JOINTLY OPTIMAL NEAR-END AND FAR-END MULTI-MICROPHONE SPEECH INTELLIGIBILITY ENHANCEMENT BASED ON MUTUAL INFORMATION

Seyran Khademi, Richard C. Hendriks and W. Bastiaan Kleijn

1. SYSTEM MODEL

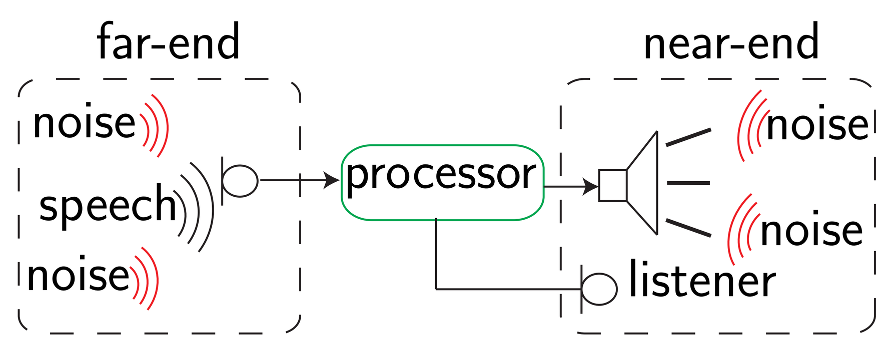

Fig.1. Speech communication system with a pre-processor for speech enhancement [1].

1. Produced : $T_{k,i} = \underbrace{S_{k,i}}_{\text{clean speech}} + \underbrace{Q_{k,i}}_{\text{production noise}}$
2. Processed : $\tilde{X}_{k,i} = \mathbf{v}_k^H \mathbf{d}_k T_k + \mathbf{v}_k{}^H \underbrace{U_{k,i}}_{\text{far-end noise}}$
3. Received : $Y_{k,i} = \underbrace{\tilde{X}_{k,i}}_{\text{processed}} + \underbrace{N_{k,i}}_{\text{near-end noise}}$
4. Interpreted : $Z_{k,i} = Y_{k,i} + \underbrace{W_{k,i}}_{\text{interpretation noise}}$

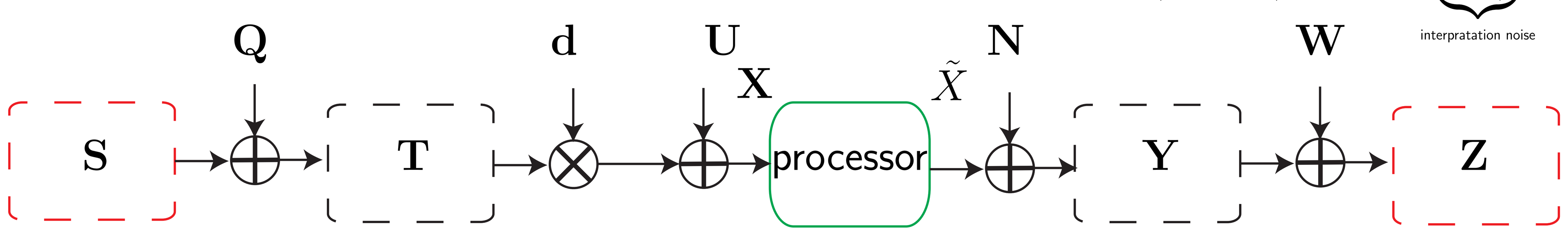

Fig.2. Proposed multi-mic. system model including the room transfer function, production and interpretation noise.

Speech vector in time domain ($\mathbf{s}$) is transformed into frequency-time matrix ($\mathbf{S}$) using short discrete Fourier transform (SDFT), where the acoustic signal at frequency-time point (k,i) is $S_{k,i}$, $k = 1,\dots,K$ and $i = 1,\dots,N$.

- $d_{k,j}$ denotes the acoustic transfer function from source to microphone j and we write $\mathbf{d}_k = [d_{k,1}, \dots, d_{k,M}]^T$.
- Far-end noise recorded by microphones: $\mathbf{u}_k = [u_{k,1}, \dots, u_{k,M}]^T$.

ASSUMPTIONS

1. Signal model follows the Markov chain model: $S \to T \to X \to \tilde{X} \to Y \to Z$.
2. The enhancement is performed by a linear time-invariant operator.
3. All processes are jointly Gaussian, stationary, and memoryless so we omit the time-frame index i for notational convenience so $\rho_{S_{k,i} Z_{k,i}} = \rho_{S_k Z_k}$.
4. Individual component signals of the vectors $\mathbf{S}_k$ and $\mathbf{Z}_k$ are independent so we can then write

$$I(\mathbf{S}_i;\mathbf{Z}_i) = I(\mathbf{S};\mathbf{Z}) = \sum_k I(S_k;Z_k) = -\frac{1}{2}\log(1-\rho_{0_k}^2\,\rho_{T_k\tilde{X}_k}^2\,\rho_{\tilde{X}_kY_k}^2);\quad \rho_{0_k}^2 = \rho_{S_kT_k}\,\rho_{Y_kZ_k}.$$

2. PROBLEM STATEMENT

$$\mathcal{P}_1: \sup_{\{\mathbf{v}_k\}\in\mathbb{C}^M} I(\mathbf{S};\mathbf{Z}) = -\tfrac{1}{2}\sum_k \log\left(1-\frac{\rho_{0_k}^2\,\mathbf{v}_k^H\mathbf{d}_k\mathbf{d}_k^H\mathbf{v}_k\sigma_{T_k}^2}{\mathbf{v}_k^H\mathbf{d}_k\mathbf{d}_k^H\mathbf{v}_k\sigma_{T_k}^2+\mathbf{v}_k^H\underbrace{\mathbf{R}_{U_k}}_{\mathbb{E}\{\mathbf{u}_k\mathbf{u}_k{}^H\}}\mathbf{v}_k+\sigma_{N_k}^2}\right)$$
$$\text{subject to } \sum_k \mathbf{v}_k^H\mathbf{d}_k\mathbf{d}_k^H\mathbf{v}_k\sigma_{T_k}^2 = \sum_k \sigma_{T_k}^2$$

$$\alpha_k \in \mathbb{R}_+ = \mathbf{v}_k^H\mathbf{d}_k\mathbf{d}_k^H\mathbf{v}_k$$
$$\mathbf{v}_k = \sqrt{\alpha_k}\mathbf{w}_k$$

$$\mathcal{P}_2: \sup_{\mathbf{w}_k\in\mathbb{C}^M,\,\alpha_k\in\mathbb{R}_+} I(\mathbf{S};\mathbf{Z}) = -\tfrac{1}{2}\sum_k \log\left(1-\frac{\rho_{0,k}^2\alpha_k\sigma_{T_k}^2}{\alpha_k\sigma_{T_k}^2+\alpha_k\mathbf{w}_k^H\mathbf{R}_{U_k}\mathbf{w}_k+\sigma_{N_k}^2}\right)$$
$$\text{subject to } \mathcal{C}_1: \sum_k \alpha_k\,\sigma_{T_k}^2 = \sum_k \sigma_{T_k}^2$$
$$\mathcal{C}_2: \mathbf{w}_k^H\mathbf{d}_k = 1,\ \forall k$$

3. SOLUTION

$$\boxed{\sup_{x,y} f(x,y) = \sup_x \sup_y f(x,y)}$$

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

$$\mathcal{P}_4: \sup_{\alpha_k\in\mathbb{R}_+} -\tfrac{1}{2}\sum_k \log\left(1-\frac{\rho_{0,k}^2\alpha_k\sigma_{T_k}^2}{\alpha_k\sigma_{T_k}^2+\alpha_k\mathbf{w}_k^{*H}\mathbf{R}_{U_k}\mathbf{w}_k^*+\sigma_{N_k}^2}\right)$$
$$\text{subject to } \sum_k \alpha_k\,\sigma_{T_k}^2 = \sum_k \sigma_{T_k}^2$$

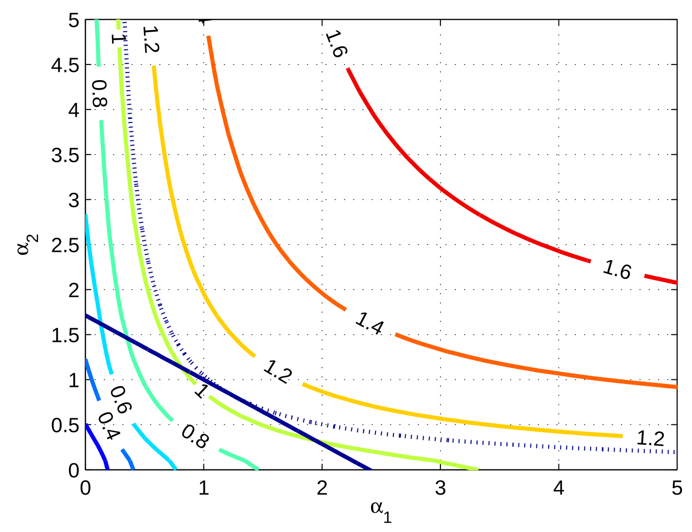

Fig.3. Contour representation of the $I(S,Z)$ for two frequency bands together with the average power constraint line.

4. EXPERIMENTAL RESULTS

- Dual microphone ($m = 2$) with 2 cm spacing, in a $3 \times 4 \times 3$ m room (Room transfer function generated using Habets room impulse response generator).
- Three correlated noise sources and one simulated uncorrelated microphone noise at 60 dB and one target source.
- 36 seconds of speech sampled at 16 kHz (SDFT with Hann window and block size of a 32 ms and 50 % overlap (k=256)).

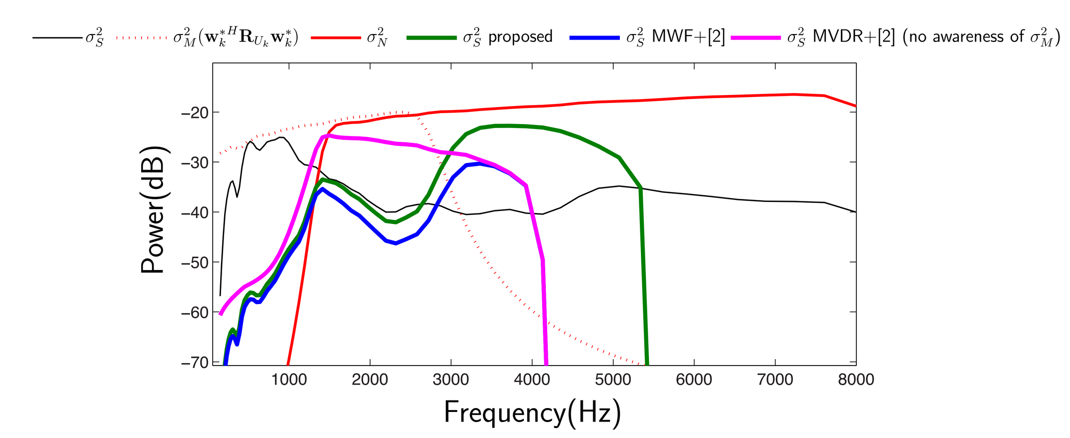

Fig.4. Average spectra for -11.1 dB SNR at the far-end reference microphone and -10 dB SNR at the near-end.

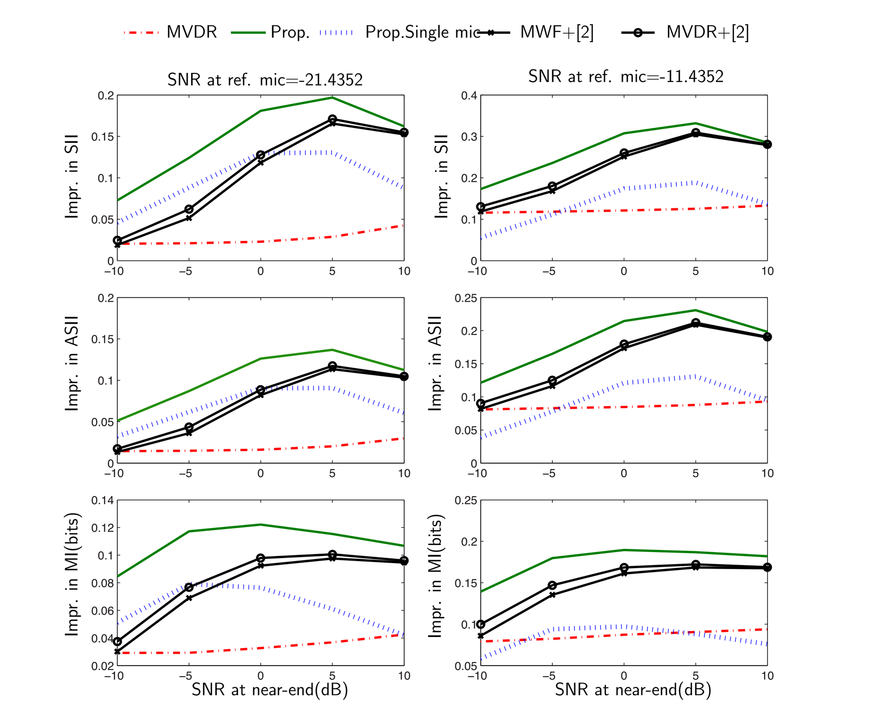

Fig.5. Predicted intelligibility in terms of MI, ASII and SII for different SNRs.

REMARKS

1. Mutual information is a general measure and a flexible model for speech enhancement.
2. Conventional independent processing of the noise at the near-end and the far-end is not optimal.
3. Processing of speech for intelligibility enhancement can be decomposed into far-end (MVDR) and near-end (post-filter) processing.
4. Near-end processing must be aware of the noise remaining from the processing performed at the far-end.
5. Considering the production and interpretation noise makes the intelligibility model more realistic and complete.

[1] C. H. Taal, J. Jensen, and A. Leijon, On optimal linear filtering of speech for near-end listening enhancement, IEEE Signal Process. Lett., vol. 20, no. 3, 2013.

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

TU Delft
Delft University of Technology