

Digital Audio and Speech Processing (EE4182)

Richard C. Hendriks

19/4/2021

1

About the Course - People

Instructor:

- Richard Hendriks (R.C.Hendriks@tudelft.nl)
 - Associate prof. Circuits & Systems group
 - Msc coordinator S&S.
 - Expertise: Signal processing for 1) audio & speech 2) Biomedical



About the Course - People

Guest stars:

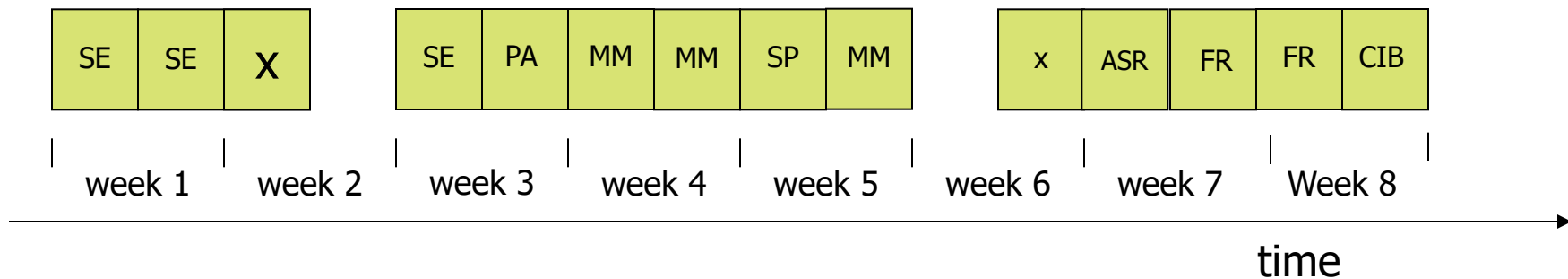
- Dr. ir. Richard Heusdens (R.Heusdens@tudelft.nl) (Defense academy)
- Dr. Odette Scharenborg (EWI, CS)
- Dr. Nikolay Gaubitch (Research manager Pindrop Security)



About the Course - Topics

The course covers different topics:

- the basics of speech production (SP)/auditory perception (AP)
- the basics of speech enhancement (multi/single microphone noise reduction) (MM/SE)
- Automatic Speech recognition (ASR)
- Applications:
 - Clock synchronization invariant beamforming (CIB)
 - Binaural spatial cue preservation for hearing aids (BC)
 - feature representation of audio (FR)



About the Course - Structure

The course consists of

- **Plenary lectures**

Goal: give easier access to reading material and highlight important points for project work

- Presence is *highly recommended* !

- One **mini project**, carried out in project groups of two students.

About the Course - Structure

The course gives 6 ECTS = $6 \times 28 = 168$ hours.

- plenary lectures: $12 \times 2 = 24$ hours
- preparation/studying lectures: $12 \times 3 = 36$ hours
- preparation exam: 24 hours
- mini project: $6 \times 14 = 84$ hours

168 hours

About the Course - Website

<http://cas.tudelft.nl/Education/courses/in4182/index.php>

- Course Information
 - Course organization (lectures, project, evaluation)
 - Link to Study guide information (teaching goals, etc.)
- Course Documents
 - Slides
 - links to literature and papers
- Mini-project
 - Project assignments
 - Audio, speech and noise material for the projects
- Brightspace: Used for announcements, to sign up for the exam, and for report submission

About the Course - Evaluation

Online oral examination

Grade is based on oral discussion on

- Project (group) report (handed in before June 18th)
- Assignment (individual) (handed in before June 18th)

The grade will be based on the discussion and question & answer.
Not on the reports themselves.

- Sign up via brightspace before June 2nd.

Speech Enhancement: Overview

The focus of the course is on *Speech Enhancement*:

- Lecture 1: Introduction, simple systems based on short-time Fourier transform, spectral subtraction. – Sec. 1,2,3, 4 intro + 4.2
- Lecture 2: More advanced noise suppression techniques based on short-time Fourier Transform, minimum mean-square error estimators. - Sec. 4, 7.1
- Lecture 5: Techniques for estimating and tracking noise power spectral density (in presence of speech) – Sec. 6.
- Lecture 6,7 & 9: Multi-microphone speech enhancement
- Lecture 14: Clock synchronization invariant beamforming

Literature:

R. C. Hendriks, T. Gerkmann, and J. Jensen, "DFT-Domain

Based Single-Microphone Noise Reduction for Speech Enhancement: A Survey of the State of the Art",
Morgan & Claypool, 2013.

For sale at Morgan & Claypool for \$30:

<http://www.morganclaypool.com/doi/pdfplus/10.2200/S00473ED1V01Y201301SAP011>

Speech Enhancement - Project

- Project is compulsory and carried out in groups of 2 students
- Q&A during oral discussion (hand in report before June 18th, brightspace)

Project:

- Design and build at least a single-microphone speech enhancement system (for far-end noise). You are free to extend this to a multi-microphone system.
 - Use matlab (or simulink)
 - The speech enhancement system should consist of a gain function, noise PSD estimator and speech PSD estimator.
 - Perform an evaluation of the speech enhancement system

Optional:

- Implement a multi-microphone system

Individual Assignment

- Individual (2 A4s)
- Q&A during oral discussion (hand in assignment before June 18th, brightspace)
- Topic: multi-microphone speech enhancement, and the related theory.

Speech Enhancement – Why?

Reduce effect of background noise on speech communication quality.

- Speech quality ('pleasantness', listener fatigue).
- Speech intelligibility.

Application Areas:

- human-to-human communication (example: digital hearing instruments, mobile phones, public address systems, conference systems, etc.).
- human-to-machine (example: voice-controlled devices, booking services, etc.).

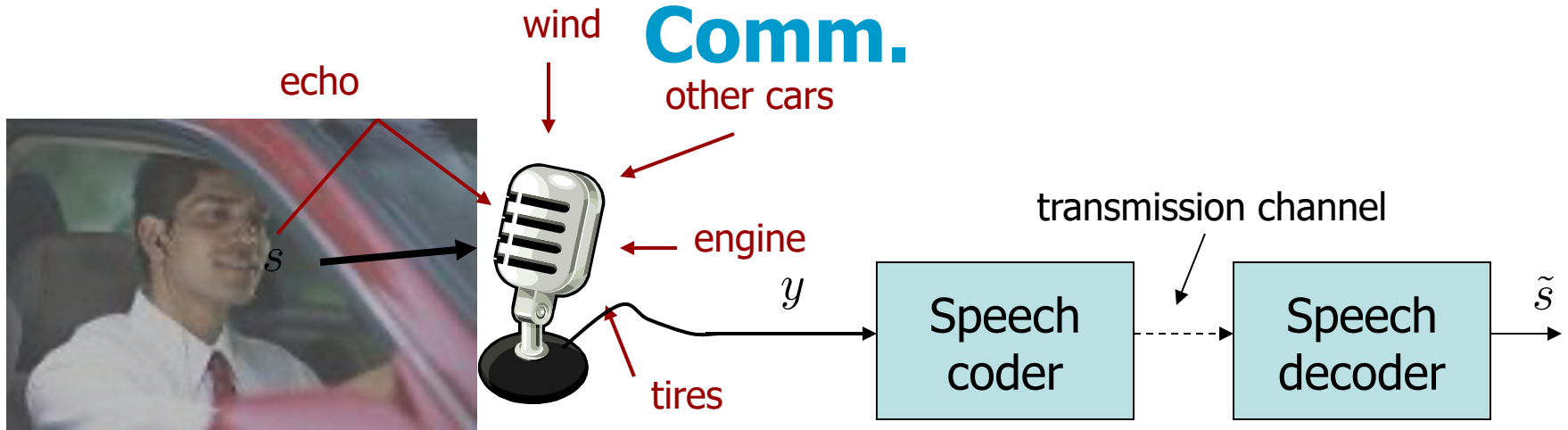
Example: Speech Enhancement for Dig. Comm.

Problem:

Generally digital speech communication systems (mobile telephony systems, automatic speech recognizers, etc.) are designed to work with relatively noise-free speech signals. If input signals to these systems are *noisy*, their performance drops since noisy speech doesn't satisfy the speech production model

- low-quality speech at receiving side of mobile phone.
- poor recognition performance.

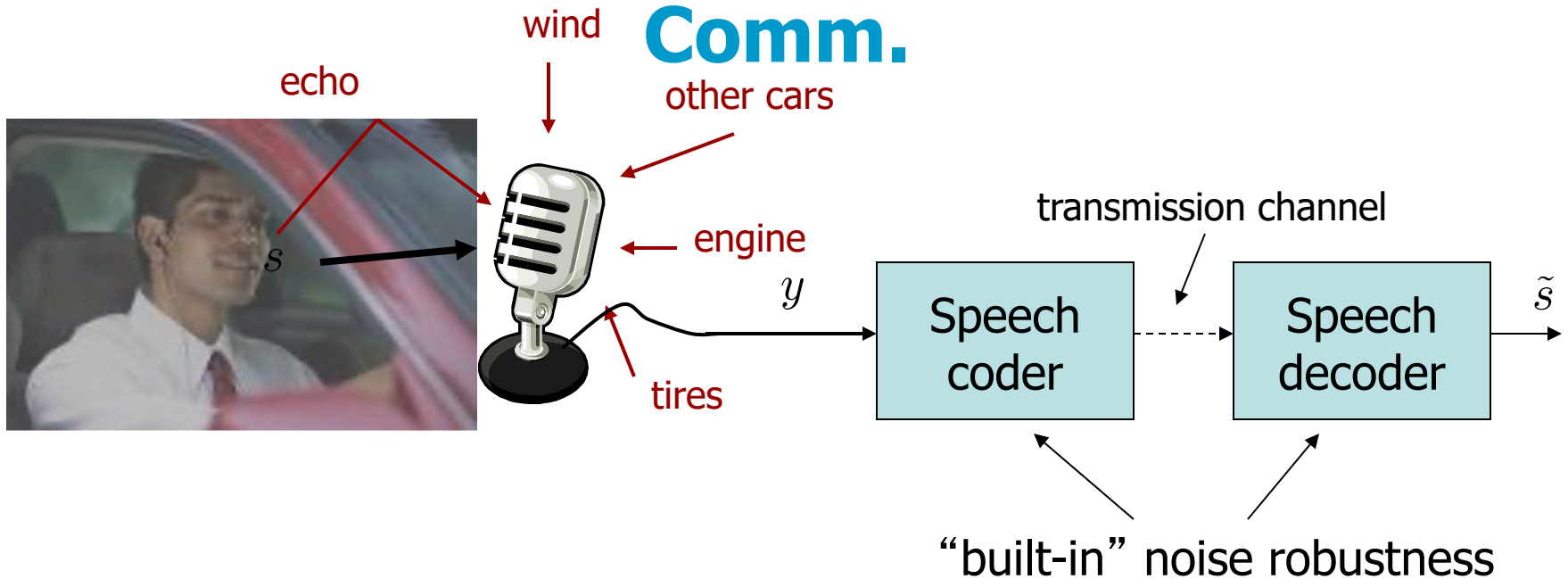
Example: Speech Enhancement for Dig. Comm.



Degradation of target due to:

- Car Noise
- Competing Speakers
- Echo
- Coding noise (modeling and quantization)
- Non-ideal channel

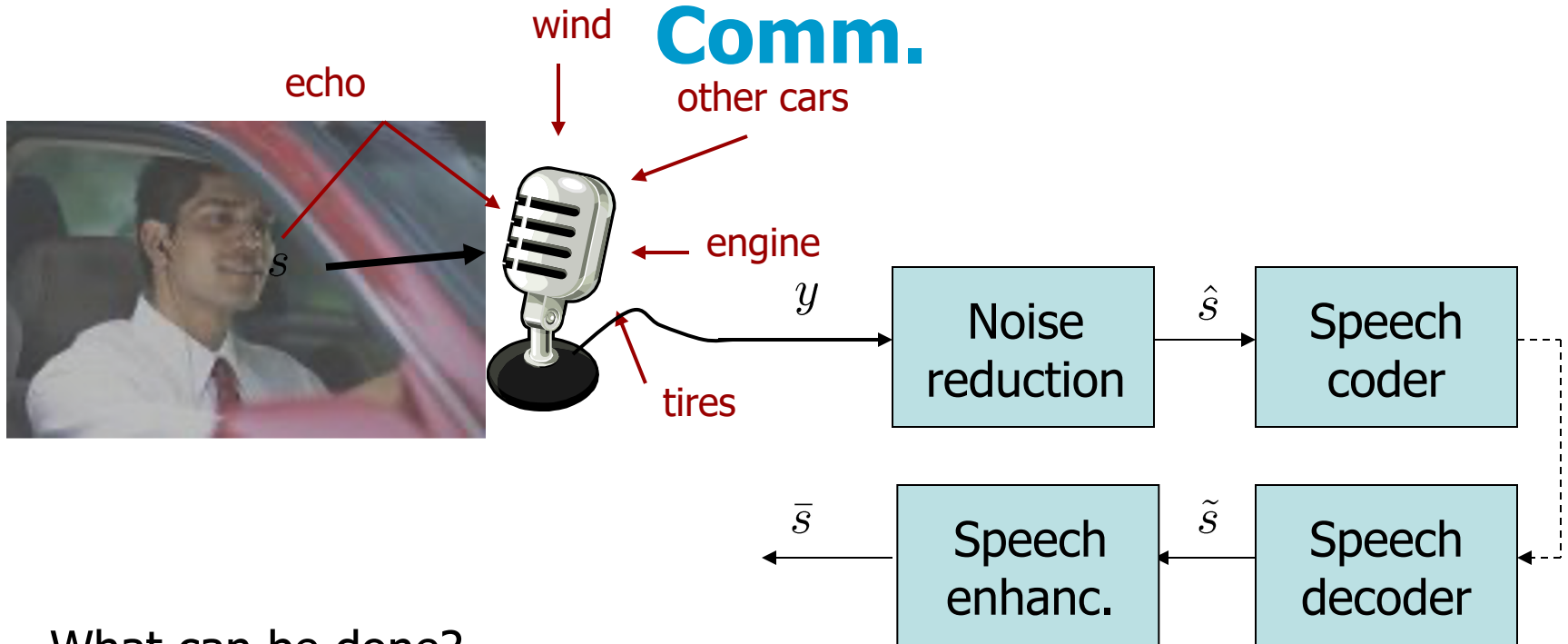
Example: Speech Enhancement for Dig. Comm.



What can be done?

- Develop new and more noise robust digital speech communication systems

Example: Speech Enhancement for Dig. Comm.



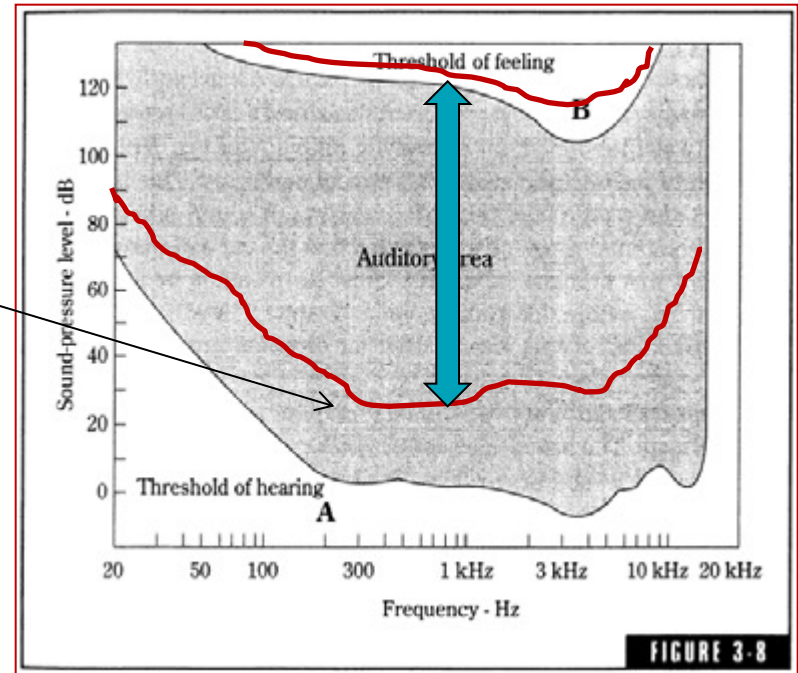
What can be done?

- Develop new and more noise robust digital speech communication systems
- Pre-process noisy signal before it enters speech communication systems

Example: Speech Enhancement for Hearing Devices

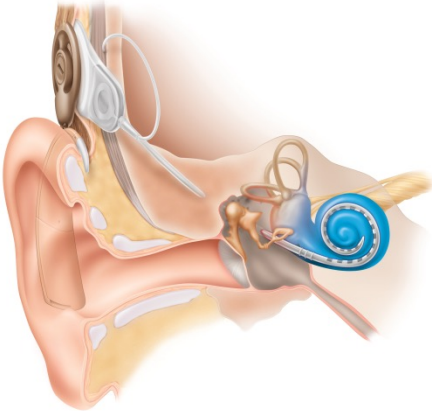
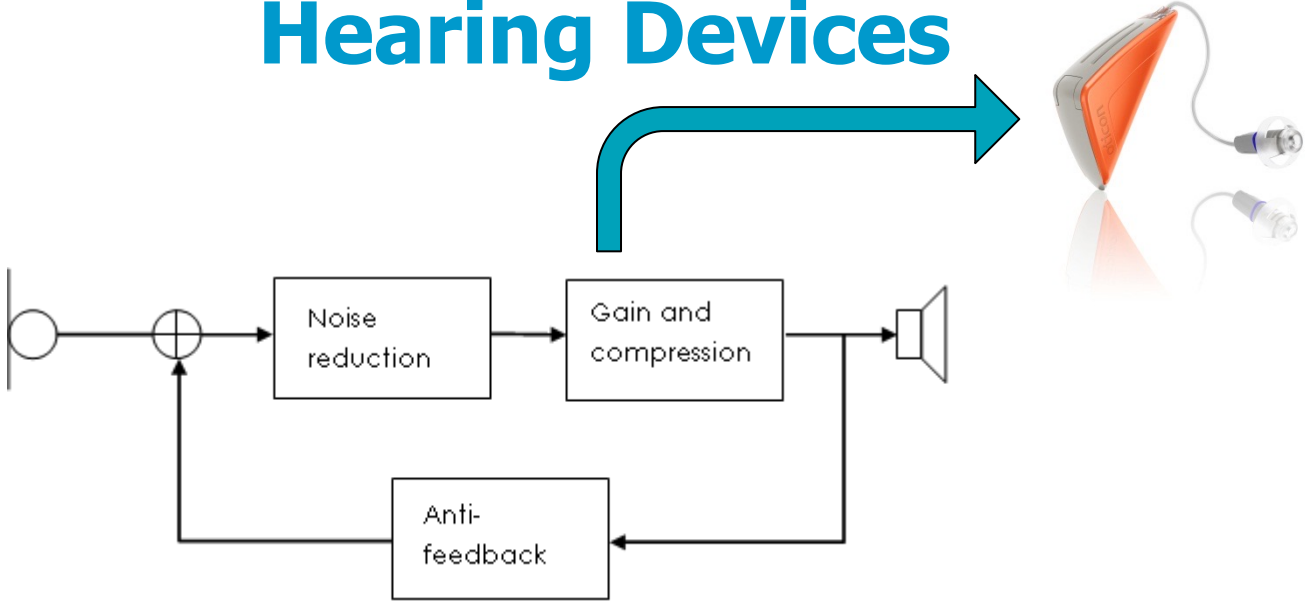


- Sensitivity
- Dynamic range
- Temporal resolution
- Frequency resolution
- Ability to exploit binaural cues



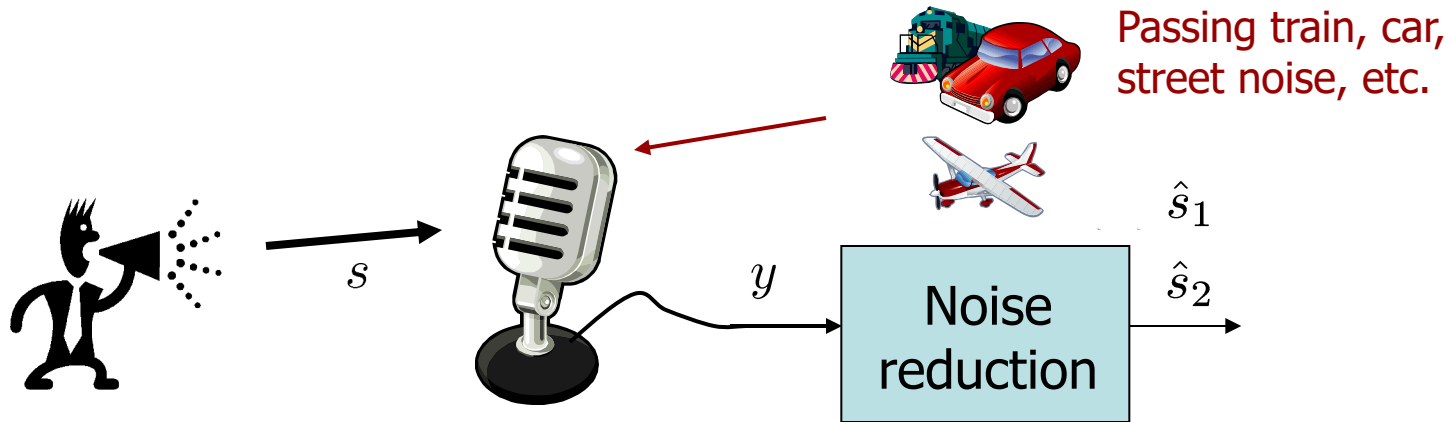
How to compensate for this?

Example: Speech Enhancement for Hearing Devices

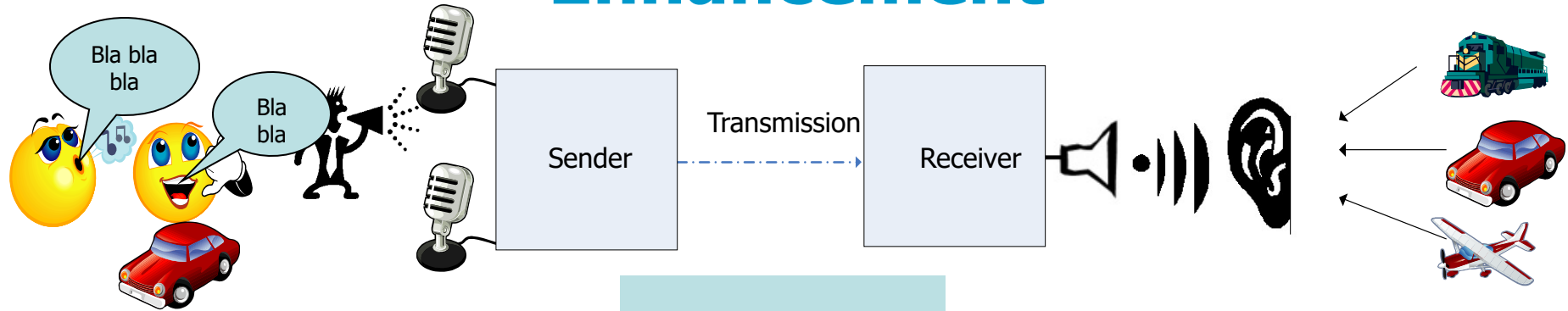


Example: Speech Enhancement for Hearing Devices

Example: single mic. noise reduction for non-stationary noise



Single and Multi-Microphone Speech Enhancement



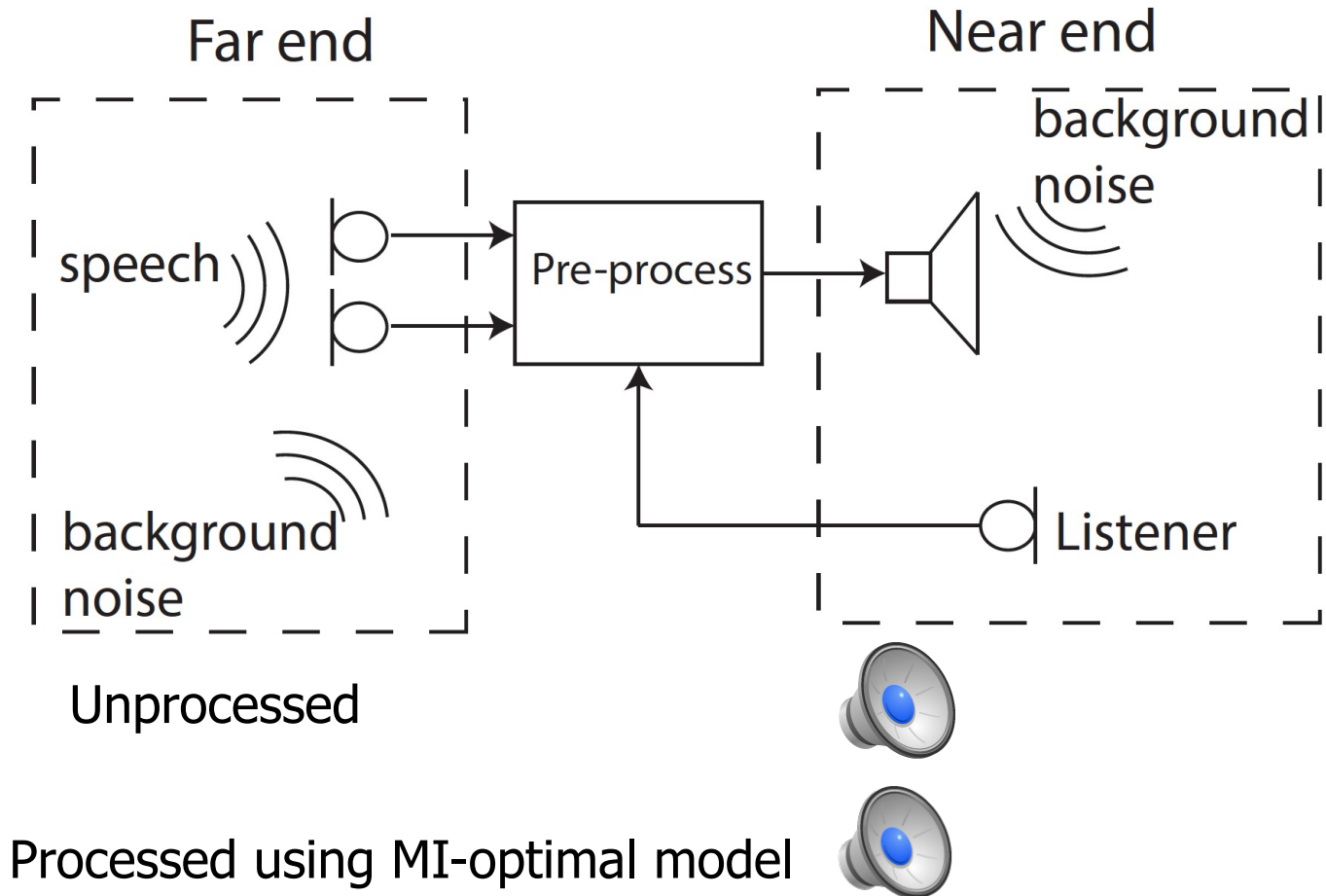
Far-end noise reduction

Applications:

- Hearing aids
- Mobile telephony
- Headsets
- Etc.

Near-end speech enhancement

Example: Near-end Speech Enhancement

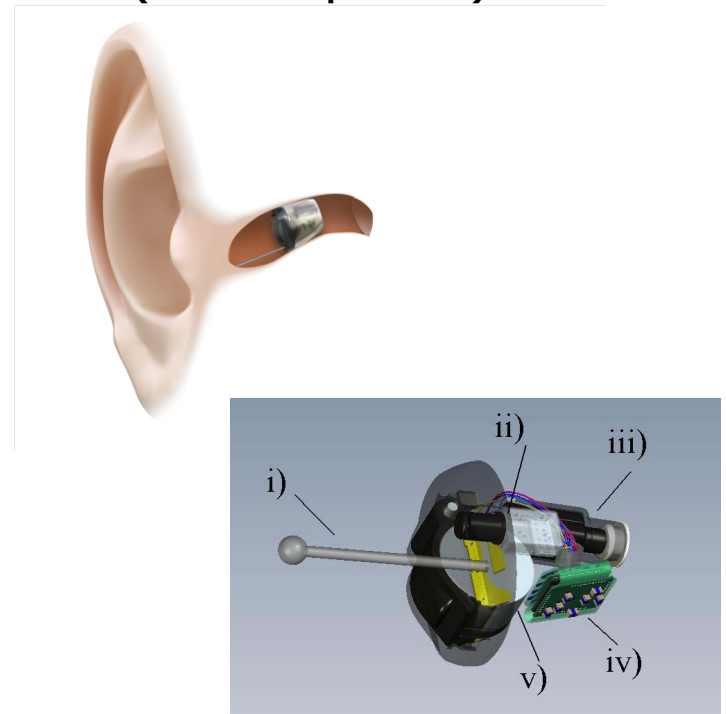


Single and Multi-Microphone Noise Reduction

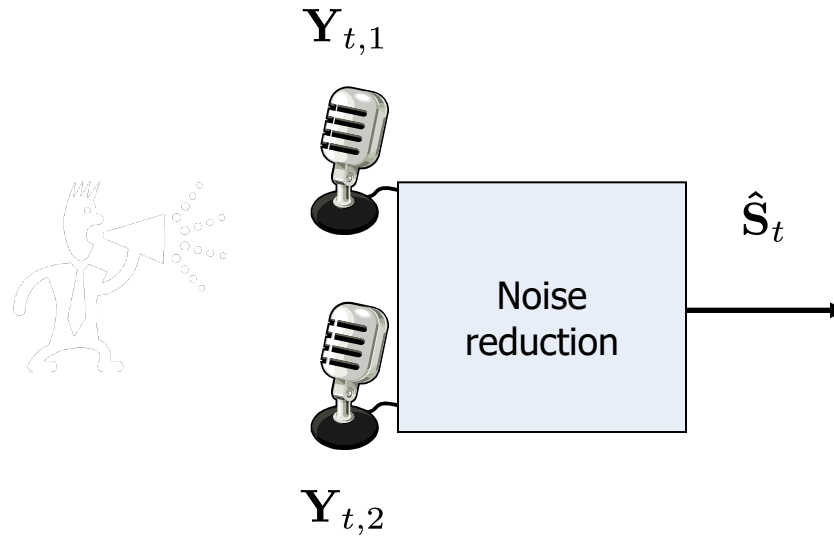
Behind the ear hearing aid
(2 microphones)



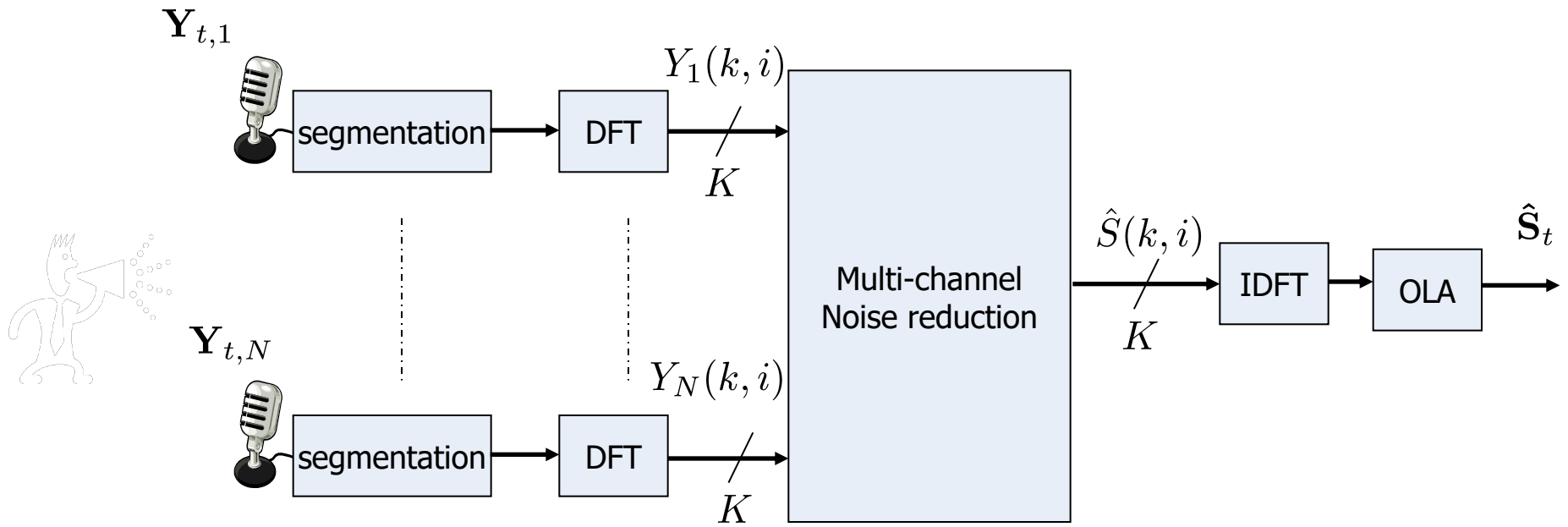
In the ear hearing aid
(1 microphone)



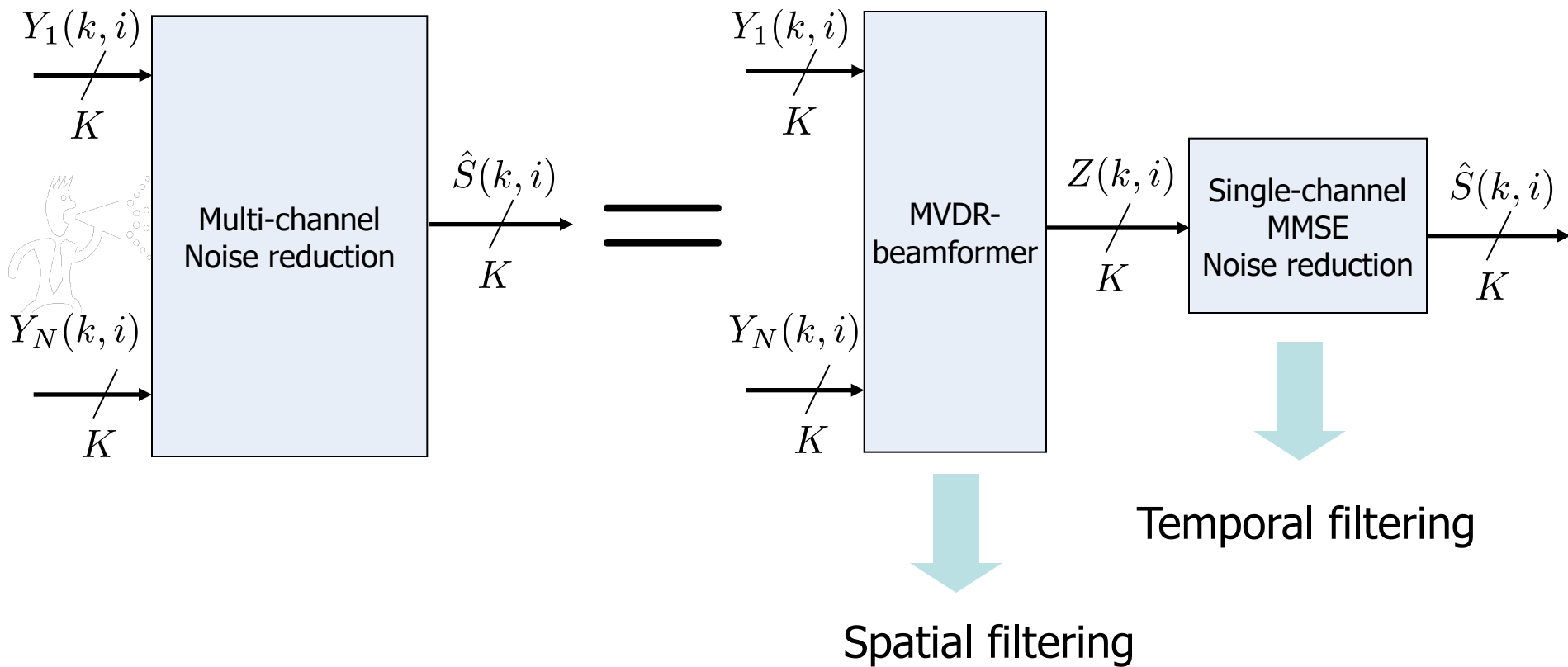
Multi-Microphone Noise Reduction



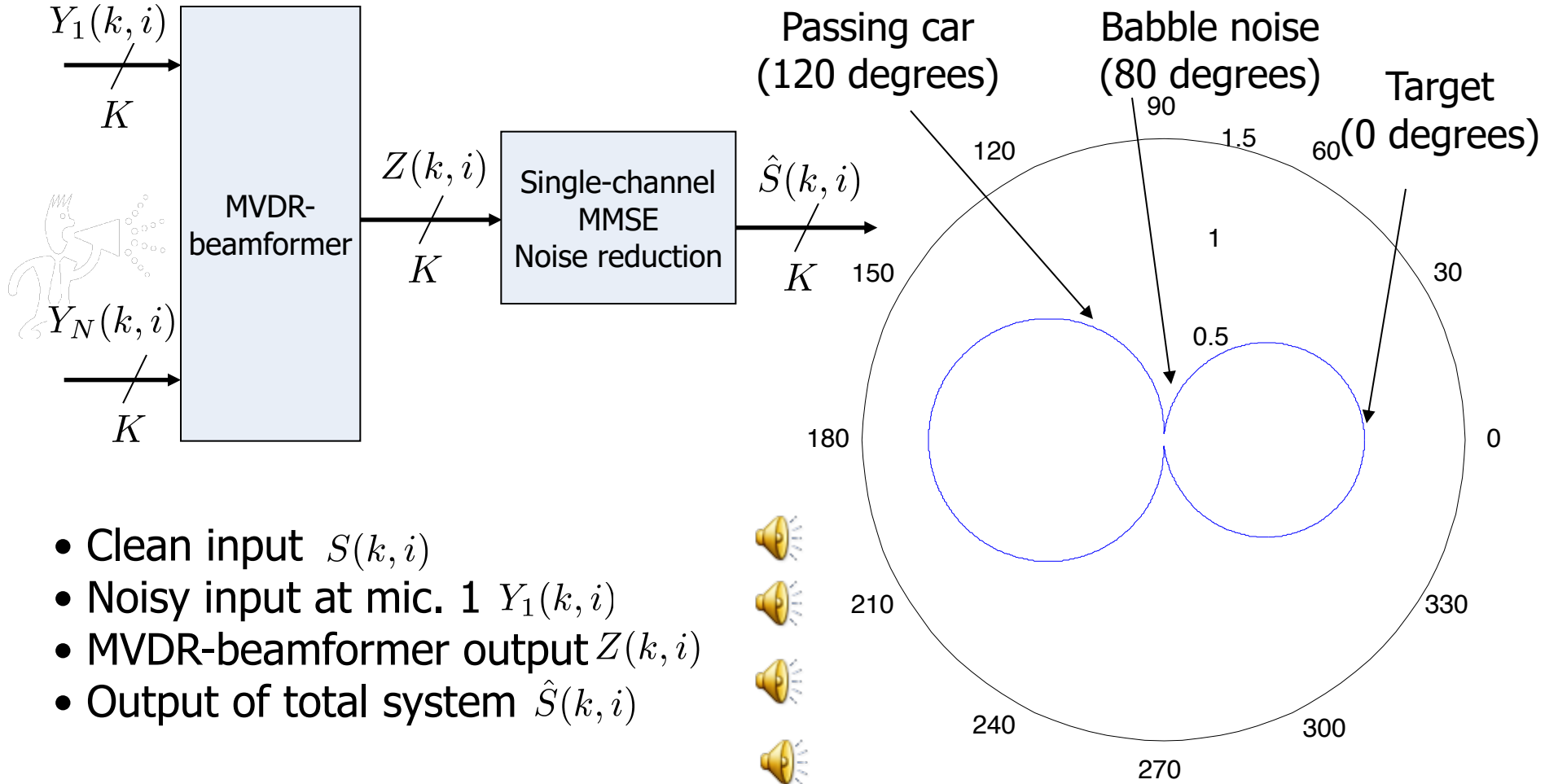
Multi-Microphone Noise Reduction



Multi-Microphone Noise Reduction



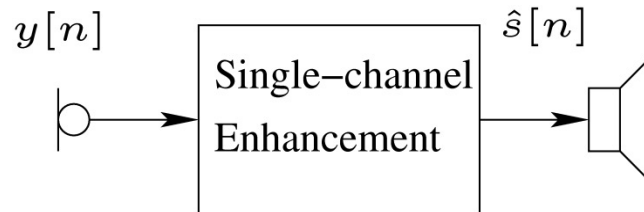
Example: Multi-Channel Noise Reduction



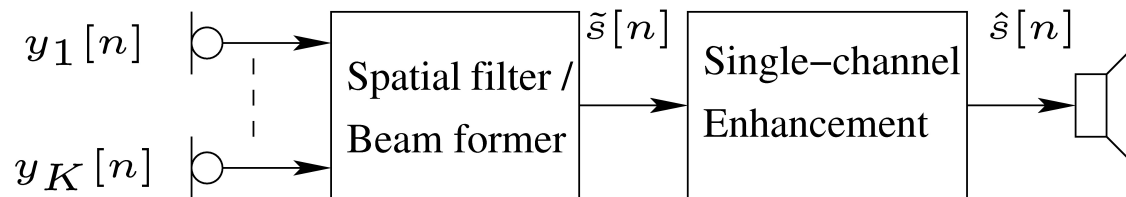
- Clean input $S(k, i)$
- Noisy input at mic. 1 $Y_1(k, i)$
- MVDR-beamformer output $Z(k, i)$
- Output of total system $\hat{S}(k, i)$



Is Single-Channel Speech Enhancement Still Relevant?



Yes, as it is not only used in the single-channel application, but also in the multi-channel application as a post-processor



Speech Enhancement – What?

Design criteria:

- Delay/latency, i.e., real-time vs. off-line processing
- Complexity (e.g. for battery-driven applications, hearing aids, mobile phones, etc., power and thus complexity is limited)
- Trade-off speech distortion vs. noise suppression
 - “pleasantness”
 - intelligibility

Speech Enhancement – What?

Design criteria (continued):

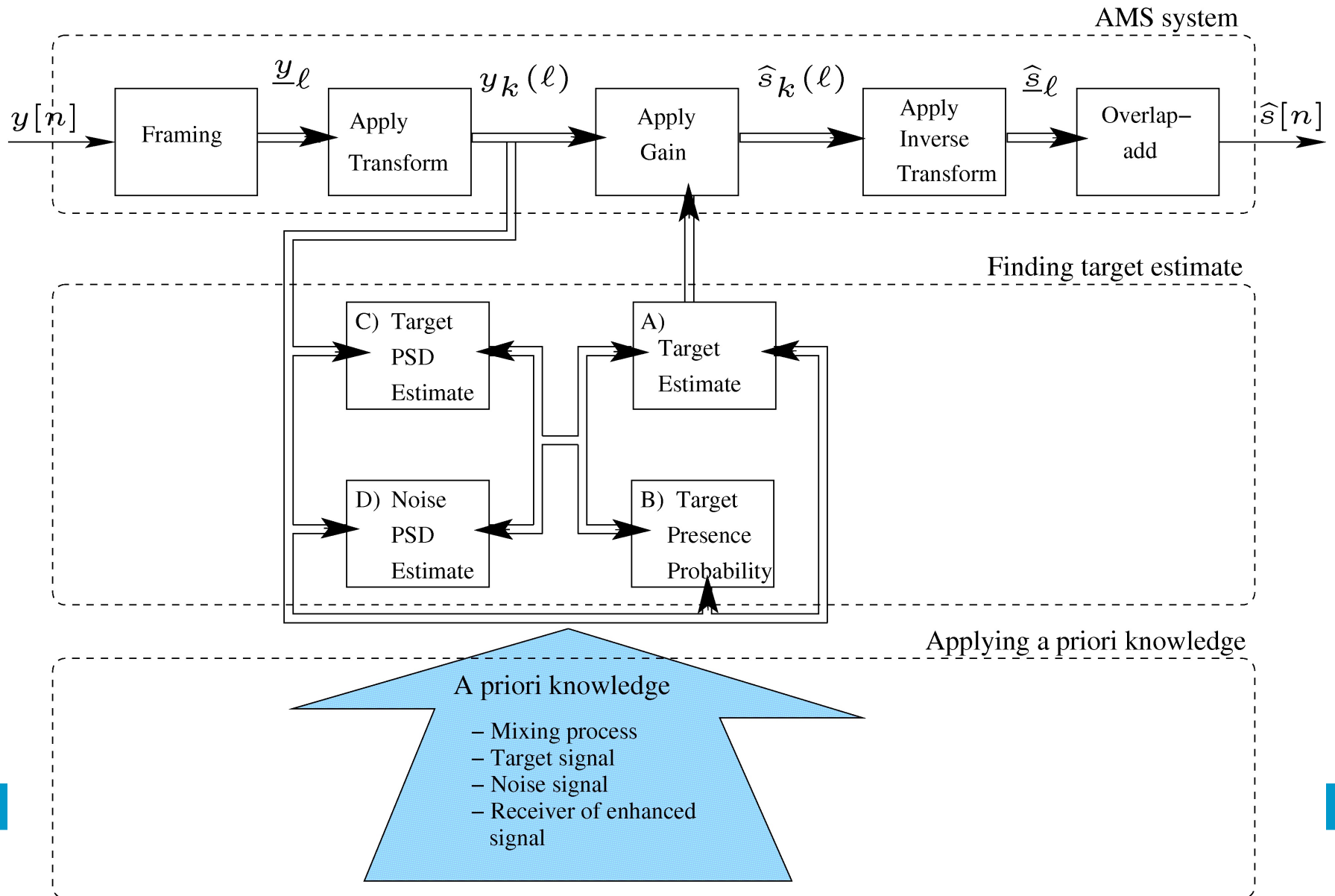
- multi- vs. single-sensor solutions. Multi-sensor solutions are in principle superior in terms of performance, but they
 - require more space.
 - are more complexity/power consuming.
 - are more expensive (hardware).

Speech Enhancement – Our focus

We focus on solutions which are applicable in typical mobile communication applications:

- Single/multi-microphone (additive) noise suppression
- Real-time algorithms
- Relatively low complexity
- Speech quality (i.e., 'pleasantness') over intelligibility
- FFT-based enhancement schemes

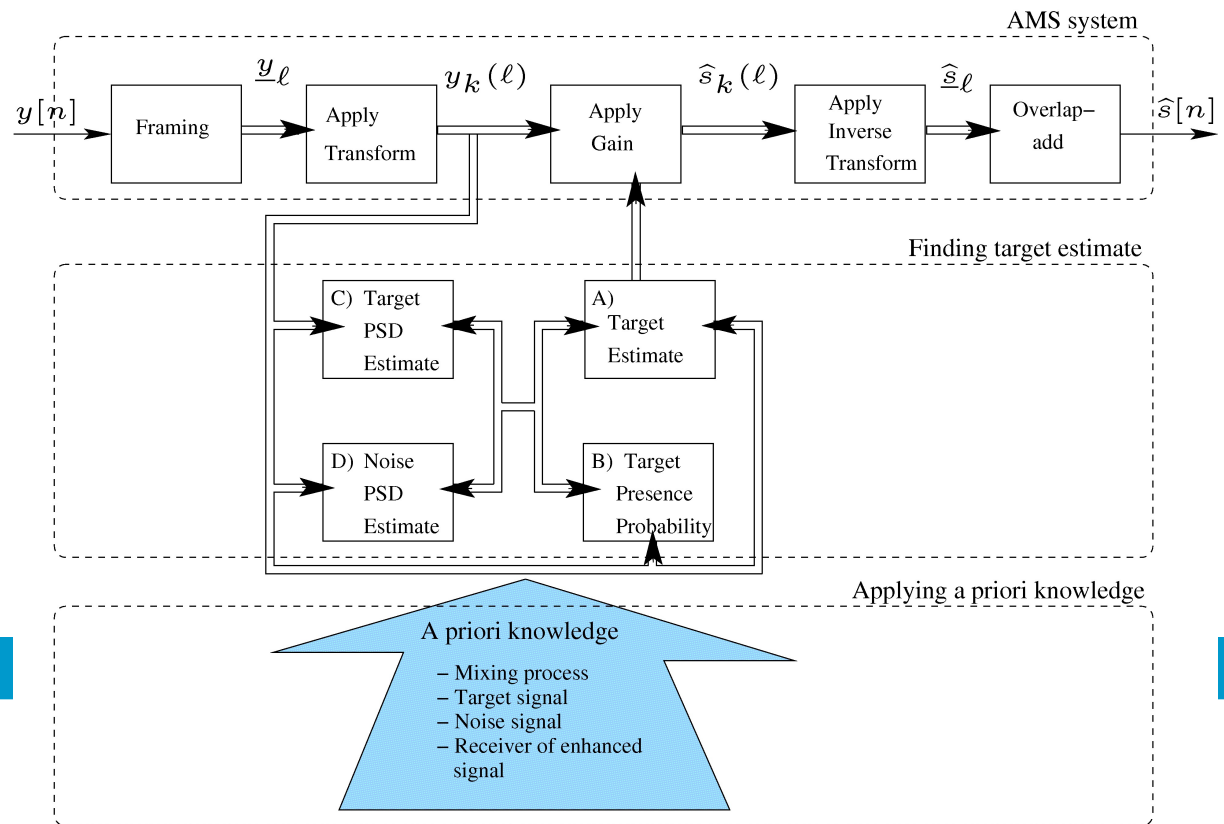
Overview of single-channel NR algorithm



The Transform

Generally, enhancement is performed on the spectral coefficients by first applying a spectral transform to the noisy time-frame.

Why?



The Transform

The transform makes the spectral coefficients uncorrelated (or even independent)

⇒ process spectral coefficients independently across frequency.

Which transform?

- Karhunen-Loeve Transform (KLT): is optimal, but data dependent and very complex.
- Short-time Fourier Transforms using a discrete Fourier transform (DFT).

The Transform

The DFT is most popular because:

1. Very efficient and low complex implementation
2. Delivers approximately uncorrelated spectral coefficients
3. Speech enhancement performance usually similar compared to KLT transform.
4. Allows an easy interpretation of the spectral signal content (useful to link the spectra to knowledge about speech production and perception).

Therefore, during the next lectures we focus on the DFT.

The Transform

Additional often argued advantage: Applying the DFT is often argued to make the coefficients more Gaussian.

Remember the central limit theorem: The distribution of a sum of independent random variables tend to approach a Gaussian distribution.

The DFT transform applied to a frame of speech

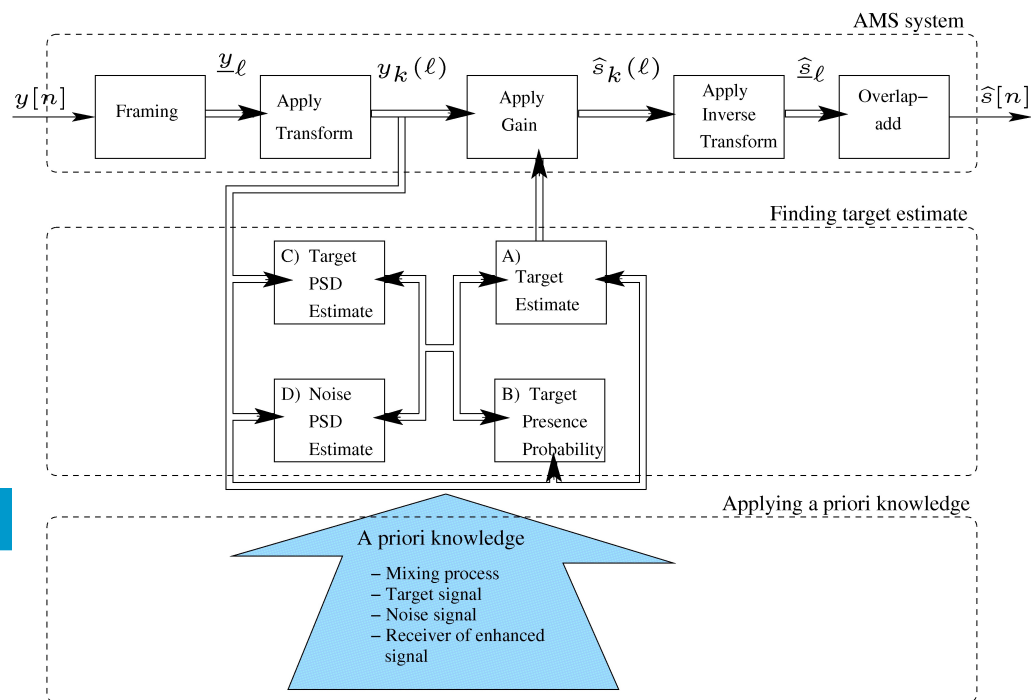
$$S_k(l) = \sum_{n=0}^{N-1} w(n)S(n)\exp\left(\frac{-j2\pi nk}{N}\right)$$

is a sum of samples. Is it also a sum of *independent* samples? We will come back to this in three lectures.

Taking A Priori Knowledge Into Account

The more a priori knowledge is taken into account, the better the performance will generally be (if the assumption on the a priori knowledge is right!)

Hence: Assuming that the input is speech, while it is music, might degrade quality.



Taking A Priori Knowledge Into Account

What kind of a priori knowledge can be used?

- Mixing process
- Taking into account that the target signal is speech. For example:
 - Statistical distribution of speech
 - Take knowledge of speech production process into account (speech can be modelled as an AR process)
 - Make sure that output signal shows speech-like characteristics (across time and/or frequency)
 - Use HMMs or codebooks trained on speech

Taking A Priori Knowledge Into Account

What kind of a priori knowledge can be used? (continued)

- Taking noise signal characteristics into account. For example:
 - Statistical distribution
 - Assume that noise process changes slower than the speech process
 - Use codebooks or HMMs trained on noise

Taking A Priori Knowledge Into Account

What kind of a priori knowledge can be used? (continued)

- Knowledge of the human auditory system. For example:
 - Minimize a distortion measure that reflects perceptual aspects.
 - Take simple models of the auditory system into account.

Speech Enhancement – Basic Assumptions and notation

For random variables we use upper case letters: S

For realizations we use lower case letters: s

- time-domain random variable: S_t
- time-domain realization: s_t
- Frequency-domain random variable: S
- Frequency-domain realization: s

Speech Enhancement – Basic Assumptions and notation

Additive Noise:

$$Y_t(n) = S_t(n) + N_t(n)$$

- S_t : stochastic process representing clean speech signal (our target – cannot be observed).
- N_t : stochastic process representing noise. (cannot be observed).
- Y_t : stochastic process representing noisy signal. *Can* be observed (one realization).
- n : discrete-time index.

Speech Enhancement – Basic Assumptions and notation

For the DFT coefficients this means:

$$\mathcal{F}\{Y_t(n)\} = Y_k(l) = S_k(l) + N_k(l)$$

- Y , S and N : Noisy speech, speech and noise DFT coefficients
- k : frequency bin index.
- l : time frame index.

Due to the fact that the gain function is assumed to be applied independently per time-frame and frequency bin, we usually neglect the indices k and l .

Speech Enhancement – Basic Assumptions and notation

Speech and noise independent and uncorrelated:

$$R_{S_t N_t}(n, m) = E \{S_t(n)N_t(m)\} = 0,$$

$$Y_t(n) = S_t(n) + N_t(n)$$

Consequently, we get:

$$\begin{aligned} R_{Y_t Y_t}(n, m) &= E \{Y_t(n)Y_t(m)\} \\ &= E \{(S_t(n) + N_t(n))(S_t(m) + N_t(m))\} \\ &= R_{S_t S_t}(n, m) + R_{N_t N_t}(n, m) + \underbrace{2R_{S_t N_t}(n, m)}_0 \end{aligned}$$

Speech Enhancement – Basic Assumptions and notation

(Short-time) wide-sense stationarity (wss):

- Mean: $E \{S_t(n)\} = m_s (= 0) \forall n$
- Correlations:

$$\begin{aligned}R_{S_t S_t}(n, m) &= E\{S_t(n)S_t(m)\} = R_{S_t S_t}(m - n). \\R_{N_t N_t}(n, m) &= E\{N_t(n)N_t(m)\} = R_{N_t N_t}(m - n) \\R_{Y_t Y_t}(n, m) &= R_{S_t S_t}(n, m) + R_{N_t N_t}(n, m) \\ &= R_{S_t S_t}(m - n) + R_{N_t N_t}(m - n).\end{aligned}$$

We see that $R_{Y_t Y_t}(n, m)$ also only depends on $m - n$.

Speech Enhancement – Basic Assumptions and notation

Assumptions in summary:

- Speech and noise are additive
- Speech and noise are uncorrelated
- Speech and noise are assumed to be wide sense stationary

Speech Enhancement – Basic Assumptions and notation

Validity of assumptions:

- Additivity: This assumption holds in many situations. However, when the noise is due to reverberation we deal with convolutional noise.
- Uncorrelatedness: in many cases noise originates from process independent of speech production (e.g. car). However, sometimes there is acoustic feedback from noise to speaking person, i.e., speaker may adapt speaking style to combat noise (Lombardt effect).
- Stationarity: speech typically assumed short-time wss, i.e., across time intervals of 20-30 ms.
Noise anything is possible!

Power Spectral Densities

The power spectral density (psd) $P_{YY,k}$ of a stochastic process $Y_t(n)$ is defined as the Fourier transform of the corresponding auto-correlation sequence $R_{Y_t Y_t}$.

$$P_{YY,k} = \lim_{L \rightarrow \infty} \sum_{m=-L/2}^{L/2} R_{Y_t Y_t}(m) e^{-j2\pi \frac{km}{K}}.$$

Since the Fourier transform is a linear transform, the uncorrelatedness of $S_t(n)$ and $N_t(n)$ leads to

$$\begin{aligned} P_{YY,k} &= \lim_{L \rightarrow \infty} \sum_{m=-L/2}^{L/2} R_{S_t S_t}(m) e^{-j2\pi \frac{km}{K}} + \lim_{L \rightarrow \infty} \sum_{m=-L/2}^{L/2} R_{N_t N_t}(m) e^{-j2\pi \frac{km}{K}} \\ &= P_{SS,k} + P_{NN,k} \end{aligned}$$

Power Spectral Densities

The PSD can also be written as a function of the DFT coefficients:

$$\begin{aligned} P_{YY,k}(l) &= \frac{1}{L} E [|Y_k(l)|^2] \\ &= \frac{1}{L} E [|S_k(l)|^2 + |N_k(l)|^2 + 2\Re (S_k(l)N_k(l))] \\ &= \frac{1}{L} E [|S_k(l)|^2] + \frac{1}{L} E [|N_k(l)|^2] \end{aligned}$$

Another often used notation is using

$$E [|Y_k(l)|^2] = \sigma_{Y,k}^2(l) = \sigma_{S,k}^2(l) + \sigma_{N,k}^2(l),$$

as $E [|Y_k(l)|^2]$, $E [|S_k(l)|^2]$ and $E [|N_k(l)|^2]$ are the variances of the noisy speech, speech and noise DFT coefficients.

Estimating Power Spectral Densities

Hence, the psd is an *expected* value:

$$P_{YY,k}(l) = \frac{1}{L} E [|Y_k(l)|^2]$$

In practice it must be estimated from available data. Let $Y_t(n)$, $n = 0, \dots, L - 1$ denote an L sample observation vector.

The periodogram estimator (Schuster, 1899) is then defined as:

$$\hat{P}_{YY,k}^P(l) = \frac{1}{L} \left| \sum_{n=0}^{L-1} Y_t(n) e^{-j2\pi kn/L} \right|^2 = \frac{1}{L} |Y_k(l)|^2.$$

Estimating Power Spectral Densities

- It can be shown that for the periodogram

$$\lim_{L \rightarrow \infty} E\{\hat{P}_{YY,k}^P(l)\} = P_{YY,k}(l)$$

i.e., the periodogram is asymptotically unbiased.

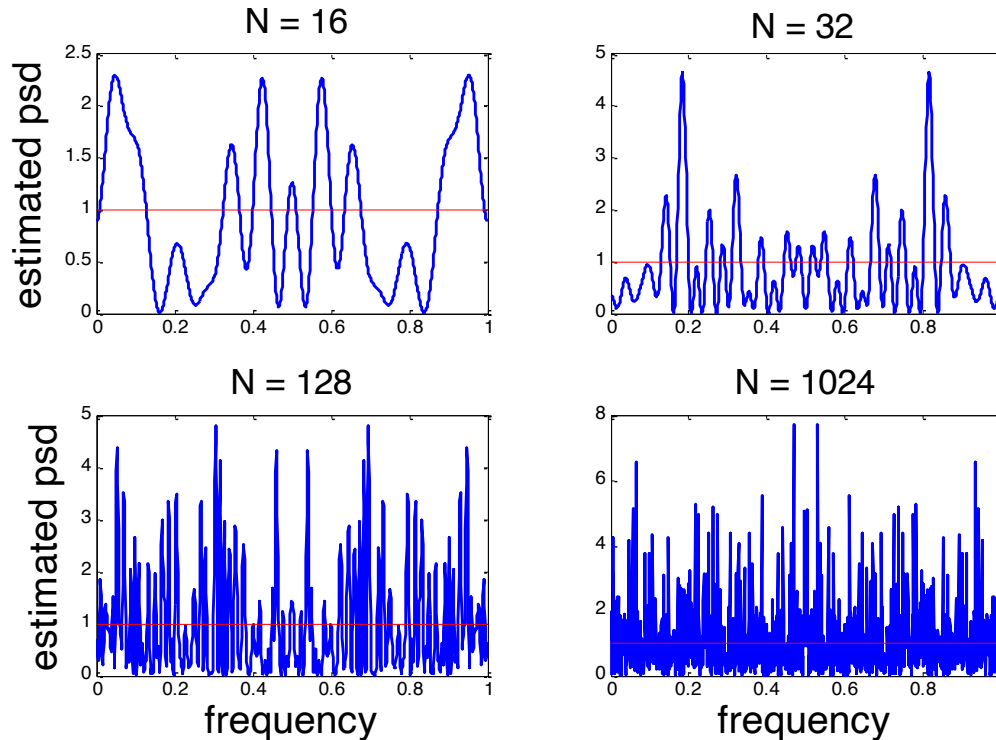
- Also, it can be shown

$$\lim_{L \rightarrow \infty} Var\{\hat{P}_{YY,k}^P(l)\} = P_{YY,k}^2(l)$$

i.e., the variance does not decrease for longer data records.

Estimating Power Spectral Densities

Estimated psd for a Gaussian process with $N \in \{16, 32, 128, 1024\}$.



We see that indeed the variance of the periodogram does *not* decrease for longer data records.

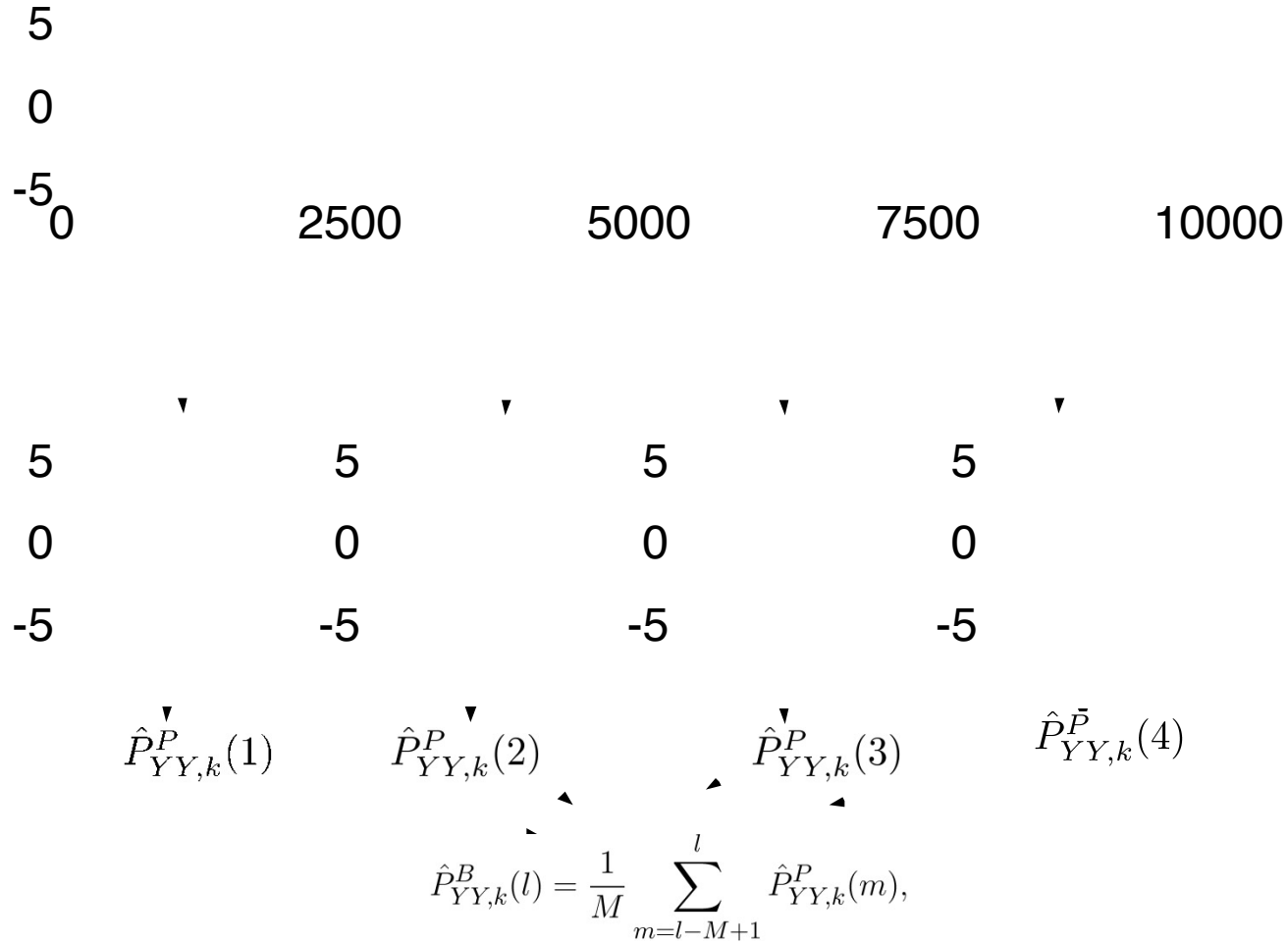
Estimating Power Spectral Densities

The Bartlett estimate is computed as the average of several periodograms, taken from *different* signal segments:

$$\hat{P}_{YY,k}^B(l) = \frac{1}{M} \sum_{m=l-M+1}^l \hat{P}_{YY,k}^P(m),$$

where $\hat{P}_{YY,k}^P(m)$ is the periodogram of the m 'th signal segment.

Bartlett Estimate



Bartlett Estimator

- It can be shown that for the Bartlett estimator

$$\lim_{L \rightarrow \infty} E\{\hat{P}_{YY,k}^B(l)\} = P_{YY,k}(l)$$

i.e., the Bartlett estimator is asymptotically unbiased.

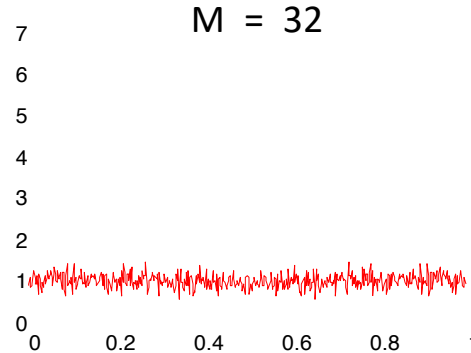
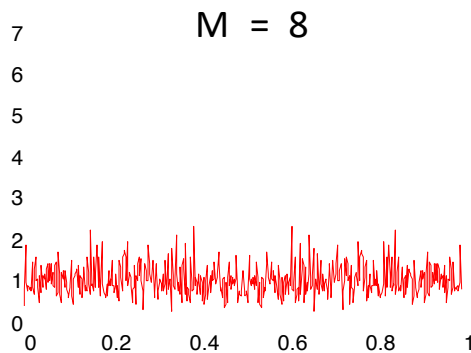
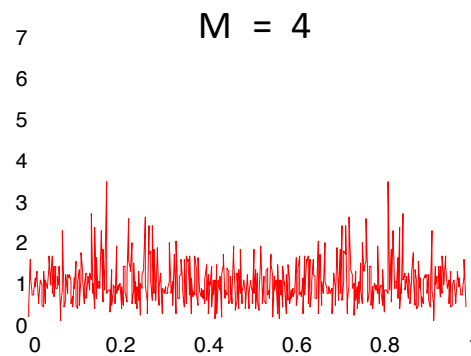
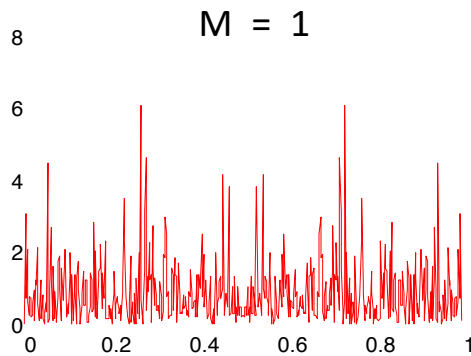
- Also, it can be shown

$$\lim_{L \rightarrow \infty} Var\{\hat{P}_{YY,k}^B(l)\} = \frac{1}{M} P_{YY,k}^2(l)$$

i.e., the variance is reduced by a factor M over the periodogram.

Bartlett Estimate

Bartlett estimate for a Gaussian process with $M \in \{1, 4, 8, 32\}$



Speech Enhancement – DFT Based

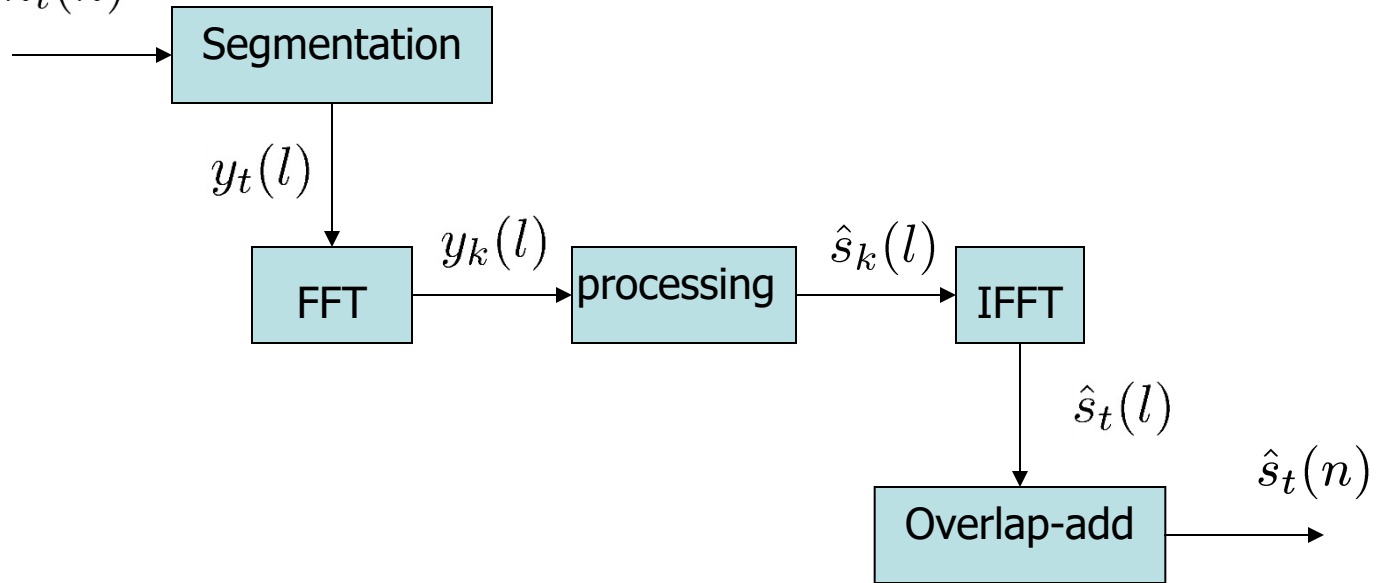
Our goal is to make an estimate $\hat{s}_t(n)$ of the clean speech signal based on a realization of the noisy speech $y_t(n)$.

To do this, we focus on methods that are based on the discrete Fourier transform (DFT). These methods can efficiently be implemented using FFTs.

Speech Enhancement – DFT Based

DFT-based enhancement schemes:

$$y_t(n) = s_t(n) + n_t(n)$$

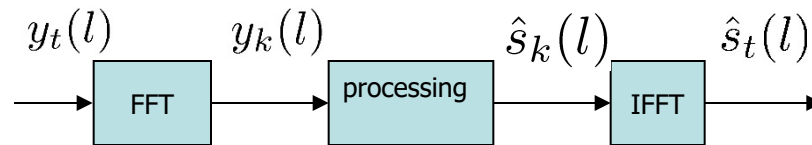
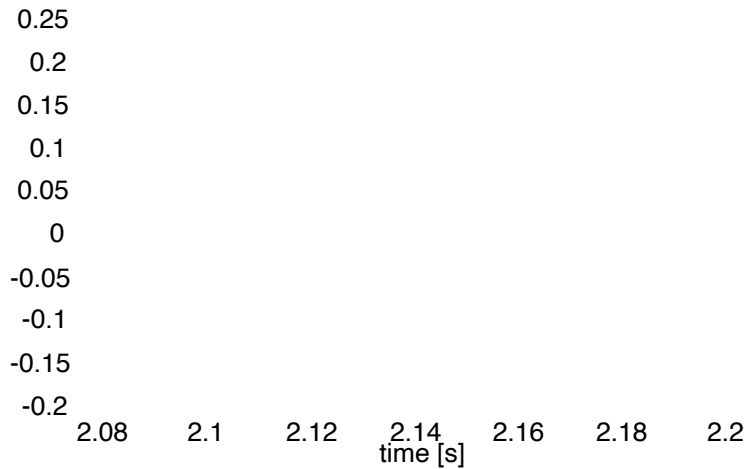


The index l indicates explicitly that we work per frame!

Speech Enhancement – DFT Based

Segmentation

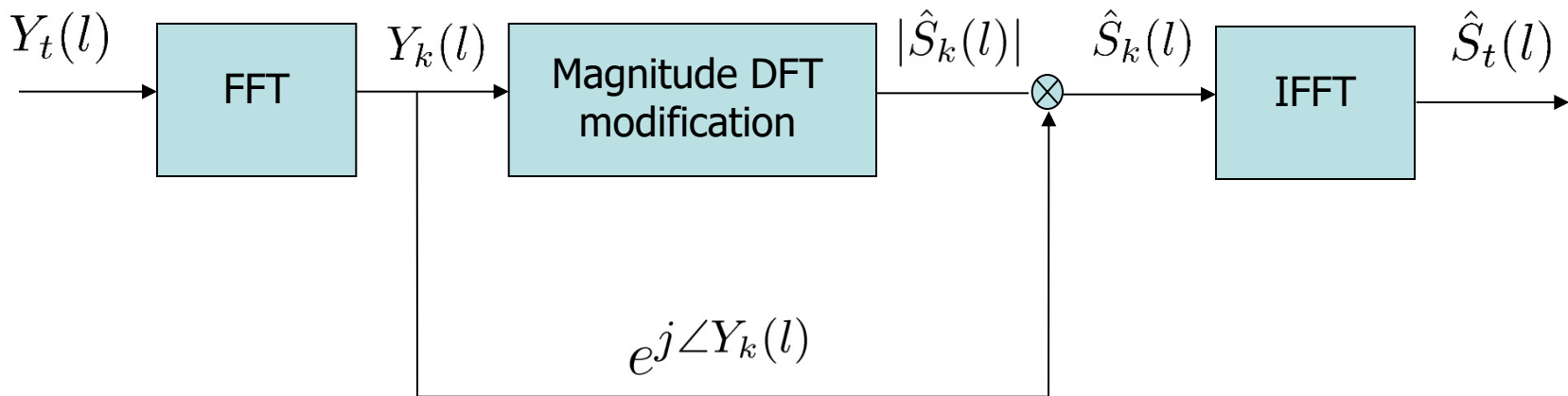
Overlap-add



Speech Enhancement – DFT Based

We will mainly consider the discrete Fourier transform (DFT) based methods for speech enhancement.

Most speech enhancement methods only modify the magnitude of the DFT coefficients:



Maintaining the noisy phase

Most noise reduction techniques for speech enhancement aim at modifying the short-term magnitude spectrum of the clean signal, while maintaining the phase spectrum of the noisy signal.

$$\angle \hat{S}_k(l) = \angle Y_k(l)$$

Question: Is it reasonable to pay more attention to the restoration of the magnitude spectrum over the phase spectrum?


Magnitude or Phase spectrum?


Question: Is it reasonable to pay more attention to the restoration of the magnitude spectrum over the phase spectrum?

Example: Clean speech degraded with white noise at -20 dB SNR.

 $S_k(l)$ clean speech

 $Y_k(l)$ noisy speech

 $\hat{S}_k(l) = |Y_k(l)|e^{j\angle S_k(l)}$ clean phase, noisy magnitude

 $\hat{S}_k(l) = |S_k(l)|e^{j\angle Y_k(l)}$ noisy phase, clean magnitude

Magnitude or Phase spectrum?

Conclusion:

Apparently the *magnitude* of speech conveys more information than the *phase*.

Is it optimal to use the noisy phase?

- It can be shown that using the noisy phase is optimal in mean squared-error (MSE) sense. (To prove this is outside the scope of this course)
- In the next lecture we will derive optimal MSE estimators. It turns out that these estimators are optimal when using the noisy phase.

However, notice that some information that is important for intelligibility is contained in the phase spectrum!

Power Spectral Subtraction

One of the most simple (and heuristic...) methods for noise reduction is power spectral subtraction.

Remember our assumptions:

- Speech and noise are additive
- Speech and noise are uncorrelated



$$P_{YY,k}(l) = P_{SS,k}(l) + P_{NN,k}(l)$$

$$P_{SS,k}(l) = P_{YY,k}(l) - P_{NN,k}(l)$$

Using expectation operators: $E[|S_k(l)|^2] = E[|Y_k(l)|^2 - E[|N_k(l)|^2]]$

Leaving out expectations: $\widehat{|S_k(l)|^2} = |Y_k(l)|^2 - |N_k(l)|^2$

Power Spectral Subtraction

Taking the square-root then gives:

$$|\widehat{S(\omega)}| = \sqrt{|Y_k(l)|^2 - |N_k(l)|^2}$$

Notice that we do have realizations of $|Y_k(l)|^2$, but not of $|N_k(l)|^2$.

Therefore, instead of $|N_k(l)|^2$ the noise power spectral density is used:

$$|\widehat{S_k(l)}| = \sqrt{|Y_k(l)|^2 - E[|N_k(l)|^2]}$$

How to estimate $E[|N_k(l)|^2]$ will be discussed in one of the following lectures.

Power Spectral Subtraction

Problem 1: Negative psd estimates

The expected value

$$E[|S_k(l)|^2] = E[|Y_k(l)|^2] - E[|N_k(l)|^2],$$

is always non-negative (as one would expect from a psd).
However, our estimate (when using realizations) can become *negative* for some frequencies:

$$\widehat{|s_k(l)|^2} = |y_k(l)|^2 - E[|N_k(l)|^2]$$

What to do?

Power Spectral Subtraction

Solution 1: Negative psd estimates

Often, the estimate is modified as follows:

$$\widehat{|s'_k(l)|}^2 = \begin{cases} |y_k(l)|^2 - E[|N_k(l)|^2] & \text{for } |y_k(l)|^2 \geq E[|N_k(l)|^2] \\ 0 & \text{otherwise} \end{cases}$$

which leads to

$$\widehat{|s'_k(l)|} = \left(\max \left\{ 1 - \frac{E[|N_k(l)|^2]}{|y_k(l)|^2}, 0 \right\} \right)^{\frac{1}{2}} |y_k(l)|.$$

Power Spectral Subtraction

Problem 2: Noise variations

Recall:

$$\widehat{|s'_k(l)|^2} = \begin{cases} |y_k(l)|^2 - E[|N_k(l)|^2] & \text{for } |y_k(l)|^2 \geq E[|N_k(l)|^2] \\ 0 & \text{otherwise} \end{cases}$$

$$\widehat{|S_k(l)|^2} = \underbrace{|S_k(l)|^2}_{\text{what we want}} + \underbrace{(|N_k(l)|^2 - E[|N_k(l)|^2])}_{\text{noise variations}} + \underbrace{S_k^*(l)N_k(l) + S_k(l)N_k^*(l)}_{\text{cross terms}}$$

The noise variations are due to the variance of the instantaneous power spectral density estimate $|N_k(l)|^2$.

How to reduce it?

Power Spectral Subtraction

Solution 2: Noise variations

Replace $|Y_k(l)|^2$ with time-averaged version (**Bartlett estimate**):

$$\overline{|Y_k(l)|^2} = \frac{1}{L} \sum_{m=l-L+1}^l |Y_k(m)|^2.$$

We get:

$$|\hat{S}''_k(l)|^2 = \underbrace{\overline{|S_k(l)|^2}}_{\text{smoothed target}} + \underbrace{\left(\overline{|N_k(l)|^2} - E[|N_k(l)|^2] \right)}_{\text{reduced noise variations}} + \underbrace{\overline{S_k^*(l)N_k(l) + S_k(l)N_k^*(l)}}_{\text{cross terms}}.$$

Power Spectral Subtraction

Solution 2: Noise variations

The noise variations are reduced due to the reduced variance of the Bartlett estimate $\overline{|N_k(l)|^2}$.

Question: Why can't we just pick L high enough to reduce the noise variations sufficiently?

Reducing Residual Noise

Residual (“Musical”) Noise

Consider spectral region where speech energy is low, i.e.,
 $S_k(l) \approx 0$.

Combining solutions 1 and 2 from before, our estimator becomes:

$$|\widehat{S''_k(l)}|^2 \approx \max(\overline{|N_k(l)|^2} - E[|N_k(l)|^2], 0)$$

Plugging in realizations we have either:

$$\overline{|n_k(l)|^2} > E[|N_k(l)|^2]$$

(which leads to non-zero estimator)

Reducing Residual Noise

or:

$$\overline{|n_k(l)|^2} \leq E [|N_k(l)|^2]$$

(in which case the estimator is zero).

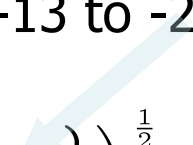
Reducing Residual Noise

Residual (“Musical”) Noise

As a result, the estimated signal contains short-duration tonal (residual) signal components occurring at seemingly random frequencies and time instances.

The *perceptual* effect is “musical” noise, which may be perceptually very disturbing


One way to reduce the effect of musical noise is to constrain the minimum gain to for example -13 to -20 dB:


$$|\widehat{s'_k(l)}| = \left(\max \left\{ 1 - \frac{E[|N_k(l)|^2]}{|y_k(l)|^2}, 0.2 \right\} \right)^{\frac{1}{2}} |y_k(l)|.$$



Examples of Power Spectral Subtraction

Speech degraded with white noise at 10 dB SNR

 noisy


$$\widehat{|s'_k(l)|} = \left(\max \left\{ 1 - \frac{E[|N_k(l)|^2]}{|y_k(l)|^2}, 0 \right\} \right)^{\frac{1}{2}} |y_k(l)| \text{ with } L = 1.$$


$$\widehat{|s'_k(l)|} = \left(\max \left\{ 1 - \frac{E[|N_k(l)|^2]}{|y_k(l)|^2}, 0 \right\} \right)^{\frac{1}{2}} |y_k(l)| \text{ with } L = 3.$$


$$\widehat{|s'_k(l)|} = \left(\max \left\{ 1 - \frac{E[|N_k(l)|^2]}{|y_k(l)|^2}, 0.2 \right\} \right)^{\frac{1}{2}} |y_k(l)|, \text{ with } L = 3.$$

Power Spectral Subtraction

Algorithm Summary

For each noisy signal frame $y_t(l)$ where speech is present:

1. Apply window and compute DFT (FFT) $\rightarrow y_k(l)$
2. Compute $\overline{|y_k(l)|^2} = \frac{1}{L} \sum_{m=l-L+1}^l |y_k(l)|^2$.
3. Compute $\widehat{|s_k''(l)|^2} = \max(\overline{|y_k(l)|^2} - E[|N_k(l)|^2], 0.2)$.
4. Append noisy phase $\hat{s}_k''(l) = \widehat{|s_k''(l)|} \cdot e^{j\angle y_k(l)}$.
5. Compute enhanced time-domain frame using inverse DFT (IFFT) of $\hat{s}_k''(l)$.

Power Spectral Subtraction – Maximum Likelihood Estimate

Instead of the previous heuristic derivation, Power spectral subtraction can also be derived as a maximum likelihood (ML) estimator:

Remember: $Y_k(l) = S_k(l) + N_k(l)$

Let us assume that the speech and noise DFT coefficients are complex-Gaussian distributed, i.e., $S_k(l) \sim N(0, \sigma_S^2)$ and $N_k(l) \sim N(0, \sigma_N^2)$.

Power Spectral Subtraction – maximum likelihood

The noisy speech DFT coefficients are then also complex-Gaussian distributed, that is, $Y_k(l) \sim N(0, \sigma_{Y,k}^2(l))$ with $\sigma_{Y,k}^2(l) = \sigma_{S,k}^2(l) + \sigma_{N,k}^2(l)$

$$p_Y(y; \sigma_{S,k}^2(l), \sigma_{N,k}^2(l)) = \frac{1}{\pi(\sigma_{S,k}^2(l) + \sigma_{N,k}^2(l))} \exp \left[-\frac{|y_k(l)|^2}{\sigma_{S,k}^2(l) + \sigma_{N,k}^2(l)} \right]$$

The maximum likelihood estimate of $\sigma_{S,k}^2(l)$ is then given as

$$\widehat{\sigma_{S,k}^2(l)}_{ml} = \arg \max_{\sigma_{S,k}^2(l)} p_Y(y; \sigma_{S,k}^2(l), \sigma_{N,k}^2(l))$$

Power Spectral Subtraction – maximum likelihood

The maximum likelihood estimate of $\sigma_{S,k}^2(l)$ is then given as

$$\begin{aligned}\widehat{\sigma_{S,k}^2(l)}_{ml} &= \arg \max_{\sigma_{S,k}^2(l)} p_Y(y_k(l); \sigma_{S,k}^2(l), \sigma_{N,k}^2(l)) \\ &= \arg \max_{\sigma_{S,k}^2(l)} \log p_Y(y_k(l); \sigma_{S,k}^2(l), \sigma_{N,k}^2(l))\end{aligned}$$

$$\frac{d \log p_Y(y; \sigma_{S,k}^2(l), \sigma_{N,k}^2(l))}{d \sigma_{S,k}^2(l)} = -\frac{1}{(\sigma_{S,k}^2(l) + \sigma_{N,k}^2(l))} + \frac{|y_k(l)|^2}{(\sigma_{S,k}^2(l) + \sigma_{N,k}^2(l))^2} = 0$$

$$\sigma_{S,k}^2(l) = |y_k(l)|^2 - \sigma_{N,k}^2(l)$$

Power Spectral Subtraction – maximum likelihood

Power spectral subtraction essentially leads thus to an estimate of the speech PSD $\sigma_{S,k}^2(l)$.

Is it a good estimator?

$$E \left[\widehat{\sigma_{S,k}^2(l)}_{ml} \right] = E \left[|y_k|^2 \right] - \sigma_{N,k}^2(l) = \sigma_{S,k}^2(l).$$

It is thus an unbiased estimate of the speech PSD.

However, using it as an estimator for the clean speech magnitude $|S_k(l)|$ it will be biased.

Power Spectral Subtraction

Power spectral subtraction is a very simple method for noise reduction of speech signals.

However

- It is rather heuristic.
- It does not optimize a distortion measure.
- Noise reduced speech contains a lot of musical noise
- Can also be derived as a maximum likelihood estimator.

Next week:

- More advanced estimators that are
 - mathematically more solid
 - Derived under distributional assumptions that match distribution of speech DFT coefficients.
- An alternative way for PSD estimation that reduces musical noise and increases quality.