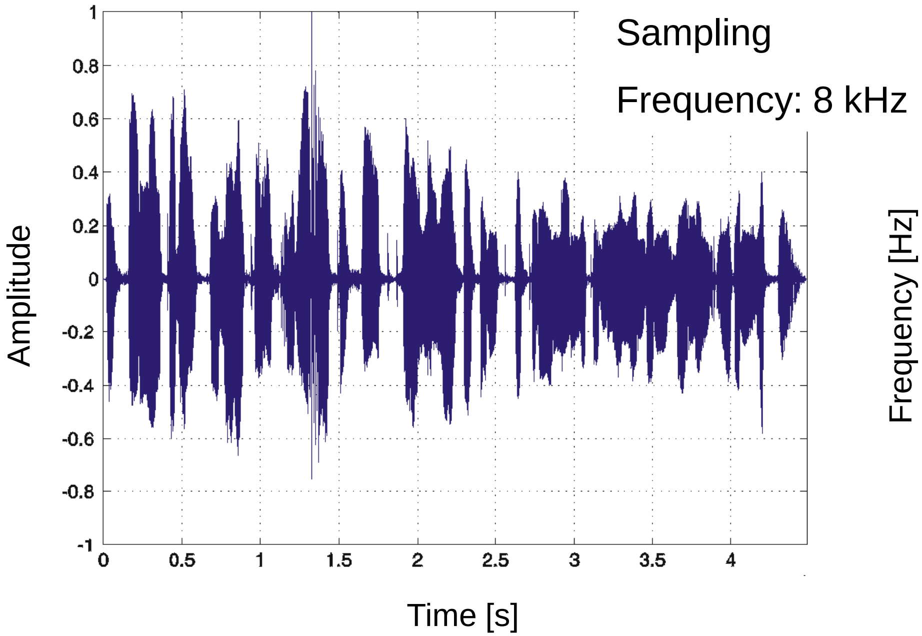
- **Speech generation**

- **Speech signal propagation:**
  - direct path and reflections in the acoustic environment
  - head shading and different time of arrival at the human ears

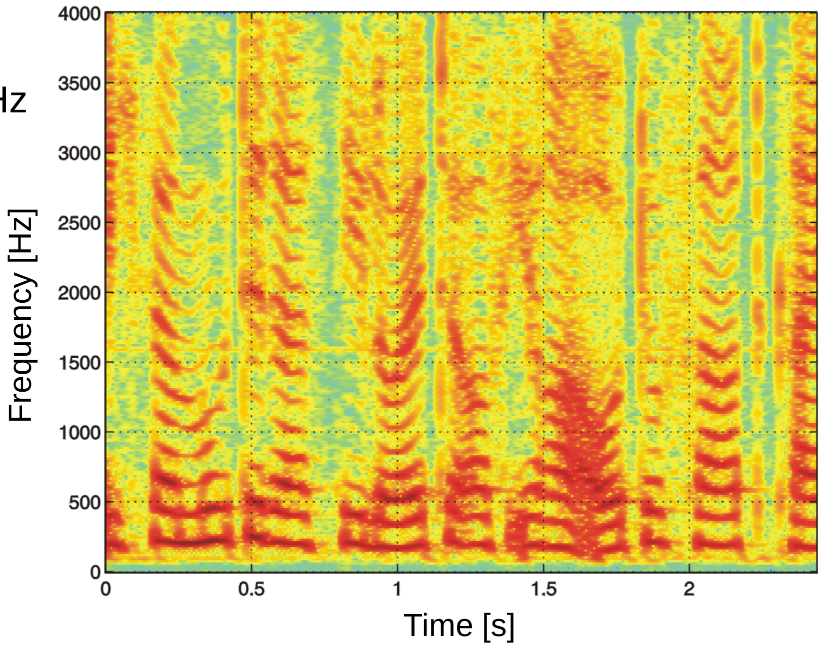
- **Speech perception:**
  - outer, middle and inner ear effects

Speech signals

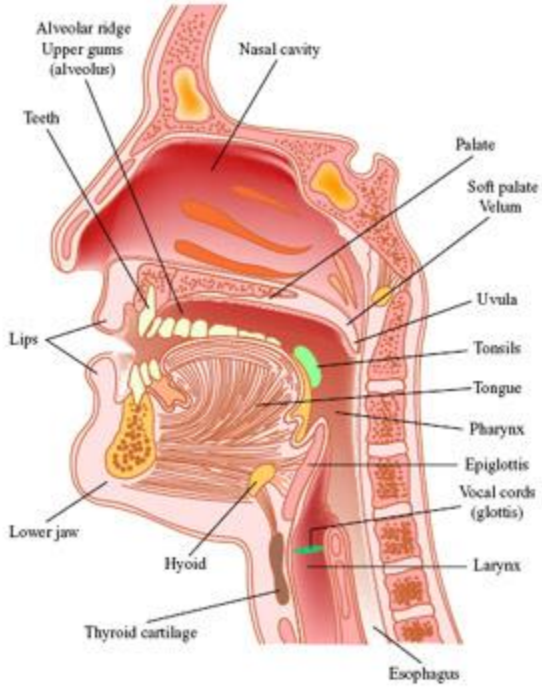
- **Sampled signals**



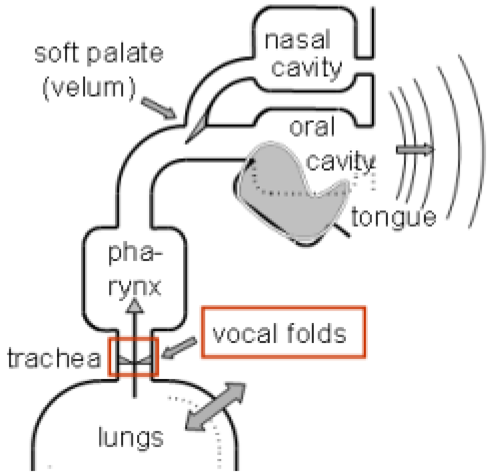
- **Spectrogram: Time / frequency analysis**



Human speech generation system

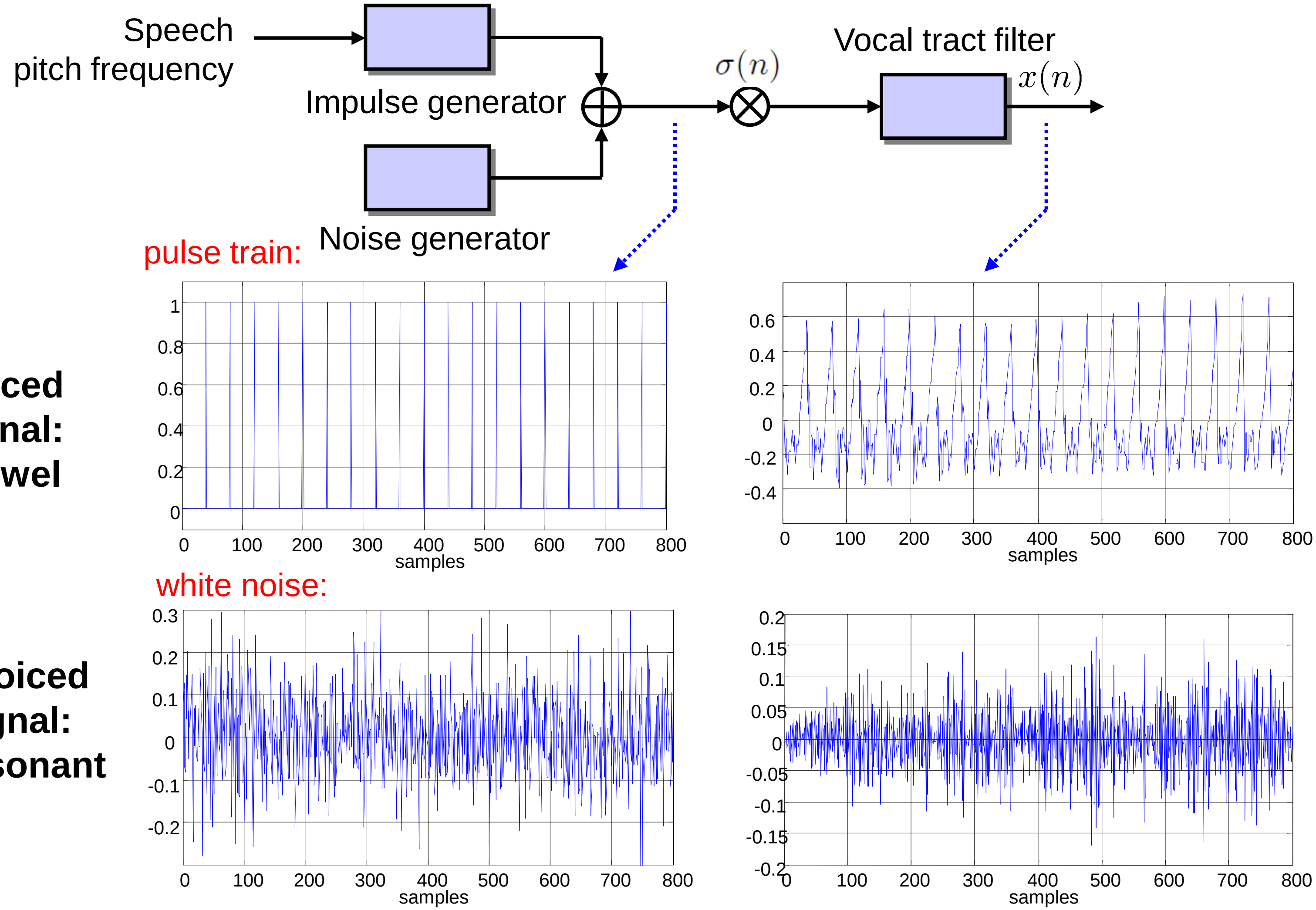


©MIT Open CourseWare



Source: Austrian Academy of Sciences
- Acoustics Research Institute

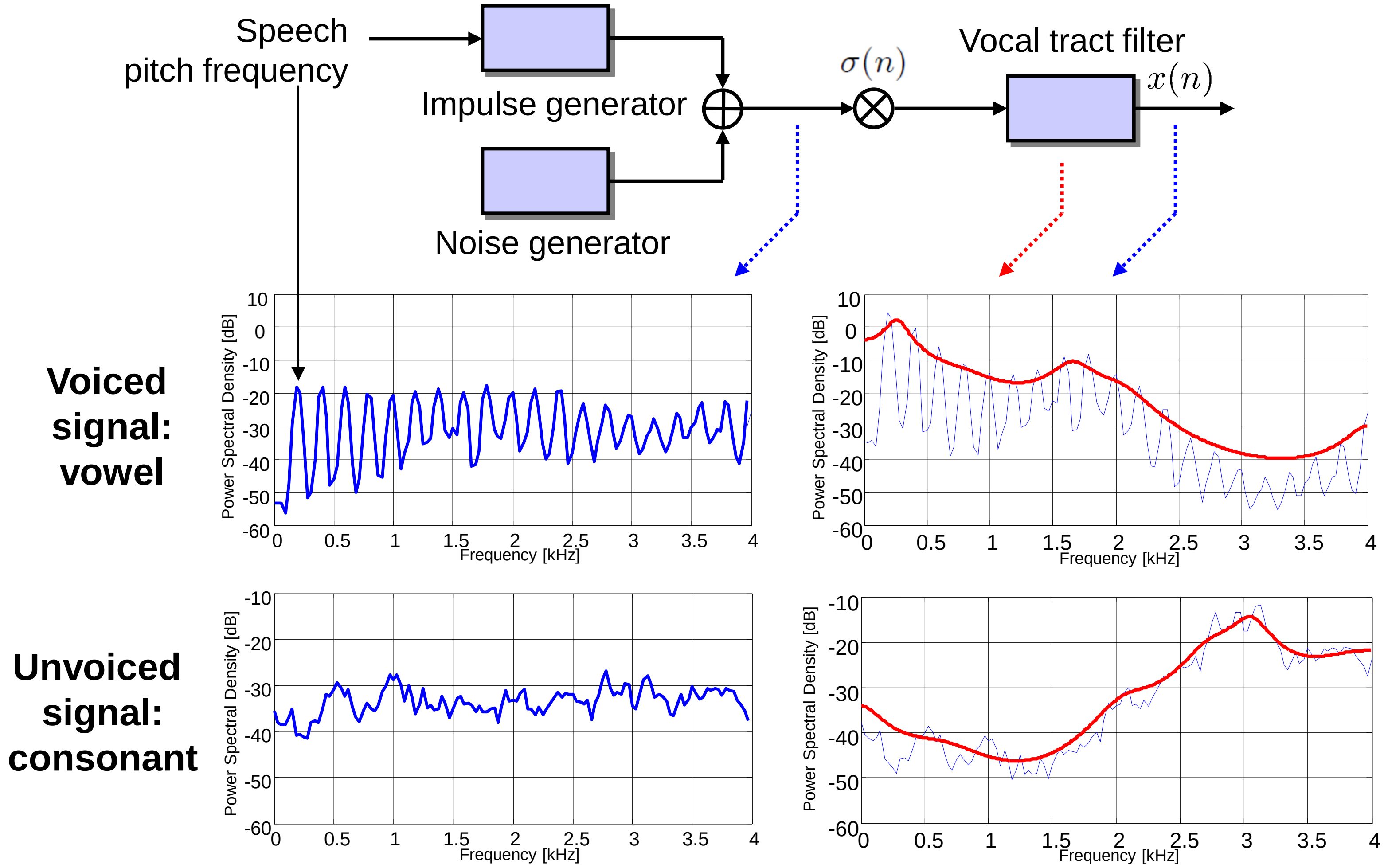
Speech models: Time domain



Voiced signal: vowel

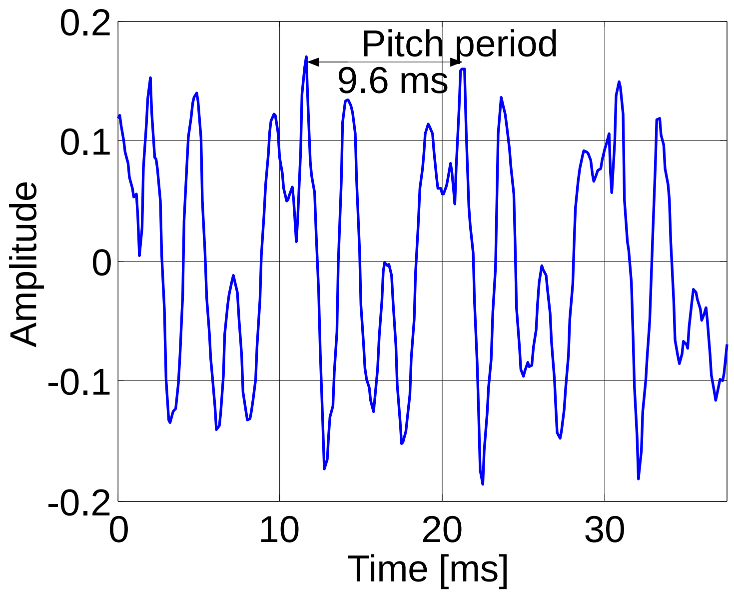
Unvoiced signal: consonant

Speech models: Frequency domain

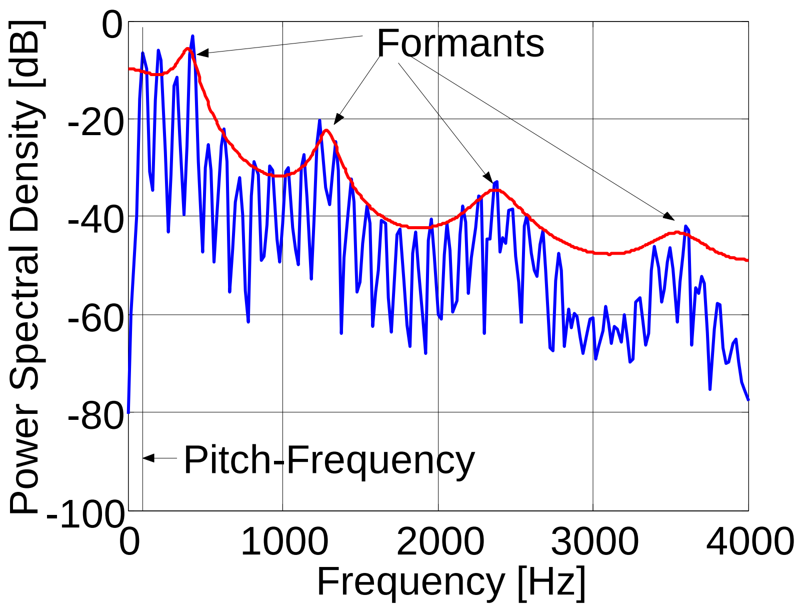


Voiced speech frames

- **Pitch period and pitch frequency or Fundamental period / frequency**



- **Formants**



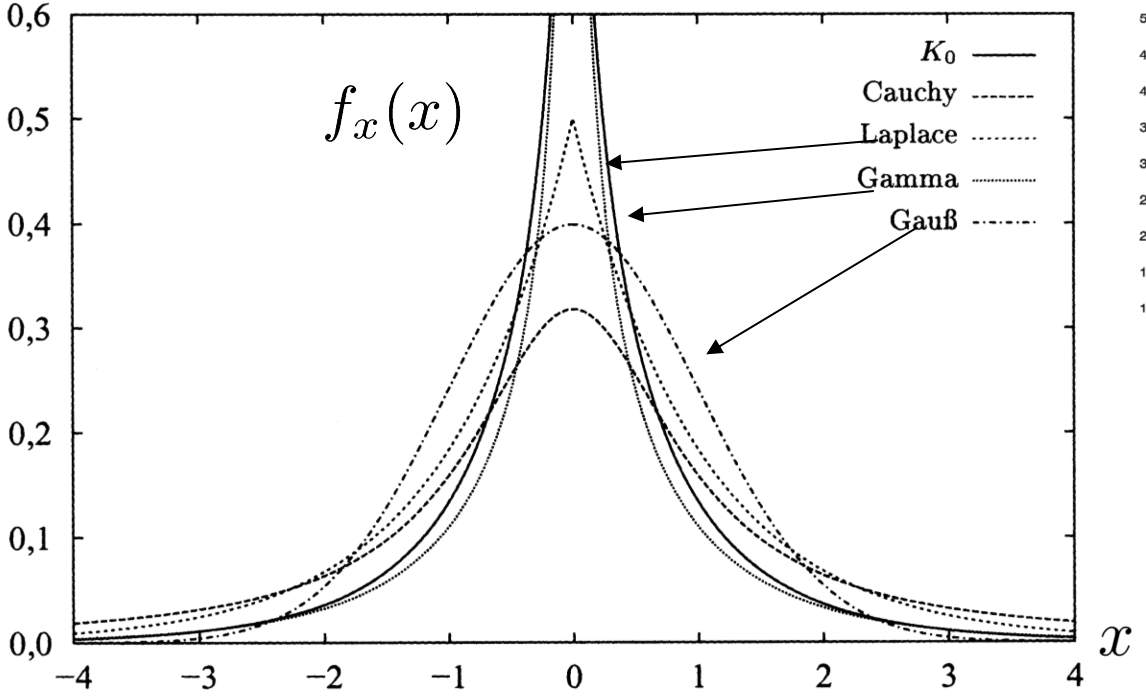
Special Random Processes

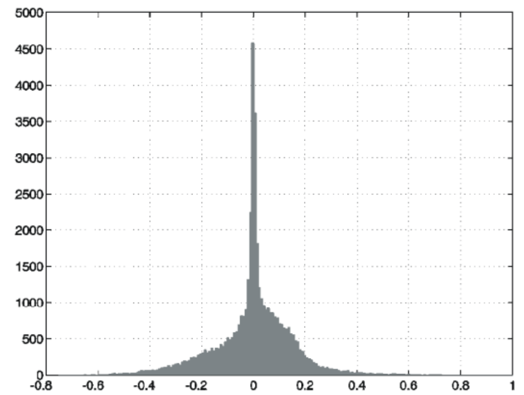
Random variables generated by combinations of statistically independent unbiased Gaussian random variables x, z und w with equal variances σ^2:

K_0	$y = x\,z$	$\dfrac{1}{\pi\,\sigma^2}\,\mathrm{K}_0\left(\dfrac{\lvert y\rvert}{\sigma^2}\right)$
Cauchy	$y = x/z$	$\dfrac{1}{\pi}\,\dfrac{1}{1+y^2}$
Rayleigh	$y = \sqrt{x^2+z^2}$	$\dfrac{y}{\sigma^2}\,e^{-\frac{y^2}{2\,\sigma^2}}\,,\quad y \geq 0$
Laplace	$y = x\,\sqrt{z^2+w^2}$	$\dfrac{1}{2\,\sigma^2}\,e^{-\frac{\lvert y\rvert}{\sigma^2}}$
Gamma	$y = x^2\,\mathrm{sgn}\,x$	$\dfrac{1}{\sqrt{8\,\pi\,\sigma^2\,\lvert y\rvert}}\,e^{-\frac{\lvert y\rvert}{2\,\sigma^2}}$

Probability Densities of Various Random Processes

- Speech is typically modeled as a Laplace or Gamma distribution:
  "Super Gaussian" distributions: Higher peak and higher values for high x-values

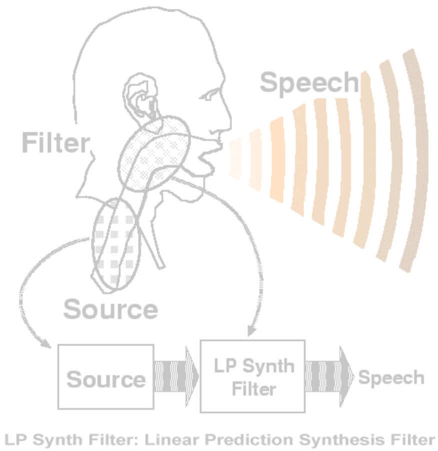


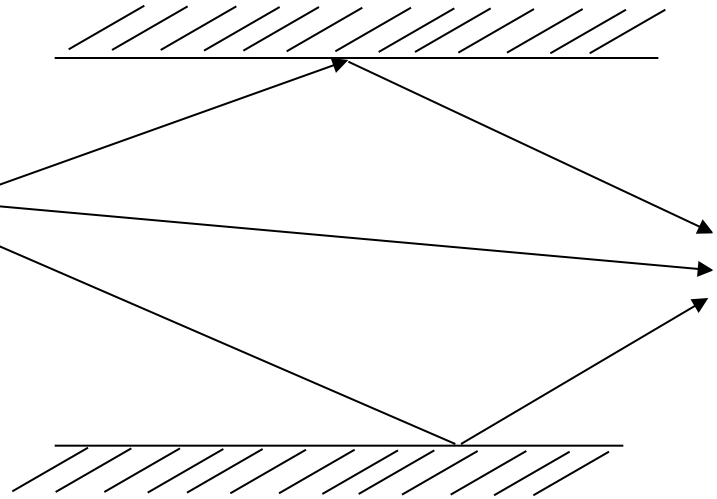


Histogram of speech amplitudes

from D. Wolf: Signaltheorie – Modelle und Strukturen. Springer, 1999

Speech signal analysis



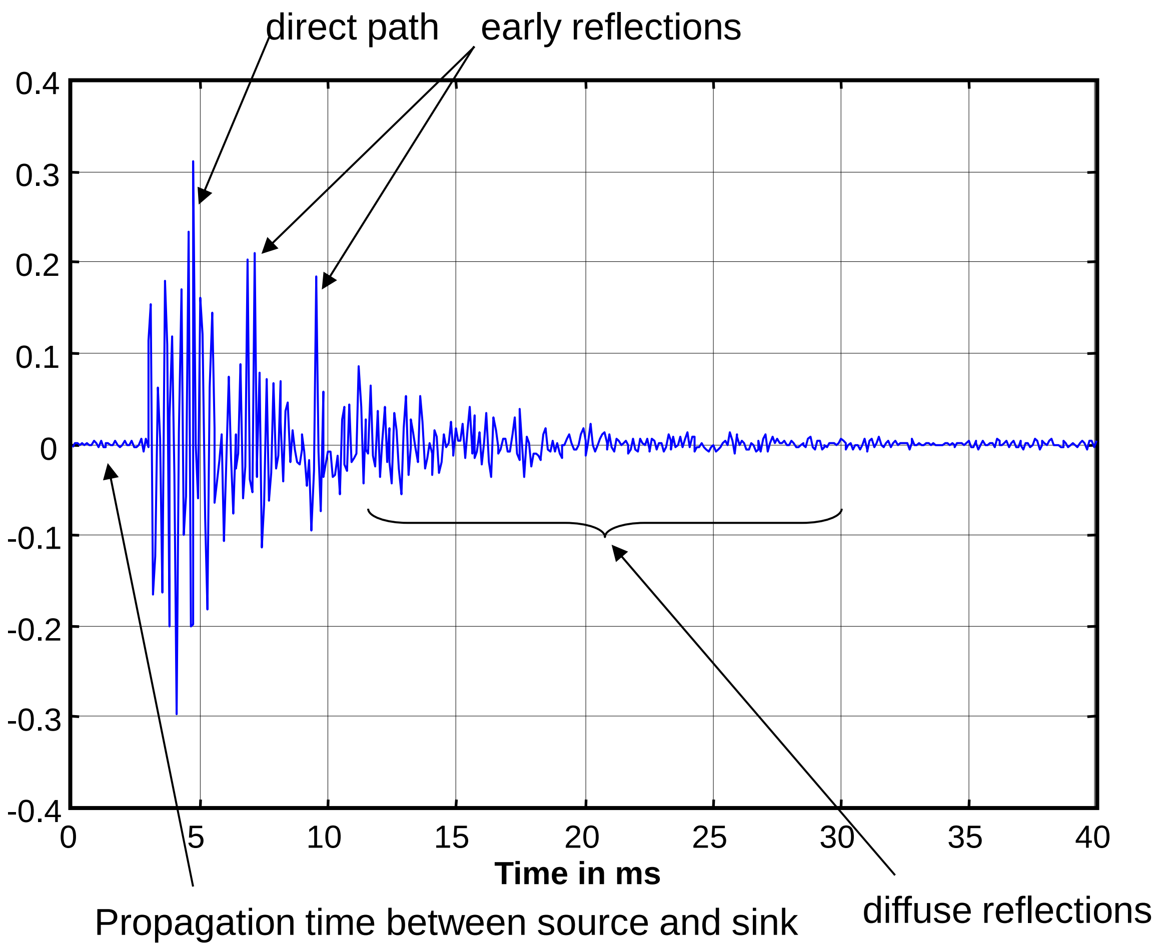


- Speech generation

- **Speech signal propagation:**
  - direct path and reflections in the acoustic environment
  - head shading and different time of arrival at the human ears

- **Speech perception:**
  - outer, middle and inner ear effects

Sound propagation in rooms



- **Typical room impulse response of a car cabin**

Reverberation time

o que é e? erro ou excitação?
a atenuação é boa? se estiver se referindo ao erro e eu quero reduzir ao máximo
Ou estpa dizendo que meu sinal é mto pequeno mesmo com poucas amostras pq fica dificil fazer as coisas?

- Reverberation after a time t = N*Ts

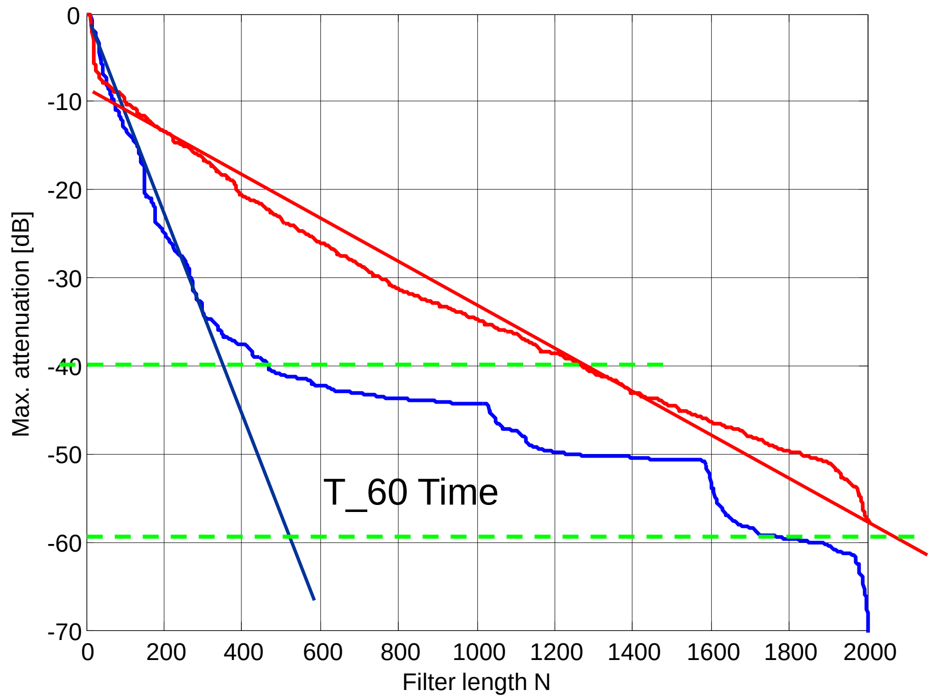
$$att_{max} = \frac{\sigma_e^2(N)}{\sigma_y^2} = \frac{\sum_{v=N}^{\infty} h_v^2}{\sum_{v=0}^{\infty} h_v^2}$$

40 dB attenuation:
N = 450 for a car cabin (example)
N = 1250 for a office room (example)

- Determine reverberation time
  T_60 is a value which typically characterizes the reverberation:

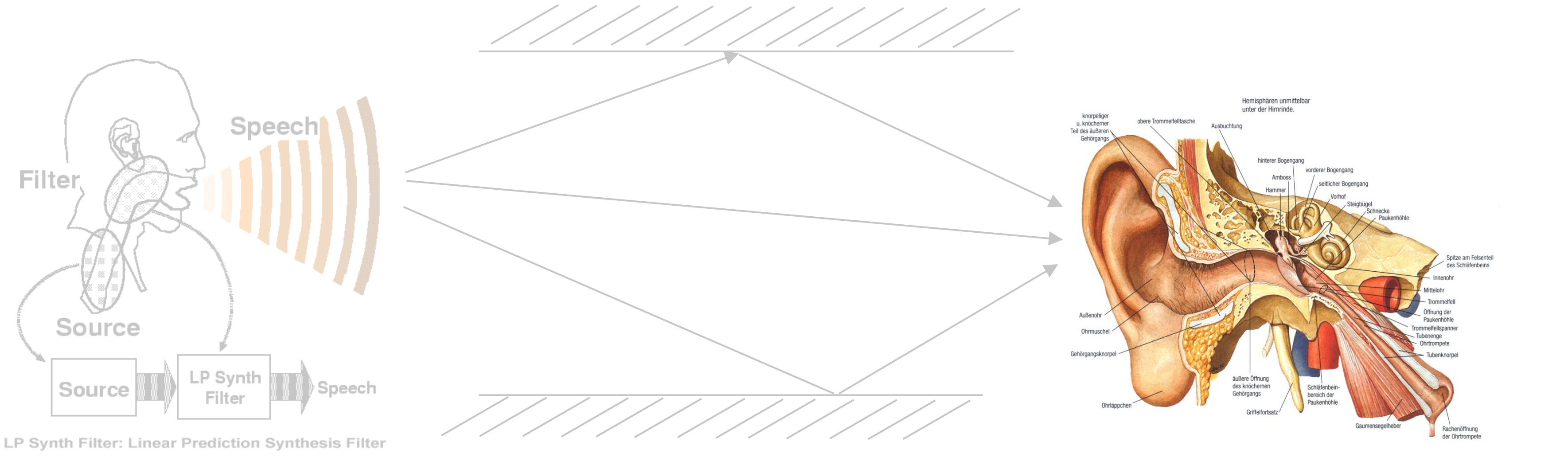
  - Set att_max to 60 dB and calculate corresponding N, or t.

Attenuation in dependence of N



red: office room
blue: car cabin

Speech signal analysis

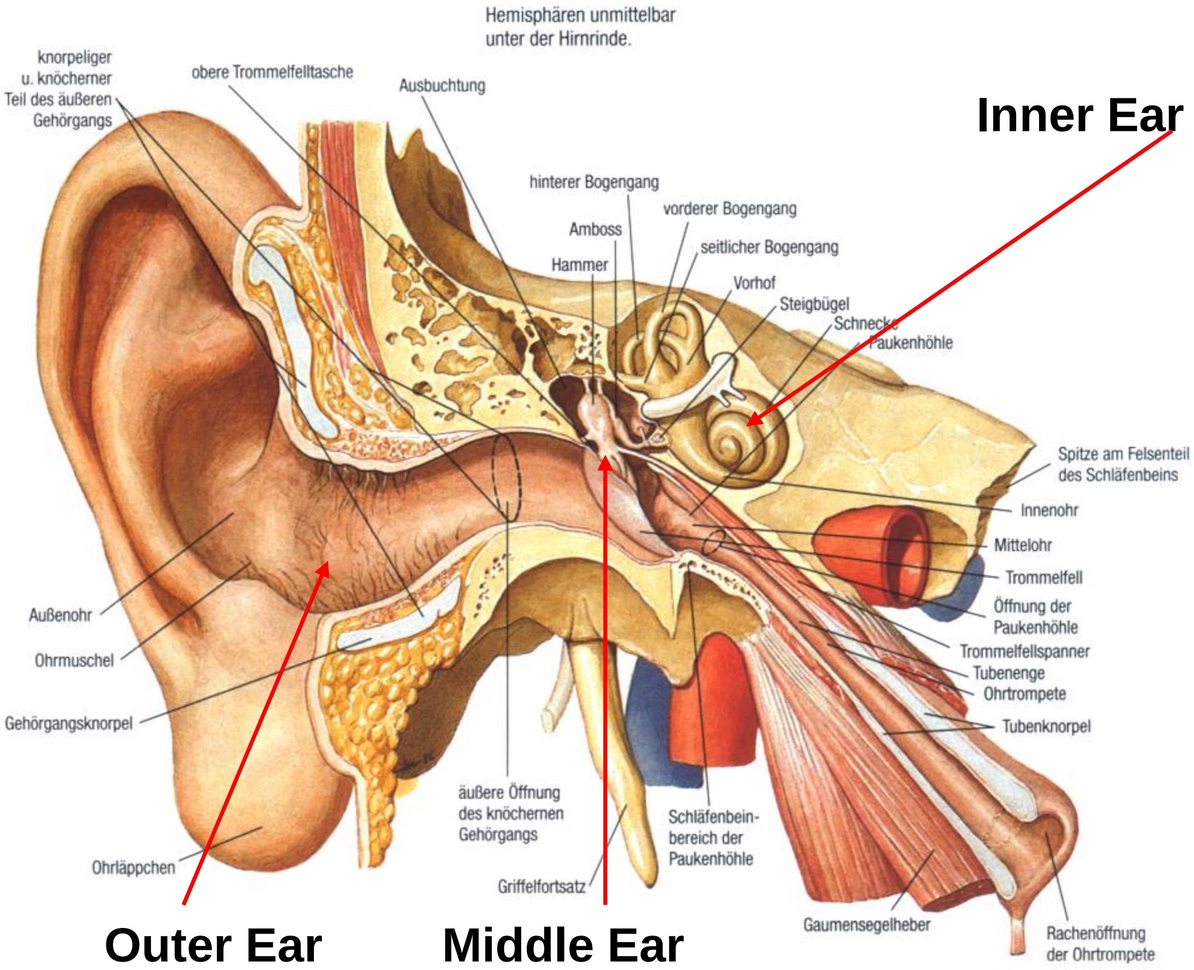


- ❑ **Speech generation**

- ❑ **Speech signal propagation:**
  - direct path and reflections in the acoustic environment
  - head shading and different time of arrival at the human ears

- ❑ **Speech perception:**
  - outer, middle and inner ear effects

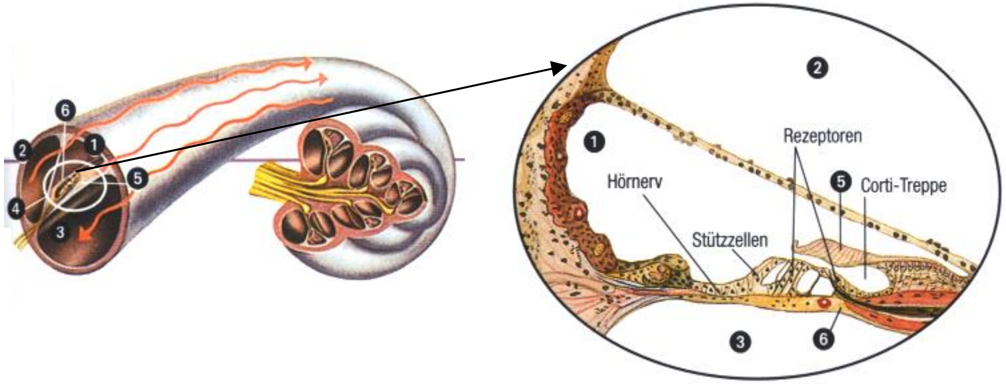
The Ear


Hemisphären unmittelbar unter der Hirnrinde.
knorpeliger u. knöcherner Teil des äußeren Gehörgangs
obere Trommelfelltasche
Ausbuchtung
Inner Ear
hinterer Bogengang
vorderer Bogengang
Amboss
seitlicher Bogengang
Hammer
Vorhof
Steigbügel
Schnecke
Paukenhöhle
Spitze am Felsenteil des Schläfenbeins
Innenohr
Mittelohr
Trommelfell
Öffnung der Paukenhöhle
Trommelfellspanner
Tubenenge
Ohrtrompete
Tubenknorpel
Außenohr
Ohrmuschel
Gehörgangsknorpel
äußere Öffnung des knöchernen Gehörgangs
Schläfenbein-bereich der Paukenhöhle
Ohrläppchen
Griffelfortsatz
Gaumensegelheber
Rachenöffnung der Ohrtrompete
Outer Ear
Middle Ear

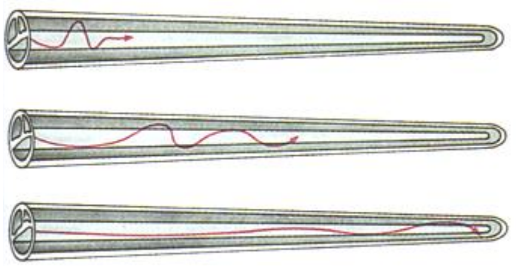
Anatomy of the ear

TECHNISCHE UNIVERSITÄT DARMSTADT

- **Inner ear: Cochlear**

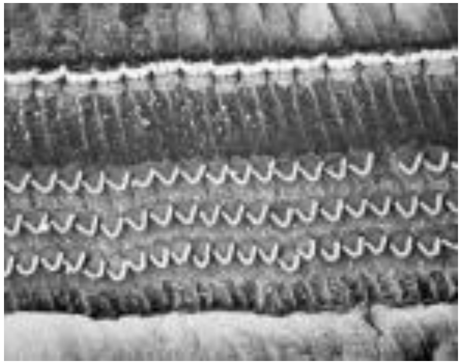


- **Cochlear:**
  frequency-location transformation:

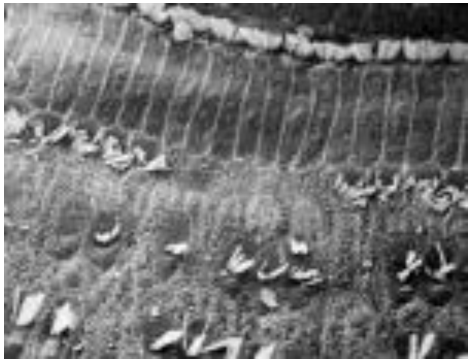


- **Hair cells**

Normal:

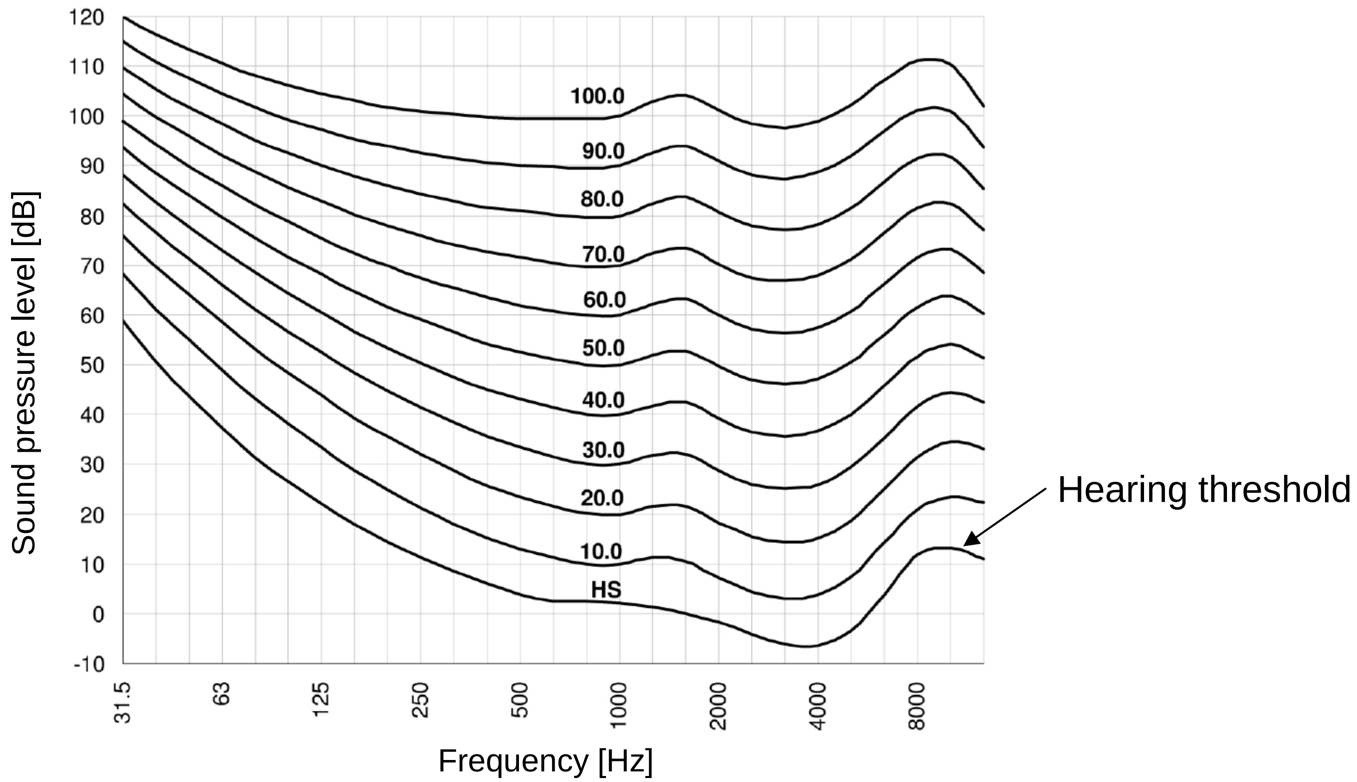


Hearing impaired:



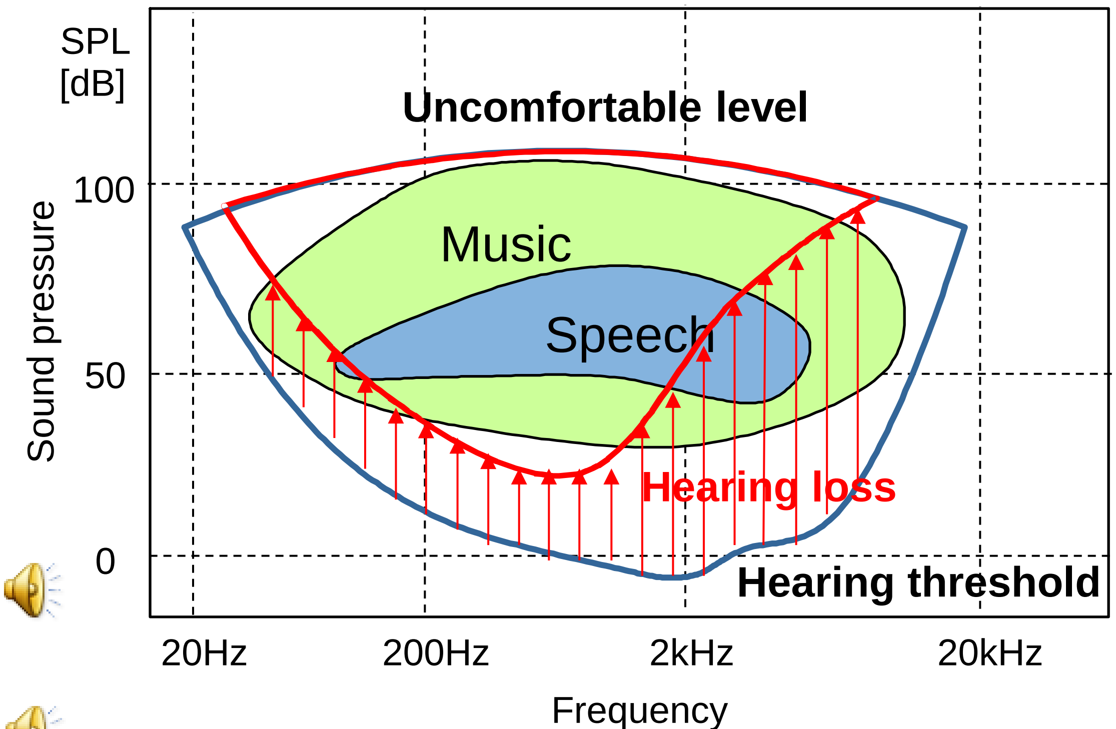
Hearing threshold

- Hearing threshold and curves of the same loudness:



The human auditory system

- **Human auditory system:**
  covers a large dynamic (> 100 dB) and a large frequency range (20 Hz – 16 kHz)



Normal hearing:

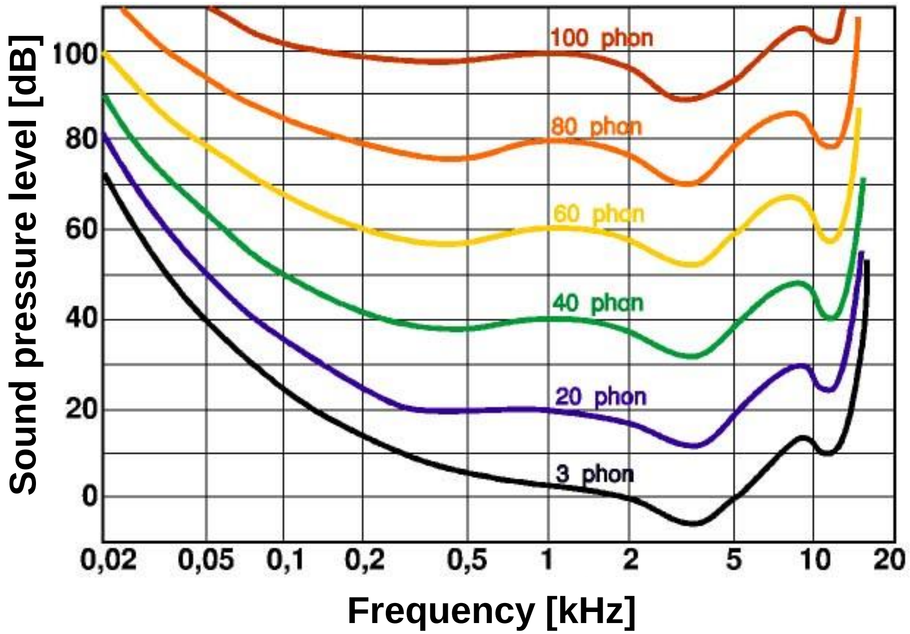
Hearing impaired:

Sound pressure level (SPL) / Schalldruckpegel [in dB] vs. Loudness level / Lautstärkepegel [in phone]

- **Definition of “sound pressure level” (SPL), “Schalldruckpegel”:**

$$L = 20\log_{10}(p/p_0) \qquad\qquad p_0 = 20\mu\text{Pa} = 2\cdot 10^{-5}\text{N/m}^2$$

- **Loudness level: defined at reference frequency 1 kHz**
  - At 1 kHz the sound level in “phone” is equivalent to the sound pressure level in “dB”
  - The values at other frequencies are subjectively evaluated by equivalent loudness to the reference tone at 1 kHz.



Consider the differences:

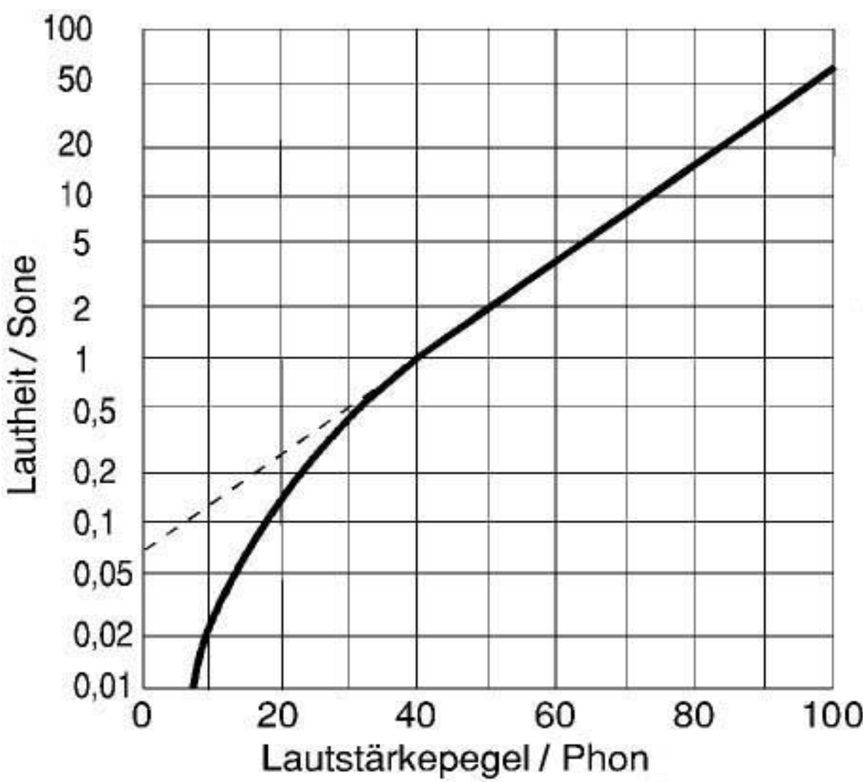
Sound pressure level:
objective measure

Loudness:
subjective measure

Loudness / Lautheit [in sone]

Esse sone é o SPL?

- **Mapping loudness level [phone] to loudness [sone]:**
  An increase of 10 phone is most often perceived as doubling the loudness



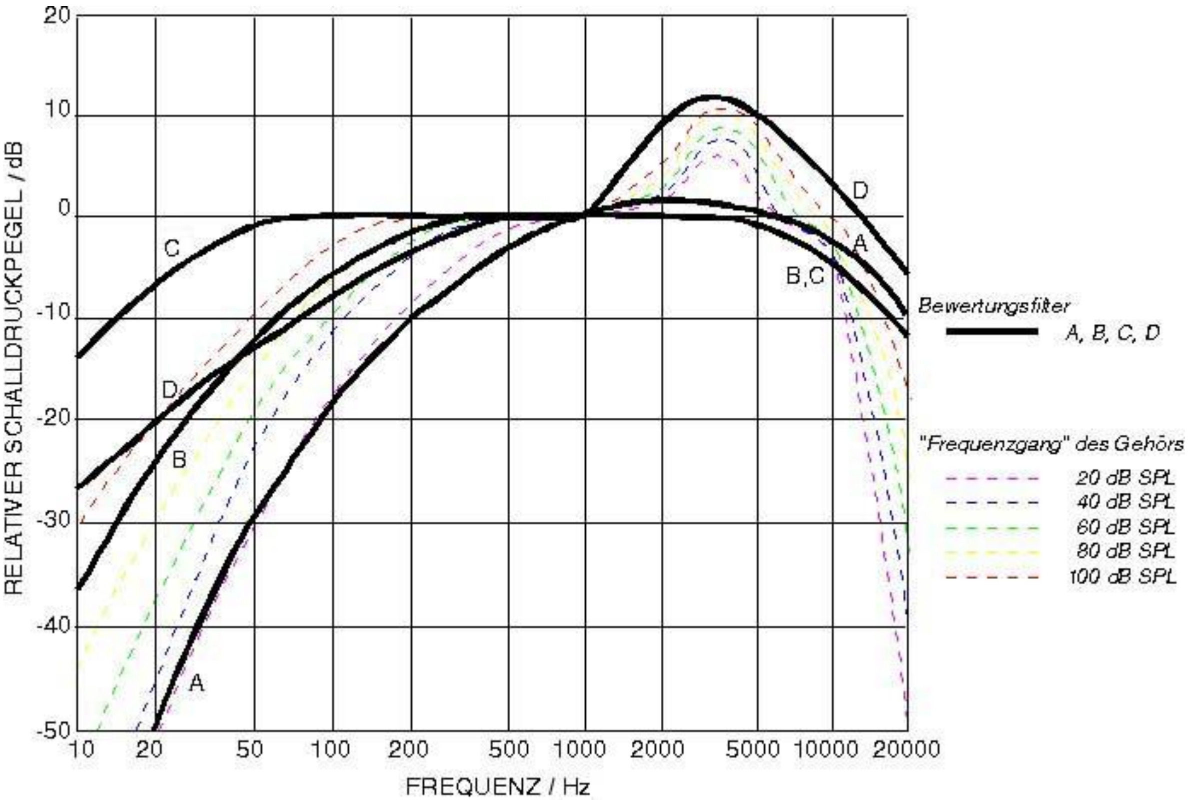
Lautheit / Sone	Pegel /Phon
64	100
32	90
16	80
8	70
4	60
2	50
1	40
1/2	32
1/4	25
1/8	19
1/16	14
1/32	11
1/64	9

Loudness (N) in dependence of loudness level (L_N)

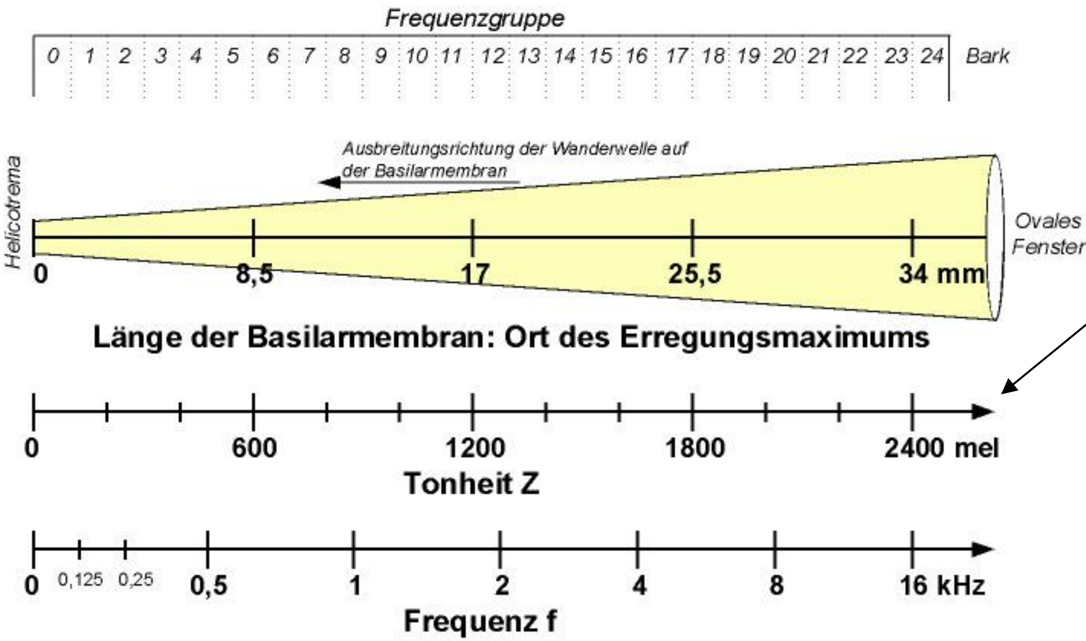
$$N = \left(10^{\frac{L_N - 40}{10}}\right)^{0.30103} \approx 2^{\frac{L_N-40}{10}}$$

Measurement of sound pressure level

- **Weighting filters:** Calculating dBA values:
  mean weighted sound pressure level considering the subjective loudness perception at different frequencies.



Critical band rate / Tonheit



- Definition according to Zwicker: normalized at the tone c: 131 Hz

f = 131 Hz
=> z = 1.31 Bark; with 1 Bark = 100 Mel

$$z = 13\arctan(0.00076\,f) + 3.5\arctan((f/7500)^2)$$

Definition according to Stanley Smith Stevens:

$$m = 2595\,\mathrm{Mel}\,\log_{10}\left\{\frac{f}{700Hz} + 1\right\}$$

normalized at 1000 Hz:
f = 1000Hz => m = 1000 Mel

A sound which is perceived as double as high as a reference sound => double tone value („Tonheit“)
A sound which is perceived as half as high as a reference sound => half tone value („Tonheit“)

Speech intelligibility index (SII) / Speech intelligibility (SI)

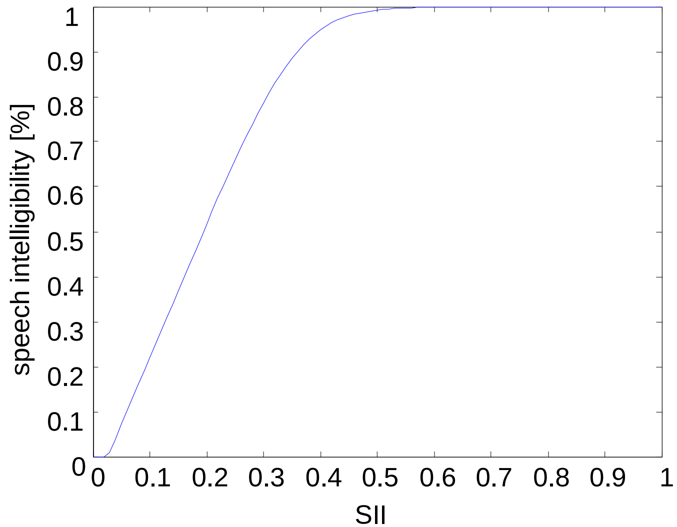
- Defined by ANSI, S3.5 1997, to predict speech intelligibility in stationary noise.
- SII: - weighted SNR sum over N = 18, bark scale frequency bands.
  - value between 0 and 1

$$SII = \frac{1}{30}\sum_{i=1}^{N} w_i(SNR_i + 15) \qquad \text{with: } SNR_i \in [-15\,\mathrm{dB}, 15\,\mathrm{dB}]; N = 18$$

- Speech intelligibility in %:

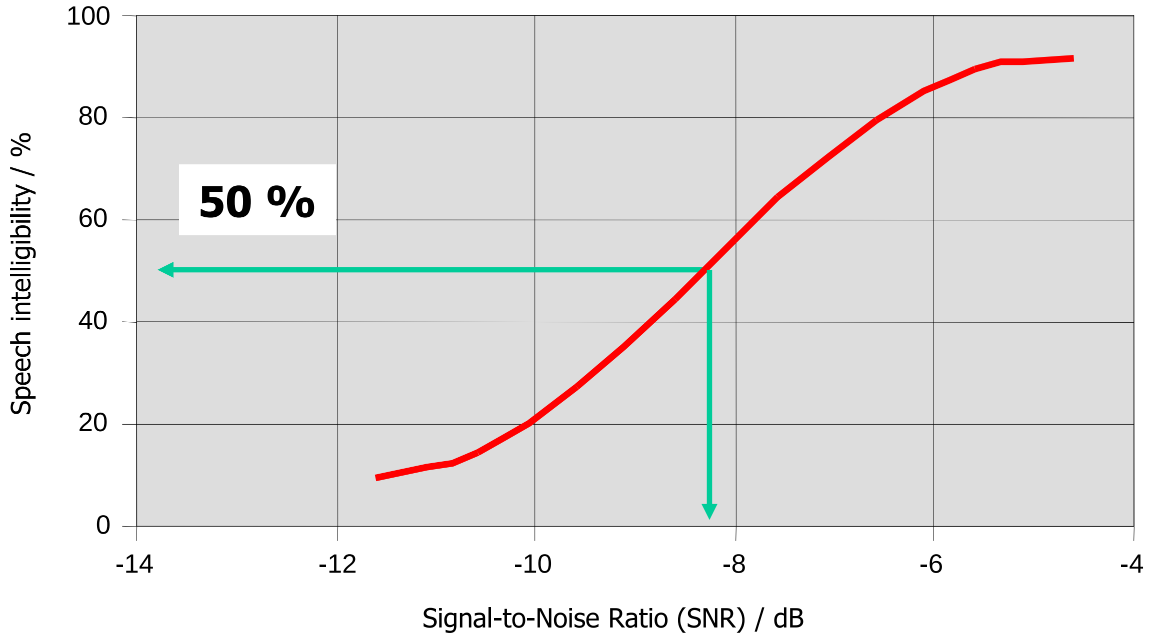
$$SI[\%] = \log_{10}\left[10^{[a*SSI-k]/Q} + 10^{1/Q}\right]^{Q} \qquad \text{with: } a = 3.15; k = 0.0802; Q = -0.3339$$

STI-Index SII-Index	Sprachverständlichkeit	Alcons
0 bis 0,3	Nicht akzeptierbar, unverständlich	100 % bis 33 %
0,3 bis 0,45	Schlecht	33 % bis 15 %
0,45 bis 0,6	Genügend	15 % bis 7 %
0,6 bis 0,75	Gut	7 % bis 3 %
0,75 bis 1,0	Ausgezeichnet	3 % bis 0 %



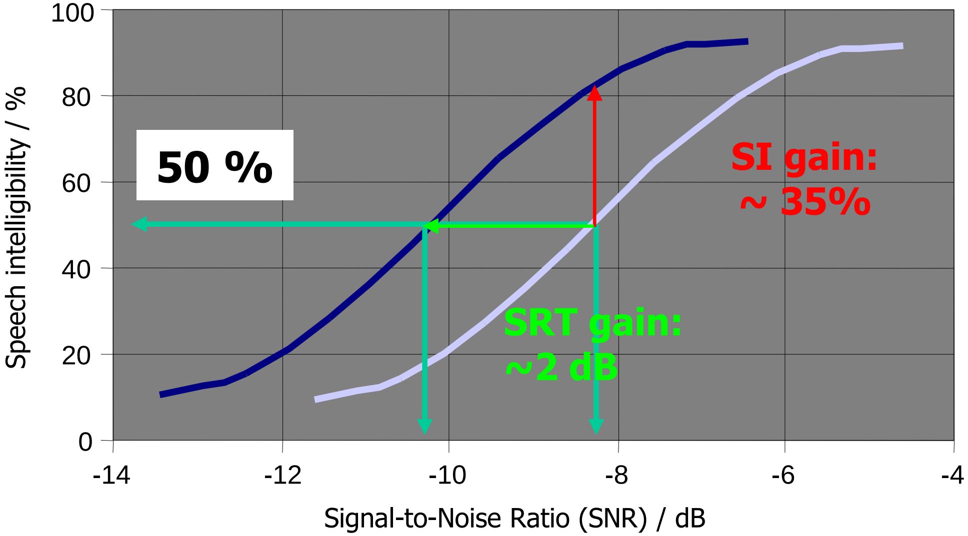
Speech intelligibility: relation to the SNR

- Typically rather sharp discrimination curves: < 4 dB SNR difference between 20% and 80% speech intelligibility.



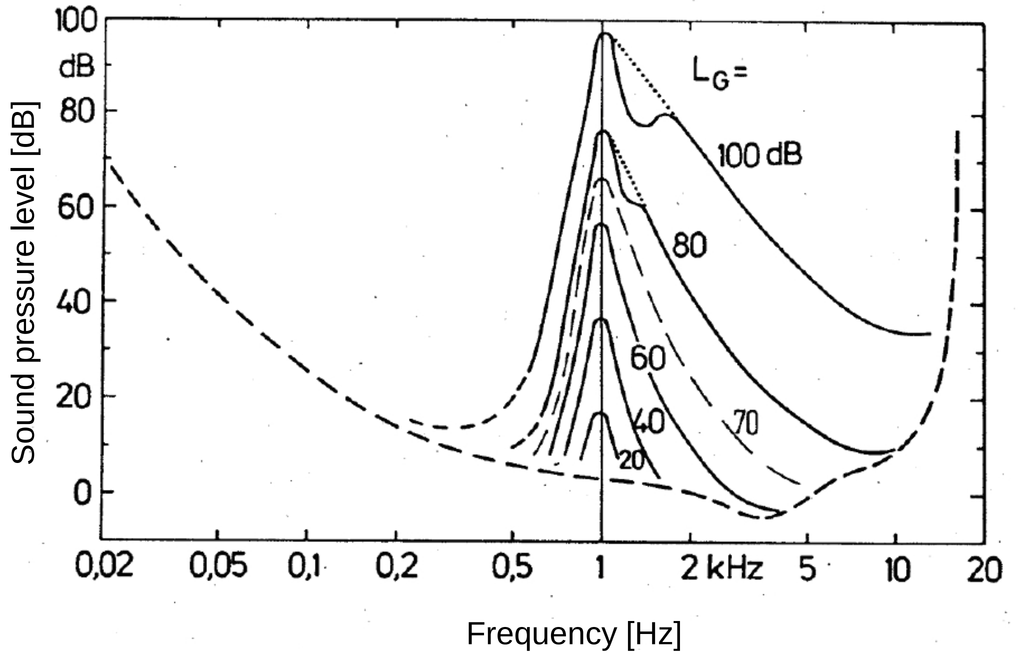
SRT (Speech reception threshold) Measurements

- Evaluation of noise suppression algorithms, such as beamforming, noise reduction, etc.
- Adjustment of the input SNR such that a speech intelligibility of 50 % is obtained => this value: SRT value
- SRT Gain: comparison of the SNRs with activated and deactivated processing



Masking (I): Frequency Masking

- Sound are masked below the “masking” threshold, dependent on the level and the frequency of the masker.

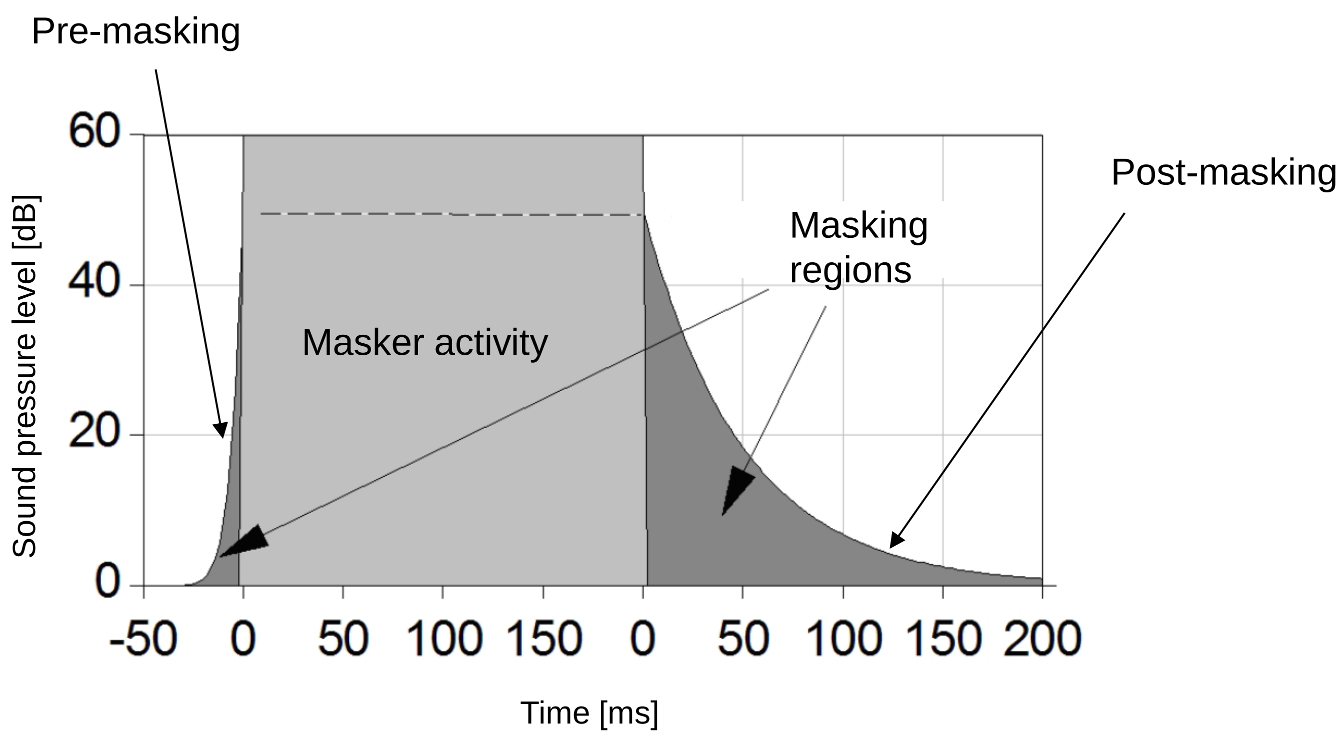


- Reminder: Definition of “sound pressure level” (SPL), “Schalldruckpegel”:

$$L = 20\log_{10}(p/p_0) \qquad p_0 = 20\mu\mathrm{Pa} = 2\cdot 10^{-5}\mathrm{N/m}^2$$

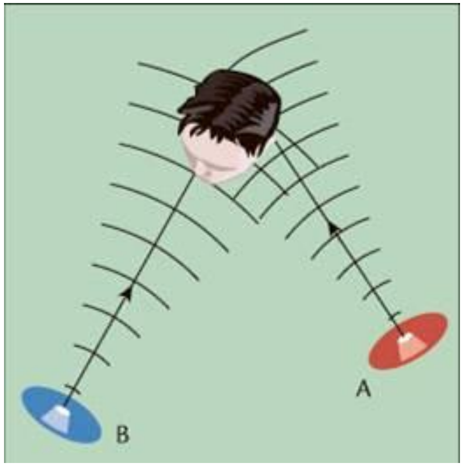
Masking (II): Time Masking

- Sound are masked slightly before and after a sound incidence.



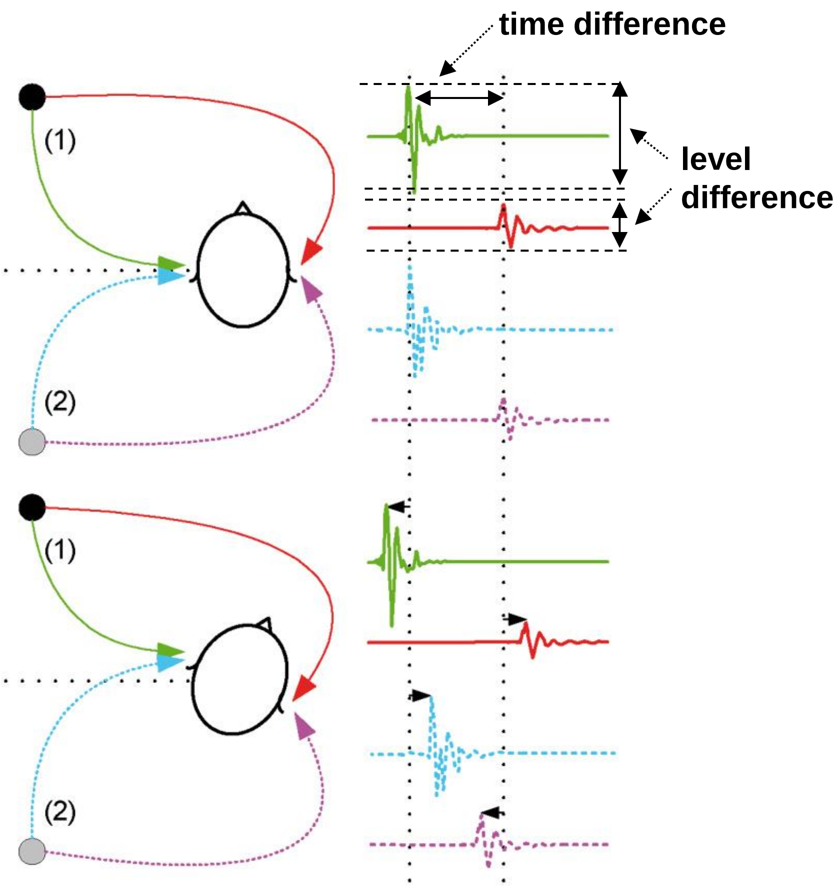
Localization (I): Binaural cues

- Binaural cues describe the level and time differences of signals received at both ears
- These binaural cues allow the localization of signals in the acoustic environment
- Localization is mainly based on
  - Interaural time differences (ITD) below 800 Hz and
  - Interaural level differences (ILD) above 1600 Hz



© Elisa Setmire

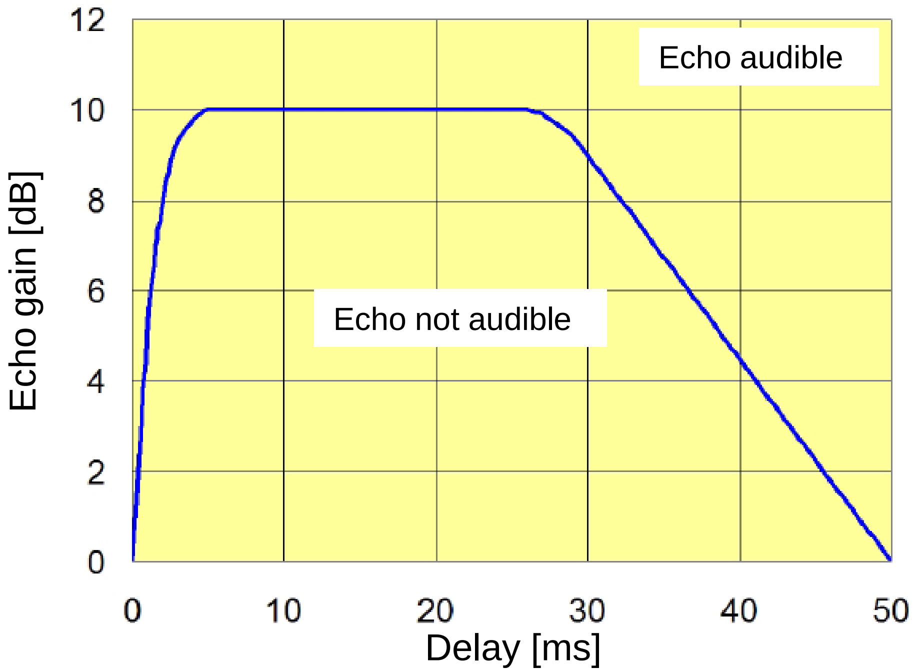
head movements help to localize frontal and back sounds



© Institute of Technical Acoustics
RWTH Aachen University

Localization (II): Precedence effect

- Precedence effect (Haas effect): Law of the first wavefront.
- In case the delay between the first wave front and reflections is **between approx. 2 and 30 to max. 50 ms**, sound is **localized at the direction of the first wave front**.
- **Below approx. 2 ms**, in case two loudspeakers play the same signal, the sound is **localized between the position of the sources.** dependent on the level difference of the sources.
- **Above approx. 50 ms**, an **echo signal is perceived** with a dedicated direction.

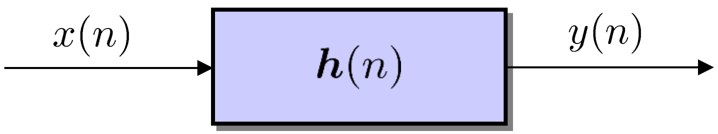


Processing methods

Sample-based vs. block-based processing

- **Sample-based processing:**
  get one input sample and process one output sample
  e.g., by time domain filtering

$$y(n) = \sum_{i=0}^{N-1} x(n-i)\, b_i(n) - \sum_{i=1}^{M-1} y(n-i)\, a_i(n)$$



- Latency of the processing determined by the group delay of the applied filters

- Sometimes it makes sense to look some samples “ahead”, e.g., in order to detect transient noise signals and attenuate them appropriately.

Sample-based vs. block-based processing

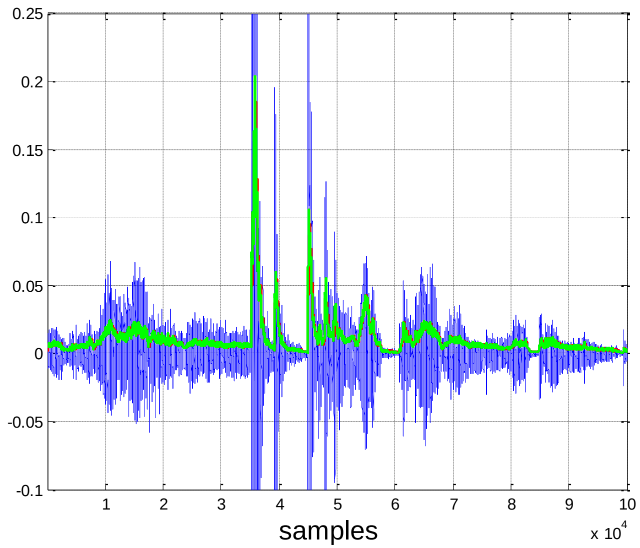
- Smoothed magnitude of input samples in order to detect raising signals:

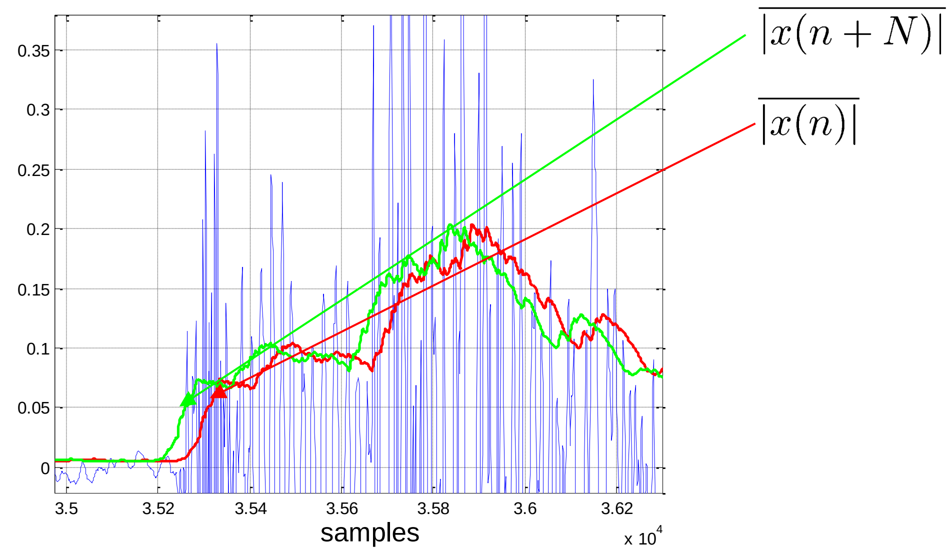
$$\overline{|x(n)|} = \alpha\,\overline{|x(n-1)|} + (1-\alpha)\,|x(n)|$$

- Look ahead allows to detect and attenuate raising signal slopes efficiently.
  => Introduction of a delay which is equivalent to the “look ahead”.

$$att(n) = F(\overline{|x(n+N)|}) \longleftarrow$$ attenuation a function of the „looked ahead“ smoothed signal

  - Non-causal solution: $y(n) = x(n)\,att(n) = x(n)\,F(\overline{|x(n+N)|})$
  - Causal solution: $y_{\text{kausal}}(n) = y(n-N) = x(n-N)\,F(\overline{|x(n)|})$ with delay!

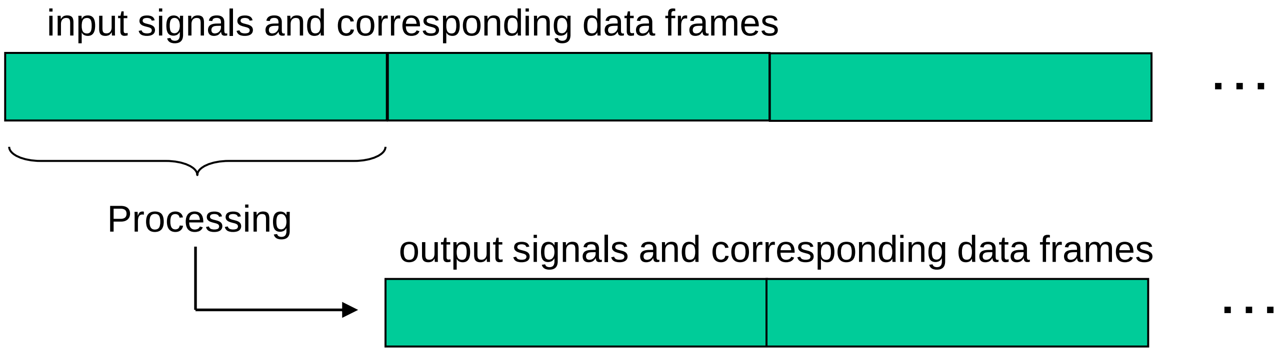




Sample-based vs. block-based processing

- **Block-based processing:**
  - divide the input signal data stream into consecutive (overlapping) blocks
  - perform a processing of the data samples of the blocks

  => introduces a processing delay equal or larger than the frame shift of the data blocks



- Output of the first sample can start after the complete data of the block was received and processed.

  => **minimum processing latency: one block frame**

  (in case the processing of all the block data can be performed in one sample)

Real-time vs. batch signal processing

- **Batch processing:**
  - The complete input signal is known
  - Analysis can be performed based on the complete input signal, e.g. mean value can be determined by summing over all samples

$$\overline{|x|} = \frac{1}{N} \sum_{n=1}^{N} |x(n)|$$

  - example: record data and evaluate, also possible on a server
    => Shazam – App: Data is analyzed and send to a server
    Dragon – App: Some words are spoken and then recognized on a server

- **Real-time processing:**
  - The processing is based on the current and a limited amount (memory!) of past data.
  - A certain look ahead is possible (=> introduces latency!)
  - For each input sample an output sample has to be processed typically at the same sample rate.

  - The computational complexity of the algorithms cannot exceed the hardware performance. For an input signal at 8 kHz sampling rate, the processing of 8000 samples should be possible in a less than 1 sec on the processor.

Signal analysis in the time domain: Power

- Short-term power

$$\overline{x^2(n)} = \alpha\, \overline{x^2(n-1)} + (1-\alpha)\, |x(n)|^2$$

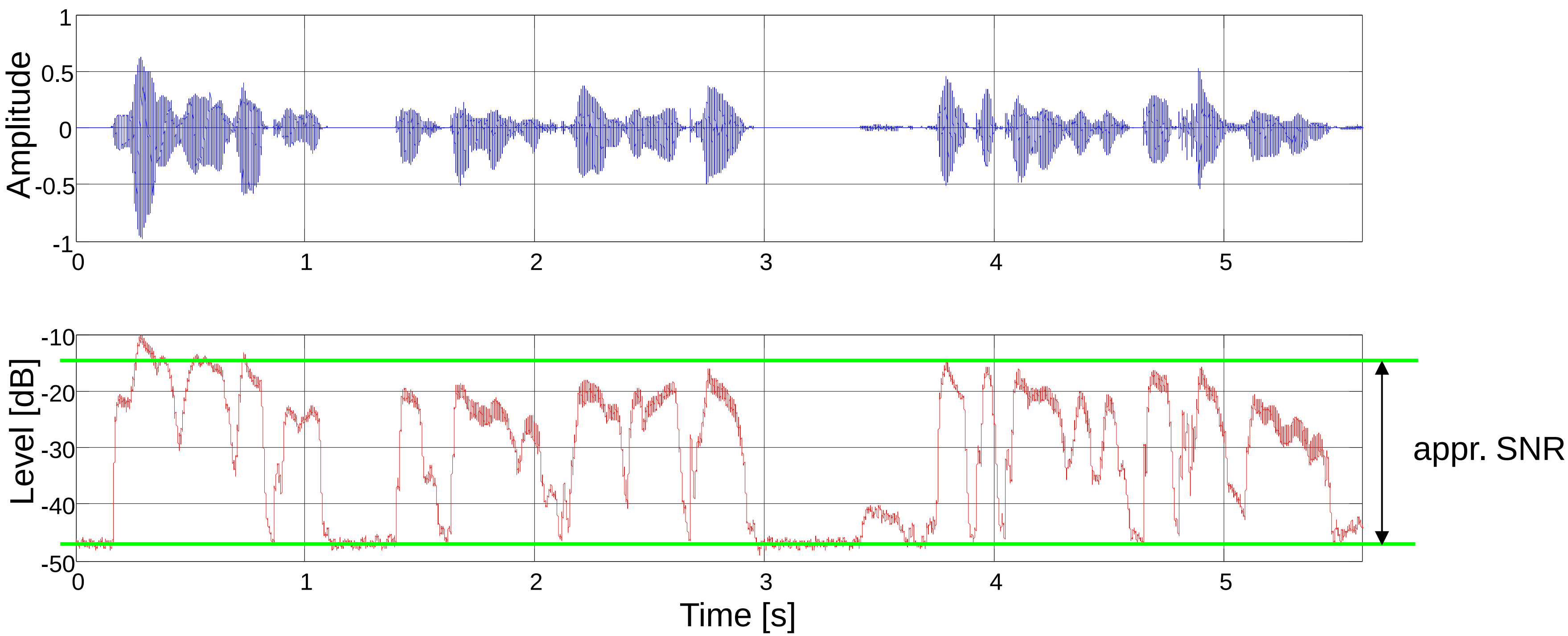
- Short term magnitude

$$\overline{|x(n)|} = \alpha\, \overline{|x(n-1)|} + (1-\alpha)\, |x(n)|$$

- The smoothing constant α should be in the interval: $0 \ll \alpha < 1$
- The magnitude smoothing allows the calculation with a reduced dynamic (important in case of fixed-point processing)
- For complex values, the magnitude are complicated to be calculated, sometimes an approximation with $|x(n)| = |\mathrm{Re}\{x(n)\}| + |\mathrm{Im}\{x(n)\}|$ is used.

Short-term signal power estimates

- Results for speech signals:
  Signal samples & corresponding signal level

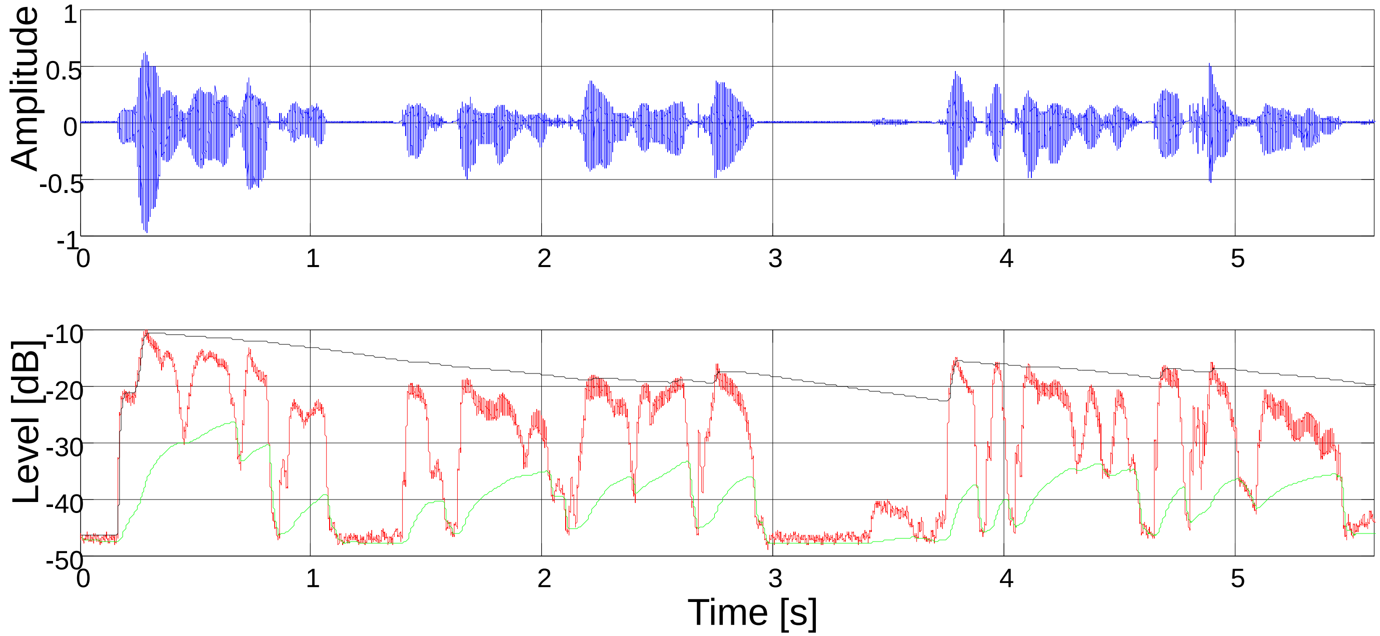


Non-linear smoothing

- Non-linear smoothing with different constants for raising & falling signal slopes:

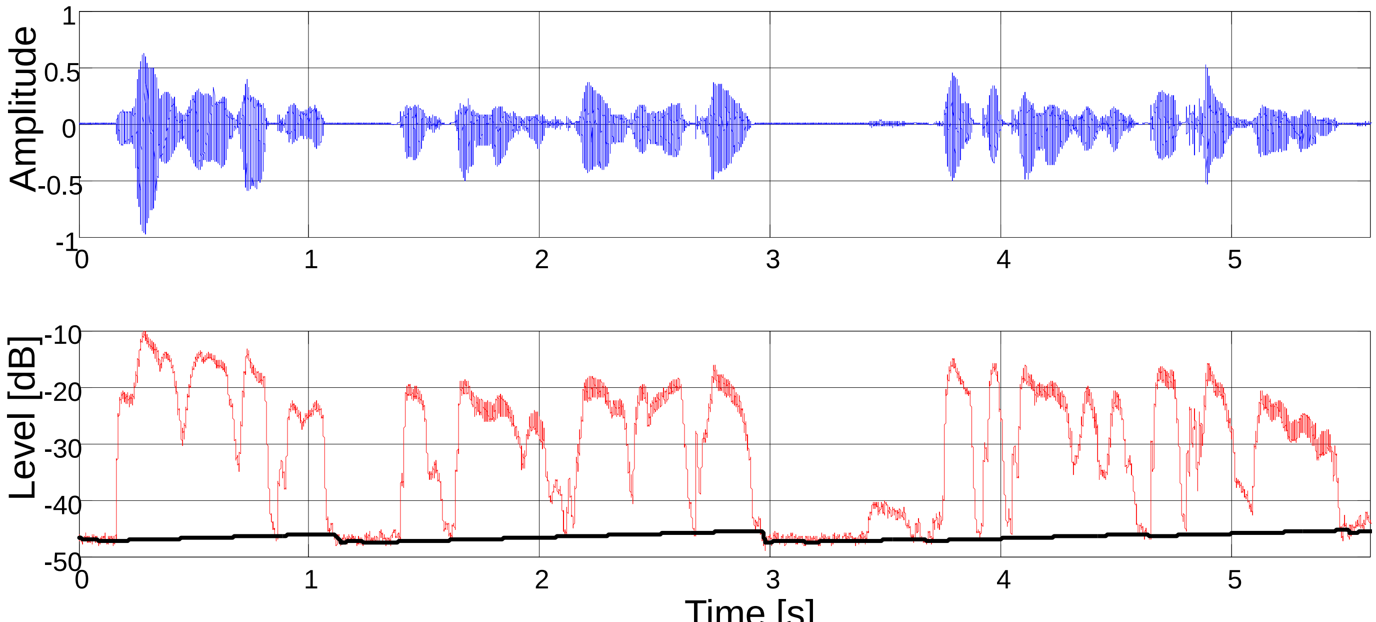
$$\overline{|x(n)|} = \begin{cases} \alpha_r \, \overline{|x(n-1)|} + (1-\alpha_r)\,|x(n)| & : \quad |x(n)| > \overline{|x(n-1)|} \\ \alpha_f \, \overline{|x(n-1)|} + (1-\alpha_f)\,|x(n)| & : \quad \text{else} \end{cases}$$

- Maximum tracker: $\alpha_r < \alpha_f$ Minimum tracker: $\alpha_r > \alpha_f$



Minimum estimator for noise power estimation

- Two step procedure for a simple background noise estimation procedure:
  1) Smoothing:
  $$\overline{|x(n)|} = \alpha\,\overline{|x(n-1)|} + (1-\alpha)\,|x(n)|$$
  2) Minimum value, with a slight increase to avoid a freezing of the estimate:
  $$\overline{|b(n)|} = \min\left\{\overline{|x(n)|}, \overline{|b(n-1)|}\right\}(1+\epsilon) \qquad \text{with: } \epsilon << 1$$
  ϵ : determines the tracking capabilities of the estimator

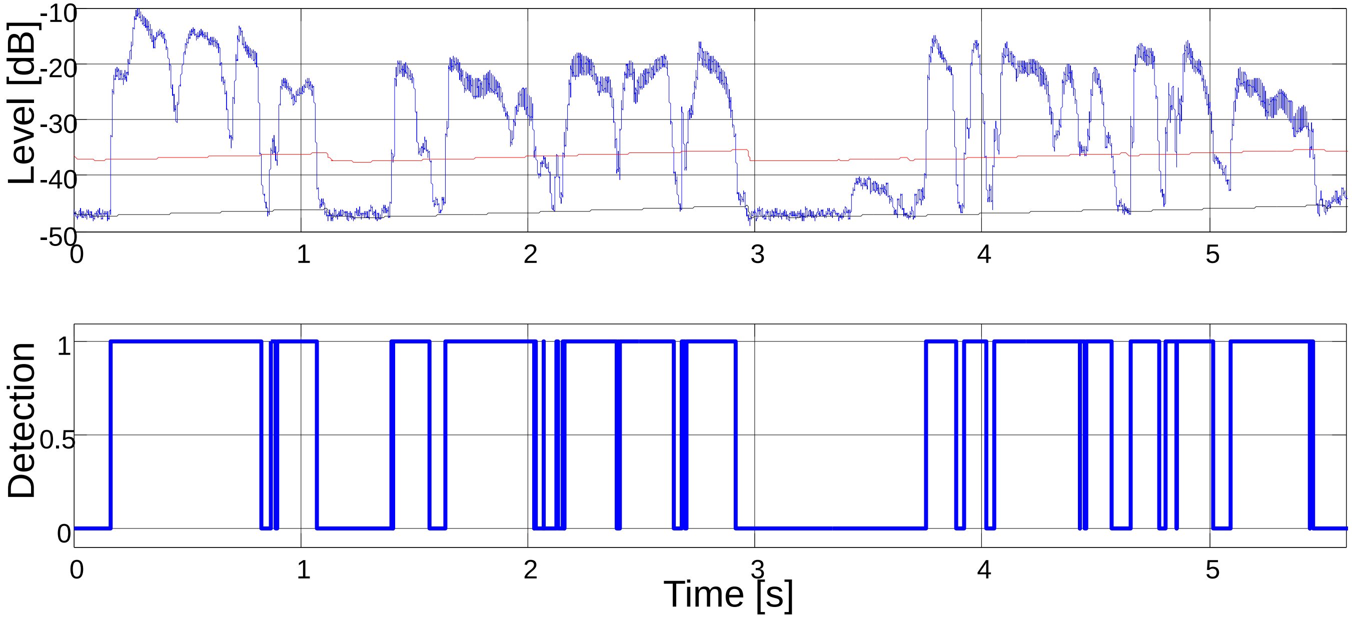


Speech activity detection

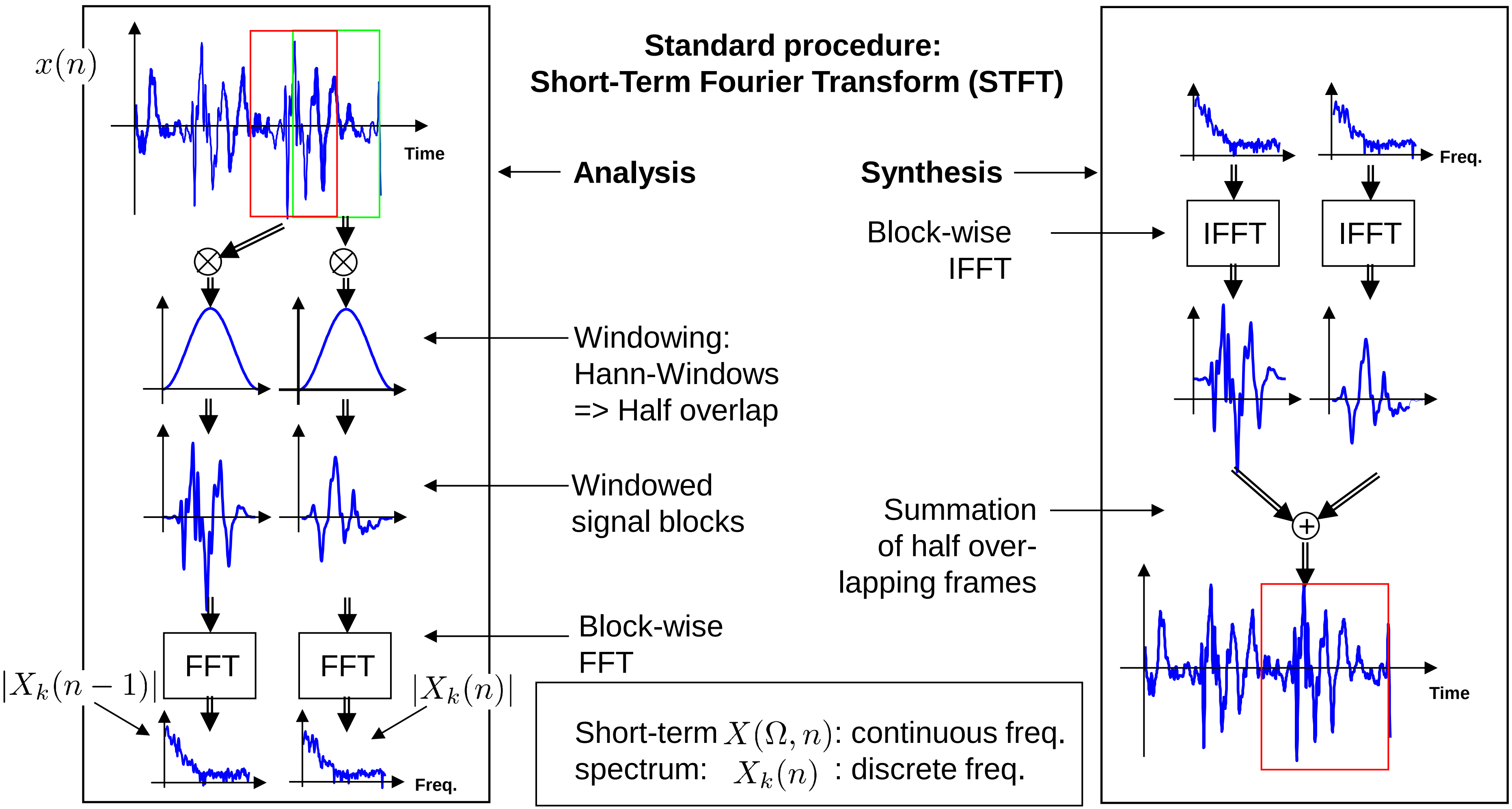
- Simple procedure for speech activity detection:
  - Compare the short-term power estimate with the (raised) noise power estimate:

$$D(n) = \begin{cases} 1 & : \quad \overline{|x(n)|} > K_b \, \overline{|b(n)|} \\ 0 & : \quad \text{else} \end{cases}$$

where K_b has been chosen to 10 dB, equals 3,16



Frequency domain processing: STFT



STFT: Formal description

- Extraction of a signal block of length *N* (~20-30 msec):

$$\boldsymbol{x}(n) = \left[x(n),\, x(n-1),\, ...,\, x(n-N+1)\right]^{\mathrm{T}}$$

- Windowing of the block:

$$\boldsymbol{x}_{\mathrm{F}}(n) = \left[x(n)\, h_0,\, x(n-1)\, h_1,\, ...,\, x(n-N+1)\, h_{N-1}\right]^{\mathrm{T}}$$

- Definition of a window matrix:

$$\boldsymbol{H} = \begin{bmatrix} h_0 & 0 & \dots & 0 \\ 0 & h_1 & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \dots & 0 & h_{N-1} \end{bmatrix}$$

STFT: Formal description

- Alternative notation of the windowing:

$$
\begin{aligned}
\boldsymbol{x}_{\mathrm{F}}(n) &= \begin{bmatrix} h_0 & 0 & \dots & 0 \\ 0 & h_1 & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \dots & 0 & h_{N-1} \end{bmatrix} \begin{bmatrix} x(n) \\ x(n-1) \\ \vdots \\ x(n-N+1) \end{bmatrix} \\
&= \boldsymbol{H}\,\boldsymbol{x}(n)
\end{aligned}
$$

- Fourier transform of the windowed signal:

$$
X\left(e^{j\Omega_\mu}, n\right) = \sum_{k=0}^{N-1} x(n-k)\, h_k\, e^{-j\frac{2\pi}{N}k\mu}
$$

STFT: Formal description

- Fourier transform of the windowed signal:

$$X\left(e^{j\Omega_\mu}, n\right) = \sum_{k=0}^{N-1} x(n-k)\, h_k\, e^{-j\frac{2\pi}{N}k\mu}$$

- Vector notation:

$$\boldsymbol{X}(e^{j\Omega}, n) = \left[X(e^{j\Omega_0}, n), X(e^{j\Omega_1}, n), \cdots, X(e^{j\Omega_{N-1}}, n)\right]^T$$

- DFT matrix:

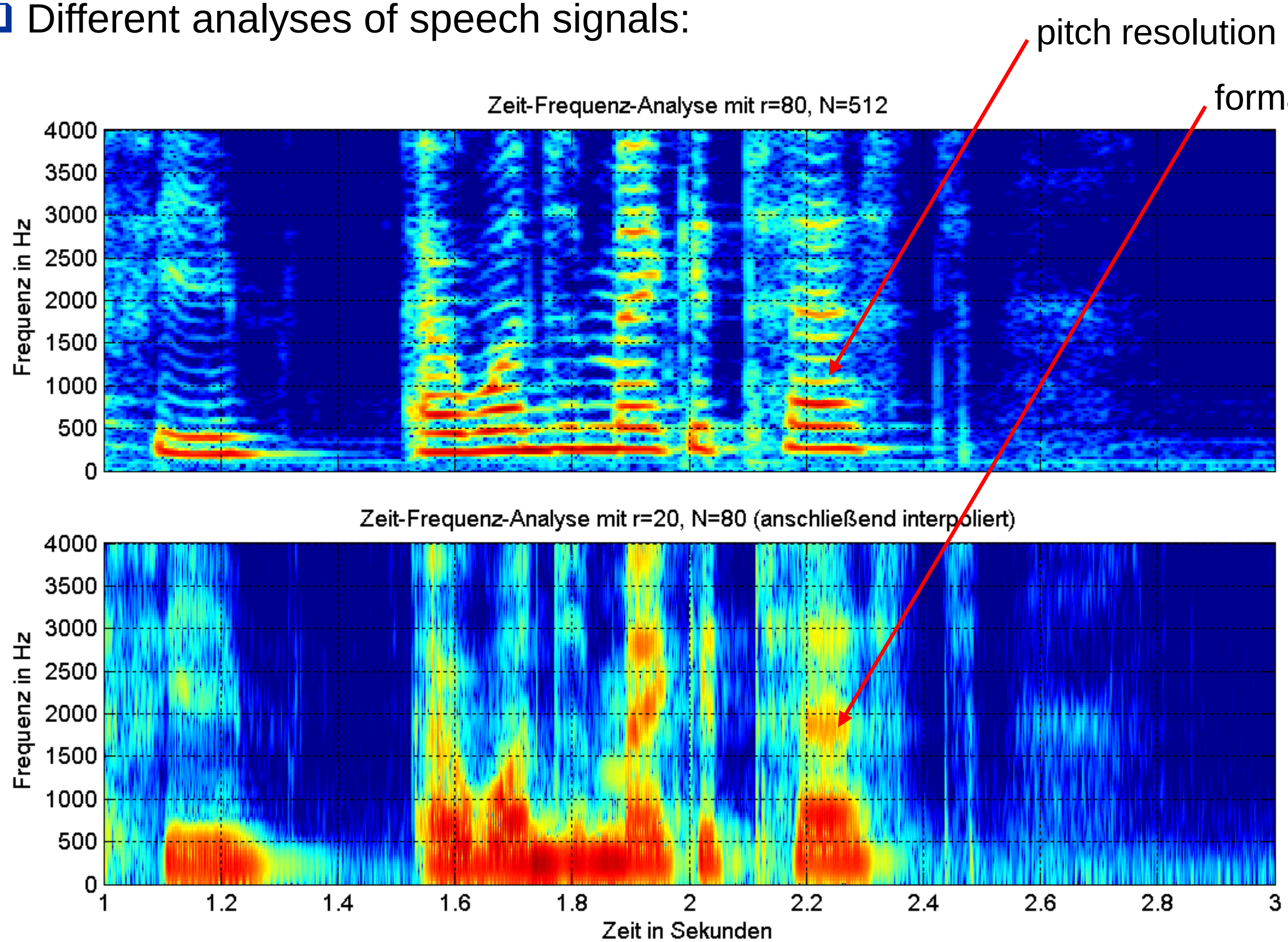
$$\boldsymbol{T} = \begin{bmatrix} e^{-j\frac{2\pi}{N}0} & e^{-j\frac{2\pi}{N}0} & \cdots & e^{-j\frac{2\pi}{N}0} \\ e^{-j\frac{2\pi}{N}0} & e^{-j\frac{2\pi}{N}1} & \ddots & \vdots \\ \vdots & \ddots & \ddots & e^{-j\frac{2\pi}{N}(N-2)(N-1)} \\ e^{-j\frac{2\pi}{N}0} & \cdots & e^{-j\frac{2\pi}{N}(N-1)(N-2)} & e^{-j\frac{2\pi}{N}(N-1)(N-1)} \end{bmatrix}$$

- Matrix vector notation:

$$\boldsymbol{X}(e^{j\Omega}, n) = \boldsymbol{T}\boldsymbol{H}\boldsymbol{x}(n)$$

Time-frequency analysis

- Different analyses of speech signals:



- high frequency resolution
- low time resolution

- low frequency resolution
- high time resolution

Estimation of the power spectral densities (PSD)

- The power spectral density is defined as the FT of the autocorrelation function.

- The auto-correlation function may be estimated based on the autocorrelation method:

$$\hat{r}_{xx}(\nu, n) = \frac{1}{N} \sum_{l=0}^{N-1-\nu} x(n+l)\, x(n+l+\nu), \quad \text{for } \nu = 0, 1, \ldots, N-1$$

$$x_w(n) = [\ldots, 0, 0, x(n), x(n+1), \ldots, x(n+N-1), 0, 0, \ldots]$$

$$\hat{r}_{xx}(\nu, n) = \frac{1}{N} x_w(n) \underbrace{*}_{\text{convolution}} x_w(-n) = \sum x(n)\, x(-(n-\nu)) = \sum x(n)\, x(n+\nu)$$

- The PSD can be calculated based on the periodogram which is the FT of the windowed signal:

$$\hat{S}_{xx,per}(\Omega_\mu, n) = \frac{1}{N} \left| X(e^{j\Omega_\mu}, n) \right|^2 \qquad X\left(e^{j\Omega_\mu}, n\right) = \sum_{k=0}^{N-1} x(n-k)\, h_k\, e^{-j\frac{2\pi}{N}k\mu}$$

with: $\Omega_\mu = \frac{2\pi}{N}\mu$

Here, h_k is a rectangular window

Estimation of the power spectral densities (PSD)

- For the estimation, typically smoothed periodograms are used. The smoothing is either performed with a rectangular

$$\hat{S}_{xx}(\Omega_\mu, n) = \frac{1}{N_p} \sum_{\nu=0}^{N_p-1} \hat{S}_{xx,per}(\Omega_\mu, n-\nu)$$

or an exponential window:

$$\hat{S}_{xx}(\Omega_\mu, n) = (1-\lambda) \sum_{\nu=0}^{\infty} \lambda^\nu \hat{S}_{xx,per}(\Omega_\mu, n-\nu)$$

and the equivalent recursive calculation:

$$\hat{S}_{xx}(\Omega_\mu, n) = \lambda \, \hat{S}_{xx}(\Omega_\mu, n-1) + (1-\lambda) \, \hat{S}_{xx,per}(\Omega_\mu, n)$$

Summary & Outlook

- Notations
- Speech signal analysis
  - Human speech generation, Acoustic signal propagation
    Acoustic signal perception => The human ear
- Sample-based vs. block-based processing
- Basic processing schemes
  - Power estimation, Non-linear smoothing, Minimum power / noise power estimation, Speech activity detection
- Next lecture: Prediction & Codebook based processing