Optimizing Speech Intelligibility in Noisy Environments Using a Simple Model of Communication

Richard C. Hendriks

work done in collaboration with
Bastiaan Kleijn (TUDelft, Victoria), Seyran Khademi (TUDelft) and
Steven van Kuijk (Victoria)

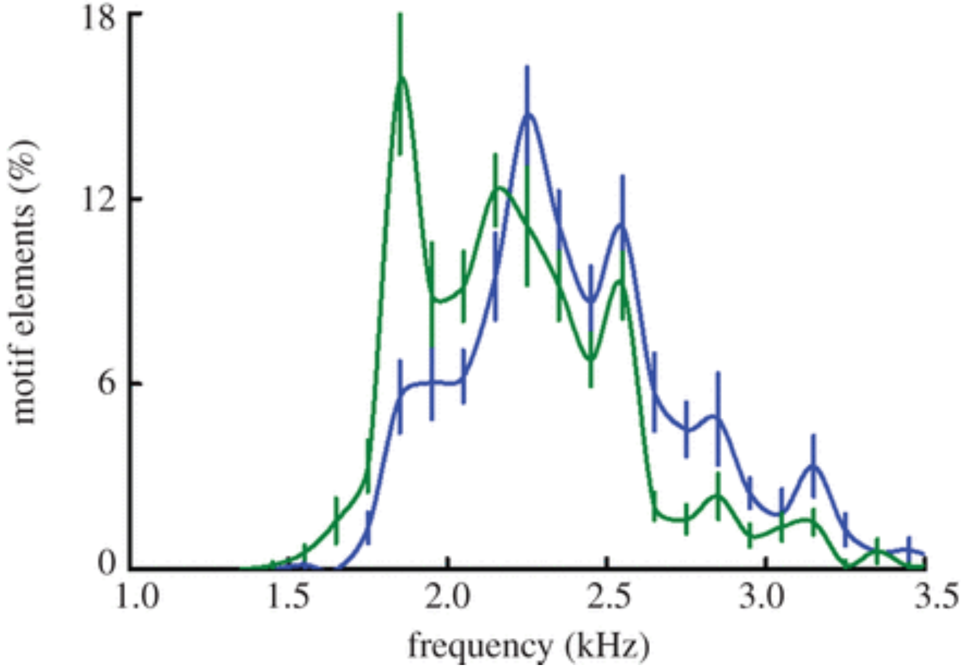
Circuits and Systems
Department of Microelectronics

TU Delft
Delft University of Technology

Blackbirds Know it...

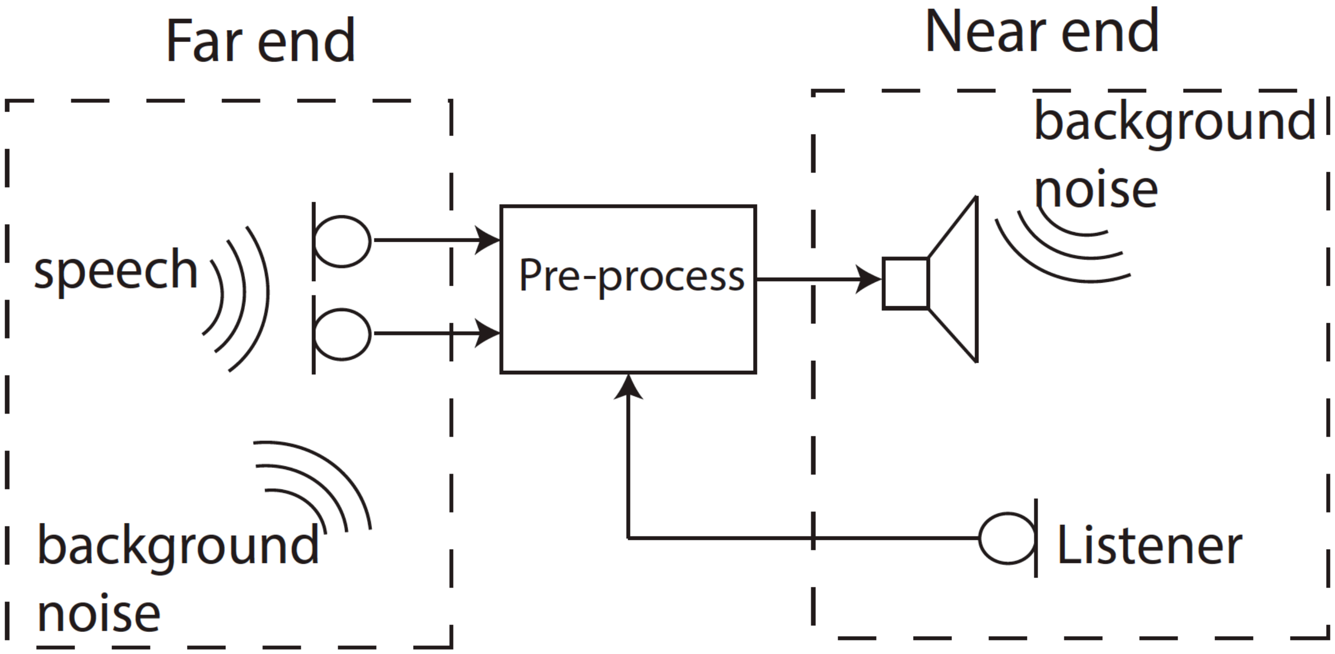
Average percentage of peak frequency values in 100 Hz intervals for

- 16 Viennese city blackbirds (blue)
- 17 Viennese forest blackbirds (green).



Taken from Nemeth et al. [1].

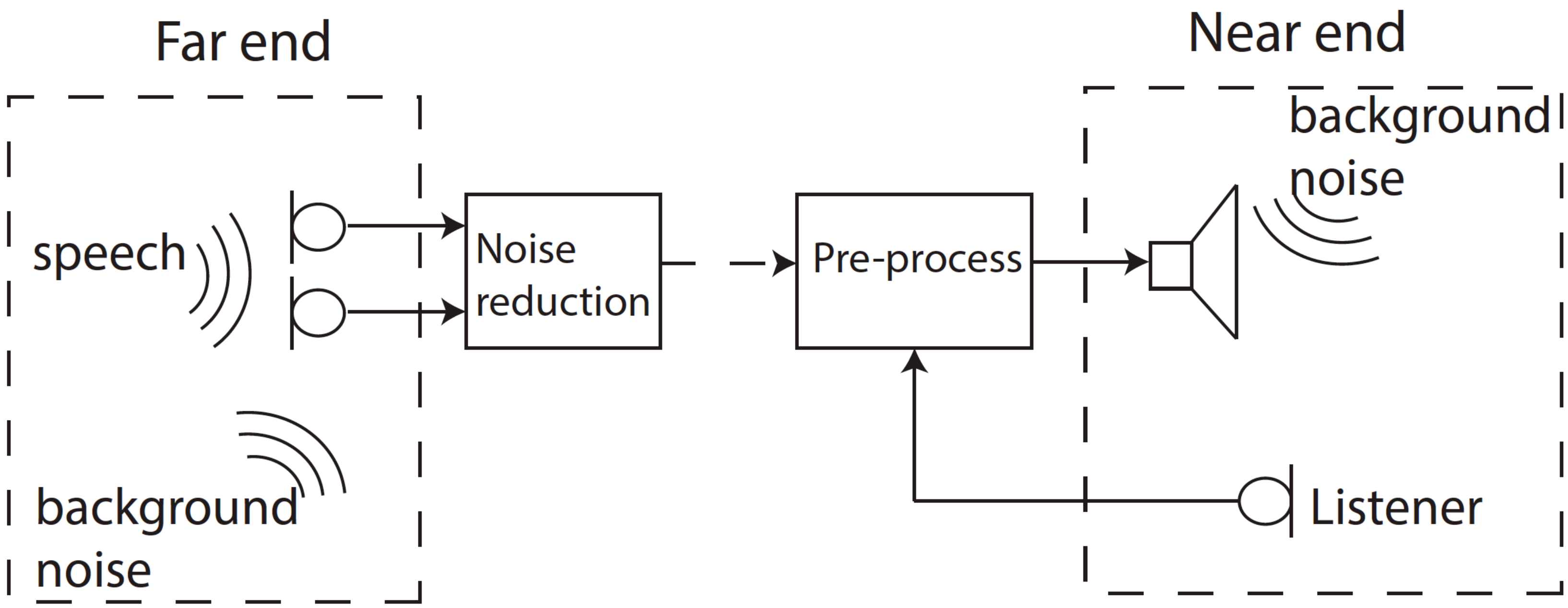
Typical Application Scenario



- How to maximize the intelligibility at the near-end?

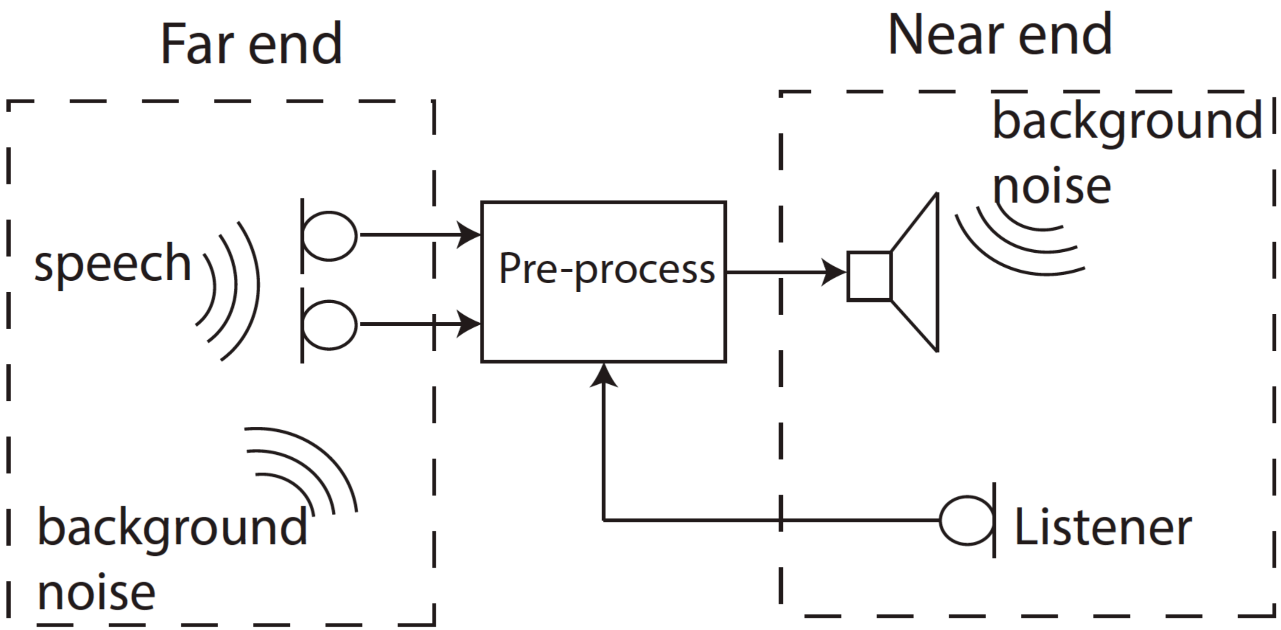
Circuits and Systems
Department of Microelectronics
TU Delft

Typical Application Scenario



- Typically independent processing with respect to noise at the near-end and noise at the far-end.
- Is this optimal in any sense?
- Which far-end information is needed for optimal processing?

Typical Application Scenario



Key questions:

- How to maximize near-end intelligibility using a processor that is jointly optimal with respect to the noise at far-end and near-end?

- How to model speech intelligibility?

Modelling Intelligibility

Classical Measures of Intelligibility

The Articulation Index (AI) and the Speech Intelligibility Index (SII):

- general structure:

$$\sum_{k \in \kappa} I_k A_k(\xi_k).$$

- I_k: maximum contribution of frequency band to intelligibility (band importance function)

- A_k: fraction to which a frequency band contributes to the intelligibility (band audibility).

$$\begin{aligned} A_k^{AI}(\xi_k) &= \min(\max(10\log_{10}\xi_k, 0), 30)/30 \\ A_k^{SII}(\xi_k) &= \max(\min(10\log_{10}\xi_k, 15), -15)/30 + 1/2 \end{aligned}$$

Classical Measures of Intelligibility

The Articulation Index (AI) and the Speech Intelligibility Index (SII):

- Functions I_k and A_k (including the different constants) are determined empirically using listening experiments.
- Roots date back to 1920, before information theory...
- ...can be interpreted as a measure of the rate of information (Allen [2]).

Circuits and Systems
Department of Microelectronics
TU Delft

Classical Measures of Intelligibility

Notice that the function A_k is not smooth and not concave, which complicates optimization:

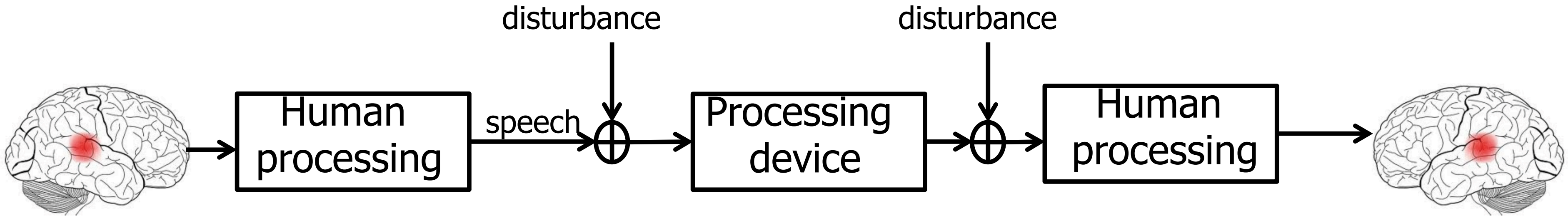
$$
\begin{aligned}
A_k^{AI}(\xi_k) &= \min(\max(10\log_{10}\xi_k, 0), 30)/30 \\
A_k^{SII}(\xi_k) &= \max(\min(10\log_{10}(\xi_k), 15), -15)/30 + 1/2
\end{aligned}
$$

A recent approximation of A_k was proposed by Taal et al. [3]:

$$A_k^{ASII}(\xi_k) = \frac{\xi_k}{\xi_k + 1}.$$

The similarity of this approximation and A_k^{AI}/A_k^{SII} is well within the precision of reasoning used to derive the AI and SII.

A Simple Model for Communication



What determines intelligibility?

- How well is the message from the talker's brain received by the listener's brain?
- Speech intelligibility: Transfer of information over a noisy channel.

Motivates the use of an information theoretical approach.

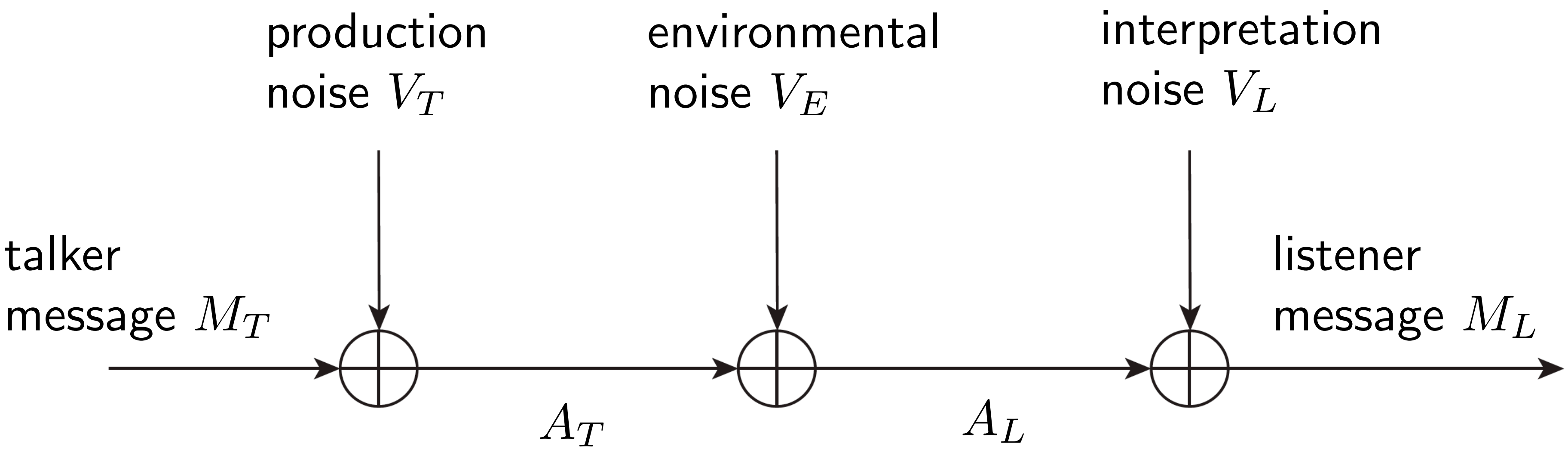
A Simple Model for Communication

- Define talker message and listener message by M_T and M_L.
- Define talker and listener acoustic equivalents as A_T and A_L.
- Define Markov chain: $M_T \to A_T \to A_L \to M_L$

$$\begin{aligned} A_T &= M_T + V_T \\ A_L &= A_T + V_E \\ M_L &= A_L + V_L \end{aligned}$$

- V_T: production noise
- V_L: interpretation noise
- V_E: environmental noise

A Simple Model for Communication



Consequence of production and interpretation noise:
The intelligibility will saturate when environmental noise decreases.

Or is the production noise multiplicative? $\otimes$

Production and Interpretation Noise

Production noise:

- Speech production is a probabilistic process.
- A speech sound shows variability for a single speaker, certainly across speakers.
- Variability is independent of the production level: The production SNR $\frac{\sigma^2_{M_T}}{\sigma^2_{V_T}}$ is scale independent.
- Consequence: correlation coefficient $\rho_{M_T A_T}$ is fixed.

Production and Interpretation Noise

Interpretation noise:

- In a similar way we could argue that certain aspects of the interpretation of the message is scale invariant.
- The interpretation SNR $\frac{\sigma^2_{A_L}}{\sigma^2_{V_L}}$ is fixed.
- Consequence: correlation coefficient $\rho_{A_L M_L}$ is fixed.

Consequence of fixed production/interpretation SNR: Only little benefit to have a frequency band with channel SNR $\xi_k = \frac{\sigma^2_{A_{T,k}}}{\sigma^2_{V_{E,k}}}$ above the production/interpretation SNR.

Usefulness of a channel saturates near production/interpretation SNR!

Mutual Information Between Talker and Listener

- Consider know a time-frequency representation, with frame index i and frequency bin index k.
- We assume all processes jointly Gaussian, stationary (omit time index i) and memoryless
- Independence across frequency channels:

$$I(M_T, M_L) = \sum_k I(M_{T_k}, M_{L_k})$$

- Let $\rho_{0,k} = \rho_{M_T A_T,k} \rho_{A_L M_L,k}$ and $\xi_k = \frac{\sigma^2_{A_{T,k}}}{\sigma^2_{V_{E,k}}}$

Circuits and Systems
Department of Microelectronics
TU Delft

Mutual Information Between Talker and Listener

Mutual information between M_L and M_T:

$$I(M_T; M_L) = -\sum_{k \in \kappa} \frac{1}{2} \log \left(\frac{(1-\rho_{0,k}^2)\xi_k + 1}{\xi_k + 1} \right)$$
$$= \sum_{k \in \kappa} I_k A_k(\xi_k)$$

with

$$A_k(\xi_k) = \frac{\log \frac{(1-\rho_{0,k}^2)\xi_k+1}{\xi_k+1}}{\log(1-\rho_{0,k}^2)} \quad \text{and} \quad I_k = -\frac{1}{2}\log(1-\rho_{0,k}^2)$$

Remember the AI and the SII!

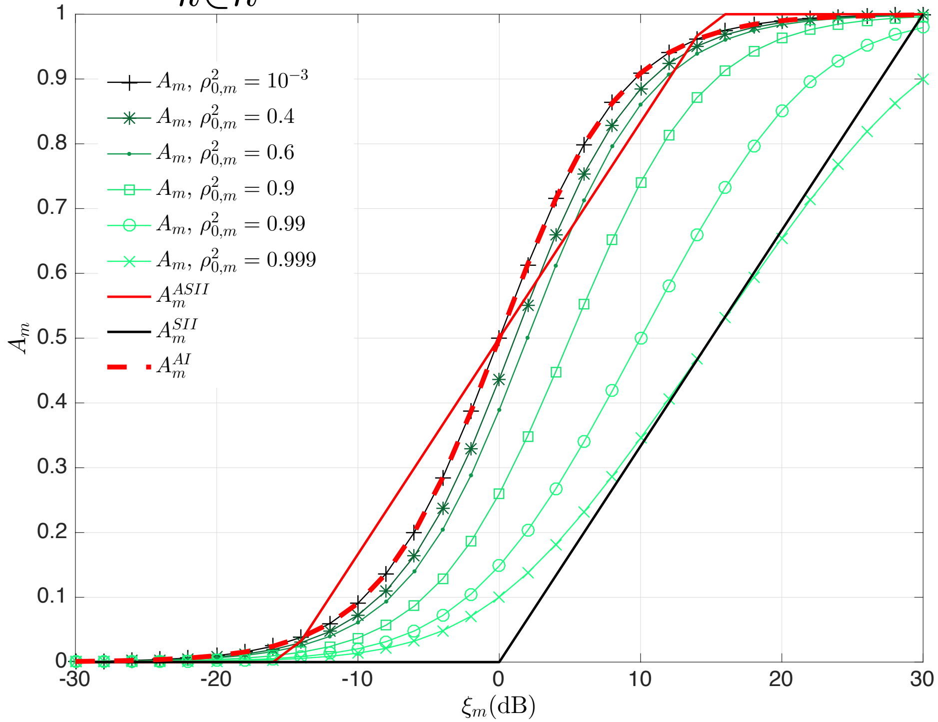
Circuits and Systems
Department of Microelectronics
TU Delft

Comparing to Classical models

SII: $$\sum_{k\in\kappa} I_k A_k^{\mathsf{SII}}, \quad A_k^{\mathsf{SII}} = \frac{\max(\min(10\log_{10}\xi_k, 15), 15)}{30} + \frac{1}{2}$$

ASII: $$\sum_{k\in\kappa} I_k A_k^{\mathsf{ASII}}, \quad A_k^{\mathsf{ASII}} = \frac{\xi_k}{\xi_k + 1}$$

prop.: $$\sum_{k\in\kappa} I_k A_k(\xi_k), \quad A_k(\xi_k) = \frac{\log \frac{(1-\rho_{0,k}^2)\xi_k+1}{\xi_k+1}}{\log(1-\rho_{0,k}^2)} \quad \text{and} \quad I_k = -\frac{1}{2}\log(1-\rho_{0,k}^2)$$



Although proposed A_k differs from the AI/SII A_k, it is well within the precision of reasoning used for AI/SII.

The Speech Production Uncertainty

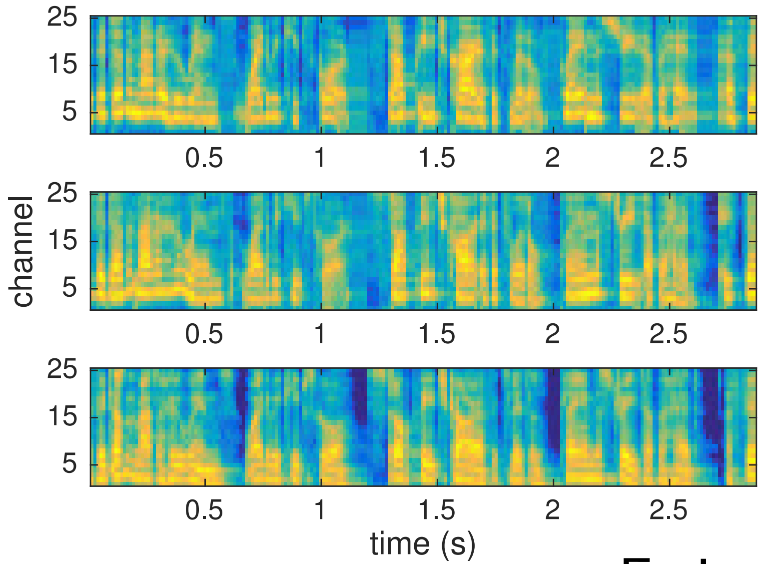
Can we measure the production noise?

- Many talkers producing the same sentence.
- Dynamic time warping to align signals.
- Ensemble average is the message.
- Production noise can then be estimated by considering the variability of each TF unit over the ensemble.

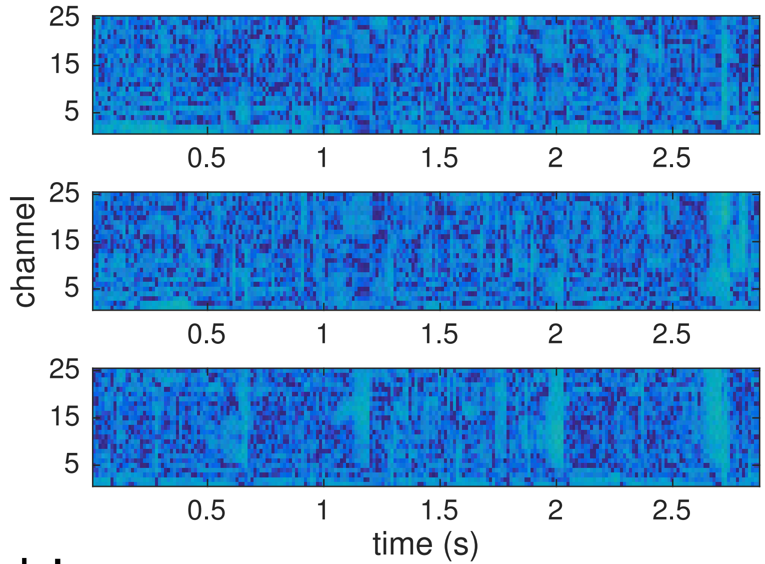
Notice: In this presentation I model the production noise as being additive, however, it is more likely to be multiplicative as we believe the production noise has its origin in variations in the envelope.

The Speech Production Uncertainty

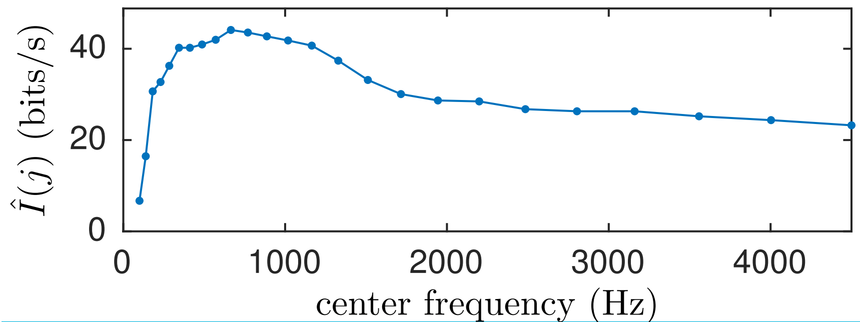
Aligned/warped signals



Variations among the signals



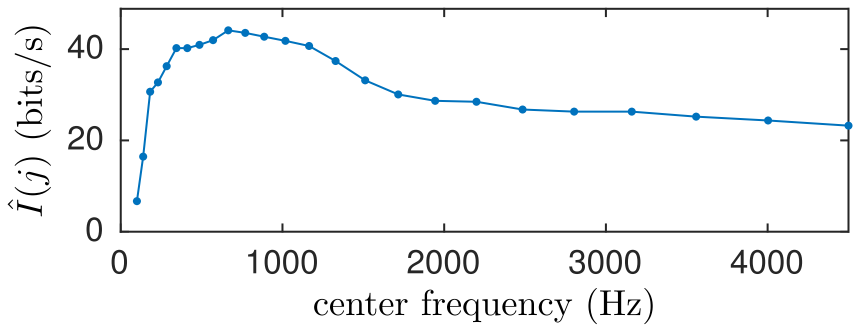
Estimate of band importance:



$$\hat{I}(k) = -\frac{R}{2}\log\left(1-\rho^2_{M_T A_T}(k)\right)$$

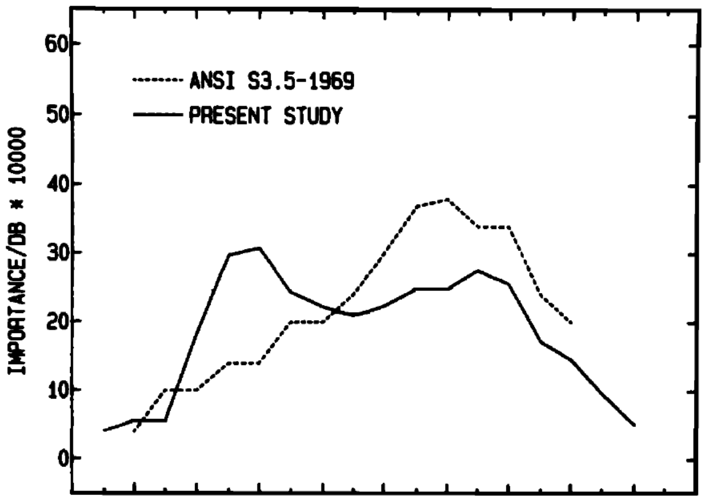
The Speech Production Uncertainty

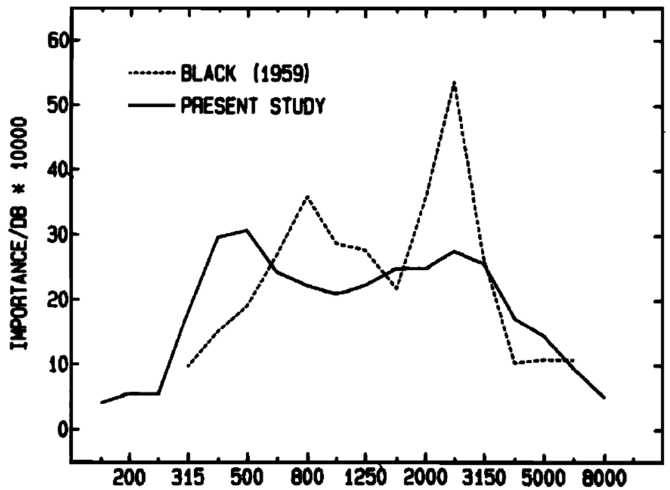
Estimate of band importance:



$$\hat{I}(k) = -\frac{R}{2}\log\left(1-\rho^2_{M_T A_T}(k)\right)$$

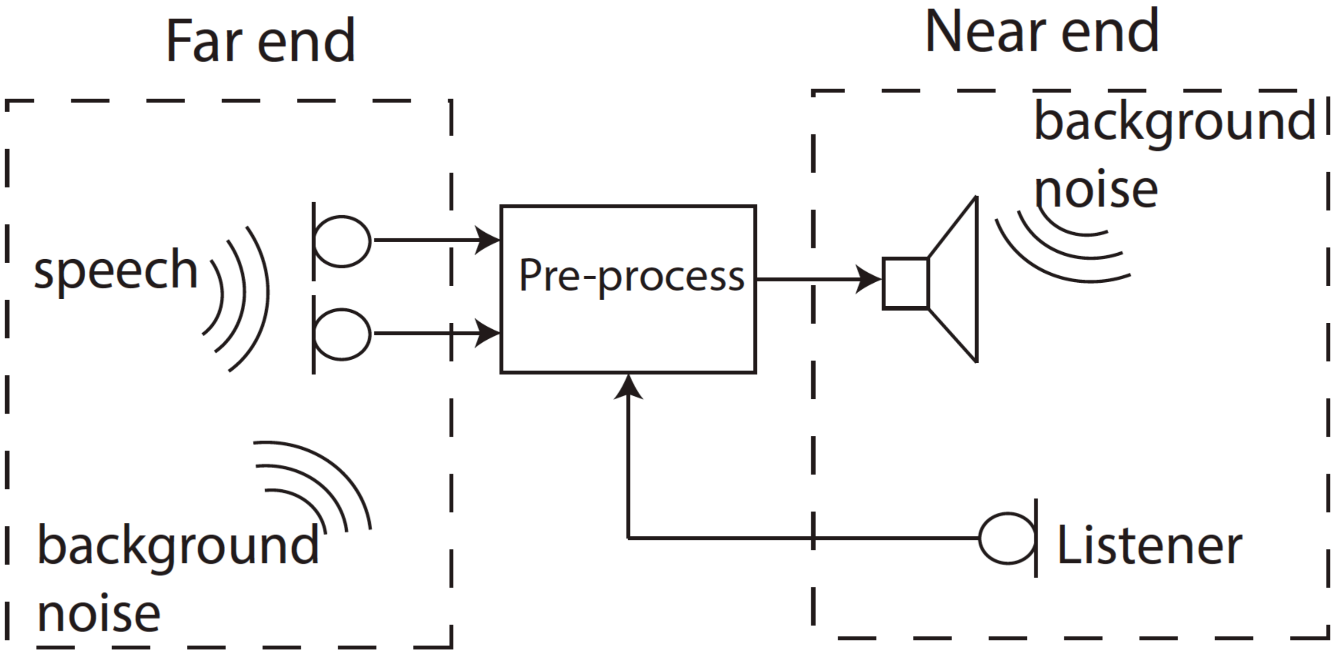
For comparison, some band importance functions published in Studebaker (1986).





Optimizing for Intelligibility

Scenario



Far end
Near end
speech
background noise
Pre-process
background noise
Listener

Assumptions

1. All processes are jointly Gaussian, stationary, and memoryless (we omit the time-frame index i for notational convenience)

2. Signal model follows the Markov chain model: $S \to T \to X \to \tilde{X} \to Y \to Z$.

3. Enhancement is performed by a linear time-invariant operator, $\mathbf{v}_k$.

4. Individual component signals of the vectors $\mathbf{s}_k$ and $\mathbf{z}_k$ are independent so the total mutual information is

$$I(S_i; Z_i) = \sum_k I(S_{k,i}; Z_{k,i})$$

Signal Model – Multi-Mic.

Multi-mic. Setup:

1. Produced signal : $T_k = \underbrace{S_k}_{\text{clean speech}} + \underbrace{V_k}_{\text{production noise}}$
2. Multi-mic. Rec. : $\mathbf{X}_k = \mathbf{d}_k T_k + \underbrace{\mathbf{U}_k}_{\text{far-end noise}}$
3. process. signal: $\tilde{X}_k = \mathbf{v}_k^H \mathbf{X}_k$
4. Received signal : $Y_k = \underbrace{\tilde{X}_k}_{\text{processed}} + \underbrace{N_k}_{\text{near-end noise}}$
5. Interpreted signal : $Z_k = Y_k + \underbrace{W_k}_{\text{interpratation noise}}$

With acoustic transfer function $\mathbf{d}_k = [d_{k,1}, ..., d_{k,M}]^T$ and far-end noise $\mathbf{U}_k = [U_{k,1}, ..., U_{k,M}]^T$.

Mutual Information

- In Markov chain the overall correlation coefficient (ρ) is the product of all coefficients:

$$\rho_{S_k Z_k} = \rho_{S_k T_k}\, \rho_{T_k \tilde{X}_k}\, \rho_{\tilde{X}_k Y_k}\, \rho_{Y_k Z_k}$$

- The mutual information:

$$I(S;Z) = \sum_k -\frac{1}{2}\log(1-\rho^2_{S_k Z_k}) = \sum_k -\frac{1}{2}\log(1-\rho^2_{0,k}\rho^2_{T_k \tilde{X}_k}\rho^2_{\tilde{X}_k Y_k})$$

with fixed $\rho_{0,k} = \rho_{S_k T_k}\rho_{Y_k Z_k}$.

- Hence, $\rho^2_{S_k T_k} = \frac{1}{1+\frac{\sigma^2_{V_k}}{\sigma^2_{S_k}}}$

Mutual Information

- For linear processing with $\mathbf{v}_k$ such that $\tilde{X}_k = \mathbf{v}_k^H \mathbf{X}_k$, we have $\rho^2_{T_k\tilde{X}_k} = \frac{1}{1+\frac{\mathbf{v}_k^H \mathbf{R}_{U_k}\mathbf{v}_k}{|\mathbf{v}_k^H \mathbf{d}_k|^2 \sigma^2_{T_k}}}$ and $\rho^2_{\tilde{X}_k Y_k} = \frac{1}{1+\frac{\sigma^2_{N_k}}{\mathbf{v}_k^H \mathbf{R}_{X_k}\mathbf{v}_k}}$

- Notice that for single-microphone processing with $\mathbf{v}_k = \sqrt{\alpha}$ we thus have
$\rho^2_{T_k\tilde{X}_k} = \frac{1}{1+\frac{\sigma^2_{U_k}}{\sigma^2_{T_k}}}$ and $\rho^2_{\tilde{X}_k Y_k} = \frac{1}{1+\frac{\sigma^2_{N_k}}{\alpha\sigma^2_{X_k}}}$

- With single-microphone processing we can thus only change the correlation coefficient with respect to the near-end noise.

Optimization for Intelligibility 1

$$\mathcal{P}_1: \quad \max_{\{\mathbf{v}_k\} \in \mathbb{C}^M} \quad -\frac{1}{2}\sum_k \log\left(1 - \frac{\rho_{0,k}^2\, \mathbf{v}_k^H \mathbf{d}_k \mathbf{d}_k^H \mathbf{v}_k \sigma_{T_k}^2}{\mathbf{v}_k^H \mathbf{d}_k \mathbf{d}_k^H \mathbf{v}_k \sigma_{T_k}^2 + \mathbf{v}_k^H \underbrace{\mathbf{R}_{U_k}}_{\mathbb{E}\{\mathbf{U}_k \mathbf{U}_k{}^H\}} \mathbf{v}_k + \sigma_{N_k}^2}\right)$$

$$s.t. \quad \sum_k \mathbf{v}_k^H \mathbf{d}_k \mathbf{d}_k^H \mathbf{v}_k \sigma_{T_k}^2 = \sum_k \sigma_{T_k}^2$$

Variable Change: $\mathbf{v}_k = \sqrt{\alpha_k}\mathbf{w}_k, \quad \alpha_k \in \mathbb{R}_+ = |\mathbf{v}_k^H \mathbf{d}_k|^2$

Hence, this implies: $\mathbf{v}_k = \sqrt{\alpha_k}\mathbf{w}$ with $\mathbf{w}^H \mathbf{d} = 1$.

Circuits and Systems
Department of Microelectronics

TU Delft

Optimization for Intelligibility 2

$$\mathcal{P}_2: \quad \max_{\mathbf{w}_k \in \mathbb{C}^M, \alpha_k \in \mathbb{R}_+} \quad I(\alpha_k, \mathbf{w}_k)$$

$$s.t. \quad \begin{array}{l} \mathcal{C}_1 : \sum_k \alpha_k \, \sigma_{T_k}^2 = \sum_k \sigma_{T_k}^2 \\ \mathcal{C}_2 : \mathbf{w}_k^H \mathbf{d}_k = 1, \ \forall k \end{array}$$

$$I(\alpha_k, \mathbf{w}_k) = -\frac{1}{2} \sum_k \log \left(1 - \frac{\rho_{0,k}^2 \alpha_k \sigma_{T_k}^2}{\alpha_k \sigma_{T_k}^2 + \alpha_k \mathbf{w}_k^H \mathbf{R}_{U_k} \mathbf{w}_k + \sigma_{N_k}^2} \right)$$

$$\boxed{\max_{x,y} \; f(x,y) = \max_x \max_y f(x,y)}$$

$$\mathcal{P}_3: \quad \max_{\alpha_k \in \mathbb{R}_+, \mathcal{C}_1} \quad \max_{\mathbf{w}_k \in \mathbb{C}^M, \mathcal{C}_2} \quad I(\alpha_k, \mathbf{w}_k)$$

Optimization for Intelligibility 3

$$\mathcal{P}_3 : \max_{\alpha_k \in \mathbb{R}_+, \mathcal{C}_1} \underbrace{\max_{\mathbf{w}_k \in \mathbb{C}^M, \mathbf{w}_k^H \mathbf{d}_k = 1, \ \forall k}}_{\mathbf{w}_k^* = \frac{\mathbf{R}_{U_k}^{-1} \mathbf{d}_k}{\mathbf{d}_k^H \mathbf{R}_{U_k}^{-1} \mathbf{d}_k}} I(\alpha_k, \mathbf{w}_k)$$

Using $\mathbf{w}_k^*$, the outer maximization is over α_k

$$\mathcal{P}_4 : \begin{array}{ll} \max\limits_{\alpha_k \in \mathbb{R}_+} & -\frac{1}{2} \sum_k \log\left(1 - \frac{\rho_{0,k}^2 \alpha_k \sigma_{T_k}^2}{\alpha_k \sigma_{T_k}^2 + \alpha_k \sigma_{M_k}^2 + \sigma_{N_k}^2}\right) \\ s.t. & \sum_k \alpha_k \sigma_{T_k}^2 = \sum_k \sigma_{T_k}^2 \end{array}$$

$$\boxed{\sigma_{M_k}^2 = \mathbf{w}_k^{*H} \mathbf{R}_{U_k} \mathbf{w}_k^*}$$

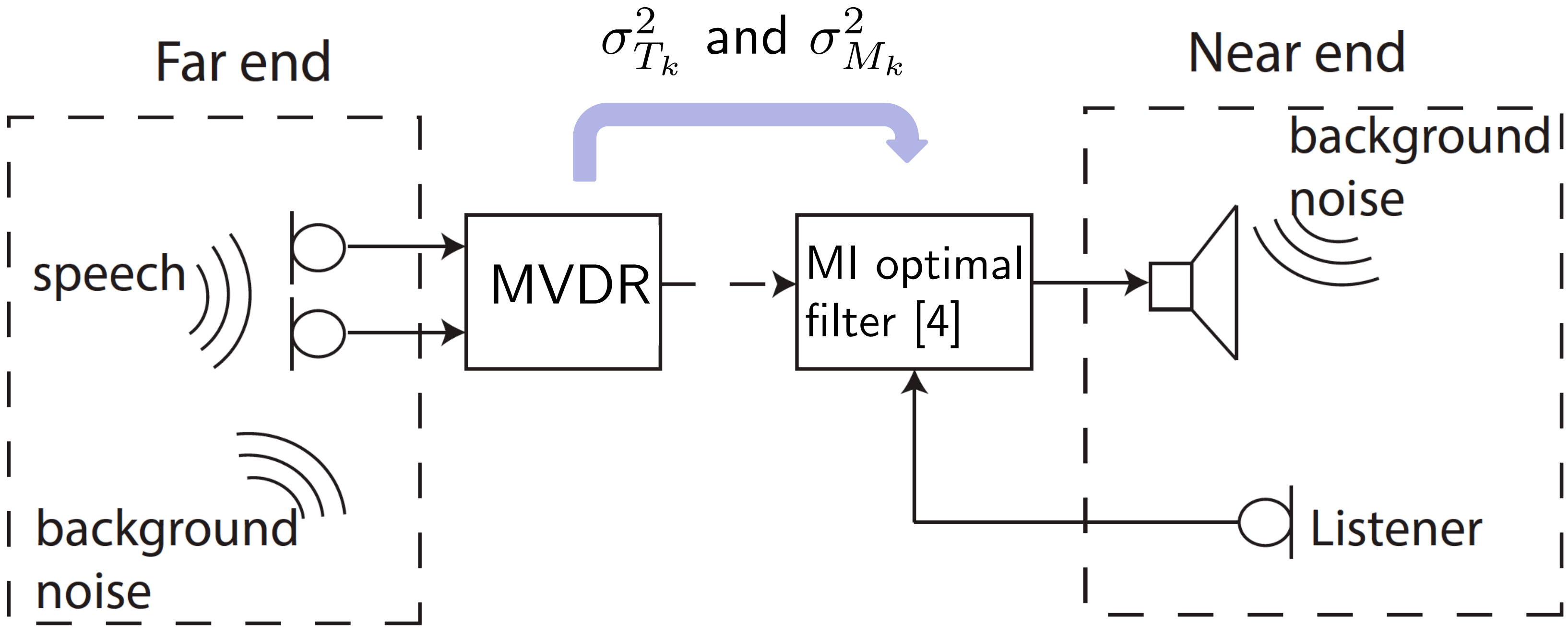
KKT Conditions

1. $\frac{\partial \mathcal{L}(\{\alpha_k\},\lambda,\{\mu_k\})}{\partial \alpha_k} = \frac{1}{2}\frac{\sigma_{T_k}^2+\sigma_{M_k}^2}{\alpha_k(\sigma_{T_k}^2+\sigma_{M_k}^2)+\sigma_{N_k}^2} - \frac{1}{2}\frac{(1-\rho_{0,k}^2)(\sigma_{T_k}^2+\sigma_{M_k}^2)}{\alpha_k(1-\rho_{0,k}^2)(\sigma_{T_k}^2+\sigma_{M_k}^2)+\sigma_{N_k}^2} + \lambda\sigma_{T_k}^2 - \mu_k = 0$
2. $\mu_k\alpha_k = 0, \forall k$ (complementary slackness)
3. $\alpha_k\sigma_{T_k}^2 \geq 0, \quad \mu_k \geq 0, \forall k$ (primal and dual feasibility)
4. $\sum_k \alpha_k\sigma_{T_k}^2 - \sum_k \sigma_{T_k}^2 = 0$ (equality constraint)

$$a_k\alpha_k^2 + b_k\alpha_k + c_k = 0\,, \quad \alpha_k = \frac{-b_k \pm \sqrt{b_k^2 - 4a_kc_k}}{2a_k}$$

$$\begin{aligned}
a_k &= -(\sigma_{T_k}^2 + \sigma_{M_k}^2)((1-\rho_{0,k}^2)\sigma_{T_k}^2 + \sigma_{M_k}^2)\lambda \\
b_k &= -((2-\rho_{0,k}^2)\sigma_{T_k}^2 + 2\sigma_{M_k}^2)\sigma_{N,k}^2\lambda \\
c_k &= \frac{1}{2}\rho_{0,k}^2\sigma_{N_k}^2 - \sigma_{N_k}^4\lambda
\end{aligned}$$

Optimal Filter



The optimal strategy can thus be decomposed into

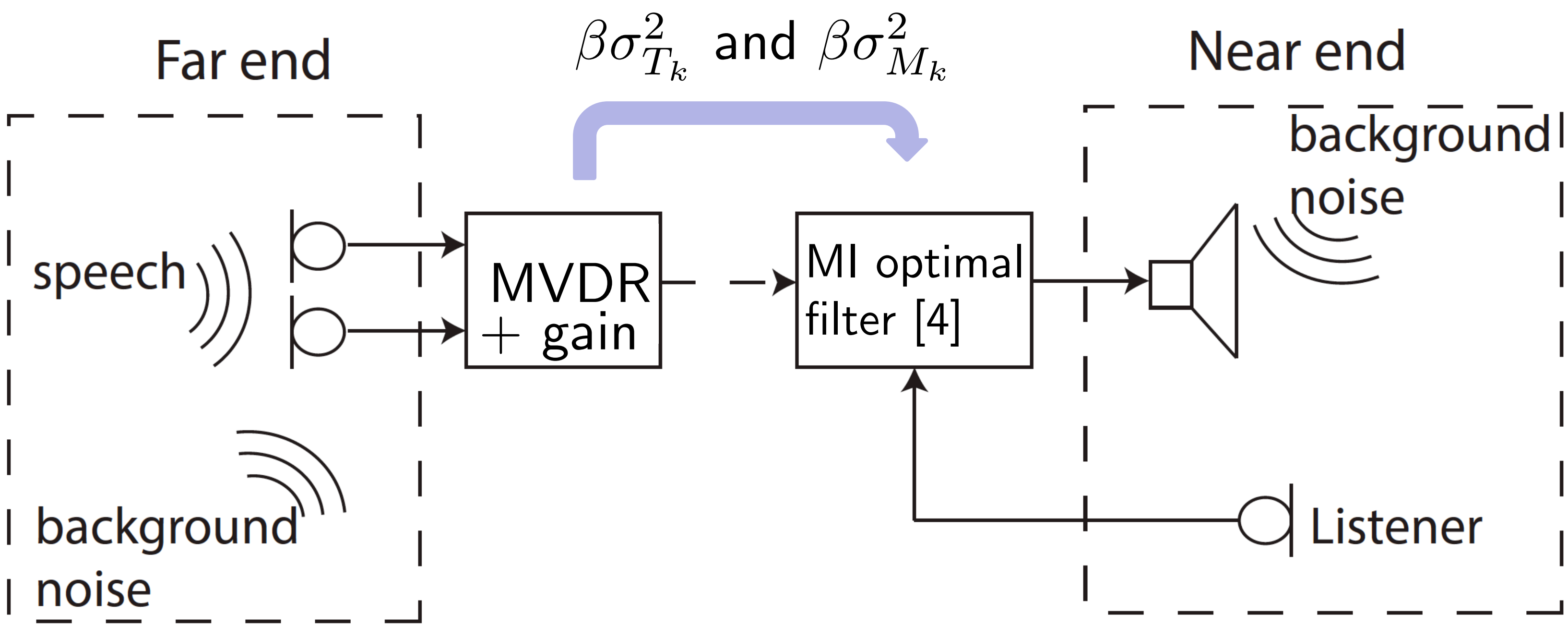
- MVDR
- Single channel MI optimal filter from [4], taking the remaining noise from the far-end into account.

Reference Methods

Circuits and Systems
Department of Microelectronics

TUDelft

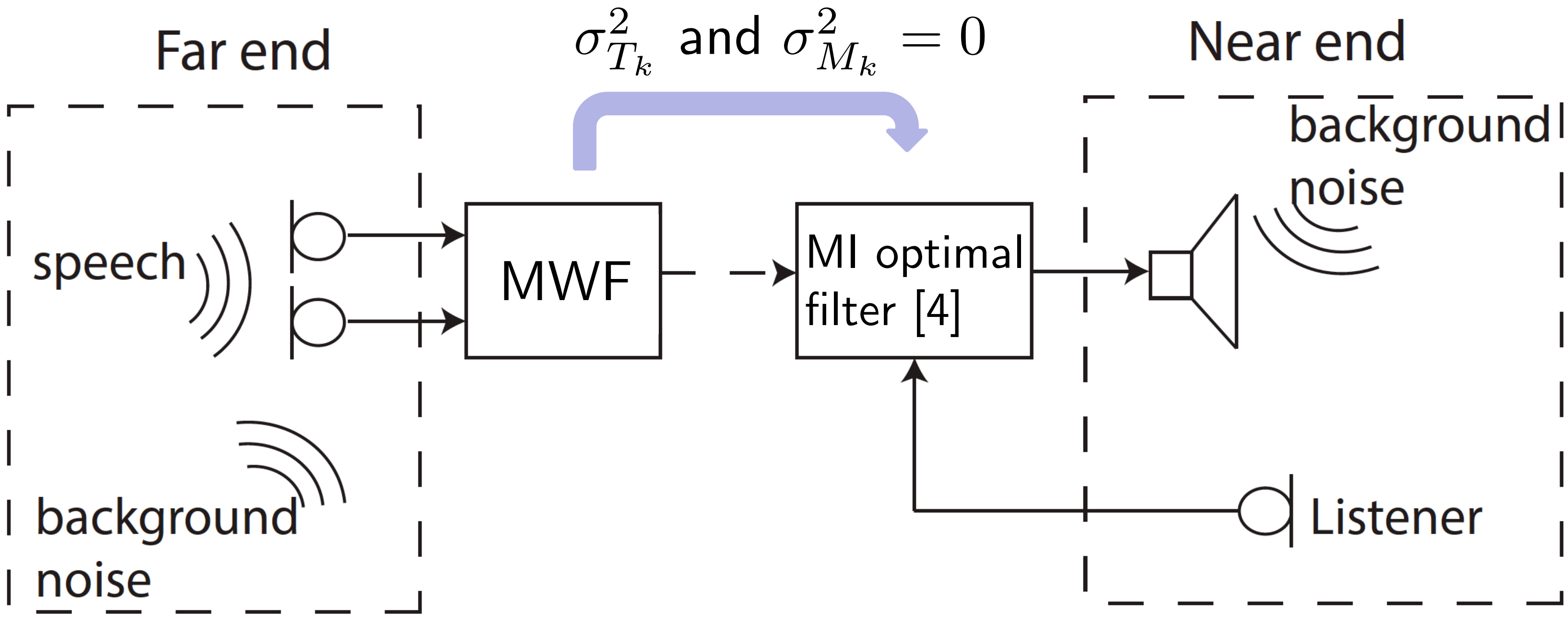
Optimal Filter



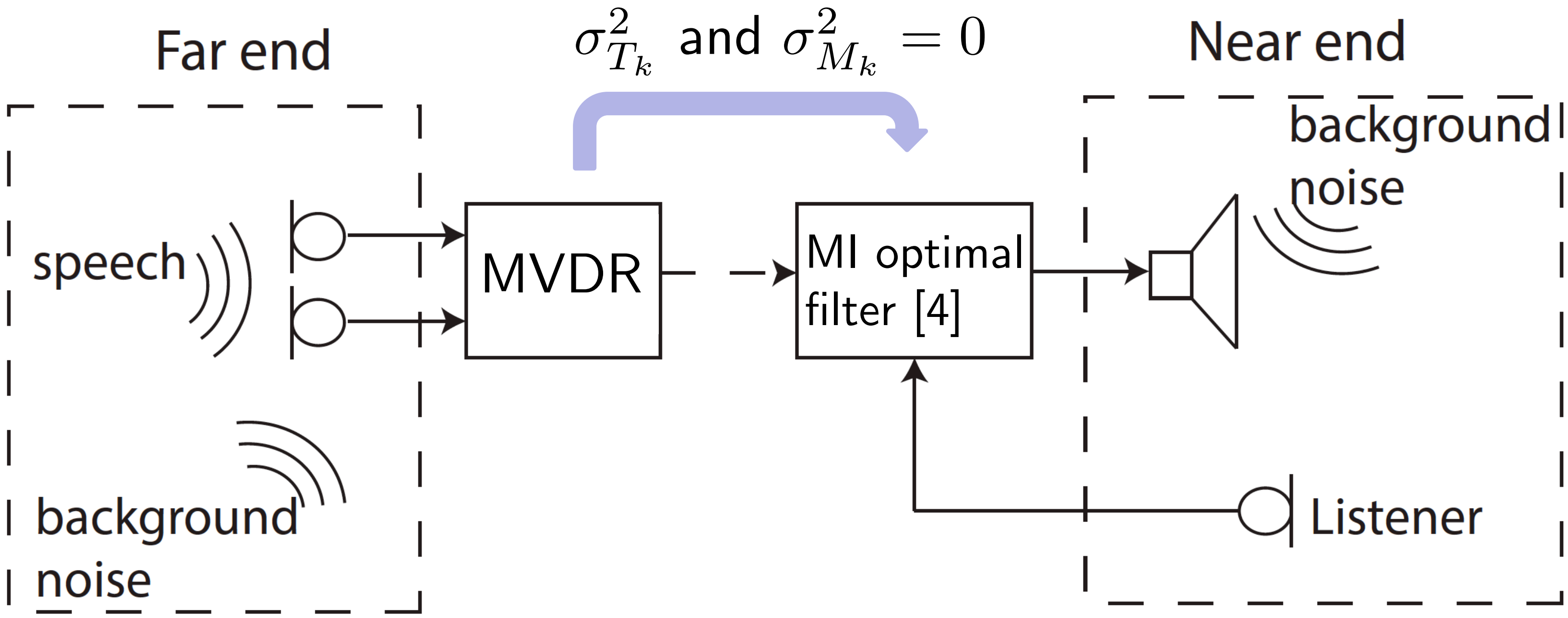
What if far-end processor has applied additional linear processing?

- Far-end Processing: MVDR + linear gain $\sqrt{\beta}$
- If near-end processor is informed, (linear) MI optimal gain: $\frac{\sqrt{\alpha}}{\sqrt{\beta}}$. Hence, the additional processing is completely compensated.

Non-Optimal Reference Methods



Non-Optimal Reference Methods

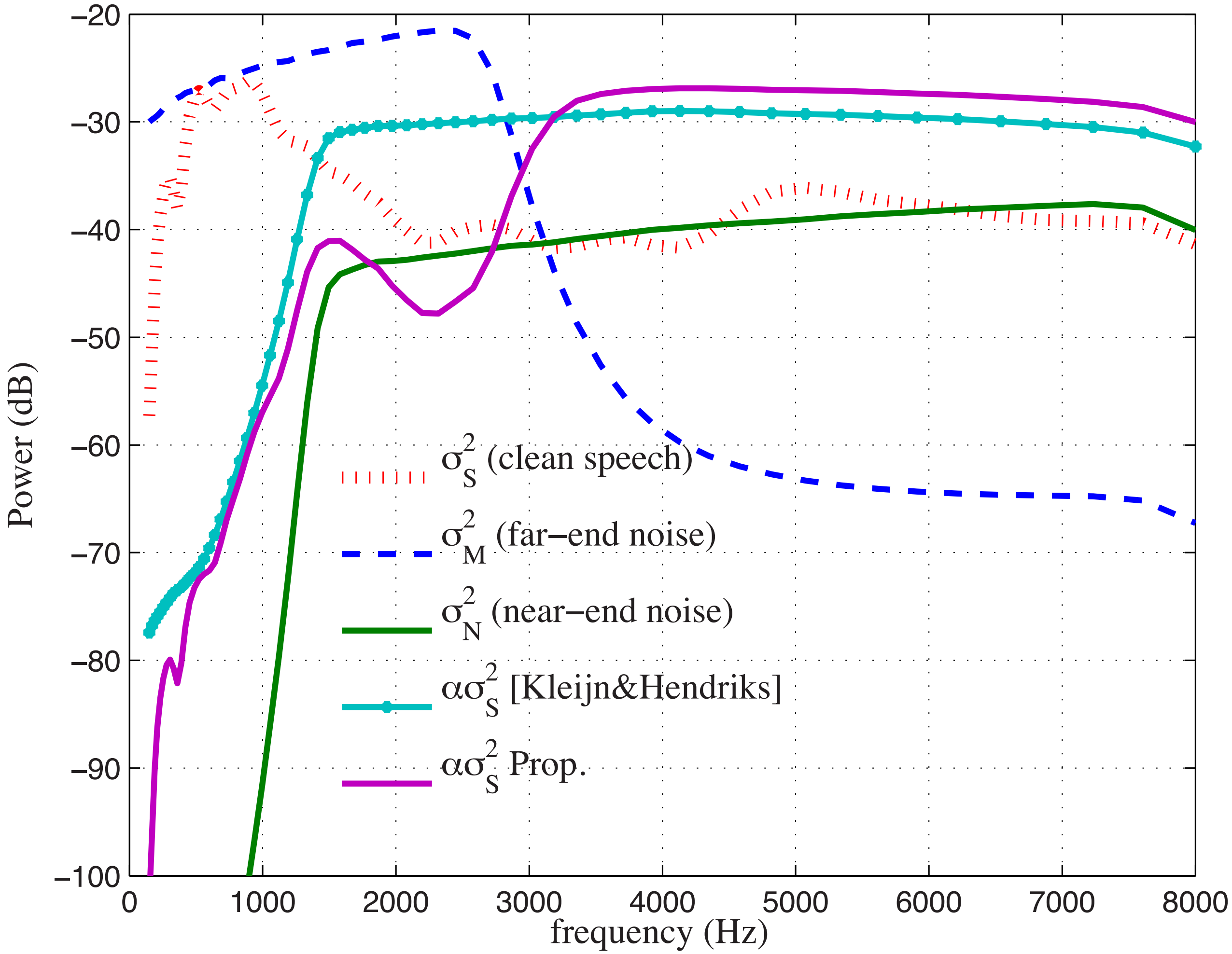


Circuits and Systems
Department of Microelectronics
TU Delft

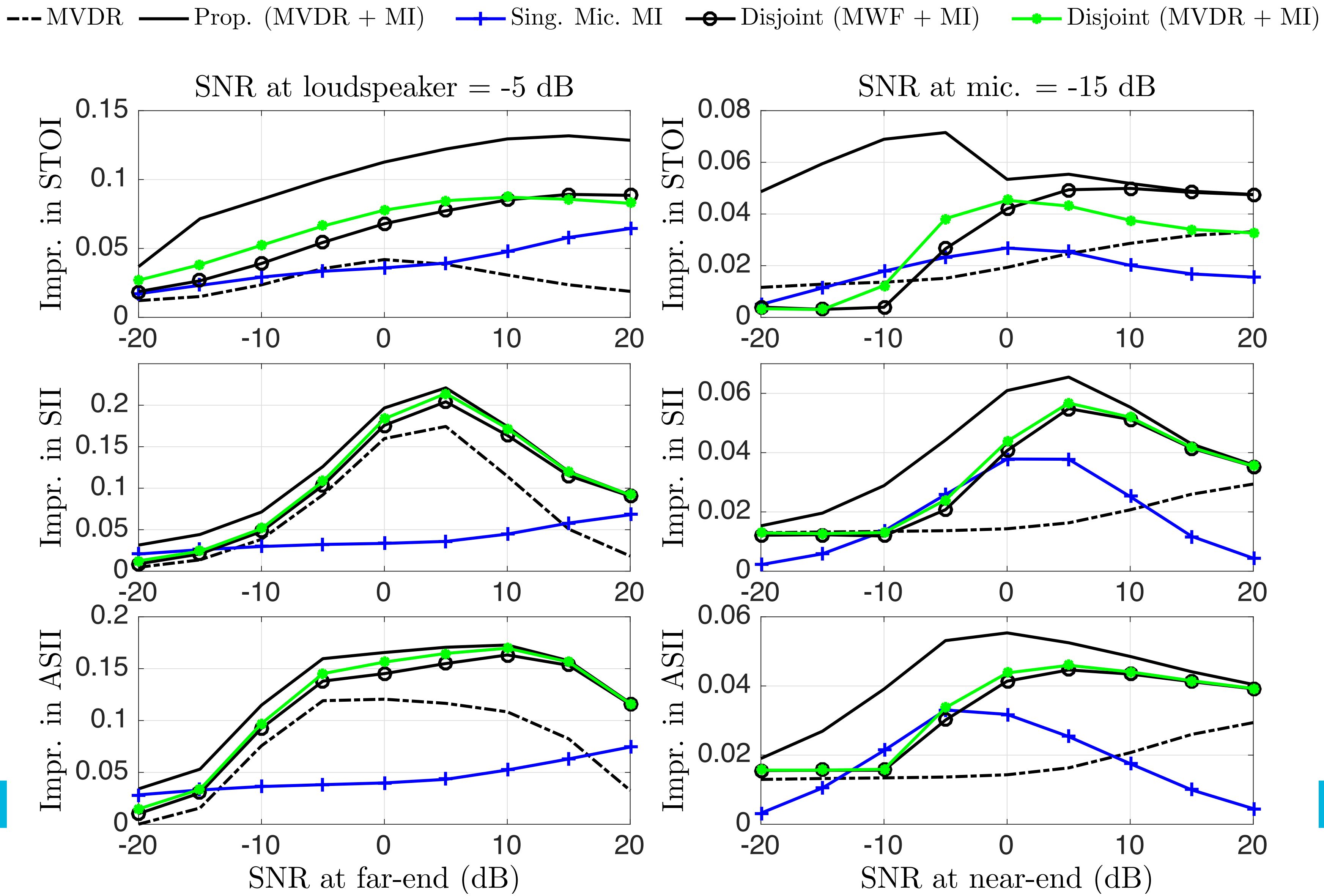
Simulation Setup

- Dual microphone ($m = 2$) with 2 cm spacing, in a $3 \times 4 \times 3$ m room with one target source.
- Far-end noise: Three correlated noise sources and simulated uncorrelated microphone noise at 60 dB.
- 36 seconds of speech sampled at 16 kHz.
- Simulated Room transfer function [Habets]
- Far-end and near-end noise sources with an overlapping region from 1.5 kHz till 3 kHz.
- Short-time DFT with square-root-Hann window and block size of a 32 ms and 50 % overlap ($K = 256$).

Simulation Results



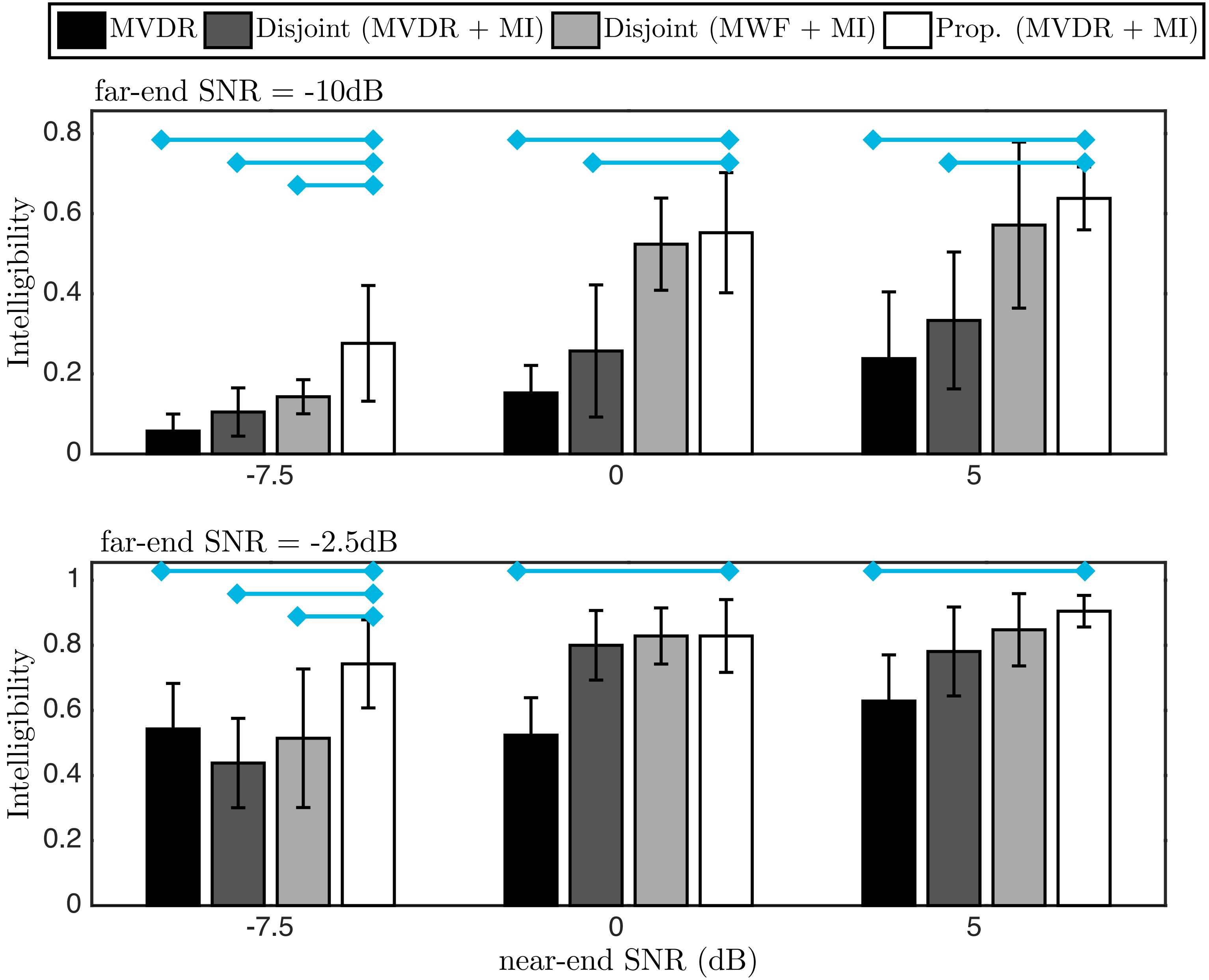
Simulation Results - Instrumental



Simulation Results - Intelligibility Test

- Dutch matrix test (closed) with seven participants
- Far-end noise: -10 dB and -2.5 dB
- Near-end noise: -7.5 dB, 0 dB and 5 dB
- Reference algorithms:
  - MVDR
  - Disjoint (MVDR + MI)
  - Disjoint (MWF + MI)
  - Prop. Jointly optimal (MVDR + MI)
- The values for ρ_0^2 in these experiments are based on the band importance functions from the SII.

Simulation Results - Intelligibility Test



Summary

- Model of speech communication was presented based on speech production uncertainty.
- Although derived from a different viewpoint, the presented model shows strong similarities with classical intelligibility models.
- Conventional independent processing of near-end noise and far-end noise is not optimal.
- The optimal processor of speech can be separated into a far-end and near-end processor.
- Near-end processing must be aware of the processing performed at the far-end.

References

1. E. Nemeth et al. Bird song and anthropogenic noise: vocal constraints may explain why birds sing higher-frequency songs in cities," Proceedings of the Royal Society of London B: Biological Sciences, vol. 280, no. 1754, 2013.

2. J. B. Allen. The Articulation Index is a Shannon channel capacity. Auditory Signal Processing. Springer New York, 2005. 313-319.

3. C. H. Taal et al. On optimal linear filtering of speech for near-end listening enhancement, IEEE Signal Process. Lett. , vol. 20, no. 3, pp. 225 - 228, 2013.

4. W. B. Kleijn and R. C. Hendriks, "A simple model of speech communication and its application to intelligibility enhancement," IEEE Signal Process. Lett., 2014.

5. G. A. Studebaker et al, "A frequency importance function for continuous discourse", Jasa, 1987.

6. W. B. Kleijn, J. B. Crespo, R. C. Hendriks, P. Petkov, B. Sauert and P. Vary. "Optimizing Speech Intelligibility in a Noisy Environment: A unified view", IEEE Signal Processing Magazine, 2015.

7. S. Khademi, R. C. Hendriks and W. B. Kleijn. "Jointly optimal near-end and far-end multi-microphone speech intelligibility enhancement based on mutual information" ICASSP, 2016.