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)
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?
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:**
  \[ \sum_{k \in K} I_k A_k(\xi_k). \]

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

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

\[
A_{k}^{AI}(\xi_k) = \min(\max(10 \log_{10} \xi_k, 0), 30)/30
\]

\[
A_{k}^{SII}(\xi_k) = \max(\min(10 \log_{10} \xi_k, 15), -15)/30 + 1/2
\]
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]).
Classical Measures of Intelligibility

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

$$A_k^{AI}(\xi_k) = \min(\max(10 \log_{10} \xi_k, 0), 30)/30$$

$$A_k^{SII}(\xi_k) = \max(\min(10 \log_{10}(\xi_k), 15), -15)/30 + 1/2$$

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: $M_T \rightarrow A_T \rightarrow A_L \rightarrow M_L$

\[A_T = M_T + V_T\]
\[A_L = A_T + V_E\]
\[M_L = A_L + V_L\]

- $V_T$: 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? ☠️

production noise $V_T$

environmental noise $V_E$

interpretation noise $V_L$
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 $\frac{\sigma_{MT}^2}{\sigma_{VT}^2}$ is scale independent.
- Consequence: correlation coefficient $\rho_{MTAT}$ 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 $\frac{\sigma_{AL}^2}{\sigma_{VL}^2}$ is fixed.

- Consequence: correlation coefficient $\rho_{AL,ML}$ is fixed.

Consequence of fixed production/interpretation SNR: Only little benefit to have a frequency band with channel SNR $\xi_k = \frac{\sigma_{AT,k}^2}{\sigma_{VE,k}^2}$ above the production/interpretation SNR.

Usefulness of a channel saturates near production/interpretation SNR!
Mutual Information Between Talker and Listener

- Consider known 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 $\rho_{0,k} = \rho_{MT AT_k} \rho_{AT_k M_L} \rho_{ML_k}$ and $\xi_k = \frac{\sigma^2_{AT_k}}{\sigma^2_{VE_k}}$.
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

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

Remember the AI and the SII!
Comparing to Classical models

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

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

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

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:

$$\hat{I}(k) = -\frac{R}{2} \log \left( 1 - \rho_{MTA_T}^2(k) \right)$$
The Speech Production Uncertainty

Estimate of band importance:

\[
\hat{I}(k) = -\frac{R}{2} \log \left( 1 - \rho_{MTAT}^2(k) \right)
\]

For comparison, some band importance functions published in Studebaker (1986).
Optimizing for Intelligibility
Scenario

Far end

- speech
- background noise

Pre-process

Near end

- 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: \(S \rightarrow T \rightarrow X \rightarrow \tilde{X} \rightarrow Y \rightarrow Z\).

3. Enhancement is performed by a linear time-invariant operator, \(v_k\).

4. Individual component signals of the vectors \(s_k\) and \(z_k\) 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: \( T_k = S_k + V_k \)
   clean speech                   production noise

2. Multi-mic. Rec.: \( X_k = d_k T_k + U_k \)
                        far-end noise

3. process. signal: \( \tilde{X}_k = v_k^H X_k \)

4. Received signal: \( Y_k = \tilde{X}_k + N_k \)
                           processed                          near-end noise

5. Interpreted signal: \( Z_k = Y_k + W_k \)
                          interpretation noise

With acoustic transfer function \( d_k = [d_{k,1}, ..., d_{k,M}]^T \) and far-end noise \( U_k = [U_{k,1}, ..., U_{k,M}]^T \).
Mutual Information

• In Markov chain the overall correlation coefficient ($\rho$) is the product of all coefficients:

$$\rho_{S_k, Z_k} = \rho_{S_k, T_k} \rho_{T_k, X_k} \rho_{X_k, Y_k} \rho_{Y_k, Z_k}$$

• The mutual information:

$$I(S; Z) = \sum_k -\frac{1}{2} \log(1 - \rho_{S_k, Z_k}^2) = \sum_k -\frac{1}{2} \log(1 - \rho_{0,k}^2 \rho_{S_k, T_k}^2 \rho_{X_k, Y_k}^2)$$

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

• Hence, $\rho_{S_k, T_k}^2 = \frac{1}{\sigma_{S_k}^2} \frac{1}{1 + \frac{\sigma_{V_k}^2}{\sigma_{S_k}^2}}$
Mutual Information

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

- Notice that for single-microphone processing with $\mathbf{v}_k = \sqrt{\alpha}$ we thus have
  \[
  \rho_{T_k \tilde{X}_k}^2 = \frac{1}{1 + \frac{\sigma_{U_k}^2}{\sigma_{T_k}^2}} \quad \text{and} \quad \rho_{\tilde{X}_k Y_k}^2 = \frac{1}{1 + \frac{\sigma_{N_k}^2}{\alpha \sigma_{X_k}^2}}
  \]

- With single-microphone processing we can thus only change the correlation coefficient with respect to the near-end noise.
Optimization for Intelligibility 1

\[ \max \quad -\frac{1}{2} \sum_k \log \left( 1 - \frac{\rho_{0,k}^2 v_k^H d_k d_k^H v_k \sigma_{T_k}^2}{v_k^H d_k d_k^H v_k \sigma_{T_k}^2 + v_k^H R_{U_k} v_k + \sigma_{N_k}^2} \right) \]

\[ P_1 : \quad \{v_k\} \in \mathbb{C}^M \]

s.t. \[ \sum_k v_k^H d_k d_k^H v_k \sigma_{T_k}^2 = \sum_k \sigma_{T_k}^2 \]

Variable Change: \[ v_k = \sqrt{\alpha_k} w_k, \quad \alpha_k \in \mathbb{R}_+ = |v_k^H d_k|^2 \]

Hence, this implies: \[ v_k = \sqrt{\alpha_k} w \text{ with } w^H d = 1. \]
Optimization for Intelligibility 2

\[ P_2 : \quad \max_{w_k \in \mathbb{C}^M, \alpha_k \in \mathbb{R}_+} I(\alpha_k, w_k) \]

s.t.

\[ C_1 : \sum_k \alpha_k \sigma^2_{T_k} = \sum_k \sigma^2_{T_k} \]

\[ C_2 : w_k^H d_k = 1, \forall k \]

\[ I(\alpha_k, w_k) = -\frac{1}{2} \sum_k \log \left( 1 - \frac{\rho_{0,k}^2 \alpha_k \sigma^2_{T_k}}{\alpha_k \sigma^2_{T_k} + \alpha_k w_k^H R_{U_k} w_k + \sigma^2_{N_k}} \right) \]

\[
\max_{x,y} f(x,y) = \max_{x} \max_{y} f(x,y)
\]

\[ P_3 : \quad \max_{\alpha_k \in \mathbb{R}_+, C_1} \max_{w_k \in \mathbb{C}^M, C_2} I(\alpha_k, w_k) \]
Optimization for Intelligibility 3

\[
\begin{align*}
\max_{\alpha_k \in \mathbb{R}_+, C_1} \max_{\mathbf{w}_k \in \mathbb{C}^M, \mathbf{w}_k^H \mathbf{d}_k = 1, \forall k} I(\alpha_k, \mathbf{w}_k) \\
\mathcal{P}_3: \\
\mathbf{w}_k^* = \frac{\mathbf{R}^{-1}_U \mathbf{d}_k}{\mathbf{d}_k^H \mathbf{R}^{-1}_U \mathbf{d}_k}
\end{align*}
\]

Using \( \mathbf{w}_k^* \), the outer maximization is over \( \alpha_k \)

\[
\max_{\alpha_k \in \mathbb{R}_+} -\frac{1}{2} \sum_k \log(1 - \frac{\rho_{0,k}^2 \alpha_k \sigma_{T_k}^2}{\alpha_k \sigma_{T_k}^2 + \alpha_k \sigma_{M_k}^2 + \sigma_{N_k}^2})
\]

\( \mathcal{P}_4: \)

\[
\begin{align*}
\max_{\alpha_k \in \mathbb{R}_+} & -\frac{1}{2} \sum_k \log(1 - \frac{\rho_{0,k}^2 \alpha_k \sigma_{T_k}^2}{\alpha_k \sigma_{T_k}^2 + \alpha_k \sigma_{M_k}^2 + \sigma_{N_k}^2}) \\
\text{s.t.} & \sum_k \alpha_k \sigma_{T_k}^2 = \sum_k \sigma_{T_k}^2 \\
& \sigma_{M_k}^2 = \mathbf{w}_k^*^H \mathbf{R}_U \mathbf{w}_k^*
\end{align*}
\]
KKT Conditions

1. \[
\frac{\partial \mathcal{L}(\{\alpha_k\}, \lambda, \{\mu_k\})}{\partial \alpha_k} = \frac{1}{2} \left( \sigma^2_{T_k} + \sigma^2_{M_k} \right) - \frac{1}{2} \left( \sigma^2_{T_k} + \sigma^2_{M_k} \right) + \lambda \sigma^2_{T_k} - \mu_k = 0
\]

2. \( \mu_k \alpha_k = 0, \forall k \) (complementary slackness)

3. \( \alpha_k \sigma^2_{T_k} \geq 0, \quad \mu_k \geq 0, \forall k \) (primal and dual feasibility)

4. \( \sum_k \alpha_k \sigma^2_{T_k} - \sum_k \sigma^2_{T_k} = 0 \) (equality constraint)

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

\[
a_k = -(\sigma^2_{T_k} + \sigma^2_{M_k})((1 - \rho^2_{0,k})\sigma^2_{T_k} + \sigma^2_{M_k}) \lambda
\]

\[
b_k = -((2 - \rho^2_{0,k})\sigma^2_{T_k} + 2\sigma^2_{M_k})\sigma^2_{N,k} \lambda
\]

\[
c_k = \frac{1}{2} \rho^2_{0,k} \sigma^2_{N,k} - \sigma^4_{N,k} \lambda
\]
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
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

\[ \sigma^2_{T_k} \text{ and } \sigma^2_{M_k} = 0 \]

Far end

speech

background noise

Near end

background noise

Listener

MWF

MI optimal filter [4]
Non-Optimal Reference Methods

$\sigma^2_{T_k}$ and $\sigma^2_{M_k} = 0$
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

![Graph showing simulation results]

- **$\sigma_S^2$ (clean speech)**
- **$\sigma_M^2$ (far-end noise)**
- **$\sigma_N^2$ (near-end noise)**
- **$\alpha \sigma_S^2$ [Kleijn & Hendriks]**
- **$\alpha \sigma_S^2$ Prop.**

**Axes:**
- Y-axis: Power (dB)
- X-axis: Frequency (Hz)
Simulation Results - Instrumental

- MVDR
- Prop. (MVDR + MI)
- Sing. Mic. MI
- Disjoint (MWF + MI)
- Disjoint (MVDR + MI)

**SNR at loudspeaker = -5 dB**

**SNR at mic. = -15 dB**

- **Impr. in STOI**
- **Impr. in SII**
- **Impr. in ASII**

**SNR at far-end (dB)**

**SNR at near-end (dB)**
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 $\rho^2_0$ in these experiments are based on the band importance functions from the SII.
Simulation Results - Intelligibility Test

far-end SNR = -10dB

far-end SNR = -2.5dB
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


