Microphone Array Processing

Introduction to

- **Some applications**
- **Speech signals**
- **Introduction to Beamforming for microphone arrays**

Richard C. Hendriks

TU Delft

Delft University of Technology

Speech Enhancement - Project

- Project is compulsory and carried out in groups of 2 students
- Q&A during oral discussion (hand in report before June 21st 2023, brightspace)

Project:

- Design and build a multi-microphone speech enhancement/beamforming system for far-end noise reduction.
- Use matlab
- Generate signals according to the signal model discussed in class using the audio files and impulse responses (see website).
- Perform an evaluation of the speech enhancement system.

Microphone arrays

Can be used for (spatial) processing to improve speech intelligibility and reduce the effect of background noise on speech communication quality.

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

Application Areas:

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

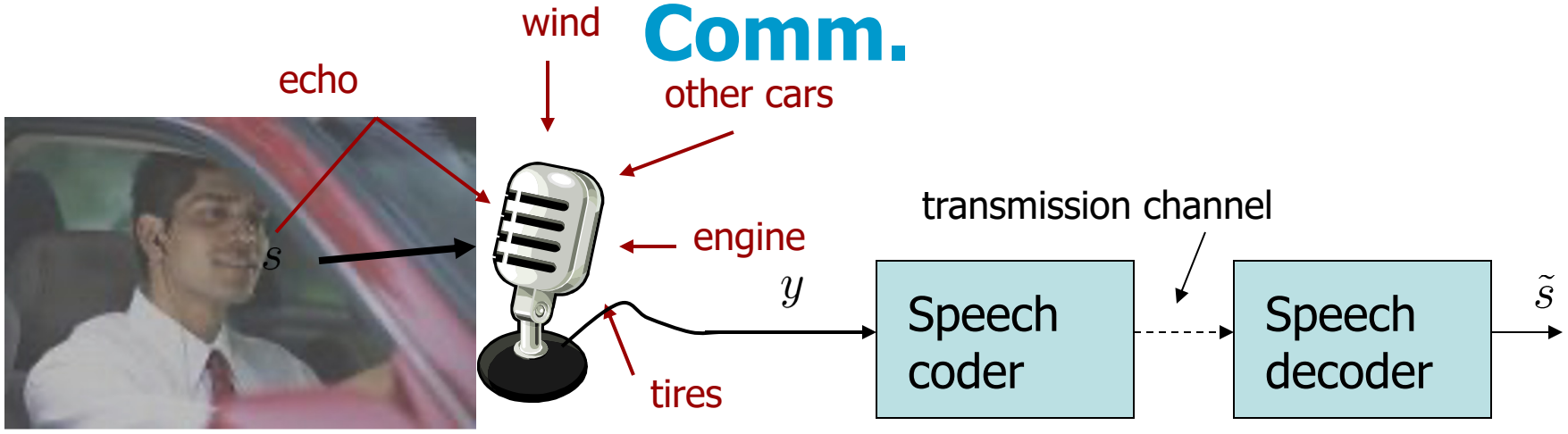
Example: Speech Enhancement for Dig. Comm.

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.

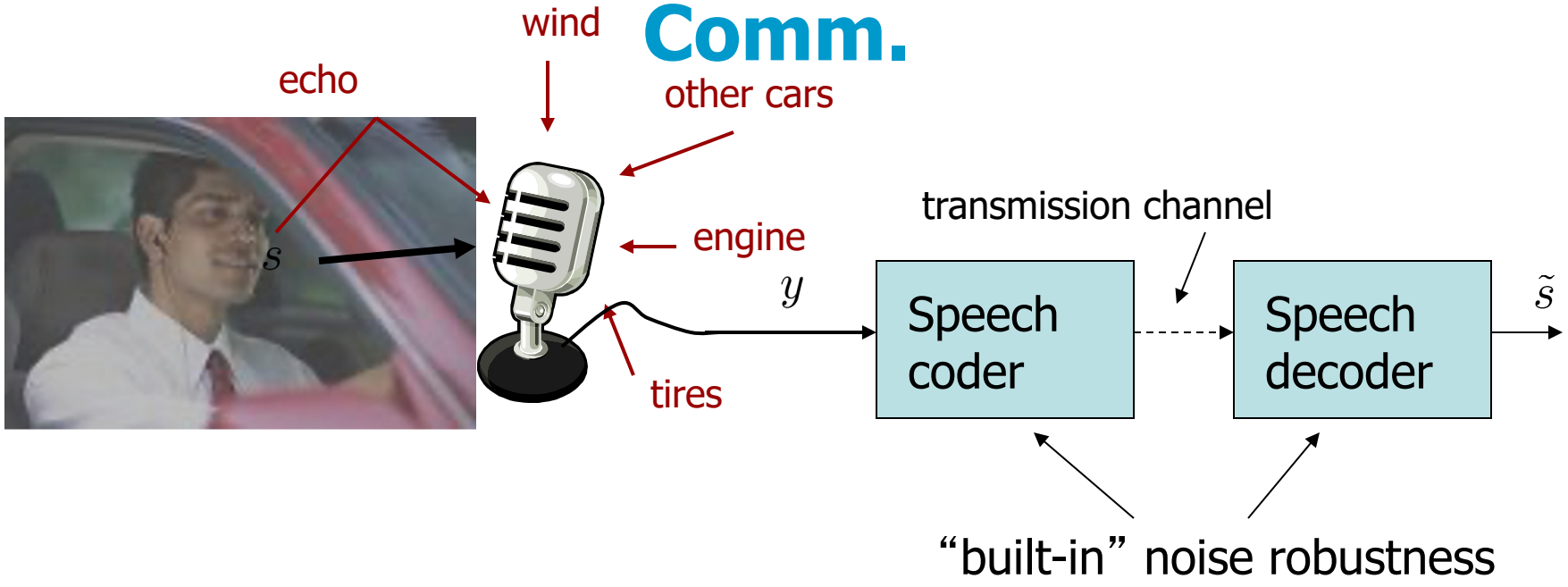
Example: Speech Enhancement for Dig. Comm.



Degradation of target due to:

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

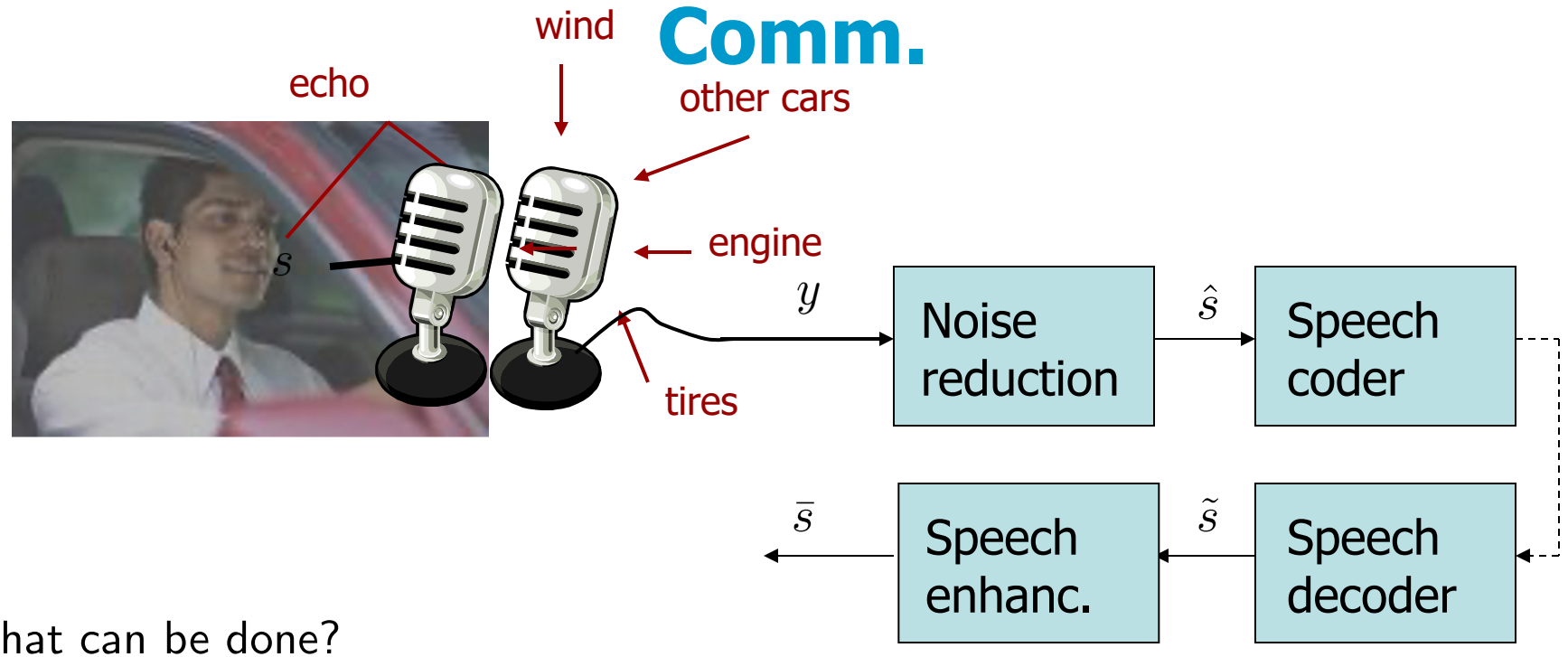
Example: Speech Enhancement for Dig. Comm.



What can be done?

- Develop new and more noise robust digital speech communication systems

Example: Speech Enhancement for Dig. Comm.



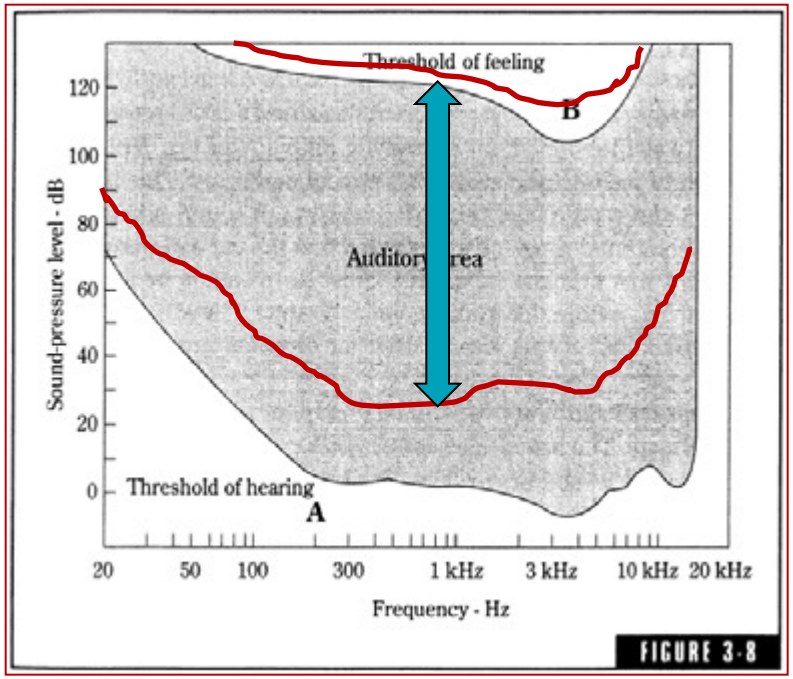
What can be done?

- 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

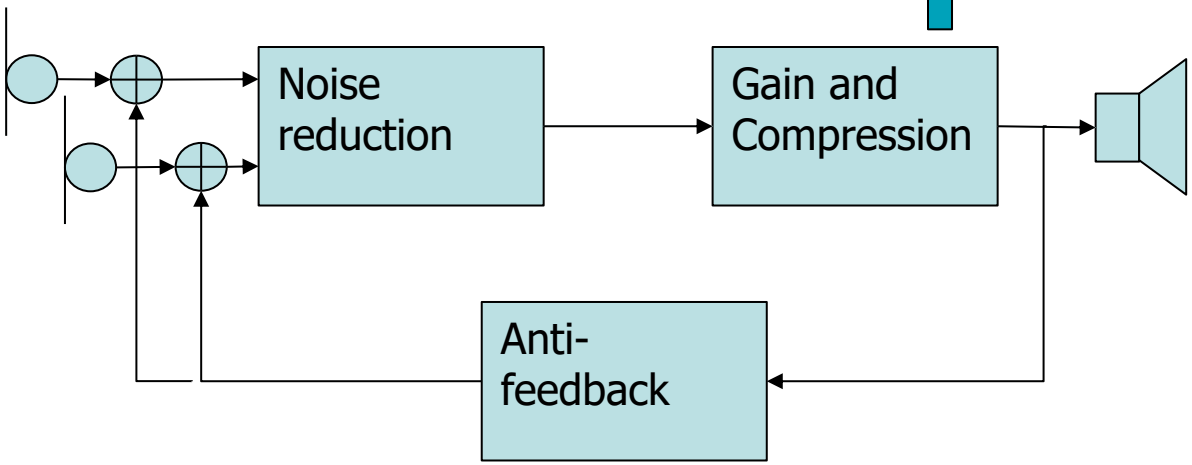
- Reduced sensitivity and reduced dynamic range
- Temporal resolution
- Frequency resolution
- Inability to exploit spatial cues

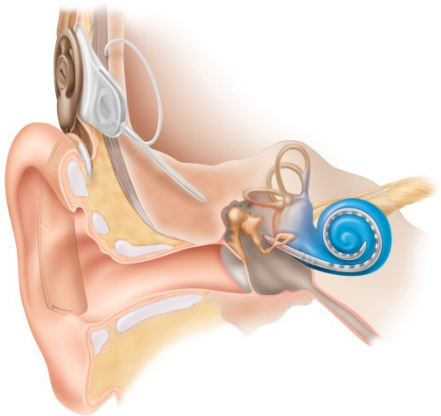
How to compensate for this?



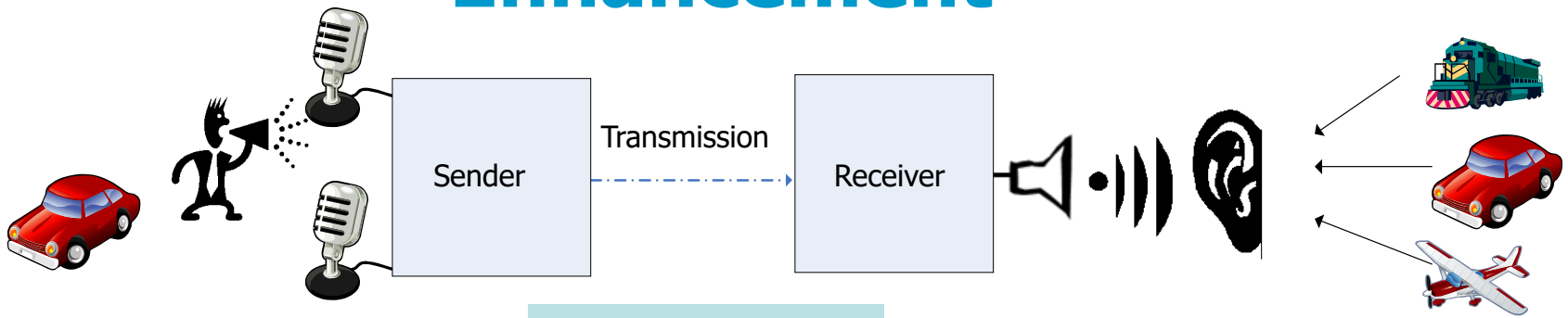
TU Delft

Example: Speech Enhancement for Hearing Devices





Single and Multi-Microphone Speech Enhancement



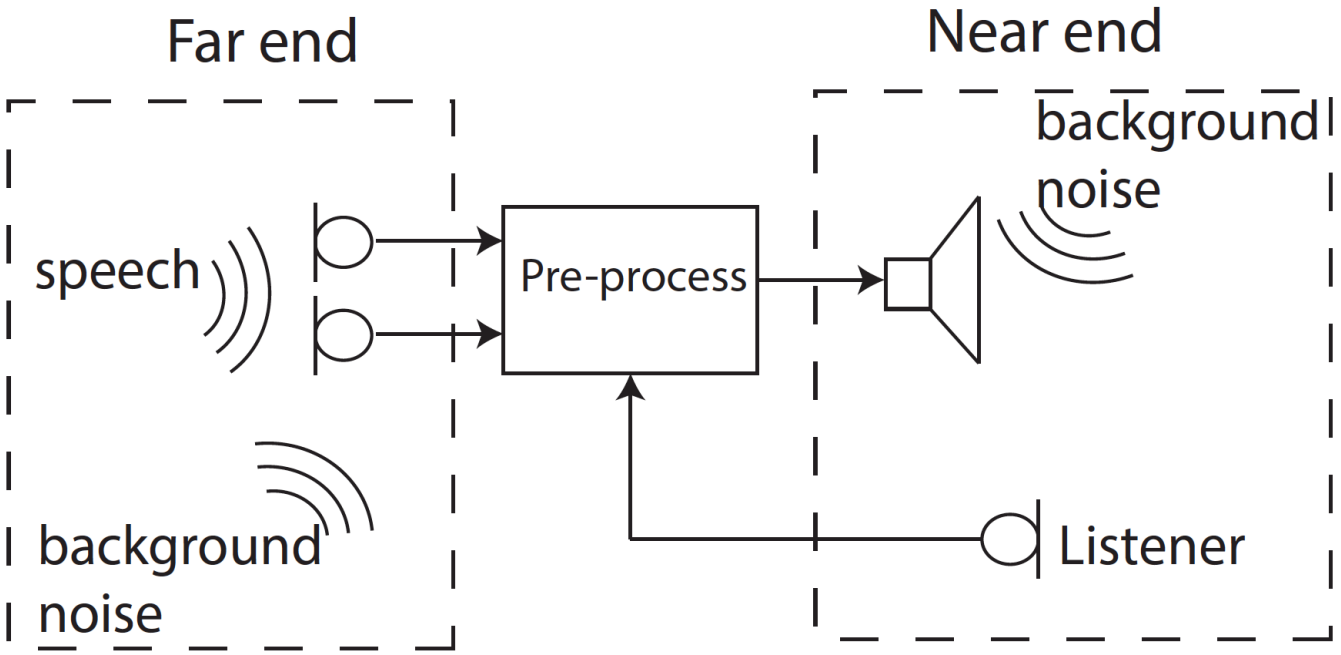
Far-end noise reduction

Applications:

- Hearing aids
- Mobile telephony
- Headsets
- Etc.

Near-end speech enhancement

Example: Near-end Speech Enhancement



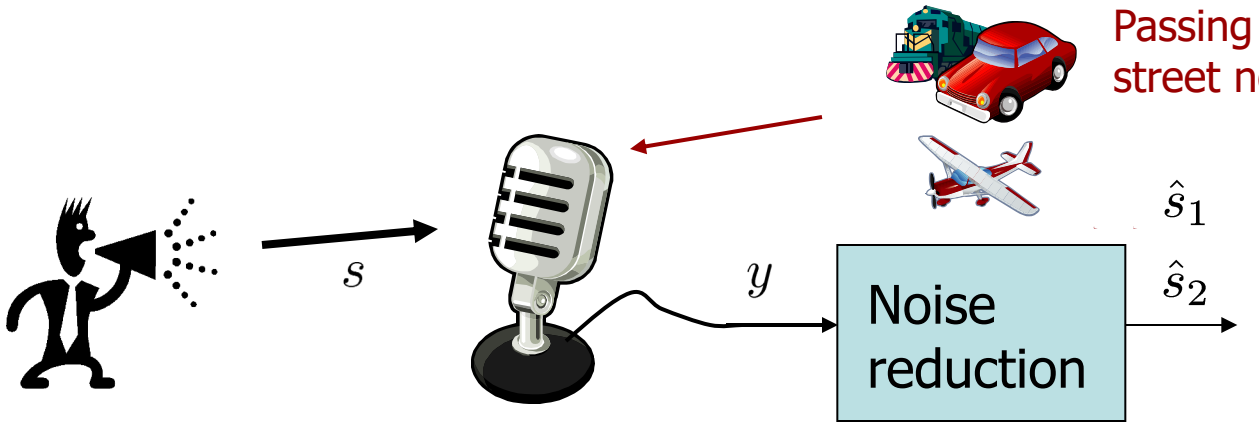
Unprocessed

Processed using MI-optimal model

TU Delft

Example: Far-end Speech Enhancement

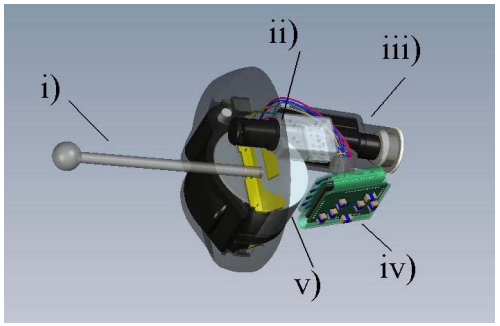
Example: single mic. noise reduction for non-stationary noise



Single and Multi-Microphone Noise Reduction

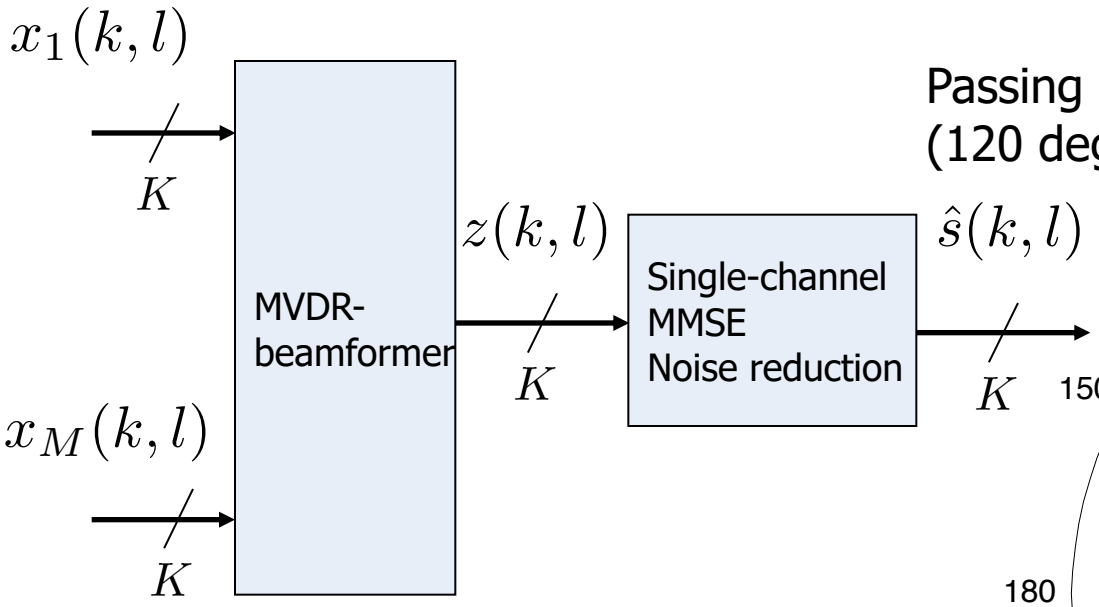
Behind the ear hearing aid
(2 microphones)

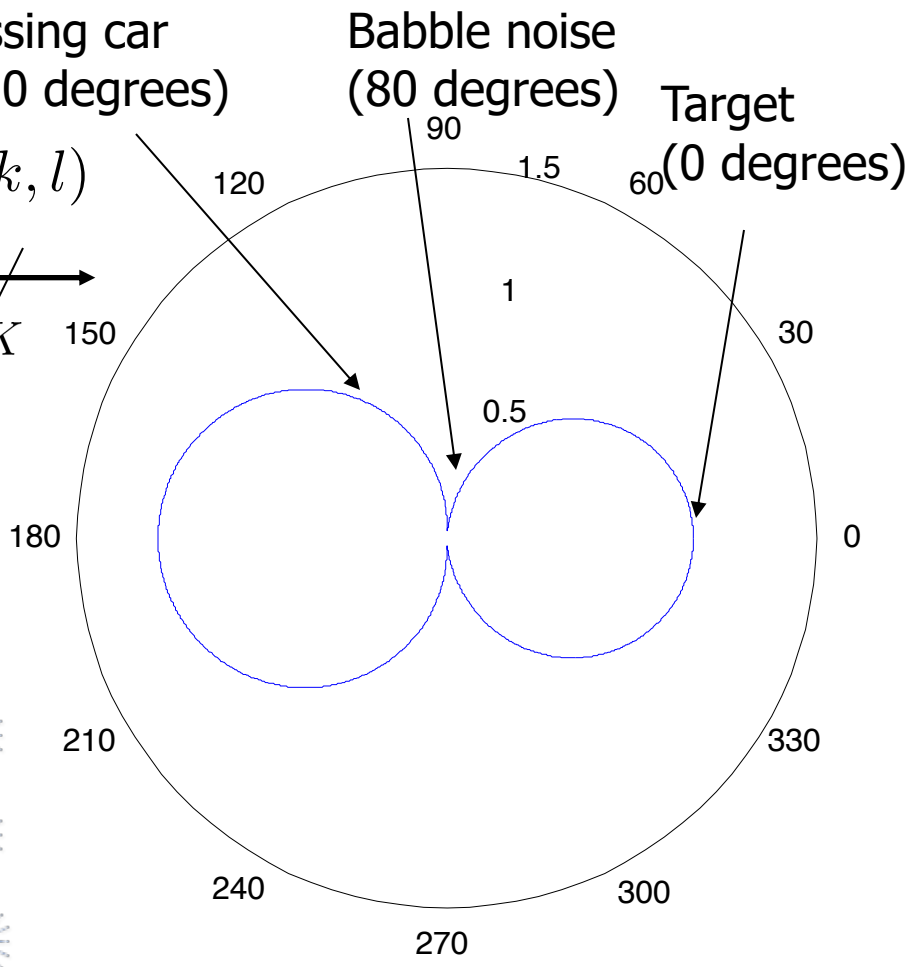
In the ear hearing aid
(1 microphone)



TU Delft

Example: Multi-Channel Noise Reduction





- Clean input $s(k,l)$
- Noisy input at mic. 1 $x_1(k,l)$
- MVDR-beamformer output $z(k,l)$
- Output of total system $\hat{s}(k,l)$

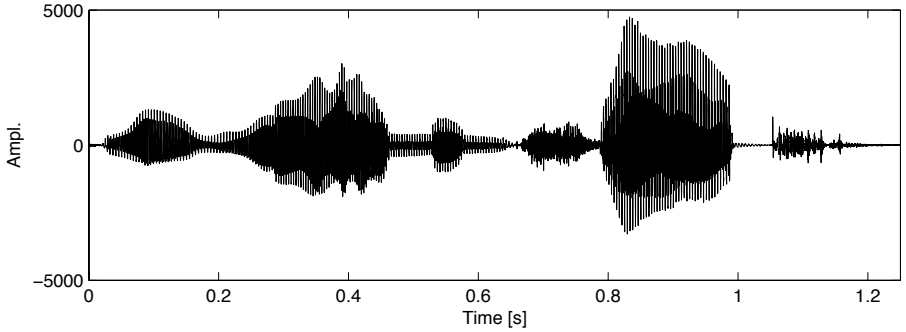
Focus – Microphone Array Processing

today

- Speech signals: The look and feel
- Microphone array signal model
- Beamforming
  - Optimal beamformers (Wiener, MVDR, LCMV)
  - Relations between optimal beamformers
  - The acoustic transfer function (ATF)
- The EVD & GEVD
- Estimating the ATF
- Estimating $\mathbf{R}_\mathbf{s}$
- ATF Estimation and Cramér-Rao lower bounds

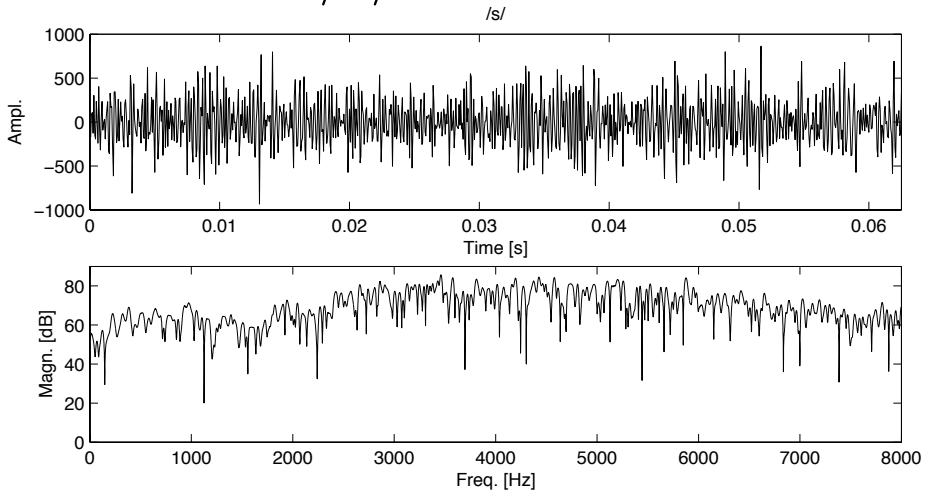
Speech Signals - A First Encounter

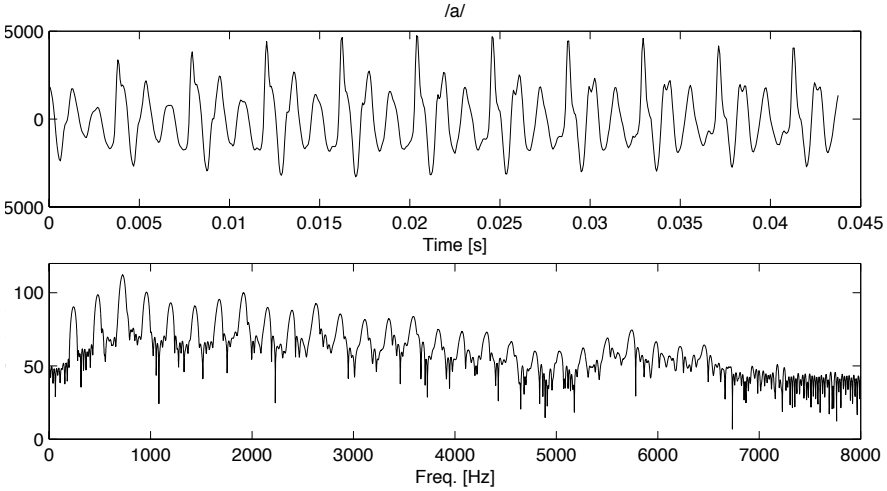
Characteristics of speech change across time due to changing production system:



uvoiced: /s/

voiced: /a/



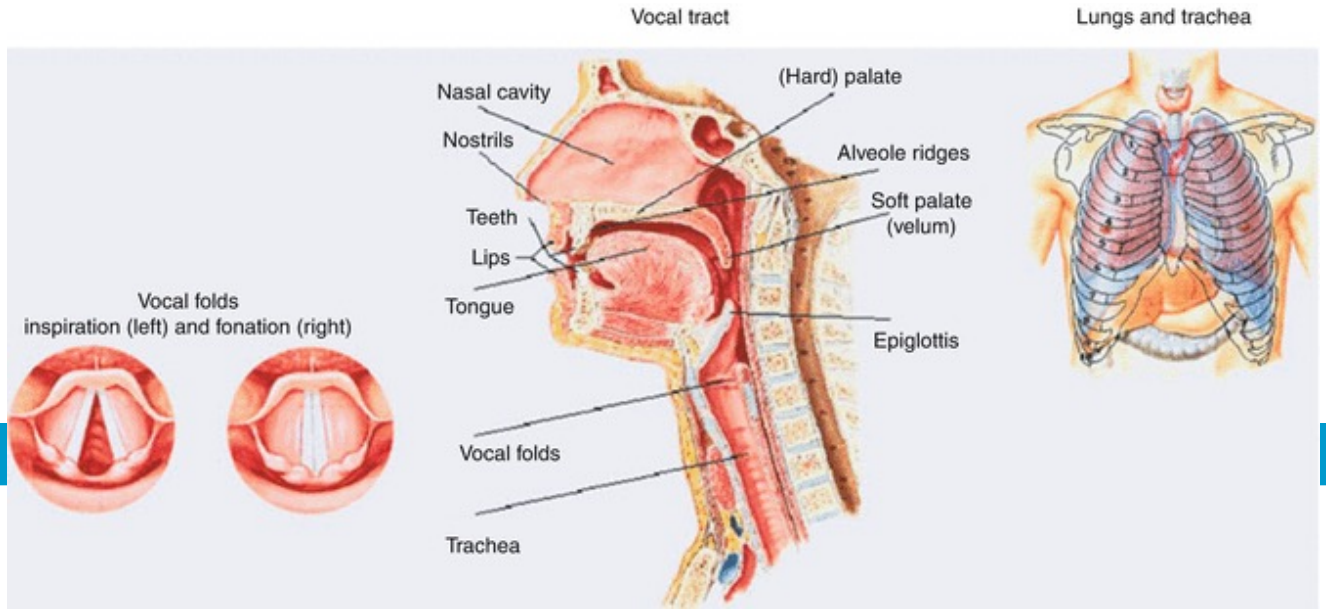


TU Delft

Speech Production - Anatomy

Overview of speech production system:

- Lungs
- Larynx (organ of voice production).
- Vocal Tract
  - throat (pharyngeal cavity).
  - oral+nasal cavity.

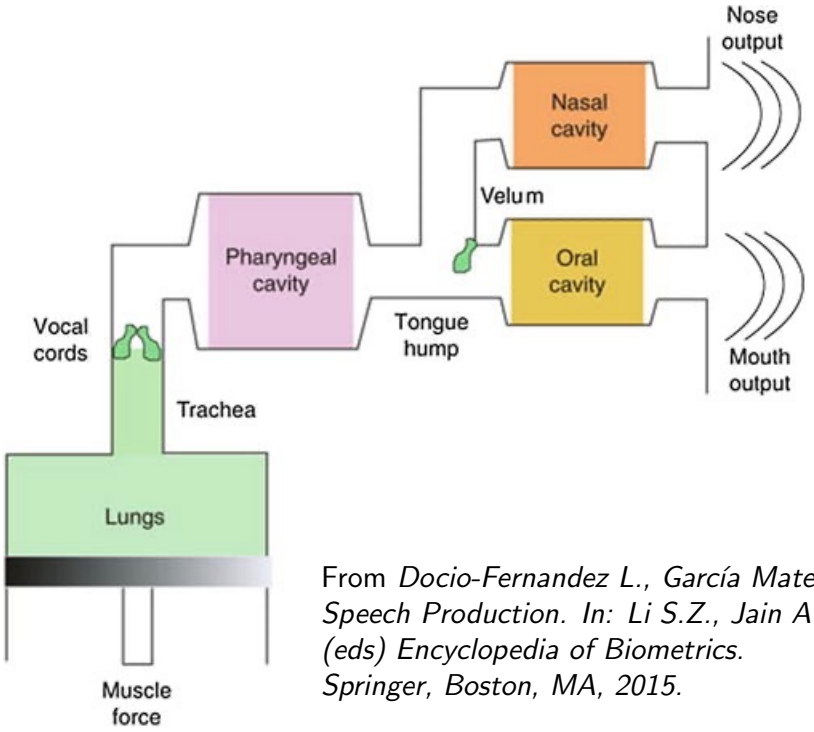


From *Docio-Fernandez L., García Mateo C. Speech Production. In: Li S.Z., Jain A.K. (eds) Encyclopedia of Biometrics. Springer, Boston, MA, 2015.*

Speech Production - Anatomy

Acoustic filter model:

- Lungs+vocal folds: Excitation.
- Cavities: Main acoustic filter.
- Velum: "switch" for nasal sounds.



From *Docio-Fernandez L., García Mateo C. Speech Production. In: Li S.Z., Jain A.K. (eds) Encyclopedia of Biometrics. Springer, Boston, MA, 2015.*

TU Delft

Speech Production - Excitation

Excitation signal: The air stream signal that enters the paryngeal cavity (throat), i.e., after vocal folds.

Types of excitation:

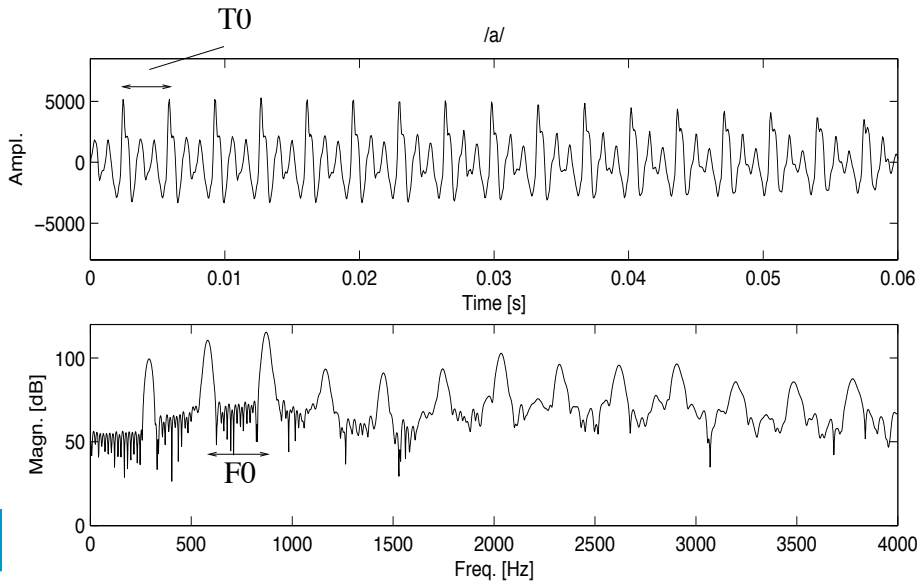
- *Voiced*: Air pushed through glottis which oscillate, generating quasi-periodic puffs of air (e.g. vowels /a/, /i/, etc.).
- *Unvoiced*: Air forced through constriction somewhere along vocal tract (e.g. /s/, /f/).
- *Mixed*: Quasi-periodic excitation but with constriction along vocal tract (e.g. /z/).
- *Plosive*: Complete closure of vocal tract, build-up of air pressure + release (e.g. /p/, /t/).

Speech Production - Excitation Signal

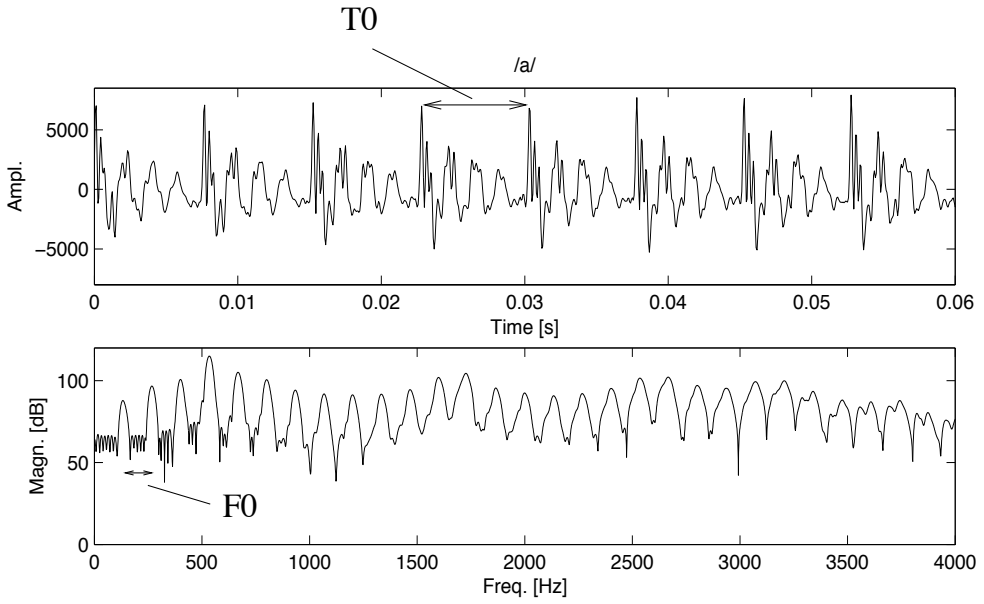
Voicing:

The fundamental period/frequency is evident in the time domain as well as the frequency domain representations of speech.

Female speaker



Male speaker

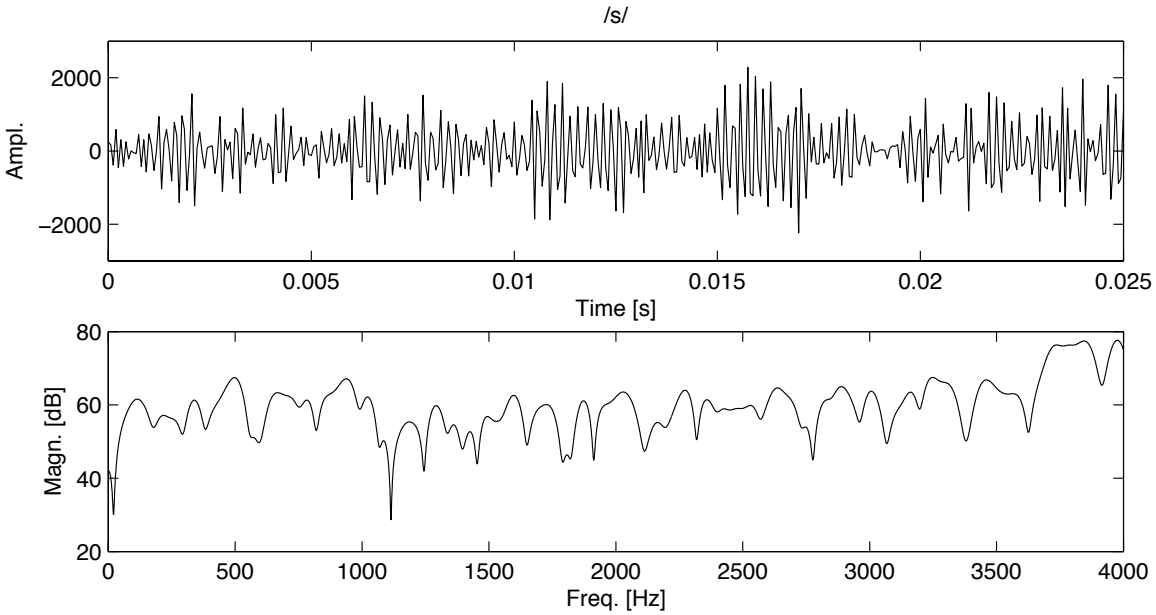


TU Delft

Speech Production - Excitation Signal

Unvoiced regions:

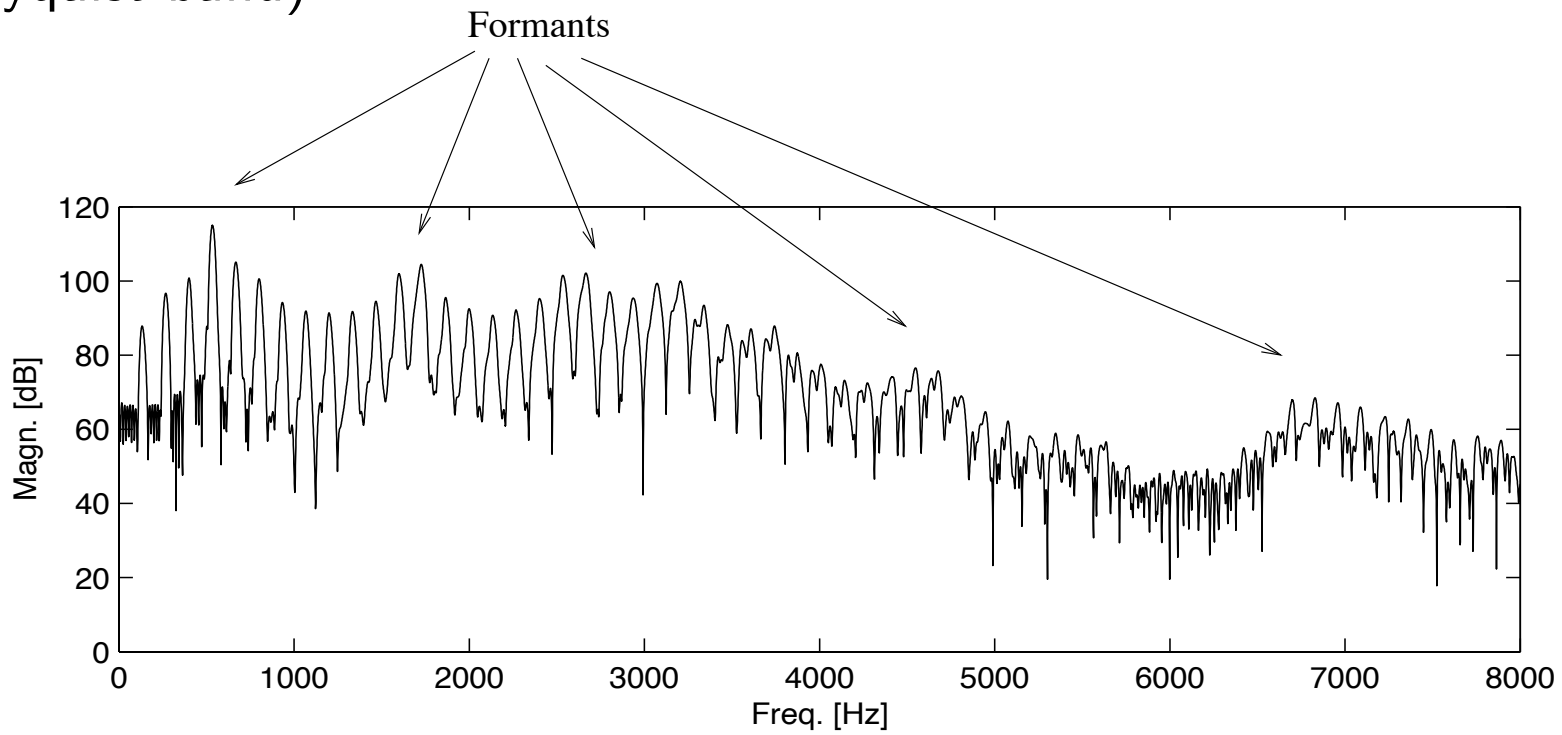
In unvoiced regions, the excitation signal is noise-like (i.e., without the periodicity that characterizes voiced signals.)



TU Delft

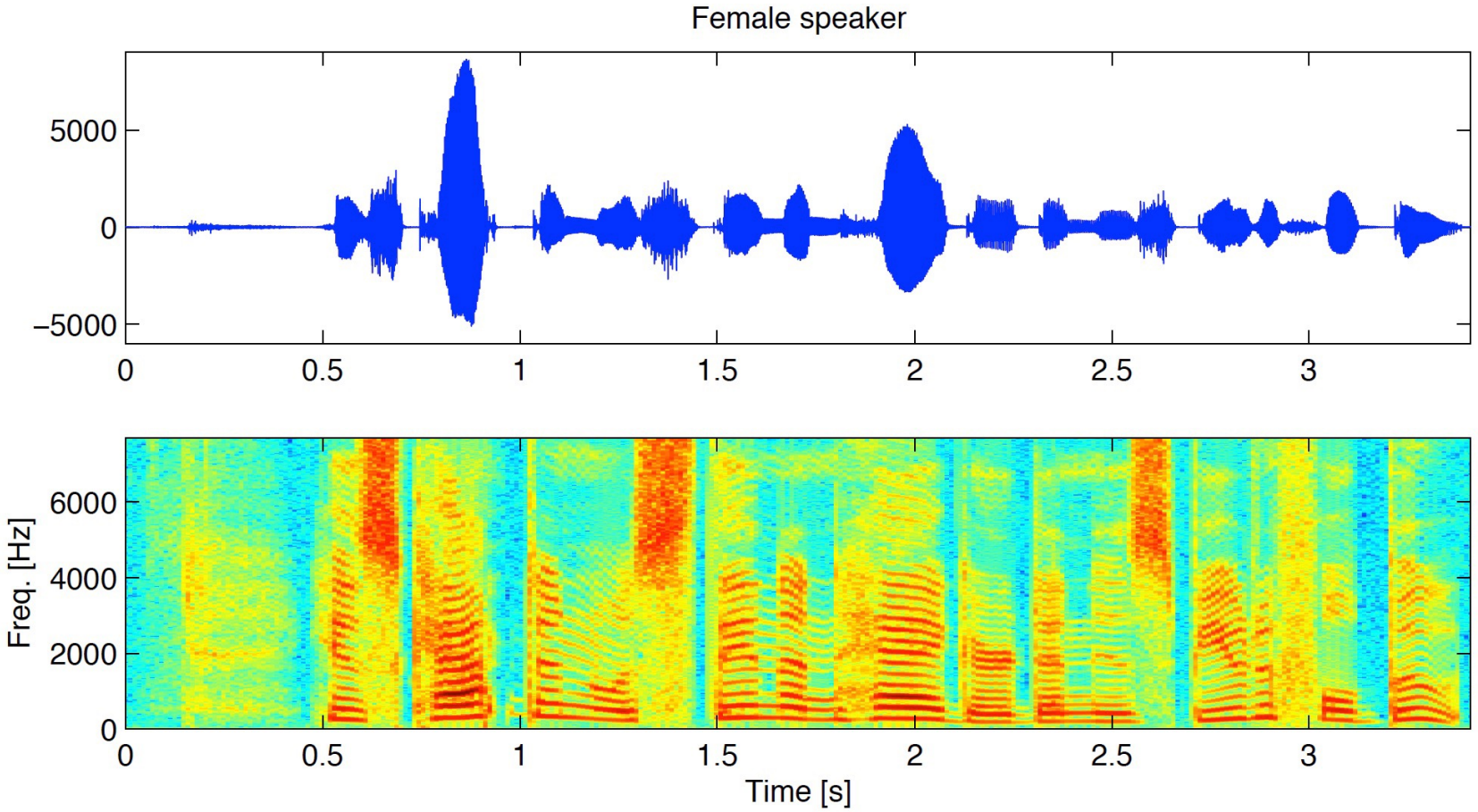
Speech Production - The Vocal Tract

- Configuration of vocal tract "shapes" excitation to generate specific speech sound, i.e., overall spectral characteristic determined by vocal tract.
- Resonance frequencies of vocal tract system give rise to peaks in overall spectrum $\sim$ *formants* (3-5 formants within Nyquist band).



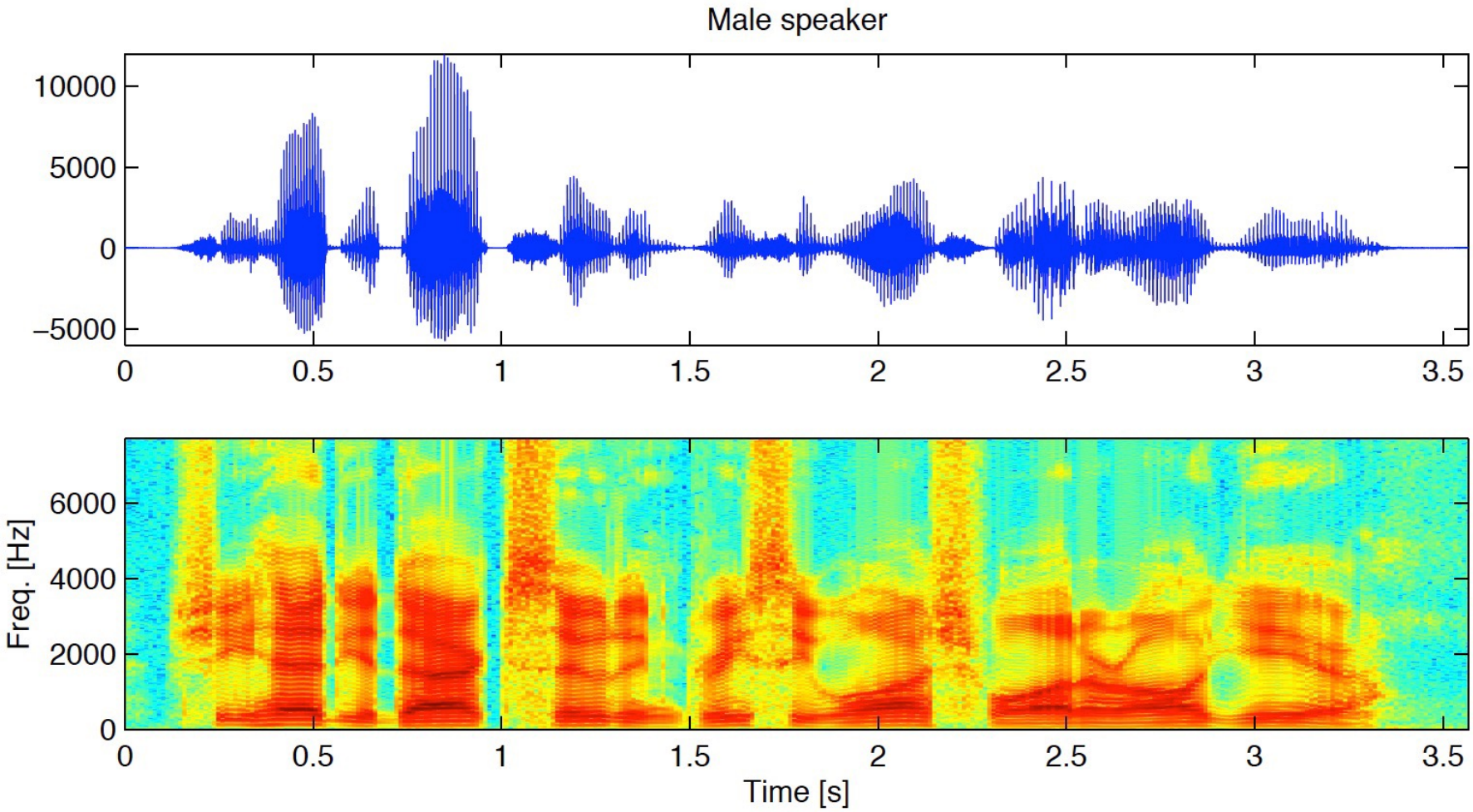
Spectrograms

Spectrogram: Time-vs-Freq-vs-Spectral Magnitude (no phase!).
"His captain was thin and haggard and his beautiful boots..."



TU Delft

Spectrograms

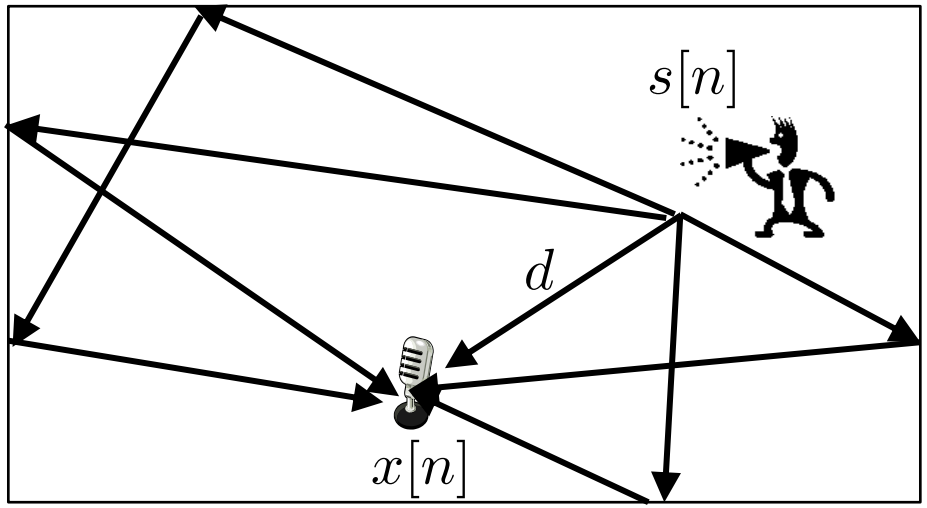


TU Delft

Speech Production - The Vocal Tract

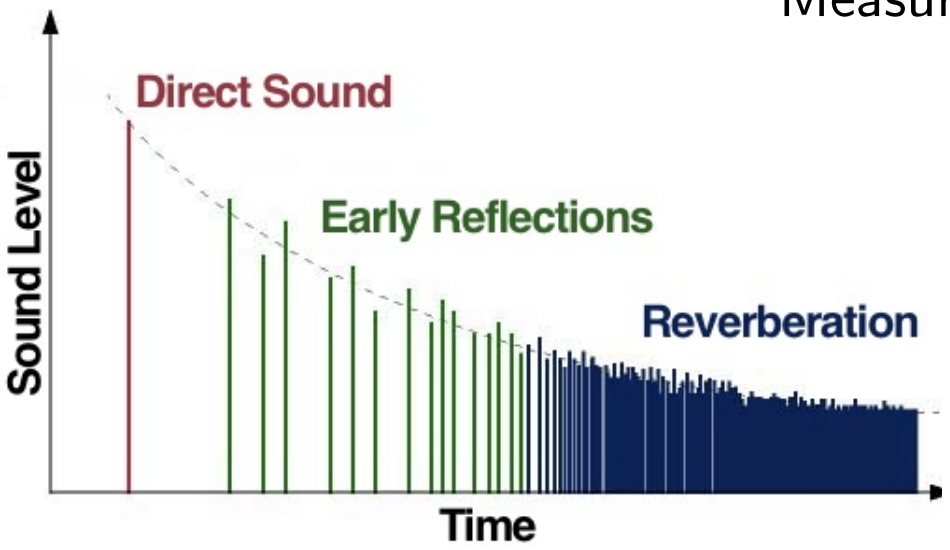
- Speech signals can be decomposed into two components: Vocal tract filter and the excitation (input) of this filter.
- Vocal tract system changes over time $\Rightarrow$ spectral/temporal characteristics of the speech waveform are *time-varying* $\Rightarrow$ only short segments of speech waveform can be assumed to have similar acoustic properties (*"non-stationarity" vs "short-term stationarity"*).
- Speech is considered a stochastic process (excitation signal is realization of random process).
- Speech signals typically assumed stationary over 20-30 ms time frames.
- Typical maximum speech bandwidth 7-8 kHz.

Microphone Measurement Model

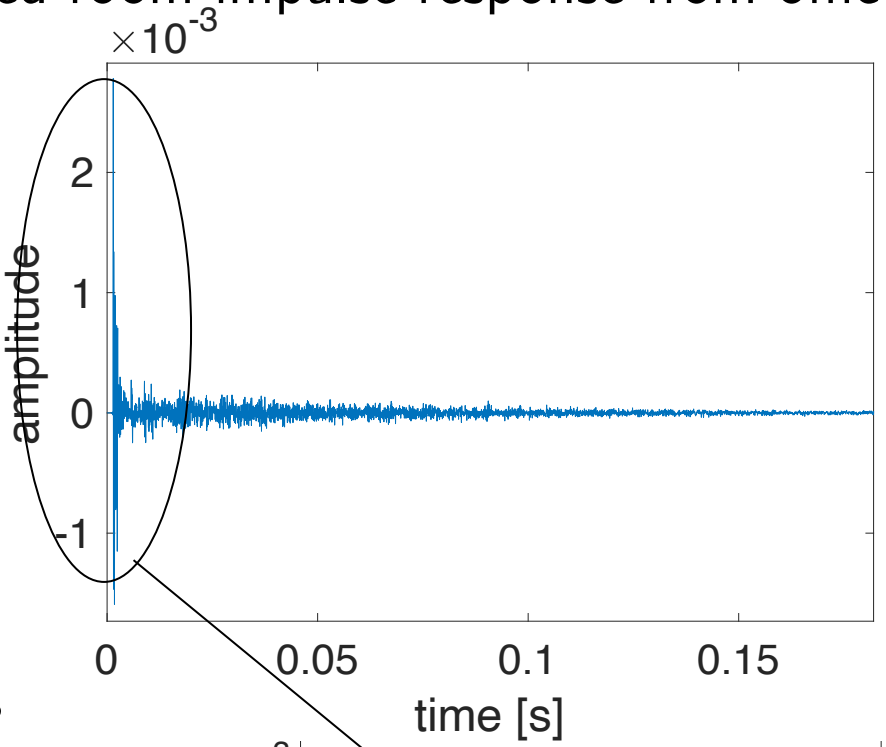


- Direct path: $x[n] = a(d)s[n - \tau(d)]$
- Reflections, modelled with room impulse response.
- $x[n] = (h * s)[n]$

Microphone Measurement Model

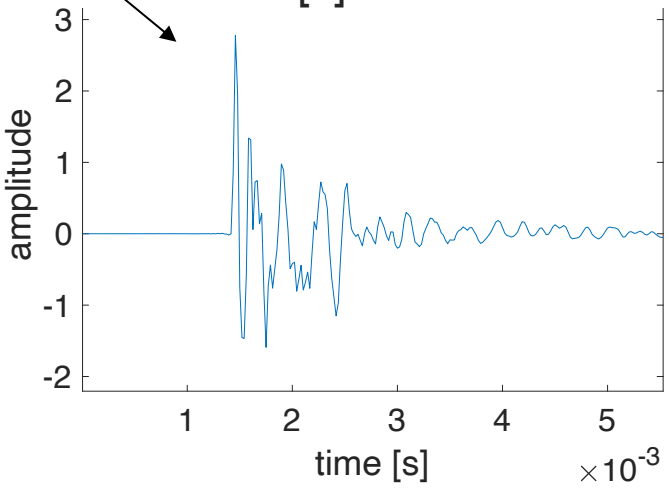


Measured room impulse response from office.



Notice

- Direct path and early reflections determine intelligibility.
- Late reflections (reverb) typically degrades intelligibility.
- Notice the long duration of h compared to typical frame size (20 ms).



Microphone Measurement Model

Single microphone model:

$$x[n] = \sum_{i=1}^{d} (h_i * s_i)[n] + n[n]$$

- d Point sources s_i
- Room impulse responses h_i from source position i to microphone.
- n models microphone self noise and often also other diffuse noise components (e.g., late reverberation).

Microphone Measurement Model

Single microphone model:

$$x[n] = \sum_{i=1}^{d} (h_i * s_i)[n] + n[n]$$

Assumptions: Sources are assumed to be

- Additive
- zero-mean and mutually uncorrelated, i.e., $E[s_i] = 0$, $E[n] = 0$, $E[s_i s_j] = 0 \, \forall i, j$ and $E[s_i n] = 0 \, \forall i$.
- short-time stationary.

Validity of these assumptions?

Microphone Measurement Model

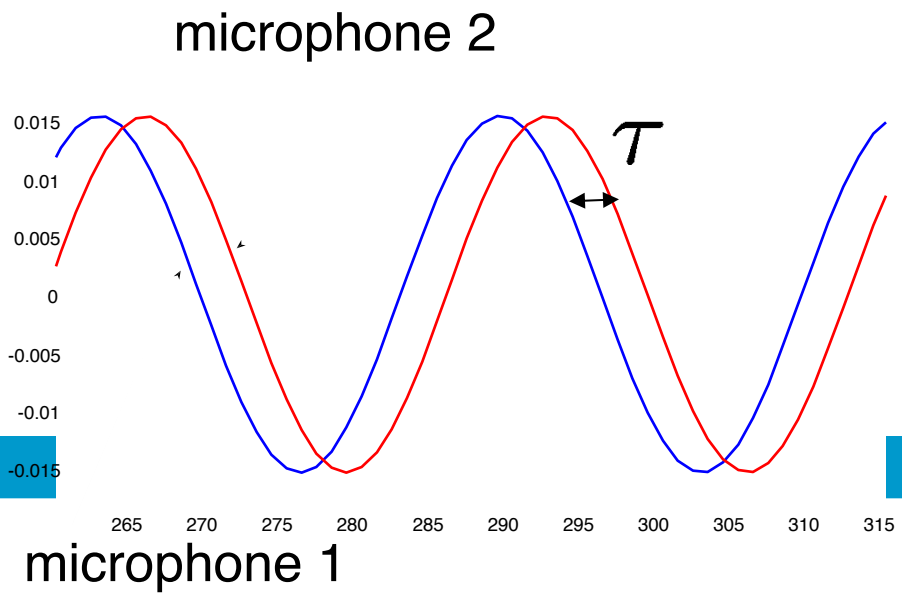
The impulse h_i response is often much longer than a time frame. Therefore h_i is often split in early (desired) and late (disturbing) components.

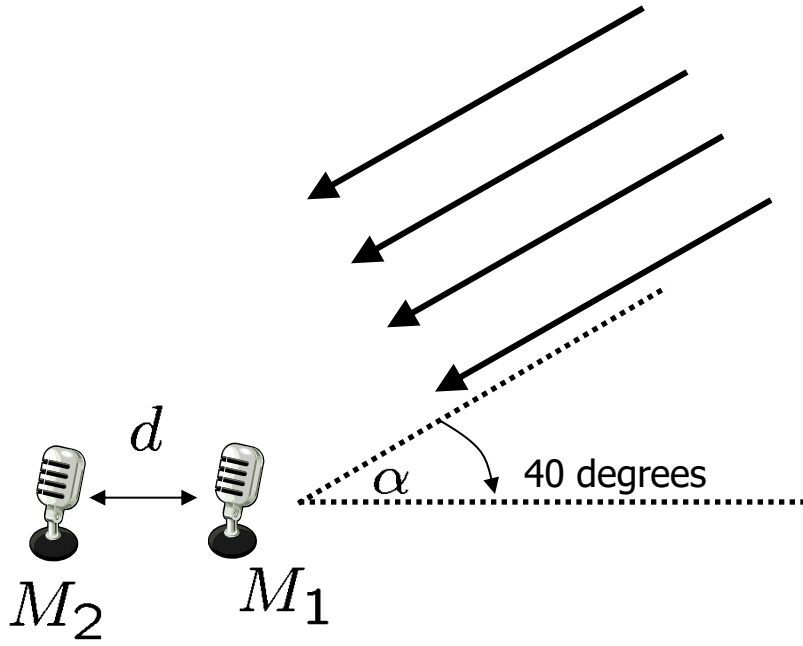
- Strictly speaking, early and late components are correlated via the source s_i.
- Often a known structure is assumed for the spatial correlation function of the late reflections (diffuse components), with a scaling depending on the variance of source s_i.

$$x[n] = \underbrace{\sum_{i=1}^{d} (h_e * s_i)[n]}_{\text{Early refl.}} + \underbrace{\sum_{i=1}^{d} (h_l * s_i)[n]}_{\text{Late refl.}} + n[n]$$

Concept of Beamforming

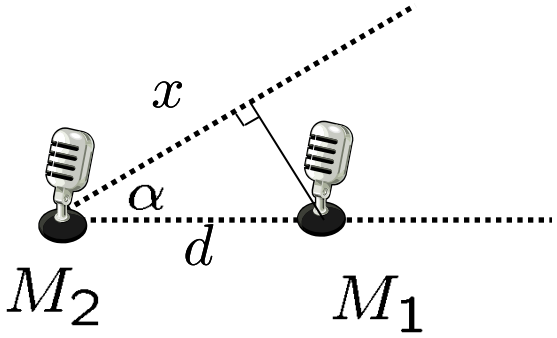
- Consider a sinusoidal source at 40 degrees of a dual microphone array (d=0.17 m).
- The sound source is in the far field (sound waves can be considered planar)





TU Delft

Concept of Beamforming



$$x = \frac{\tau}{f_s} c$$

$$\tau = \cos(\alpha) \frac{d}{c} f_s \quad \Rightarrow \quad \tau = 3.06 \text{ samples}$$

- $\alpha = 40$ degrees
- $f_s = 8000$ Hz
- $d = 0.17$ m
- $c = 340$ m/s

35

Concept of Beamforming

- Non-integer shifts: Use time domain interpolation or frequency domain phase change.
- The narrowband assumption: $z(t) = \text{real}\{s(t)e^{j\omega_0 t}\}$
  - The narrowband assumption: If $B\tau \ll 2\pi$ ($W\tau \ll 1$), then

    $$z_\tau(t) = z(t-\tau) = \text{real}\{s(t-\tau)e^{j\omega_0(t-\tau)}\} \approx \text{real}\{s(t)e^{j\omega_0(t-\tau)}\}$$

  - $W\tau \ll 1 \Rightarrow \tau_{max} \ll \frac{1}{W} = T_s$
  - Narrowband condition: The maximal delay τ_{max} across the array is less than the sampling period T_s.
  - with T_s in the order of $T_s = 1/8000$ this does not hold for audio.
  - Having $\omega_0 = 0$ we would have an instantanuous model, $s_\tau(t) = s(t)$, which is obviously incorrect.

Concept of Beamforming – Freq. domain

Due to non-integer shifts, processing thus done in frequency domain

- To satisfy narrowband assumption, processing per frequency band assuming narrowband assumption is satisfied per band.
  - e.g., using a DFT of size 512, $f_s = 8000$ Hz, $W\tau \ll 1 \Rightarrow \tau \ll \frac{512}{8000} = 0,064$
  - In this example, 3 samples delay is about 0.375 ms, hence narrowband assumpion is satisfied.
- Phase shifts become thus frequency dependent, and thus the beamformer response is frequency dependent.
- We have to deal with spatial aliasing (the equivalent of temporal aliasing): $d < \frac{1}{2}\lambda_{min} < \frac{1}{2}\frac{c}{\frac{1}{2}f_s} = \frac{c}{f_s}$.

Concept of Beamforming – Freq. domain

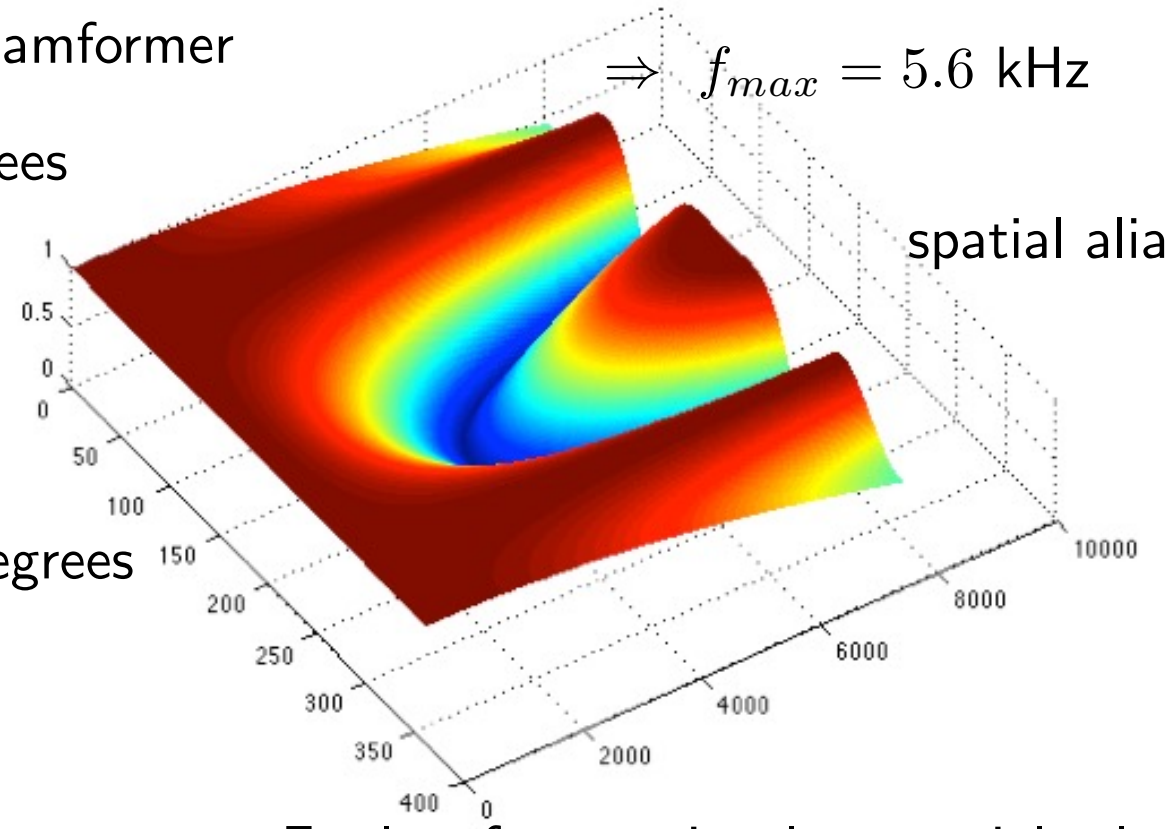
Example:

$$d = 0.03 < \frac{1}{2}\lambda_{min} = \frac{1}{2}\frac{c}{f_{max}}.$$

- Sum and delay beamformer
- Target at 60 degrees
- two microphones
- $d = 0.03$

$$\Rightarrow \quad f_{max} = 5.6 \text{ kHz}$$

spatial aliasing

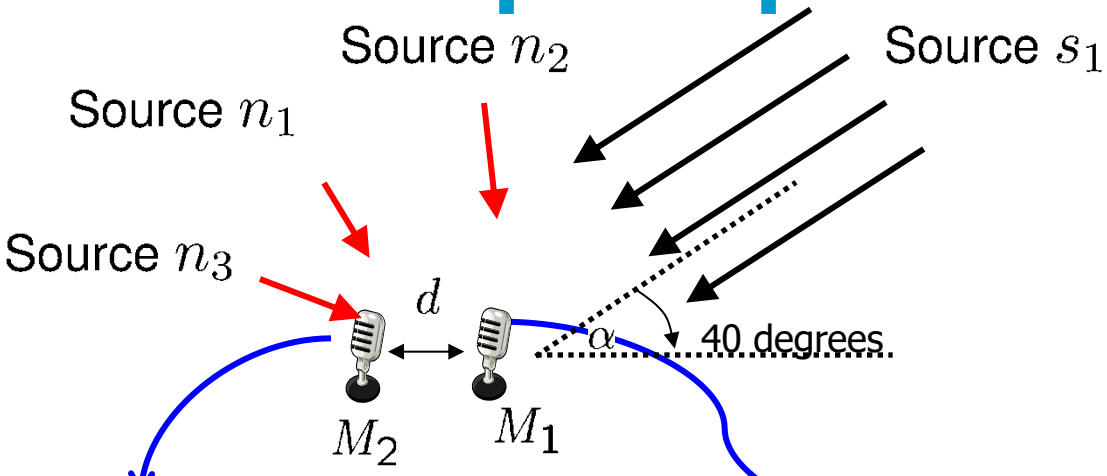


response of 1 at 60 degrees

For low frequencies, less spatial selectivity

TU Delft

How to exploit spatial filtering?



$$x_2(k,l) = s(k,l)e^{-j2\pi\frac{k\tau}{N}} + n_2(k,l)$$

$$x_1(k,l) = s(k,l) + n_1(k,l)$$

How to obtain an estimate $\hat{s}(k,l)$?

Given that direction α is known (i.e., τ) compensate for delay:

$$\hat{s}(k,l) = \frac{x_1(k,l)+x_2(k,l)e^{j2\pi\frac{k\tau}{N}}}{2}$$

$$= \frac{s(k,l)+n_1(k,l)+S(k,l)e^{-j2\pi\frac{k\tau}{N}}e^{j2\pi\frac{k\tau}{N}}+N_1(k,l)e^{j2\pi\frac{k\tau}{N}}}{2} = S(k,l)+\frac{N_1(k,l)+N_2(k,l)e^{j2\pi\frac{k\tau}{N}}}{2}$$

39

TU Delft

How to exploit spatial filtering?

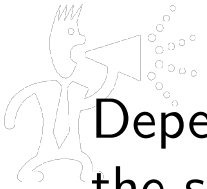
$$\hat{s}(k,l) = \frac{x_1(k,l) + x_2(k,l)e^{j2\pi\frac{k\tau}{N}}}{2} = S(k,l) + \frac{N_1(k,l) + N_2(k,l)e^{j2\pi\frac{k\tau}{N}}}{2}$$

- If the noise sources come from different angles as the speech source, the noise DFT coefficients $N_{1,k}(l)$ and $N_{2,k}(l)$ will be added destructively.
- If the noise is uncorrelated across microphones, i.e., $E[N_{1,k}(l)N_{2,k}^*(l)] = 0$, this operation involving two microphones will reduce the variance with a factor 2 (or three dB).
- This beam former is called the "delay and sum beamformer", after the two operations that are applied.

Signal models – near field

When sources travel to the microphones, the distance from source to each microphone influences the experienced damping and phase of the measured signal:

$$s(k,l) \Rightarrow s(k,l)a(d)e^{-j2\pi\frac{k\tau(d)}{N}}.$$

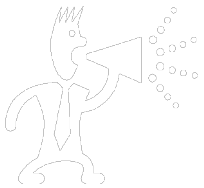


Depending on the size of the array and the distance of the array to the source, this gives rise to two different signal models:

- Near-field (and free field):
  - The source is close to the center of the array. The experienced damping is therefore different for every microphone.
  - Damping (a inversely proportional with distance) and phase differences τ are taken into account.

Signal models – far field

- Far-field (and free field):
  - The source is far away from the center of the array. The waves travel therefore parallel. The microphones experience no difference in damping.
  - Only phase differences τ are taken into account.



$$s(k,l) \Rightarrow s(k,l)e^{-j2\pi\frac{k\tau(d)}{N}}.$$

- Free field
  - No reflections, only direct path
- Typically one takes the early part of the room impulse response into account (i.e., all early reflections).

Short-Time Frequency Transform

Processing is often done in the so-called short-time frequency domain, i.e., FFT on short windowed time frames.

- Time frames should obey Short time WSS assumption.
- STFT makes convolutive model (approximately) multiplicative AND helps to satisfy narrowband assumption.
- $x(k,l) = \sum_{i=1}^{d} a_i(k,l)s_i(k,l) + n(k,l)$
- For M microphones using stacked vector notation:

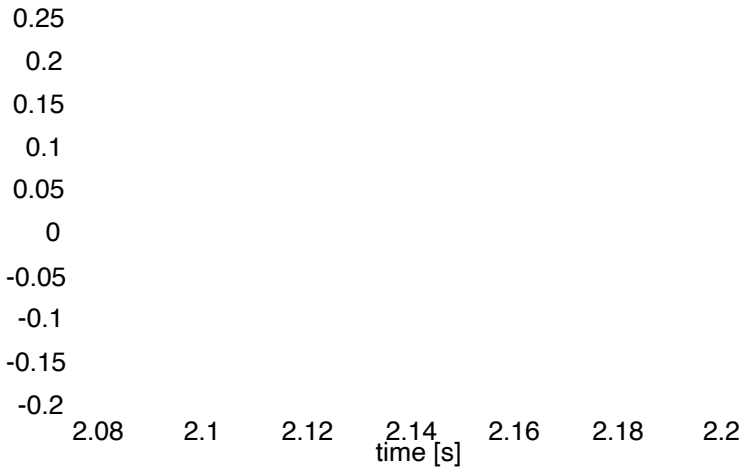
$$\mathbf{x}(k,l) = \sum_{i=1}^{d} \mathbf{a}_i(k,l)s_i(k,l) + \mathbf{n}(k,l)$$

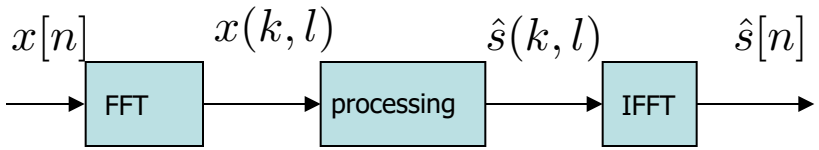
- Notice: As all processing is often done independently per frequency band and time frame, time and frequency indices are usually neglected.

Short-Time Frequency Transform

Segmentation

Overlap-add





TU Delft

Problem formulation

$$\mathbf{x}(k,l) = \sum_{i=1}^{d} \mathbf{a}_i(k,l) s_i(k,l) + \mathbf{n}(k,l)$$

- Assuming a single target and considering remaining point sources as interferers, abusing notation we can write

$$\mathbf{x}(k,l) = \underbrace{\mathbf{a}_1(k,l) s_1(k,l)}_{target} + \underbrace{\sum_{i=2}^{d} \mathbf{a}_i(k,l) s_i(k,l) + \mathbf{n}'(k,l)}_{interferers+noise}$$

$$= \mathbf{a}(k,l) s(k,l) + \mathbf{n}(k,l)$$

- Goal: Estimate $s(k,l)$ given $\mathbf{x}(k,l)$: e.g. $\hat{s}(k,l) = E[s(k,l)|\mathbf{x}(k,l)]$
- 1) Derive beamformers assuming $\mathbf{a}(k,l)$ is known.
- 2) estimation of $\mathbf{a}(k,l)$

The (Relative) Acoustic transfer function

$$\mathbf{x}(k,l) = \mathbf{a}(k,l)s(k,l) + \mathbf{n}(k,l)$$

- Notice that $\mathbf{a}(k,l)$ is the (Short Time) Fourier transform of the room impulse response per frequency, stacked across microphones
- Often $\mathbf{a}(k,l)$ is normalized with respect to the reference microphone, referred to as the relative transfer function (RTF).

$$\mathbf{x}(k,l) = \bar{\mathbf{a}}(k,l) \underbrace{a_1(k,l)s(k,l)}_{s_1(k,l)} + \mathbf{n}(k,l)$$

  Using the RTF
  - significantly shortens the length of the response.
  - implies we estimate the target at the reference microphone.

- Notice that the room impulse response (in the order of 100ms - 1 s) is typically much longer than the frame size used (20 ms).

Delay & Sum Beamformer

Assuming free and near-field, and choosing the first microphone as the reference, we have

$$\mathbf{x}(k,l) \quad = \quad s_1(k,l)\mathbf{a}(k,l) + \mathbf{n}(k,l).$$

with

$$\mathbf{a}(k,l) = [1, \frac{a_2 e^{-j2\pi \frac{k\tau_2}{N}}}{a_1}, ..., \frac{a_M e^{-j2\pi \frac{k\tau_M}{N}}}{a_1}]^T.$$

For the general case (non-free field) $\mathbf{a}(k,l)$ just models the complete ATF. Knowing $\mathbf{a}(k,l)$, we can calculate the delay and sum beamformer

$$\hat{s}(k,l) = \mathbf{w}^H(k,l)\mathbf{x}(k,l) = \frac{\mathbf{a}^H(k,l)\mathbf{x}(k,l)}{\mathbf{a}^H(k,l)\mathbf{a}(k,l)}.$$

Near and free field: $\mathbf{w}(k,l) = \frac{\mathbf{a}(k,l)}{\mathbf{a}^H(k,l)\mathbf{a}(k,l)}$

Far and free field: $\mathbf{w}(k,l) = \frac{1}{M}\mathbf{a}(k,l)$, with $\mathbf{a}(k,l)$ as defined above.

General case: $\mathbf{w}(k,l) = \frac{\mathbf{a}(k,l)}{\mathbf{a}^H(k,l)\mathbf{a}(k,l)}$ with $\mathbf{a}(k,l)$ the ATF.

Delay & Sum Beamformer

Delay and sum

- preserves the target.
- does not take explicit knowledge on the noise field into account.
- reduces the noise variance in most ideal case (uncorrelated noise across microphones) with a factor
  $\frac{1}{M} = \frac{1}{2^p} \Rightarrow -p10\log_{10}(2) \approx -3p\ dB$

MVDR - beamformer

More advanced beamformers not only exploit position of target, but position of noise sources as well. A well-known adaptive beamformer is the "minimum variance distortionless response" (MVDR) beamformer

- Constrains the beamformer to have no change of magnitude and phase in direction of target source.
- Minimizes the variance of the beamformer output in all other directions.

MVDR - beamformer

Cost function: $J(\mathbf{w}(k,l)) = \mathbf{w}^H(k,l)\mathbf{R}_\mathbf{x}(k,l)\mathbf{w}(k,l)$

$$\min_{\mathbf{w}(k,l)} J(\mathbf{w}(k,l))$$

$$s.t. \mathbf{w}(k,l)^H \mathbf{a}(k,l) = 1.$$

$$\frac{d}{d\mathbf{w}^H(k,l)} \left\{ J(\mathbf{w}(k,l)) + \lambda(\mathbf{w}^H(k,l)\mathbf{a}(k,l) - 1) \right\} = \mathbf{R}_\mathbf{x}(k,l)\mathbf{w}(k,l) + \lambda\mathbf{a}(k,l)$$

$$\mathbf{R}_\mathbf{x}(k,l)\mathbf{w}(k,l) + \lambda\mathbf{a}(k,l) = 0 \Rightarrow, \mathbf{w}(k,l) = -\left(\mathbf{R}_\mathbf{x}(k,l)\right)^{-1} \lambda\mathbf{a}(k,l)$$

MVDR - beamformer

Use the constraint: $\mathbf{a}^H(k,l)\mathbf{w}(k,l) = 1 = -\mathbf{a}^H(k,l)\left(\mathbf{R}_\mathbf{x}(k,l)\right)^{-1}\lambda\mathbf{a}(k,l)$

$$\Rightarrow \quad \lambda = -\frac{1}{\mathbf{a}^H(k,l)\left(\mathbf{R}_\mathbf{x}(k,l)\right)^{-1}\mathbf{a}(k,l)} \quad \Rightarrow$$

$$\mathbf{w}(k,l) = \frac{\left(\mathbf{R}_\mathbf{x}(k,l)\right)^{-1}\mathbf{a}(k,l)}{\mathbf{a}^H(k,l)\left(\mathbf{R}_\mathbf{x}(k,l)\right)^{-1}\mathbf{a}(k,l)}$$

MVDR - beamformer

$$\mathbf{w}(k,l) = \frac{\left(\mathbf{R_x}(k,l)\right)^{-1}\mathbf{a}(k,l)}{\mathbf{a}^H(k,l)\left(\mathbf{R_x}(k,l)\right)^{-1}\mathbf{a}(k,l)}$$

The MVDR beamformer can also be written using the noise correlation matrix $\mathbf{R_n}(k,l)$ based on the matrix inversion lemma:

$$(\mathbf{A}+\mathbf{uv}^T)^{-1} = \mathbf{A}^{-1} - \frac{\mathbf{A}^{-1}\mathbf{uv}^T\mathbf{A}^{-1}}{1+\mathbf{v}^T\mathbf{A}^{-1}\mathbf{u}}$$

Matrix $\mathbf{R_x}(k,l)$ can be written as $\mathbf{R_x}(k,l) = \mathbf{R_n}(k,l) + \mathbf{a}(k,l)\mathbf{a}^H(k,l)\sigma_s^2(k,l)$

$$\mathbf{w}(k,l) = \frac{\mathbf{R_n}(k,l)^{-1}\mathbf{a}(k,l)\left(1-\frac{\mathbf{a}(k,l)^H\mathbf{R_n}(k,l)^{-1}\mathbf{a}(k,l)\sigma_s^2(k,l)}{1+\mathbf{a}^H\mathbf{R_n}(k,l)^{-1}\mathbf{a}(k,l)\sigma_s^2(k,l)}\right)}{\mathbf{a}^H\mathbf{R_n}(k,l)^{-1}\mathbf{a}(k,l)\left(1-\frac{\mathbf{a}^H(k,l)\mathbf{R_n}(k,l)^{-1}\mathbf{a}(k,l)\sigma_s^2(k,l)}{1+\mathbf{a}^H\mathbf{R_n}(k,l)^{-1}\mathbf{a}(k,l)\sigma_s^2(k,l)}\right)} =$$

$$\frac{\mathbf{R_n}(k,l)^{-1}\mathbf{a}(k,l)}{\mathbf{a}^H(k,l)\mathbf{R_n}(k,l)^{-1}\mathbf{a}(k,l)}$$

MVDR - beamformer

$$\mathbf{w}(k,l) = \frac{\mathbf{R}_{\mathbf{x}}^{-1}(k,l)\mathbf{a}(k,l)}{\mathbf{a}^{H}(k,l)\mathbf{R}_{\mathbf{x}}^{-1}(k,l)\mathbf{a}(k,l)} = \frac{\mathbf{R}_{\mathbf{n}}^{-1}(k,l)\mathbf{a}(k,l)}{\mathbf{a}^{H}(k,l)\mathbf{R}_{\mathbf{n}}^{-1}(k,l)\mathbf{a}(k,l)}$$

This holds under the assumption that 1) $\mathbf{R}_{\mathbf{s}}(k,l)$ is rank-1 2) target and noise are uncorrelated and 3) target and noise are additive

MVDR – Spatially uncorrelated noise

$$\mathbf{w}(k,l) = \frac{\mathbf{R_n}(k,l)^{-1}\mathbf{a}(k,l)}{\mathbf{a}^H(k,l)\mathbf{R_n}(k,l)^{-1}\mathbf{a}(k,l)}$$

If the noise field is spatially uncorrelated, i.e., $\mathbf{R_n}(k,l) = \sigma_N^2(k,l)\mathbf{I}_M$, the MVDR equals the delay and sum beamformer

$$\mathbf{w}(k,l) = \frac{\mathbf{R_n}(k,l)^{-1}\mathbf{a}(k,l)}{\mathbf{a}^H(k,l)\mathbf{R_n}(k,l)^{-1}\mathbf{a}(k,l)} = \frac{\mathbf{a}(k,l)}{\mathbf{a}^H(k,l)\mathbf{a}(k,l)}$$

(assuming far-field and free-field):

$$\mathbf{w}(k,l) = \frac{\mathbf{a}(k,l)}{M}$$

Optimal Linear Multi-Channel Wiener

Signal model: $\mathbf{x}(k,l) = s(k,l)\mathbf{a}(k,l) + \mathbf{n}(k,l)$

Cost function: $J_{MSE}(\mathbf{w}(k,l)) = E[||s(k,l) - \mathbf{w}^H(k,l)\mathbf{x}(k,l)||_2^2]$

$$\begin{aligned}\frac{dJ_{MSE}(\mathbf{w}(k,l))}{d\mathbf{w}^H(k,l)} &= -E[s^H(k,l)\mathbf{x}(k,l)] + \mathbf{R}_{\mathbf{x}}(k,l)\mathbf{w}(k,l) \\ &= -\sigma_s^2(k,l)\mathbf{a}(k,l) + \mathbf{R}_{\mathbf{x}}(k,l)\mathbf{w}(k,l) = 0\end{aligned}$$

$$\mathbf{w}(k,l) = \mathbf{R}_{\mathbf{x}}^{-1}(l)\sigma_{S,k}^2\mathbf{a}(k,l)$$

Optimal Linear Multi-Channel Wiener

Using again the Matrix inversion lemma, it can be shown that

$$\mathbf{w}(k,l) = R_{\mathbf{x}}^{-1}(k,l)\sigma_s^2(k,l)\mathbf{a}(k,l)$$

can be written as

$$\mathbf{w}(k,l) = \underbrace{\frac{\sigma_s^2(k,l)}{\sigma_s^2(k,l) + (\mathbf{a}^H(k,l)R_{\mathbf{n}}^{-1}(k,l)\mathbf{a}(k,l))^{-1}}}_{\text{Single-channel Wiener}} \underbrace{\frac{R_{\mathbf{n}}^{-1}(k,l)\mathbf{a}(k,l)}{\mathbf{a}^H(k,l)R_{\mathbf{n}}^{-1}(k,l)\mathbf{a}(k,l)}}_{MVDR}$$

Optimal Linear Multi-Channel Wiener

matrix inversion lemma:

$$(\mathbf{A}+\mathbf{u}\mathbf{v}^T)^{-1} = \mathbf{A}^{-1} - \frac{\mathbf{A}^{-1}\mathbf{u}\mathbf{v}^T\mathbf{A}^{-1}}{1+\mathbf{v}^T\mathbf{A}^{-1}\mathbf{u}}$$

Matrix $\mathbf{R}_\mathbf{x}(k,l)$ can be written as $\mathbf{R}_\mathbf{x}(k,l) = \mathbf{R}_\mathbf{n}(k,l) + \mathbf{a}\mathbf{a}^H\sigma_s^2(k,l)$

$$\begin{aligned}
\mathbf{R}_\mathbf{x}^{-1}(k,l)\mathbf{a}(k,l)\sigma_s^2(k,l) &= \left(\mathbf{R}_\mathbf{n}(k,l)+\mathbf{a}\mathbf{a}^H\sigma_s^2(k,l)\right)^{-1}\mathbf{a}(k,l)\sigma_s^2(k,l)\\
&= \mathbf{R}_\mathbf{n}^{-1}(k,l)\mathbf{a}(k,l)\sigma_s^2(k,l)\\
&- \mathbf{R}_\mathbf{n}^{-1}(k,l)\mathbf{a}(k,l)\frac{\sigma_s^2(k,l)\mathbf{a}^H(k,l)\mathbf{R}_\mathbf{n}^{-1}(k,l)\mathbf{a}(k,l)}{1+\sigma_s^2(k,l)\mathbf{a}^H(k,l)\mathbf{R}_\mathbf{n}^{-1}(k,l)\mathbf{a}(k,l)}\sigma_s^2(k,l)\\
&= \mathbf{R}_\mathbf{n}^{-1}(k,l)\mathbf{a}(k,l)\left(1-\frac{\sigma_s^2(k,l)\mathbf{a}(k,l)^H\mathbf{R}_\mathbf{n}^{-1}(k,l)\mathbf{a}(k,l)}{1+\sigma_s^2(k,l)\mathbf{a}^H(k,l)\mathbf{R}_\mathbf{n}^{-1}(k,l)\mathbf{a}(k,l)}\right)\sigma_s^2(k,l)
\end{aligned}$$

Optimal Linear Multi-Channel Wiener

$$
\begin{aligned}
&= \mathbf{R}_{\mathbf{n}}^{-1}(k,l)\mathbf{a}(k,l)\left(1-\frac{\sigma_s^2(k,l)\mathbf{a}(k,l)^H\mathbf{R}_{\mathbf{n}}^{-1}(k,l)\mathbf{a}(k,l)}{1+\sigma_s^2(k,l)\mathbf{a}^H(k,l)\mathbf{R}_{\mathbf{n}}^{-1}(k,l)\mathbf{a}(k,l)}\right)\sigma_s^2(k,l)\\
&= \mathbf{R}_{\mathbf{n}}^{-1}(k,l)\mathbf{a}(k,l)\left(\frac{\sigma_s^2(k,l)}{1+\sigma_s^2(k,l)\mathbf{a}^H(k,l)\mathbf{R}_{\mathbf{n}}^{-1}(k,l)\mathbf{a}(k,l)}\right)\\
&= \frac{\mathbf{R}_{\mathbf{n}}^{-1}(k,l)\mathbf{a}(k,l)}{\mathbf{a}^H(k,l)\mathbf{R}_{\mathbf{n}}^{-1}(k,l)\mathbf{a}(k,l)}\left(\frac{\mathbf{a}^H(k,l)\mathbf{R}_{\mathbf{n}}^{-1}(k,l)\mathbf{a}(k,l)\sigma_s^2(k,l)}{1+\sigma_s^2(k,l)\mathbf{a}^H(k,l)\mathbf{R}_{\mathbf{n}}^{-1}(k,l)\mathbf{a}(k,l)}\right)\\
&= \frac{\mathbf{R}_{\mathbf{n}}^{-1}(k,l)\mathbf{a}(k,l)}{\mathbf{a}^H(k,l)\mathbf{R}_{\mathbf{n}}^{-1}(k,l)\mathbf{a}(k,l)}\left(\frac{\sigma_s^2(k,l)}{\left(\mathbf{a}^H(k,l)\mathbf{R}_{\mathbf{n}}^{-1}(k,l)\mathbf{a}(k,l)\right)^{-1}+\sigma_s^2(k,l)}\right)
\end{aligned}
$$

Optimal Linear Multi-Channel Wiener

$$\mathbf{w}(k,l) = \underbrace{\frac{\sigma_s^2(k,l)}{\sigma_s^2(k,l) + (\mathbf{a}^H(k,l)R_{\mathbf{n}}^{-1}(k,l)\mathbf{a}(k,l))^{-1}}}_{\text{Single-channel Wiener}} \underbrace{\frac{R_{\mathbf{n}}^{-1}(k,l)\mathbf{a}(k,l)}{\mathbf{a}^H(k,l)R_{\mathbf{n}}^{-1}(k,l)\mathbf{a}(k,l)}}_{MVDR}$$

The multi-channel Wiener filter can thus be seen as a concatenation of two filters:

- An MVDR as spatial filter
- Single-Channel Wiener filter as post-processor where the noise variance is set to the remaining noise PSD after beamforming:

$$\mathbf{w}^H(k,l)R_{\mathbf{n}}(k,l)\mathbf{w}(k,l) = \mathbf{a}^H(k,l)R_{\mathbf{n}}^{-1}(k,l)\mathbf{a}(k,l)$$

Sufficient Statistics

- For $\mathbf{n}$ Gaussian distributed,

$$T\left(\mathbf{x}(k,l)\right) = \mathbf{w}_{\mathrm{MVDR}}^{H}(k,l)\mathbf{x}(k,l) = \frac{\mathbf{a}^{H}(k,l)R_{\mathbf{n}}^{-1}(k,l)\mathbf{x}(k,l)}{\mathbf{a}^{H}(k,l)R_{\mathbf{n}}^{-1}(k,l)\mathbf{a}(k,l)}$$

  is known to be a sufficient statistic for s.

- This means no information is lost on s by using $T\left(\mathbf{x}(k,l)\right)$ instead of $\mathbf{x}(k,l)$.

- This result holds in general for any prior distribution on $s(k,l)$ and any cost-function (e.g., quadratic (MSE), uniform (MAP), Absolute error (Median)) and any function of s (e.g., $|s|$, $|s|^2$, etc.)

Sufficient Statistics

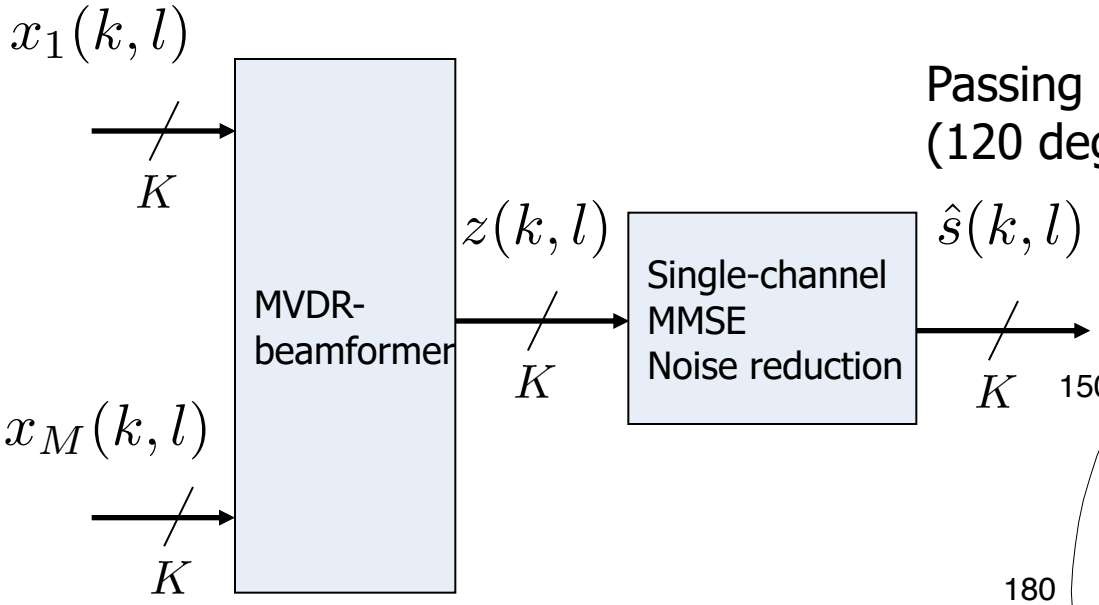
- Let $f_S(s|y)$ denote the conditional pdf of random variable S. In then holds that for a sufficient statistics $f_S(s|\mathbf{x}) = f_S\left(s|T(\mathbf{x})\right)$
- If $f_{\mathbf{x}}\left(\mathbf{x}|T(\mathbf{x}; s\right)$ is independent of s, $T(\mathbf{x})$ is a sufficient statistic for estimating s.
- Equivalent: $I\left(s; T(\mathbf{x})\right) = I\left(s; \mathbf{x}\right)$, i.e., we have equality in the data processing inequality and no information is lost.

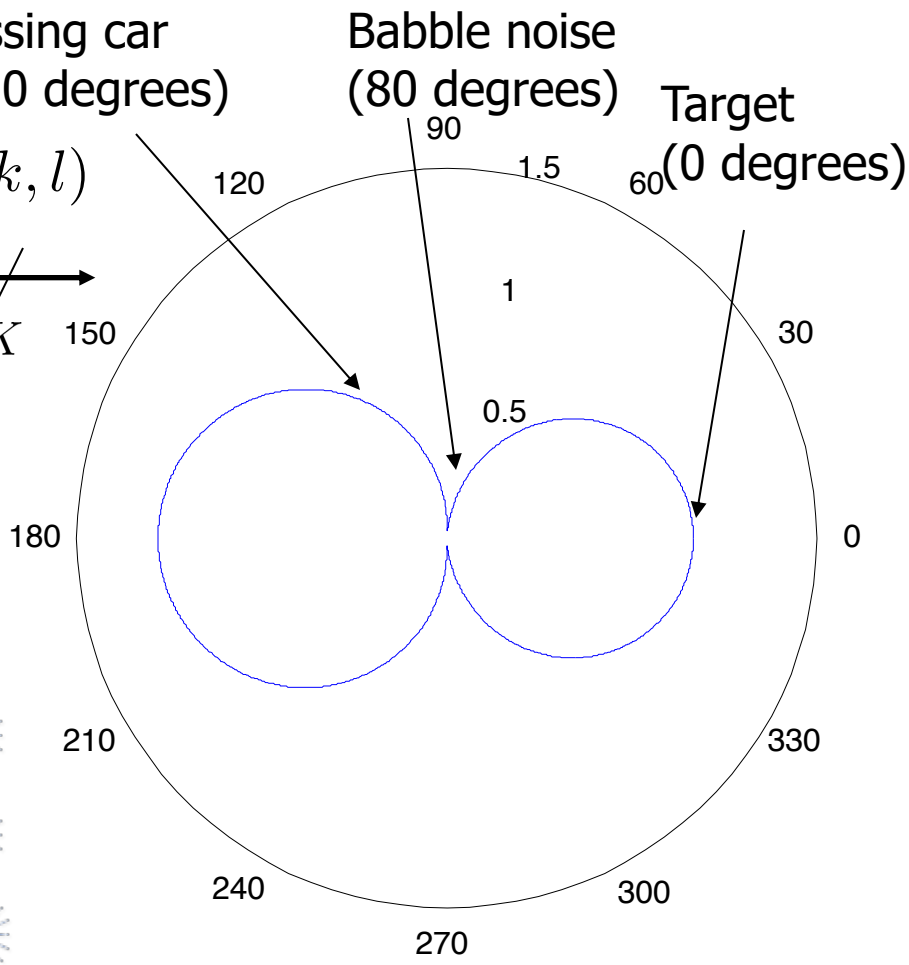
Finding a sufficient statistic: if the pdf $f_{\mathbf{x}}\left(\mathbf{x}; s\right)$ can be factorized as

$$f_{\mathbf{x}}\left(\mathbf{x}; s\right) = g(T(\mathbf{x}), s)h(\mathbf{x}),$$

then $T(\mathbf{x})$ is a sufficient statistic for s.

Example: Multi-Channel Noise Reduction





- Clean input $s(k,l)$
- Noisy input at mic. 1 $x_1(k,l)$
- MVDR-beamformer output $z(k,l)$
- Output of total system $\hat{s}(k,l)$

LCMV - beamformer

Remember the MVDR: $J(\mathbf{w}(k,l)) = \mathbf{w}^H(k,l)\mathbf{R}_\mathbf{x}(k,l)\mathbf{w}(k,l)$

$$\min_{\mathbf{w}(k,l)} J(\mathbf{w}(k,l))$$
$$s.t. \mathbf{w}(k,l)^H \mathbf{a}(k,l) = 1.$$

- The MVDR imposes one constraint.
- This can be generalised to having d constraints.

LCMV - beamformer

Cost function: $J(\mathbf{w}(k,l)) = \mathbf{w}^H(k,l)\mathbf{R}_\mathbf{x}(k,l)\mathbf{w}(k,l)$

$$\min_{\mathbf{w}(k,l)} J(\mathbf{w}(k,l))$$
$$s.t. \mathbf{w}^H(k,l)\mathbf{\Lambda}(k,l) = \mathbf{f}^H(k,l).$$

with $\mathbf{\Lambda} \in \mathbb{C}^{M \times d}$

When $d < M$, there is a closed form solution:

$$\mathbf{w}(k,l) = \mathbf{R}_\mathbf{x}^{-1}(k,l)\mathbf{\Lambda}(k,l)\left(\mathbf{\Lambda}^H(k,l)\mathbf{R}_\mathbf{x}^{-1}(k,l)\mathbf{\Lambda}(k,l)\right)^{-1}\mathbf{f}(k,l).$$

LCMV - beamformer

$$\mathbf{w}_k = \mathbf{R}_{\mathbf{x}}^{-1}(k,l)\mathbf{\Lambda}(k,l)\left(\mathbf{\Lambda}^H(k,l)\mathbf{R}_{\mathbf{x}}^{-1}(k,l)\mathbf{\Lambda}(k,l)\right)^{-1}\mathbf{f}(k,l).$$

How to use the multiple constraints?

- To steer zeros in the direction of certain noise sources.
- To maintain the signal from certain directions.
- To maintain the spatial cues of for hearing aids.

Notice that the more constraints are used, less degrees of freedom are left to control the noise reduction.

Overview of Discussed filters

- Delay and sum beamformer

$$\mathbf{w}(k,l) = \frac{\mathbf{a}(k,l)}{\mathbf{a}^H(k,l)\mathbf{a}(k,l)}$$

- MVDR beamformer

$$\mathbf{w}(k,l) = \frac{R_{\mathbf{x}}^{-1}(k,l)\mathbf{a}(k,l)}{\mathbf{a}^H(k,l)R_{\mathbf{x}}^{-1}(k,l)\mathbf{a}(k,l)} = \frac{R_{\mathbf{n}}^{-1}(k,l)\mathbf{a}(k,l)}{\mathbf{a}^H(k,l)R_{\mathbf{n}}^{-1}(k,l)\mathbf{a}(k,l)}$$

- Multi-Channel Wiener

$$\mathbf{w}(k,l) = \underbrace{\frac{\sigma_s^2(k,l)}{\sigma_s^2(k,l) + (\mathbf{a}^H(k,l)R_{\mathbf{n}}^{-1}(k,l)\mathbf{a}(k,l))^{-1}}}_{\text{Single-channel Wiener}} \underbrace{\frac{R_{\mathbf{n}}^{-1}(k,l)\mathbf{a}(k,l)}{\mathbf{a}^H(k,l)R_{\mathbf{n}}^{-1}(k,l)\mathbf{a}(k,l)}}_{MVDR}$$

Overview of Discussed filters

- LCMV beamformer
  $$\mathbf{w}(k,l) = \mathbf{R}_{\mathbf{x}}^{-1}(k,l)\mathbf{\Lambda}(k,l)\left(\mathbf{\Lambda}^{H}(k,l)\mathbf{R}_{\mathbf{x}}^{-1}(k,l)\mathbf{\Lambda}(k,l)\right)^{-1}\mathbf{f}(k,l).$$