Aalborg Universitet

AALBORG UNIVERSITY
DENMARK

Joint Far- and Near-End Speech Intelligibility Enhancement Based on the Approximated Speech Intelligibility Index

Fuglsig, Andreas Jonas; Østergaard, Jan; Jensen, Jesper; Søndergaard Bertelsen, Lars; Mariager, Peter Bank; Tan, Zheng-Hua

Published in:
ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)

DOI (link to publication from Publisher):
10.1109/ICASSP43922.2022.9746170

Publication date:
2022

Document Version
Accepted author manuscript, peer reviewed version

Link to publication from Aalborg University

Citation for published version (APA):
Fuglsig, A. J., Østergaard, J., Jensen, J., Søndergaard Bertelsen, L., Mariager, P. B., & Tan, Z-H. (2022). Joint Far- and Near-End Speech Intelligibility Enhancement Based on the Approximated Speech Intelligibility Index. In *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 7752-7756). Article 9746170 IEEE. https://doi.org/10.1109/ICASSP43922.2022.9746170

General rights
Copyright and moral rights for the publications made accessible in the public portal are retained by the authors and/or other copyright owners and it is a condition of accessing publications that users recognise and abide by the legal requirements associated with these rights.

- Users may download and print one copy of any publication from the public portal for the purpose of private study or research.
- You may not further distribute the material or use it for any profit-making activity or commercial gain
- You may freely distribute the URL identifying the publication in the public portal -

Take down policy
If you believe that this document breaches copyright please contact us at vbn@aub.aau.dk providing details, and we will remove access to the work immediately and investigate your claim.

JOINT FAR- AND NEAR-END SPEECH INTELLIGIBILITY ENHANCEMENT BASED ON THE APPROXIMATED SPEECH INTELLIGIBILITY INDEX

Andreas Jonas Fuglsig[⋆†], *Jan Østergaard*[†], *Jesper Jensen*[†], *Lars Søndergaard Bertelsen*[⋆], *Peter Mariager*[⋆], *Zheng-Hua Tan*[†]

[⋆] RTX A/S, Nørresundby, Denmark
[†]Aalborg University, Aalborg, Denmark

ABSTRACT

This paper considers speech enhancement of signals picked up in one noisy environment which must be presented to a listener in another noisy environment. Recently, it has been shown that an optimal solution to this problem requires the consideration of the noise sources in both environments jointly. However, the existing optimal mutual information based method requires a complicated system model that includes natural speech variations, and relies on approximations and assumptions of the underlying signal distributions. In this paper, we propose to use a simpler signal model and optimize speech intelligibility based on the Approximated Speech Intelligibility Index (ASII). We derive a closed-form solution to the joint far- and near-end speech enhancement problem that is independent of the marginal distribution of signal coefficients, and that achieves similar performance to existing work. In addition, we do not need to model or optimize for natural speech variations.

Index Terms— Multi-microphone, speech intelligibility enhancement, ASII, beamformer

1. INTRODUCTION

Speech communication systems, such as mobile telephony, hearing aids and intercom systems, are required to work in numerous environments. As a consequence, the user environment is often noisy which can lead to intelligibility problems.

For speech communication systems we may consider two different environments, cf. Fig. 1; the far-end environment (at the target talker) and the near-end environment (at the listener). Both the far- and near-end environment are often noisy, which leads to degradations in both speech quality and Speech Intelligibility (SI) for the listener. To remedy these effects, speech enhancement techniques may be applied at both the far- and near-end. Depending on the availability, Far-end Speech Enhancement (FSE) algorithms may utilize either a single or multiple microphones [1, 2, 3, 4]. In contrast to the far-end scenario, in Near-end Listening Enhancement (NLE)[5, 6] the interfering noise is mixed with the target speech after processing. Therefore, the noise cannot be reduced by the usual post-processing techniques of the former scenario. Instead, before playback in the noisy environment, NLE increases SI by adaptively pre-processing the FSE signal received from the far-end, while exploiting knowledge about the near-end noise.

This work is partly supported by Innovation Fund Denmark Case no. 9065-00204B.

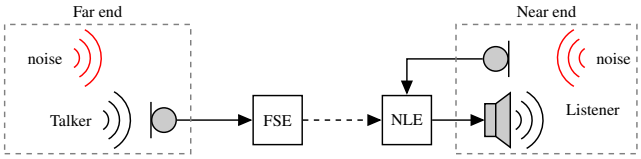

Fig. 1: Classic speech communication system with Far-end Speech Enhancement (FSE) and Near-end Listening Enhancement (NLE).

Most work on NLE assumes the signal received from the far-end is noise-free [6]. However, in many communication scenarios both the target talker and listener may be in noisy environments. Even so, until recently the processing to mitigate the effects of disturbances in the far- and near-end environments have been considered separately. However, recently, in [7, 8, 9] it was shown that optimization of SI under joint consideration of the far- and near-end noise is superior to disjoint processing. The work of [7] considers jointly controlling a single-channel noise reduction filter along with a post-filter gain for NLE designed to increase SI. In [9] a new training strategy is proposed for deep learning based single-channel enhancement given that speech has already been processed at the far-end. For joint multi-microphone FSE and NLE [8, 10] proposes to optimize the Mutual Information (MI) [11] between the clean speech and the signal received by the listener. The results of [8] are the first to show both theoretically and experimentally that joint processing, using knowledge of processing and conditions at both ends, is superior to the classic disjoint processing.

MI as a SI optimization objective provides a target that unifies heuristic views on SI and mathematically founded SI measures [8, 12]. However, solving it in closed form requires simplifying assumptions. The resulting optimization objective is an SNR-type of measure that is approximately equal to the Approximated Speech Intelligibility Index (ASII) [13, 8]. Furthermore, the method of [8] depends on the choice of the correlation of the so-called production and interpretation noise with the clean speech. With the choice made in [8], the objective function of [8] reduces fully to the ASII, whereas for more recent choices in [14] it does not.

In this paper, we illustrate how, by using a simpler well established signal model and optimizing for the ASII directly, we can derive a closed-form solution to the joint far- and near-end speech enhancement problem that is independent of the marginal distribution of signal coefficients, and without the need for introducing additional parameters in terms of a production and interpretation noise model. The proposed approach can also be seen as an extension of the ASII optimization problem [13] to joint far- and near-end optimization. Furthermore, we analyze model choices and assumptions

©2021 IEEE. Personal use of this material is permitted. Permission from IEEE must be obtained for all other uses, in any current or future media, including reprinting/republishing this material for advertising or promotional purposes, creating new collective works, for resale or redistribution to servers or lists, or reuse of any copyrighted component of this work in other works

of [8], and how these relate the approximated MI of [8] to the ASII, thus motivating our model choices. Finally, we experimentally compare the performance of [8] using the production noise model that was later derived by some of the same authors in [14, 15, 16] to our proposed ASII based optimization. We see, that our proposed method achieves similar or slightly better intelligibility in terms of ESTOI [17] when the near-end SNR is low and the far-end SNR is intermediate or high.

In summary, the contributions of this paper are: (i) A closed-form solution to ASII based joint far- and near-end speech enhancement, (ii) which is optimal independently of underlying marginal signal distributions, (iii) which does not introduce additional free parameters, e.g., in terms of production and interpretation noise, and (iv) which performs as good or slightly better than existing schemes.

2. EXISTING WORK BASED ON MUTUAL INFORMATION

MI-based methods for joint far- and near-end SI enhancement [8, 10] improve SI by maximizing the MI, $I(S;Z)$, between the clean speech, S, and the signal received by the listener, Z.

2.1. Existing model assumptions

In [12], production and interpretation noise terms are introduced to model natural variations in speakers and listeners, respectively, and adopted into the signal model of [8]. The production noise, Q, is due to the convolution of the time-domain clean speech and the vocal tract, hence, theoretically, it should be a *multiplicative* noise in the frequency domain. However, in [8] in order to simplify mathematical expressions,

1. Multiplicative production noise is modelled as *additive*.

Thus, the single-microphone signal model of [8] in the absence of processing is, in the complex short-time DFT (STFT) domain,

$$Z_{k,i} = d_{k,i}S_{k,i} + d_{k,i}Q_{k,i} + U_{k,i} + N_{k,i} + W_{k,i}, \qquad (1)$$

where k is the frequency-bin index and i the time-frame, W is the interpretation noise, U is the far-end environmental noise, N is the near-end environmental noise, and $d_{k,i}$ are the time-frequency coefficients of the room transfer function from target talker to the microphone. The work on MI [8] relies on several common signal model assumptions in the speech processing literature, e.g., that speech and noise STFT coefficients are statistically independent. However, to derive a speech enhancement procedure based on MI, [8] introduces additional assumptions and approximations;

2. The production noise, Q, and interpretation noise, W, are independent of the clean speech level and may be represented by a fixed gain (correlation), $\rho_{0,k}$, at each frequency band.
3. Critical band powers are assumed to be zero-mean independent Gaussian random variables.

In [8], the third assumption is needed since critical band powers are in-fact Chi-squared distributed and the MI between Chi-squared random variables is not expressible in closed form. However, we note that critical band powers are positive by definition, hence a zero-mean model is not necessarily appropriate.

2.2. Approximated Mutual Information vs ASII

The resulting approximated MI expression in [8] is

$$I(S;Z) \approx -\sum_j \frac{1}{2}\log\left(1-\rho_{0,j}^2\frac{\xi_j}{\xi_j+1}\right), \qquad (2)$$

where ξ_j is the SNR in critical band j. By [8, sec. VIII.A] we may take a first order Taylor approximation of the MI in $\rho_{0,j}^2$ around zero, such that we can approximate the cost function as,

$$I(S;Z) \approx \sum_j I_j \frac{\xi_j}{\xi_j+1}, \qquad (3)$$

for $\rho_{0,j}^2 \to 0^+$, where $I_j \triangleq -\frac{1}{2}\log(1-\rho_{0,j}^2)$. In the absence of a production noise model at the time, the values for $\rho_{0,j}$ where derived in [8] based on the band importance functions, γ_j, of the SII [18] such that $\rho_{0,j}^2 = 1-2^{-2\gamma_j}$. Inserting this, the cost function is

$$I(S;Z) \approx \sum_j -\frac{1}{2}\log(1-(1-2^{-2\gamma_j}))\frac{\xi_j}{\xi_j+1}. \qquad (4)$$

We recognize, (4) resembles the ASII introduced in [13] as a cost function for SI enhancement. The ASII is defined as,

$$ASII \triangleq \sum_j \gamma_j f(\xi_j), \quad f(\xi_j) \triangleq \frac{\xi_j}{\xi_j+1}, \qquad (5)$$

where the weights γ_j are the critical-band importance functions as defined in [18], and $f(\xi_j)$ is the audibility function per critical band. We notice from [18, Table 1] that band importance functions, γ_j, are in the interval of $[0.01, 0.06]$, resulting in $\rho_{0,j}^2 \in [0.0138, 0.0798]$. As shown in [10, Fig. 1] the approximation (3) holds for $\rho_{0,j}^2 \leq 0.4$. Hence, we can conclude the choice of $\rho_{0,j}^2 = 1-2^{-2\gamma_j}$ to be sufficiently close to zero for equality to hold in (4). Thus, the MI problem in [8] is equal to the ASII problem, when the parameter modelling production- and interpretation noise, $\rho_{0,j}$, is chosen according to the band importance functions of the SII.

3. SIGNAL MODEL

In this section we introduce the proposed signal model, cf. Fig. 2. The single-microphone signal model follows,

$$X_{k,i} = d_{k,i}S_{k,i} + U_{k,i},\ Y_{k,i} = vX_{k,i},\ Z_{k,i} = Y_{k,i} + N_{k,i}, \qquad (6)$$

where $X_{k,i}$ is the recorded signal in STFT domain, i.e., the clean speech, $S_{k,i}$, recorded by the microphone contaminated by the far-end noise, $U_{k,i}$. To increase SI of the received message, the noisy microphone signal, $X_{k,i}$, is linearly processed prior to playout, producing the modified signal $Y_{k,i}$. The signal received by the listener, $Z_{k,i}$, is finally contaminated by the noise in the near-end environment, $N_{k,i}$. The speech and noise processes, S, U, and N, are assumed to be stationary sequences of complex random vectors consisting of the STFT coefficients. Both the far-end noise, U, and the near-end noise, N are assumed to be independent of each other and of the target speech, S. These assumptions are similar to [8]. However, compared to [8] we do not need to model a multiplicative production noise as additive, or to introduce an additional interpretation noise. Further, we do not need assumptions on the particular marginal distributions of the signals.

3.1. Multi-Microphone Signal Model

Let us denote the acoustic transfer function from source to microphone m by $d_{k,i,m}$ with vector notation, $\boldsymbol{d}_{k,i} = [d_{k,i,1}, \ldots, d_{k,i,m}]^T$, and letting vector $\boldsymbol{U}_{k,i}$ denote far-end noise recorded by the microphones. Then the noisy microphone signals are given by

$$\boldsymbol{X}_{k,i} = \boldsymbol{d}_{k,i}S_{k,i} + \boldsymbol{U}_{k,i}. \qquad (7)$$

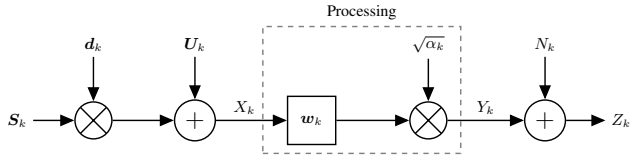

Fig. 2: Our signal model of optimal joint SI enhancement.

Denoting the linear multi-microphone processor by $\boldsymbol{v}_{k,i}$, the processed microphone signal is,

$$Y_{k,i} = \boldsymbol{v}_{k,i}^H \boldsymbol{d}_{k,i} S_{k,i} + \boldsymbol{v}_{k,i}^H \boldsymbol{U}_{k,i} \tag{8}$$

where super-script H denotes conjugate transpose.

4. OPTIMAL ASII LINEAR PROCESSOR

In this section we derive the optimal linear processor based on the ASII defined in (5). The derivation steps are similar to [8] and [13]. However, contrary to [8] we consider the ASII instead of MI. Further, we expand on [13] by including joint (multi-microphone) processing with far-end noise.

The energy of the clean speech signal within one critical band, j, and time-frame, i, is defined as

$$\mathcal{S}_{j,i}^2 \triangleq \sum_k |S_{k,i}|^2 |H_j(k)|^2, \tag{9}$$

where $H_j(k)$ is the STFT coefficients of the j'th critical band filter. Similarly, we define the critical band energy of the near-end noise, $\mathcal{N}_{j,i}^2$, and the processed far-end speech, $\widetilde{\mathcal{S}}_{j,i}^2$, and noise, $\widetilde{\mathcal{U}}_{j,i}^2$. Since we assume stationarity of the speech and noise, we can disregard the time-index, i, and let the average energy per DFT bin and critical band be based on a long-term average over several short-time frames,

$$\sigma_{S_k}^2 \triangleq \frac{1}{I} \sum_i |S_{k,i}|^2, \quad \sigma_{\mathcal{S}_j}^2 \triangleq \sum_k |H_j(k)|^2 \sigma_{S_k}^2, \tag{10}$$

where I is the total number of frames. Similar definitions hold for the noise terms U and N. The critical band filters, $H_j(k)$, are normalized such that the total energy is preserved in critical bands, i.e.,

$$\sum_j \sigma_{\mathcal{S}_j}^2 = \sum_k \sigma_{S_k}^2. \tag{11}$$

The critical band SNR at the near-end listener is then,

$$\xi_j = \frac{\sigma_{\widetilde{\mathcal{S}}_j}^2}{\sigma_{\widetilde{\mathcal{U}}_j}^2 + \sigma_{\mathcal{N}_j}^2}. \tag{12}$$

Inserting this into (5), we have that

$$f(\xi_j) = \frac{\sum_k |H_j(k)|^2 \boldsymbol{v}_k^H \boldsymbol{d}_k \boldsymbol{d}_k^H \boldsymbol{v}_k \sigma_{S_k}^2}{\sum_k |H_j(k)|^2 \left(\boldsymbol{v}_k^H \boldsymbol{d}_k \boldsymbol{d}_k^H \boldsymbol{v}_k \sigma_{S_k}^2 + \boldsymbol{v}_k^H \Sigma_{U_k} \boldsymbol{v}_k + \sigma_{N_k}^2 \right)} \tag{13}$$

$$\triangleq f_j\left(\{\boldsymbol{v}_k, \Theta_k\}\right), \tag{14}$$

where $\Theta_k = (\sigma_{S_k}^2, \Sigma_{U_k}, \sigma_{N_k}^2)$. In order to limit loudspeaker overload or unpleasant playback levels we invoke the following equal power constraint,

$$\sum_k \boldsymbol{v}_{k,i}^H \boldsymbol{d}_{k,i} \boldsymbol{d}_{k,i}^H \boldsymbol{v}_{k,i} \sigma_{S_{k,i}}^2 = \sum_k \sigma_{S_{k,i}}^2. \tag{15}$$

That is, for each time frame i the total power of the clean speech is unaltered by processing. The joint far and near-end SI enhancement problem with equal power constraint is then,

$$\begin{array}{ll} \sup\limits_{\{\boldsymbol{v}_k\} \in \mathbb{C}^M} & \sum_j \gamma_j f_j\left(\{\boldsymbol{v}_k, \Theta_k\}\right) \\ \text{subject to} & \sum_k \boldsymbol{v}_k^H \boldsymbol{d}_k \boldsymbol{d}_k^H \boldsymbol{v}_k \sigma_{S_k}^2 = \sum_k \sigma_{S_k}^2. \end{array} \tag{16}$$

We now introduce the real and positive variable, α_k, to perform a variable transformation $\boldsymbol{v}_k = \alpha_k^{1/2} \boldsymbol{w}_k$ with the additional constraint $\boldsymbol{v}_k^H \boldsymbol{d}_k \boldsymbol{d}_k^H \boldsymbol{v}_k = \alpha_k$. This leads to the equivalent problem,

$$\begin{array}{ll} \sup\limits_{\{\boldsymbol{v}_k\} \in \mathbb{C}^M, \{\alpha_k\} \in \mathbb{R}_+} & \sum_j \gamma_j f_j\left(\{\alpha_k, \boldsymbol{v}_k, \Theta_k\}\right) \\ \text{subject to} & \mathcal{C}_1 \ : \ \sum_k \alpha_k \sigma_{S_k}^2 = \sum_k \sigma_{S_k}^2, \\ & \mathcal{C}_2 \ : \ \boldsymbol{v}_k^H \boldsymbol{d}_k \boldsymbol{d}_k^H \boldsymbol{v}_k = \alpha_k, \ \forall k. \end{array} \tag{17}$$

The objective function can be rewritten in terms of α_k and $\boldsymbol{w}_k$, i.e.,

$$f_j\left(\{\alpha_k, \boldsymbol{w}_k, , \Theta_k\}\right) = \frac{\sum_k |H_j(k)|^2 \alpha_k \sigma_{S_k}^2}{\sum_k |H_j(k)|^2 \left(\alpha_k \sigma_{S_k}^2 + \alpha_k \boldsymbol{w}_k^H \Sigma_{U_k} \boldsymbol{w}_k + \sigma_{N_k}^2 \right)}.$$

We notice that $\boldsymbol{v}_k^H \boldsymbol{d}_k \boldsymbol{d}_k^H \boldsymbol{v}_k = \alpha_k \Leftrightarrow \boldsymbol{d}_k^H \boldsymbol{w}_k = 1$. Hence, writing the optimization problem in terms of $\boldsymbol{w}_k$ and α_k we have,

$$\begin{array}{ll} \sup\limits_{\{\boldsymbol{w}_k\} \in \mathbb{C}^M, \{\alpha_k\} \in \mathbb{R}_+} & \sum_j \gamma_j f_j\left(\{\alpha_k, \boldsymbol{w}_k, \Theta_k\}\right) \\ \text{subject to} & \mathcal{C}_1 \ : \ \sum_k \alpha_k \sigma_{S_k}^2 = \sum_k \sigma_{S_k}^2 \\ & \mathcal{C}_2 \ : \ \boldsymbol{d}_k^H \boldsymbol{w}_k = 1. \end{array} \tag{18}$$

We can separate (18) across the two variables [19, p. 133], i.e.,

$$\sup_{\{\alpha_k\} \in \mathbb{R}_+, \mathcal{C}_1} \ \sup_{\{\boldsymbol{w}_k\} \in \mathbb{C}^M, \mathcal{C}_2} \sum_j \gamma_j f_j\left(\{\alpha_k, \boldsymbol{w}_k, \Theta_k\}\right).$$

The inner optimization problem across $\{\boldsymbol{w}_k\}$ corresponds to the standard Minimum Variance Distortionless Response (MVDR) beamforming problem with the solution [2],

$$\boldsymbol{w}_k^* = \frac{\Sigma_{U_k}^{-1} \boldsymbol{d}_k}{\boldsymbol{d}_k^H \Sigma_{U_k}^{-1} \boldsymbol{d}_k}, \quad \forall k. \tag{19}$$

Inserting the MVDR solution into (18), the remaining problem is

$$\begin{array}{ll} \sup\limits_{\{\alpha_k\} \in \mathbb{R}_+} & \sum_j \gamma_j \left(\frac{\sum_k |H_j(k)|^2 \alpha_k \sigma_{S_k}^2}{\sum_k |H_j(k)|^2 \left(\alpha_k \sigma_{S_k}^2 + \alpha_k \sigma_{B_k}^2 + \sigma_{N_k}^2 \right)} \right) \\ \text{subject to} & \mathcal{C}_1 \ : \ \sum_k \alpha_k \sigma_{S_k}^2 = \sum_k \sigma_{S_k}^2, \end{array} \tag{20}$$

where $\sigma_{B_k}^2 \triangleq {\boldsymbol{w}_k^*}^H \Sigma_{U_k} \boldsymbol{w}_k^*$ is the residual far-end noise after processing by the MVDR beamformer.

4.1. Critical-band near-end optimization

We derive, similarly to existing work on SI enhancement [8, 7, 13, 6], the optimal near-end processor based on the assumption that all frequency gains within a critical band j are the same, i.e.,

$$\alpha_k = \alpha_{k'}, \forall k, k' \in \mathcal{K}_j, \tag{21}$$

where $\mathcal{K}_j$ is the set of frequency bins belonging to the j'th critical band. The gains are then later on converted back to DFT domain.

Starting from the optimization problem (20) we have,

$$\begin{array}{ll} \sup\limits_{\{\alpha_j\} \in \mathbb{R}_+} & \sum_j \gamma_j \frac{\alpha_j \sigma_{\mathcal{S}_j}^2}{\alpha_j \sigma_{\mathcal{S}_j}^2 + \alpha_j \sigma_{\mathcal{B}_j}^2 + \sigma_{\mathcal{N}_j}^2} \\ \text{subject to} & \mathcal{C}_1 \ : \ \sum_j \alpha_j \sigma_{\mathcal{S}_j}^2 = \sum_j \sigma_{\mathcal{S}_j}^2. \end{array} \tag{22}$$

We notice each term in the sum is concave in α_j for $\alpha_j \geq 0$. Therefore, the weighted sum of these concave functions is also concave. We describe the problem by the Lagrangian cost-function [19],

$$\mathcal{L} = \sum_j \frac{\gamma_j \alpha_j \sigma^2_{\mathcal{S}_j}}{\alpha_j \sigma^2_{\mathcal{S}_j} + \alpha_j \sigma^2_{\mathcal{B}_j} + \sigma^2_{\mathcal{N}_j}} - \nu \left(\sum_j \alpha_j \sigma^2_{\mathcal{S}_j} - r \right) + \sum_j \lambda_j \alpha_j,$$

where $r = \sum_j \sigma^2_{\mathcal{S}_j}$, and ν and λ_j are the Lagrangian multipliers for the energy constraint and inequality constraint in (22). The KKT conditions [19] for the optimization problem are,

$$r = \sum_j \alpha_j \sigma^2_{\mathcal{S}_j},\ 0 \leq \alpha_j,\ 0 \leq \lambda_j,\ 0 = \lambda_j \alpha_j, \forall j, \tag{23a}$$

$$0 = \gamma_j \frac{\sigma^2_{\mathcal{S}_j} \sigma^2_{\mathcal{N}_j}}{\left(\alpha_j \left(\sigma^2_{\mathcal{S}_j} + \sigma^2_{\mathcal{B}_j} \right) + \sigma^2_{\mathcal{N}_j} \right)^2} - \nu \sigma^2_{\mathcal{S}_j} + \lambda_j, \forall j \tag{23b}$$

Isolating λ_j in (23b), then using the complimentary slackness condition to set $\lambda_j = 0$, we solve for the non-zero α_j,

$$\alpha_j = \max \left\{ \frac{\sqrt{\sigma^2_{\mathcal{N}_j} \gamma_j}}{\sqrt{\nu} \left(\sigma^2_{\mathcal{S}_j} + \sigma^2_{\mathcal{B}_j} \right)} - \frac{\sigma^2_{\mathcal{N}_j}}{\sigma^2_{\mathcal{S}_j} + \sigma^2_{\mathcal{B}_j}}, 0 \right\},\ \forall j, \tag{24}$$

where ν is chosen such that the energy constraint in (23a) is satisfied,

$$\frac{1}{\sqrt{\nu}} = \left(r + \sum_{j \in \mathcal{J}} \frac{\sigma^2_{\mathcal{S}_j} \sigma^2_{\mathcal{N}_j}}{\sigma^2_{\mathcal{S}_j} + \sigma^2_{\mathcal{B}_j}} \right) / \left(\sum_{j \in \mathcal{J}} \frac{\sigma^2_{\mathcal{S}_j} \sqrt{\sigma^2_{\mathcal{N}_j} \gamma_j}}{\sigma^2_{\mathcal{S}_j} + \sigma^2_{\mathcal{B}_j}} \right), \tag{25}$$

here $\mathcal{J} = \{ j \in \mathbb{N} : \alpha_j > 0 \}$ denotes the set of frequency bins for which the optimal α_j are positive. We notice, that the set of indices $\mathcal{J}$ depends on the α_j[13]. Therefore, ν also depends on α_j. Hence, there is a recursive relationship between (24) and (25), which may be resolved by using, e.g., a bi-section method or evaluating (24) for a range of ν values, such that the energy constraint is satisfied [13]. Finally, to get the optimal frequency dependent gains, α^*_k, we weight the optimal, α^*_j, using the critical band filters, that is

$$\alpha^*_k = \sum_j |H_j(k)|^2 \alpha^*_j, \tag{26}$$

where the energy constraint is satisfied in both frequency and critical-band domain as per the normalization of the critical band filters in (11), i.e., $\sum_j \alpha^*_j \sigma^2_{\mathcal{S}_j} = \sum_j \sigma^2_{\mathcal{S}_j} = \sum_k \sigma^2_{S_k} = \sum_k \alpha^*_k \sigma^2_{S_k}$.

The proposed processor is summarized in Fig. 2. As in [8], the procedure consists of an MVDR beamformer, $\boldsymbol{w}_k$, followed by a frequency dependent gain, α_k. In contrast to the results of [8], the frequency bin gains, α_k, of our procedure are optimized specifically according to the ASII without approximations and assumptions on signal distributions and without introduction of additional free parameters to model natural speech variations.

5. EXPERIMENTAL EVALUATION

We have seen (Section 2) that by using the production noise model choice in [8] the MI cost function reduces to the ASII. However, this choice of production noise is due to a lack of a more accurate model at the time. Recently, some of the authors have provided an estimated model for the production noise in [14, 15, 16], where $\rho_{0,j} = 0.75\ \forall j$. We compare performance of our proposed method with [8] using the newer production noise model of [14]. We consider a Python simulation of a setup similar to that of [8, 20].

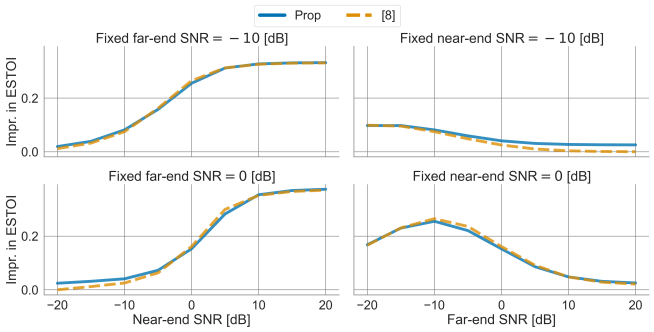

Fig. 3: ESTOI performance of the proposed method and [8], for varying near-end SNR and a fixed far-end SNR (left column), and varying far-end SNR and a fixed near-end SNR (right column).

5.1. Experimental Setup

We consider a room with dimensions $3 \times 4 \times 3\,\mathrm{m}^3$, a single target speaker located at $[1.50, 3.00, 1]\,\mathrm{m}$, and an array with two microphones spaced $2\,\mathrm{cm}$ apart at $[1.50, 2.00, 1]\,\mathrm{m}$ and $[1.50, 2.02, 1]\,\mathrm{m}$. At the far-end there are three speech shaped noise sources located at $[0.50, 1.00, 1]\,\mathrm{m}$, $[0.75, 3.00, 1]\,\mathrm{m}$ and $[3.00, 1.60, 1]\,\mathrm{m}$, respectively. The near-end noise is pink. In addition to the far-end noise each microphone is subject to a $60\,\mathrm{dB}$ SNR white noise. The source signal is speech signals from five female and five male speakers from the TIMIT-database [21] sampled at $16\,\mathrm{kHz}$. Signals were processed block-wise based on a DFT with $32\,\mathrm{ms}$ Hann windows with 50% overlap. Since we assume stationarity, the MVDR beamformer, $\boldsymbol{w}_k$ and post-filter gains, α_k, are derived based on the long-term average spectrums, leading to time-invariant processing. The long-term power of the clean speech, far-end noise, and the near-end noise are assumed to be known. Hence, we do not estimate any of these spectrums. Furthermore, the room transfer functions are assumed to be known, and generated without reverberation using [22].

5.2. Results

Fig. 3 shows improvements over 'unprocessed' in ESTOI [17], of the proposed method along with the method of [8]. The results show that generally the two methods achieve a similar performance, which is expected due to the similarity shown in Section 2. However, the proposed method is slightly better when the near-end SNR is low and the far-end SNR is intermediate or high. Thus, the production noise model choice of [14] leads to a slightly worse speech enhancement than using the model based on the SII weights [18]. ASII plots show identical performance between the two methods and are thus not reported.

6. CONCLUSION

We have derived a closed-form optimal linear processor for joint far- and near-end speech intelligibility enhancement based on ASII. The optimal processor consists of an MVDR beamformer followed by a frequency dependent post gain. The derived processor is based on a simple model without relying on assumptions and approximations of the underlying marginal signal distributions. Furthermore, we do not need to model or optimize for natural variations in speech. Finally, as a consequence, the proposed processor has comparable or slightly better ESTOI performance than existing work.

7. REFERENCES

[1] Philipos C. Loizou, *Speech Enhancement: Theory and Practice*, CRC Press, Boca Raton, FL, 2nd edition, 2013.

[2] Sharon Gannot, Emmanuel Vincent, Shmulik Markovich-Golan, and Alexey Ozerov, "A Consolidated Perspective on Multimicrophone Speech Enhancement and Source Separation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 4, pp. 692–730, Apr. 2017.

[3] Yi Hu and Philipos C. Loizou, "A comparative intelligibility study of single-microphone noise reduction algorithms," *The Journal of the Acoustical Society of America*, vol. 122, no. 3, pp. 1777–1786, Sept. 2007.

[4] Koen Eneman, Heleen Luts, Jan Wouters, Michael Büchler, Norbert Dillier, Wouter Dreschler, Matthias Froehlich, Giso Grimm, Volker Hohmann, Rolph Houben, Arne Leijon, Anthony Lombard, Dirk Mauler, Marc Moonen, Henning Puder, Michael Schulte, Ann Spriet, and Matthias Vormann, "Evaluation of signal enhancement algorithms for hearing instruments," in *2008 16th European Signal Processing Conference*, Aug. 2008, pp. 1–5.

[5] Martin Cooke, Simon King, Maëva Garnier, and Vincent Aubanel, "The listening talker: A review of human and algorithmic context-induced modifications of speech," *Computer Speech & Language*, vol. 28, no. 2, pp. 543–571, Mar. 2014.

[6] W. Bastiaan Kleijn, Joao B. Crespo, Richard C. Hendriks, Petko N. Petkov, Bastian Sauert, and Peter Vary, "Optimizing Speech Intelligibility in a Noisy Environment: A unified view," *IEEE Signal Processing Magazine*, vol. 32, no. 2, pp. 43–54, Mar. 2015.

[7] Markus Niermann, Peter Jax, and Peter Vary, "Joint Near-End Listening Enhancement and far-end noise reduction," in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. Mar. 2017, pp. 4970–4974, IEEE.

[8] Seyran Khademi, Richard C. Hendriks, and W. Bastiaan Kleijn, "Intelligibility Enhancement Based on Mutual Information," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 8, pp. 1694–1708, Aug. 2017.

[9] Ke Tan and DeLiang Wang, "Improving Robustness of Deep Learning Based Monaural Speech Enhancement Against Processing Artifacts," in *2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Barcelona, Spain, May 2020, pp. 6914–6918, IEEE.

[10] Seyran Khademi, Richard C. Hendriks, and W. Bastiaan Kleijn, "Jointly optimal near-end and far-end multimicrophone speech intelligibility enhancement based on mutual information," in *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Shanghai, Mar. 2016, pp. 654–658, IEEE.

[11] Thomas M. Cover and Joy A. Thomas, *Elements of information theory*, Wiley-Interscience, Hoboken, N.J, 2nd edition, 2006.

[12] W. Bastiaan Kleijn and Richard C. Hendriks, "A Simple Model of Speech Communication and its Application to Intelligibility Enhancement," *IEEE Signal Processing Letters*, vol. 22, no. 3, pp. 303–307, Mar. 2015.

[13] Cees H. Taal, Jesper Jensen, and Arne Leijon, "On Optimal Linear Filtering of Speech for Near-End Listening Enhancement," *IEEE Signal Processing Letters*, vol. 20, no. 3, pp. 225–228, Mar. 2013.

[14] Steven Van Kuyk, W. Bastiaan Kleijn, and Richard C. Hendriks, "An Instrumental Intelligibility Metric Based on Information Theory," *IEEE Signal Processing Letters*, vol. 25, no. 1, pp. 115–119, Jan. 2018.

[15] Steven Van Kuyk, W. Bastiaan Kleijn, and Richard C. Hendriks, "An intelligibility metric based on a simple model of speech communication," in *2016 IEEE International Workshop on Acoustic Signal Enhancement (IWAENC)*, Sept. 2016, pp. 1–5.

[16] Steven Van Kuyk, W. Bastiaan Kleijn, and Richard C. Hendriks, "On the information rate of speech communication," in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. Mar. 2017, pp. 5625–5629, IEEE.

[17] Jesper Jensen and Cees H. Taal, "An Algorithm for Predicting the Intelligibility of Speech Masked by Modulated Noise Maskers," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 24, no. 11, pp. 2009–2022, Nov. 2016.

[18] American National Standards Institute, *Methods for Calculation of the Speech Intelligibility Index*, Acoustical Society of America, ANSI S.35-1997 edition, 2017, Place: New York, N.Y.

[19] Stephen P. Boyd and Lieven Vandenberghe, *Convex optimization*, Cambridge University Press, Cambridge, UK, 2004.

[20] Seyran Khademi, Richard C. Hendriks, and W. Bastiaan Kleijn, "Joint near-end and far-end intelligibility enhancement," `https://cas.tudelft.nl/Repository/repitem.php?id=3&ti=2`, Aug. 2017, Accessed: 2021-02-15.

[21] John S. Garofolo, Lori F. Lamel, William M. Fisher, Jonathan G. Fiscus, David S. Pallett, Nancy L. Dahlgren, and Victor Zue, "TIMIT Acoustic-phonetic Continuous Speech Corpus," *Linguistic Data Consortium*, 1993.

[22] Jont Allen and David Berkley, "Image method for efficiently simulating small-room acoustics," *The Journal of the Acoustical Society of America*, vol. 65, pp. 943–950, Apr. 1979.