TU Delft

Delft University of Technology

Jointly optimal near-end and far-end multi-microphone speech intelligibility enhancement based on mutual information

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

DOI
10.1109/icassp.2016.7471756

Publication date
2016

Document Version
Accepted author manuscript

Published in
2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)

Citation (APA)
Khademi, S., Hendriks, R. C., & Kleijn, W. B. (2016). Jointly optimal near-end and far-end multi-microphone speech intelligibility enhancement based on mutual information. In M. Dong, & T. F. Zheng (Eds.), *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP): Proceedings* (pp. 654-658). IEEE . https://doi.org/10.1109/icassp.2016.7471756

Important note
To cite this publication, please use the final published version (if applicable).
Please check the document version above.

Copyright
Other than for strictly personal use, it is not permitted to download, forward or distribute the text or part of it, without the consent of the author(s) and/or copyright holder(s), unless the work is under an open content license such as Creative Commons.

Takedown policy
Please contact us and provide details if you believe this document breaches copyrights.
We will remove access to the work immediately and investigate your claim.

This work is downloaded from Delft University of Technology.
For technical reasons the number of authors shown on this cover page is limited to a maximum of 10.

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

Faculty of Electrical Engineering, Mathematics and Computer Science, TU Delft
e-mail: {s.khademi, r.c.hendriks, w.b.kleijn}@tudelft.nl.

ABSTRACT

The processing required for the global maximization of the intelligibility of speech acquired by multiple microphones and rendered by a single loudspeaker, is considered in this paper. The intelligibility is quantized, based on the mutual information rate between the message spoken by the talker and the message as interpreted by the listener. We prove that then, in each of a set of narrow-band channels, the processing can be decomposed into a minimum variance distortionless response (MVDR) beamforming operation that reduces the noise in the talker environment, followed by a gain operation that, given the far-end noise and beamforming operation, accounts for the noise at the listener end. Our experiments confirm that both processing steps are necessary for the effective conveyance of a message and, importantly, that the second step must be aware of the first step.

Index Terms— Speech intelligibility enhancement, mutual information, minimum variance distortionless response (MVDR) beamformer, multi-microphone.

1. INTRODUCTION

It is common that noise sources at the far-end of a communication system are recorded together with the target signal, leading to reduced intelligibility when the signal is finally played back at the near-end. In addition, acoustical noise sources in the near-end environment degrade the intelligibility even more. These two causes have always been considered separately in the literature. To handle the presence of noise at the far-end, standard approaches are single or multi-microphone noise reduction algorithms operating at the far-end (see *e.g.*, [1–4]), although methods operating at the near-end to remove far-end noise do exist [5].

The conventional approach is to pre-process the speech that is received from the far-end, such that when played out in the near-end environment, the intelligibility is improved or maintained see *e.g.* [6–16]. Among these near-end speech enhancement methods, there are roughly two different ways to approach the problem. The first class consists of algorithms that are based mainly on empirical considerations, and employ for example the fact that consonants and high frequencies are important for speech intelligibility, see *e.g.*, [7, 8, 17]. The second class optimizes more formal mathematical models of speech intelligibility to obtain improved speech intelligibility, see *e.g.*, [16] for an overview. Typically, these optimization procedures are carried out under an energy constraint, which can be used to satisfy average loudspeaker power constraints or to overcome hearing discomfort due to loud sounds.

Examples of classical measures that have been developed to predict intelligibility of speech in noise are the articulation index (AI) [18, 19] and the speech intelligibility index (SII) [20]. Within the near-end intelligibility enhancement context, the SII has been optimized in [10, 11]. Recently, the approximated SII (ASII) was proposed in [14] to make constrained optimization of the SII more tractable. Other metrics that have recently been used to optimize the intelligibility of speech in noise are [13, 21–23]

Another approach to quantifying speech intelligibility is to use information theory to describe the amount of information that can be transmitted through a speech communication channel. Examples of speech intelligibility predictors based on mutual information (MI) can be found in [24–27]. In [27] an effective model of human communication based on MI was derived. This model takes the noise inherent in the speech production process and the speech interpretation process into account. The resulting intelligibility predictor resembles heuristically derived classical measures of intelligibility such as the AI and the SII. The MI based measure of [27] was shown to be effective for near-end speech intelligibility enhancement.

All the existing speech intelligibility techniques treat the far-end noise reduction and near-end enhancement as two separate problems. However, optimizing intelligibility in a jointly optimal way by taking both the disturbances at the near-end and the far-end into account can exploit that a finite far-end signal-to-noise ratio (SNR) influences the effectiveness with which the intelligibility can be improved at the near-end. More specifically, depending on the far-end SNR (after far-end processing), the environmental noise at the near-end may be negligible compared to the far-end noise already present in the signal. Then increasing the near-end channel quality by boosting the power is of little benefit; it is then likely more beneficial to boost the power of channels with a high far-end SNR.

In this paper we use the MI-based model of [27] as a starting point and propose a comprehensive model accounting jointly for both noise at the far-end as well as noise at the near-end. The proposed formulation considers the presence of a microphone array at the far-end. Although the original joint problem is non-convex with respect to the beamformer variable, we prove that, within a frequency channel, the optimization problem can be decomposed into the well-known minimum variance distortionless response (MVDR) beamformer, and a scalar factor that is determined by solving a convex fractional problem through the Karush-Kuhn-Tucker (KKT) conditions. Importantly, while the processing can be done separately, the computation of the scalar factor requires knowledge of far-end noise level and the far-end processing.

2. SIGNAL MODEL AND ASSUMPTIONS

The communication model presented in [27] is extended to also include the far-end noise that is present in the environment of the microphones. In Sec. 3, we further extend the model for multiple microphones.

This work was supported in part by the Dutch Technology Foundation STW and Bosch Security Systems B.V.

The speech process S is assumed to be a sequence of complex random vectors, with each coefficient $S_{k,i}$ describing either a critical band [27] or simply a DFT time-frequency bin. Although other speech representations may exist that better characterize intelligibility and perception, we adopt this approach for the sake of mathematical tractability.

Throughout this paper we use bold upper case and lower case symbols to indicate matrices and vectors, respectively, and regular symbols (lower and upper case) for scalars. The conjugate transpose and inverse of a matrix $\mathbf{X}$ are denoted as $\mathbf{X}^H$ and $\mathbf{X}^{-1}$, respectively. As the distinctions are clear from the context, our notation does not distinguish between random variables and processes and their realizations.

2.1. Markov Chain Model for Intelligibility

The communication model from [27] takes the natural variation of communication messages (speech in the current application) into account. It is modeled by the production noise $V_{k,i}$. The acoustic signal produced at time-frequency point (k, i) is given by

$$T_{k,i} = S_{k,i} + V_{k,i}. \tag{1}$$

The variations in speech production are to a large extent independent of the presentation level. As an important consequence, the speech production SNR, $\sigma^2_{S_k}/\sigma^2_{V_k}$, remains constant.

In this work we assume that noise is present in both the near-end (listener-side) and far-end (talker-side) environments. Let the far-end noise be denoted by $U_{k,i}$. The complex coefficients of the recorded signal are

$$X_{k,i} = T_{k,i} + U_{k,i}. \tag{2}$$

Prior to rendering in the noisy near-end environment the recorded signals are processed to optimize a mutual information rate between the spoken and interpreted message under a power-preservation constraint. (Without the constraint play-back levels generally will increase.) The modified coefficients are denoted by $\tilde{\cdot}$. The signals as received by the observer in the near-end environment are

$$Y_{k,i} = \tilde{X}_{k,i} + N_{k,i}, \tag{3}$$

where $N_{k,i}$ is the near-end environmental noise.

Finally, the symbols are received by the human observer, where internal noise, the absolute hearing threshold and an increased hearing threshold can be modeled by an additional noise source, that is,

$$Z_{k,i} = Y_{k,i} + W_{k,i}, \tag{4}$$

where $W_{k,i}$ is the interpretation noise. Similarly as for the production noise, it was argued in [27] that this noise scales with the signal, and this is consistent with the notion of instantaneous masking (e.g., [28]). As a consequence, the interpretation SNR $\sigma^2_{Y_k}/\sigma^2_{W_k}$, remains constant as well.

The above signal model constitutes a Markov chain, that is, $S \to T \to X \to \tilde{X} \to Y \to Z$. Let $\mathbf{S}_i$ and $\mathbf{Z}_i$ denote K-dimensional stacked vectors of spectral coefficients in one time frame i. The mutual information rate between the original $\mathbf{S}_i$ and the received $\mathbf{Z}_i$ describes the effectiveness of the communication process. Under the assumption that the processes are memoryless, the mutual information rate is equal to the mutual information $I(\mathbf{S}_i; \mathbf{Z}_i)$. We furthermore assume that the individual component signals of the vectors $\mathbf{S}_i$ and $\mathbf{Z}_i$ are independent. We can then write

$$I(\mathbf{S}_i; \mathbf{Z}_i) = \sum_k I(S_{k,i}; Z_{k,i}). \tag{5}$$

The relation at each Markov step is described by the correlation coefficient between the corresponding variables. Using the Markov chain property we can write

$$\rho_{S_{k,i}Z_{k,i}} = \rho_{S_{k,i}T_{k,i}}\rho_{T_{k,i}\tilde{X}_{k,i}}\rho_{\tilde{X}_{k,i}Y_{k,i}}\rho_{Y_{k,i}Z_{k,i}}.$$

The fixed production and interpretation SNR imply that the corresponding correlation coefficients $\rho_{S_{k,i}T_{k,i}}$and $\rho_{Y_{k,i}Z_{k,i}}$are fixed numbers on $[0, 1]$. In addition to the production and interpretation noise, we have also introduced far-end noise into the communication model. The SNR between the far-end noise and the speech recorded at the far-end is given by $\sigma^2_{T_k}/\sigma^2_{U_k}$ and the corresponding correlation coefficient is $\rho_{T_kX_k}$. (The SNR at the near-end varies with the processing performed.)

In the following, we will assume that enhancement is performed by a linear time-invariant operator, which implies that $\rho_{T_{k,i}\tilde{X}_{k,i}} = \rho_{T_{k,i}X_{k,i}}$. We furthermore make the assumptions that all processes are jointly Gaussian, stationary, and memoryless. It is then natural to omit the time-frame index i for notational convenience. Defining $\rho_{0,k} = \rho_{S_kT_k}\rho_{Y_kZ_k}$ and $\rho_{1,k} = \rho_{T_kX_k}$, it can be shown that

$$I(S_k; Z_k) = -\frac{1}{2}\log(1 - \rho^2_{0,k}\rho^2_{1,k}\rho^2_{\tilde{X}_kY_k}). \tag{6}$$

2.2. Relation to Classical Measures of Intelligibility

We now place our intelligibility measure (6) in the context of classical measures of intelligibility. For the case with no enhancement, the overall channel SNR $\xi_k = \frac{\sigma^2_{T_k}}{\sigma^2_{N_k}+\sigma^2_{U_k}}$ can be used to write (6) as

$$I(S; Z) = -\sum_k \frac{1}{2}\log\left(\frac{(1-\rho^2_{0,k})\xi_k + 1}{\xi_k + 1}\right). \tag{7}$$

It can be seen that (7) is closely related to the classical intelligibility measures AI [18, 19], SII [20] and ASII [14] if we write it as

$$I(S; Z) = \sum_k I_k A_k(\xi_k), \tag{8}$$

with $A_k(\xi_k) = \log\frac{(1-\rho^2_{0,k})\xi_k+1}{\xi_k+1}/\log(1-\rho^2_{0,k})$ and $I_k = -\frac{1}{2}\log(1-\rho^2_{0,k})$. Comparing our measure to the classical measures, we can identify I_k as the *unnormalized* band-importance function and $A_k(\xi_k)$ as the weighting function. The unnormalized band-importance function is simply *the information rate transmitted in a band when no environmental noise is present*. This happens when $\rho^2_{1,k}\rho^2_{\tilde{X}_k,Y_k} = 1$ and, hence, ξ_k is infinite. The conventional, *normalized* band-importance function is obtained by dividing I_k by the overall mutual information rate $I(S; Z)$. The importance of a band for speech intelligibility (its maximum information rate) decreases with an increase in its interpretation and production noise.

A good intelligibility measure should have well-defined upper and lower bounds when the environmental noise is varied, to reflect that intelligibility cannot be increased below and above certain levels. The additive terms in (7) are naturally limited by the band importance value I_k and by zero (as mutual information is nonnegative).

To illustrate the relation of our measure based on mutual information to existing measures of intelligibility, we show in Fig. 1 the weighting functions for the different measures. We show the curves $A_k(\xi_k)$ for various values of $\rho^2_{0,k}$. From Fig. 1 it follows that the shape of the proposed weighting function approaches the shape of the ASII and AI weighting function for specific value of $\rho^2_{0,k}$. For $\rho^2_{0,k} \downarrow 0$, $A_k(\xi_k)$ approaches $A^{ASII}_k(\xi_k)$ and for $\rho^2_{0,k} = 1$ it approaches $A^{AI}_k(\xi_k)$. Moreover, we see that the $A^{SII}_k(\xi_k)$ appears as a linearization of the proposed $A_k(\xi_k)$ and $A^{ASII}_k(\xi_k)$. Also note that the proposed $A_k(\xi_k)$ are concave functions of the linear ξ_k, whereas $A^{AI}_k(\xi_k)$ and $A^{SII}_k(\xi_k)$ are not.

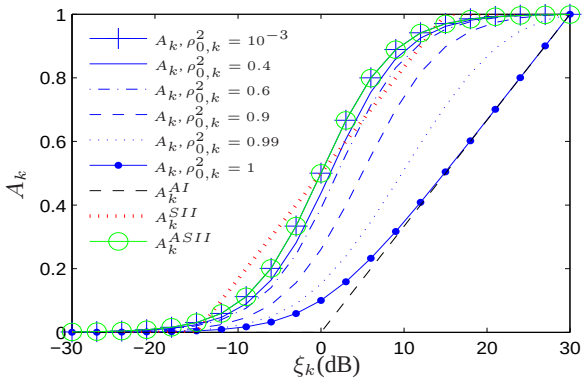


Fig. 1. Comparison of weighting functions.

3. MULTI-MICROPHONE PROBLEM FORMULATION

We aim to find, for each time-frequency bin, a linear multi-microphone processor $\mathbf{z}$ that maximizes the mutual information rate, taking both far-end and near-end noise into account. This directly follows from the model in Sec. 2.

Let $d_{k,i,m}$ denote the acoustic transfer function from source to microphone m and let us use write $\mathbf{d}_{k,i} = [d_{k,i,1}, ..., d_{k,i,M}]^T$. We denote the far-end noise recorded by the microphones by the vector $\mathbf{u}_{k,i}$. The processed noisy microphone data can then be written as

$$\tilde{X}_{k,i} = \mathbf{v}_k^H \mathbf{d}_{k,i} T_{k,i} + \mathbf{v}_k^H \mathbf{u}_{k,i}, \tag{9}$$

with $T_{k,i}$ the speech time-frequency coefficient at the source location and $\mathbf{v}$ the multi-microphone processor. (9) generalizes (2) but includes the modification operator; (3) and (4) remain unchanged.

We can now formulate the problem of finding the processor $\mathbf{v}$ that maximizes the mutual information between S_k and Z_k under a constraint on the average power:

$$\begin{array}{ll} \sup\limits_{\{\mathbf{v}_k\} \in \mathbb{C}^M} & \sum_k I(S_k; Z_k) \\ \text{subject to} & \sum_k \mathbf{v}_k^H \mathbf{d}_k \mathbf{d}_k^H \mathbf{v}_k \sigma^2_{T_k} = \sum_k \sigma^2_{T_k}, \end{array} \tag{10}$$

where the mutual information between S_k and Z_k is given by

$$I(S_k; Z_k) = -\frac{1}{2} \log \left(1 - \rho^2_{0,k} \frac{1}{1 + \frac{\mathbf{v}_k^H \mathbf{R}_{U_k} \mathbf{v}_k + \sigma^2_{N_k}}{\mathbf{v}_k^H \mathbf{d}_k \mathbf{d}_k^H \mathbf{v}_k \sigma^2_{T_k}}} \right) \tag{11}$$

with $\mathbf{R}_{U_k} = \mathbb{E}\{\mathbf{u}_k {\mathbf{u}_k}^H\}$. Expression (11) is obtained from (6) by realizing that, in the multi-microphone case,

$$\rho^2_{1,k} = \frac{1}{1 + \frac{\mathbf{v}_k^H \mathbf{R}_{U_k} \mathbf{v}_k}{\mathbf{v}_k^H \mathbf{d}_k \mathbf{d}_k^H \mathbf{v}_k \sigma^2_{T_k}}} \text{ and } \rho^2_{\tilde{X},Y} = \frac{1}{1 + \frac{\sigma^2_{N_k}}{\mathbf{v}_k^H (\mathbf{d}_k \mathbf{d}_k^H \sigma^2_{T_k} + \mathbf{R}_{U_k}) \mathbf{v}_k}}. \tag{12}$$

4. PROPOSED SOLUTION

We now solve the problem stated in (10). The objective function includes a sum of non-linear fractional terms which, in general, can not be transformed into standard convex programming framework [29]. To find an optimizer for (10), we therefore introduce slack variables α_k and an additional constraint $\mathbf{v}_k^H \mathbf{d}_k \mathbf{d}_k^H \mathbf{v}_k = \alpha_k$:

$$\begin{array}{ll} \sup\limits_{\mathbf{v}_k \in \mathbb{C}^M, \alpha_k \in \mathbb{R}_+} & -\frac{1}{2} \sum_k \log \left(1 - \frac{\rho^2_{0,k} \alpha_k \sigma^2_{T_k}}{\alpha_k \sigma^2_{T_k} + \mathbf{v}_k^H \mathbf{R}_{U_k} \mathbf{v}_k + \sigma^2_{N_k}} \right) \\ \text{subject to} & \sum_k \alpha_k \sigma^2_{T_k} = \sum_k \sigma^2_{T_k} \\ & \mathbf{v}_k^H \mathbf{d}_k \mathbf{d}_k^H \mathbf{v}_k = \alpha_k, \ \forall k. \end{array} \tag{13}$$

By also introducing new vectors $\mathbf{w}_k$ such that $\mathbf{v}_k = \alpha_k^{1/2} \mathbf{w}_k$, the last constraint can be rephrased irrespective of α_k and the objective function can be rewritten in terms of $\mathbf{w}_k$:

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

Problem (13) is then transformed into

$$\begin{array}{ll} \sup\limits_{\mathbf{w}_k \in \mathbb{C}^M, \alpha_k \in \mathbb{R}_+} & I(\alpha_k, \mathbf{w}_k) \\ \text{subject to} & \mathcal{C}_1 : \sum_k \alpha_k \sigma^2_{T_k} = \sum_k \sigma^2_{T_k} \\ & \mathcal{C}_2 : \mathbf{w}_k^H \mathbf{d}_k = 1, \ \forall k. \end{array} \tag{14}$$

The constraints $\mathcal{C}_1$ and $\mathcal{C}_2$ are now independent. Using the fact that in general $\sup_{x,y} f(x,y) = \sup_x \sup_y f(x,y)$ (see also [29, p. 133]), (14) can be rephrased as

$$\sup_{\alpha_k \in \mathbb{R}_+, \mathcal{C}_1} \ \sup_{\mathbf{w}_k \in \mathbb{C}^M, \mathcal{C}_2} I(\alpha_k, \mathbf{w}_k). \tag{15}$$

In combination with the independence of the constraints, this allows us to solve the optimization problem.

The inner maximization problem in (15) over $\mathbf{w}_k$ is the standard MVDR beamforming problem, *e.g.*, [3]. Its solution is given by

$$\mathbf{w}_k^* = \frac{\mathbf{R}_{U_k}^{-1} \mathbf{d}_k}{\mathbf{d}_k^H \mathbf{R}_{U_k}^{-1} \mathbf{d}_k}, \ \forall k.$$

Using $\mathbf{w}_k^*$, the outer maximization in (15) over α_k is

$$\begin{array}{ll} \sup\limits_{\alpha_k \in \mathbb{R}_+} & -\frac{1}{2} \sum_k \log \left(1 - \rho^2_{0,k} \frac{\alpha_k \sigma^2_{T_k}}{\alpha_k \sigma^2_{T_k} + \alpha_k \sigma^2_{M_k} + \sigma^2_{N_k}} \right) \\ \text{subject to} & \sum_k \alpha_k \sigma^2_{T_k} = \sum_k \sigma^2_{T_k} \end{array} \tag{16}$$

where $\sigma^2_{M_k} = {\mathbf{w}_k^*}^H \mathbf{R}_{U_k} \mathbf{w}_k^*$ is the far-end noise that remains after processing by the MVDR beamformer. Problem (16) is a convex problem that is of the same form as the problem for the single-microphone case in (7).

Let λ and μ_k be two Lagrangian multipliers that are non-positive and non-negative, respectively. The Lagrangian is then given by

$$\begin{aligned} \mathcal{L}(\alpha_k, \lambda, \mu_k) = & -\frac{1}{2} \sum_k \log \left(1 - \rho^2_{0,k} \frac{\alpha_k \sigma^2_{T_k}}{\alpha_k \sigma^2_{T_k} + \alpha_k \sigma^2_{M_k} + \sigma^2_{N_k}} \right) \\ & + \lambda \alpha_k \sigma^2_{T_k} + \mu_k \alpha_k. \end{aligned}$$

The map to the solution is described in [27]; here we only redefine the parameters that are changed due to the presence of the far-end noise. The optimal α_k are found by differentiation of the Lagrangian and setting the result to zero to find the stationary points. This leads to a quadratic equation to be solved, that is,

$$a \alpha_k^2 + b \alpha_k + c = 0 \tag{17}$$

where

$$a = (\sigma^2_{T_k} + \sigma^2_{M_k})((1 - \rho^2_{0,k}) \sigma^2_{T_k} + \sigma^2_{M_k})(\lambda \sigma^2_{T_k} + \mu_k) \tag{18}$$

$$b = (\sigma^2_{T_k} (2 - \rho^2_{0,k}) + 2\sigma^2_{M_k}))(\lambda \sigma^2_{T_k} + \mu_k) \sigma^2_{N,k} \tag{19}$$

$$c = \frac{1}{2} \rho^2_{0,k} \sigma^2_{N,k} \sigma^2_{T_k} + \sigma^4_{N,k} (\lambda \sigma^2_{T_k} + \mu_k). \tag{20}$$

Under gaussianity assumptions and with MI as the objective, our analysis shows that an enhancement algorithm based on a linear filtering of a multi-microphone signal for rendering over a single loudspeaker in a noisy environment naturally decomposes into a spatial processor to reduce far-end noise, followed by a post-processor to increase the speech intelligibility with respect to the near-end noise. The optimal spatial processor is a standard MVDR beamformer. Our

result is comparable to the well-known result for far-end noise reduction that the optimal multi-channel noise reduction algorithm decomposes into a spatial processor and a single-channel post-processor, see, e.g., [30].

Our result on decomposition of the optimal preprocessor into MVDR and near-end linear processor, is not surprising in light of the result of [30] that the output of the MVDR beamformer, say $G_{k,i}$, is a sufficient statistic for $S_{k,i}$ with respect to the microphone observations $\mathbf{d}_{k,I}T_{k,i}$. Mathematically this implies $p(\mathbf{d}_{k,I}T_{k,i}|S_{k,i}, G_{k,i}) = p(\mathbf{d}_{k,I}T_{k,i}|G_{k,i})$, where $p(\cdot|\cdot)$ denotes a conditional density. It then follows that $I(S_{k,i}; \mathbf{d}_{k,I}T_{k,i}) = I(S_{k,i}; G_{k,i})$, suggesting that a spatial processor consisting of an MVDR beamforming is optimal.

However, we consider a practical linear enhancement operator which does not follow the notion of $G_{i,k}$ being a sufficient statistic. In fact, the most important outcome of this work is the necessity of a transparent processor which communicates the output of the spatial beamformer to the near-end pre-processor.

5. SIMULATION RESULTS

The goal of our experimental work was to show that while our theory indicated that the optimal linear processor can be implemented by two subsequent processors, the second processor must be aware of both the far-end noise and the operation of the first processor. We first discuss our experimental setup and then our results.

5.1. Experimental Setup

We simulated a dual microphone setup with a 2 cm spacing in a 3 m $\times$ 4 m $\times$ 3 m room with one target source, three noise sources and simulated uncorrelated microphone noise at 60 dB SNR. We used 36 seconds of speech material originating from the Timit-database [31], sampled at 16 kHz.

The impulse responses were generated using [32]. The far-end and near-end noise source consisted of spectrally shaped Gaussian noise, with an overlapping region from 1.5 kHz till 3 kHz to demonstrate the effect of the different filters. Signals were processed on a block-by-block basis by applying a 32 ms Hann analysis window with 50 % overlap.

The spatial processor in all experiments was directly applied to the complex discrete Fourier transform (DFT) coefficients. The post-processors were subsequently applied per critical band to the spatial processor output. The critical band variances, e.g., $\sigma^2_{T_k}$, $\sigma^2_{M_k}$ and $\sigma^2_{N_k}$ were obtained by taking the sample-mean of the critical band energy over the entire utterance, leading to a time-invariant filter.

5.2. Results

Using simulations, we compare five approaches:

1. The system based on our solution of (10)-(11) which is referred to as *proposed*.
2. The system based on our solution of (10)-(11) applied to a single microphone is denoted as *proposed single microphone*.
3. A standard time-invariant multi-channel Wiener filter (MWF) that accounts for the far-end noise, combined with the intelligibility enhancement algorithm from [27] to account for the near-end noise (*MWF+ [27]*).
4. The near-end intelligibility enhancement algorithm of [27] applied directly to the output from the MVDR, erroneously assuming this is noise free (*MVDR+ [27]*).
5. The output of the far-end MVDR (*MVDR*).

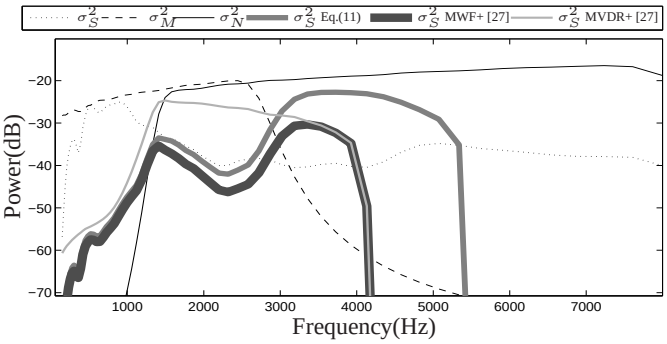


Fig. 2. Average spectra for -11.1 dB SNR at the far-end reference microphone and -10 dB SNR at the near-end.

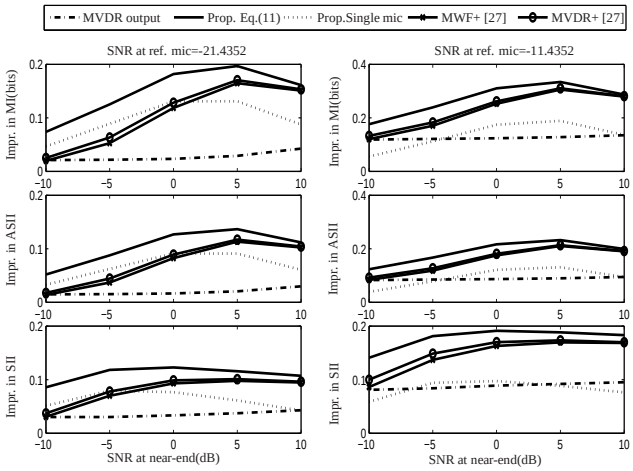


Fig. 3. Predicted intelligibility in terms of MI, ASII and SII.

Fig. 2 shows the processed speech spectra for the different algorithms, compared to the unprocessed speech spectrum, at the far-end and near-end noise spectrum σ^2_N. The use of independent processing (*MVDR+ [27]*) leads to erroneously amplified speech, as the input to the near-end algorithm is not clean, but processed noisy speech. *MWF+ [27]* amplifies the far-end noise. Instead, *proposed* correctly applies most amplification in the higher frequency bands that are not saturated by the near-end noise.

In Fig. 3 we compare the improvement in speech intelligibility prediction over the noisy reference microphone for several combinations of far-end and near-end SNR. As speech intelligibility predictors we use the MI, the ASII and the SII. As expected, *MVDR* and *proposed single microphone* perform worst. *Proposed* consistently obtains the best performance, showing a clear improvement over *MWF+ [27]* and *MVDR+ [27]* .

6. CONCLUDING REMARKS

We conclude that conventional independent processing to the noise at the near-end and the far-end is not optimal. Our theory shows that the processing can be separated into far-end and near-end processing. However, the near-end processing must be aware of the noise and processing performed at the far-end, which is not the case in conventional systems. Our experimental results clearly confirm the awareness requirement.

7. REFERENCES

[1] P. Loizou, *Speech enhancement theory and practice*. CRC Press, 2007.

[2] R. C. Hendriks, T. Gerkmann, and J. Jensen, *DFT-Domain Based Single-Microphone Noise Reduction for Speech Enhancement: A Survey of the State of the Art*. Morgan & Claypool, 2013.

[3] J. Benesty, M. M. Sondhi, and Y. Huang (Eds), *Springer handbook of speech processing*. Springer, 2008.

[4] M. Brandstein and D. Ward (Eds.), *Microphone arrays: signal processing techniques and applications*. Springer, 2001.

[5] V. Grancharov, J. Samuelsson, and W. B. Kleijn, "Noise-dependent postfiltering," in *IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, (Montreal), pp. I457–I460, 2004.

[6] J. D. Griffiths, "Optimum linear filter for speech transmission," *J. Acoust. Soc. Amer.*, vol. 43, p. 81, 1968.

[7] R. Niederjohn and J. Grotelueschen, "The enhancement of speech intelligibility in high noise levels by high-pass filtering followed by rapid amplitude compression," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 24, no. 4, pp. 277–282, 1976.

[8] C. Tantibundhit, J. R. Boston, C. C. Li, J. D. Durrant, S. Shaiman, K. Kovacyk, and A. El-Jaroudi, "New signal decomposition method based speech enhancement," *Signal Processing*, vol. 87, pp. 2607 – 2628, 2007.

[9] B. Sauert and P. Vary, "Near end listening enhancement: speech intelligibility improvement in noisy environments," in *IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, vol. 1, pp. 493–496, 2006.

[10] B. Sauert and P. Vary, "Near end listening enhancement optimized with respect to speech intelligibility index," in *EURASIP Europ. Signal Process. Conf. (EUSIPCO)*, vol. 17, pp. 1844–1848, 2009.

[11] B. Sauert and P. Vary, "Near end listening enhancement optimized with respect to speech intelligibility index and audio power limitations," in *EURASIP Europ. Signal Process. Conf. (EUSIPCO)*, pp. 1919–1923, 2010.

[12] B. Sauert and P. Vary, "Recursive closed-form optimization of spectral audio power allocation for near end listening enhancement," in *ITG-Fachtagung Sprachkommun.*, VDE VERLAG GmbH, 2010.

[13] C. H. Taal, R. C. Hendriks, and R. Heusdens, "A speech preprocessing strategy for intelligibility improvement in noise based on a perceptual distortion measure," in *IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, pp. 4061–4064, IEEE, 2012.

[14] 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, pp. 225 – 228, 2013.

[15] M. Cooke, C. Mayoe, and C. Valentini-Botinhao, "Intelligibility enhancing speech modifications: the hurricane challenge," *ISCA Interspeech*, 2013.

[16] 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 Process. Mag.*, vol. 32, no. 2, pp. 43–54, 2015.

[17] J. L. Hall and J. L. Flanagan, "Intelligibility and listener preference of telephone speech in the presence of babble noise," *J. Acoust. Soc. Amer.*, vol. 127, pp. 280–285, 2010.

[18] N. R. French and J. C. Steinberg, "Factors governing the intelligibility of speech sounds," *J. Acoust. Soc. Amer.*, vol. 19, pp. 90–119, January 1947.

[19] K. D. Kryter, "Methods for the calculation and use of the Articulation Index," *J. Acoust. Soc. Amer.*, vol. 34, pp. 1689–1697, November 1962.

[20] American National Standards Institute, *American National Standard Methods for the Calculation of the Speech Intelligibility index*. ANSI S3.5-1997 ed.

[21] M. Cooke, "A glimpsing model of speech perception in noise," *J. Acoust. Soc. Amer.*, vol. 119, no. 3, pp. 1562–1573, 2006.

[22] C. H. Taal, R. C. Hendriks, and R. Heusdens, "A low-complexity spectro-temporal distortion measure for audio processing applications," *IEEE Trans. Audio, Speech, Language Process.*, vol. 20, no. 5, pp. 1553–1564, 2012.

[23] C. H. Taal, R. C. Hendriks, and R. Heusdens, "Speech energy redistribution for intelligibility improvement in noise based on a perceptual distortion measure," *Computer Speech & Language*, 2014.

[24] J. Taghia, R. Martin, and R. C. Hendriks, "On mutual information as a measure of speech intelligibility," in *IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, pp. 65–68, IEEE, 2012.

[25] J. Taghia and R. Martin, "Objective intelligibility measures based on mutual information for speech subjected to speech enhancement processing," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 22, no. 1, pp. 6–16, 2014.

[26] J. Jensen and C. H. Taal, "Speech intelligibility prediction based on mutual information," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 22, no. 2, pp. 430–440, 2014.

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

[28] H. Fastl and E. Zwicker, *Psychoacoustics Facts and Models*. Springer, 2006.

[29] S. Boyd and L. Vandenberghe, *Convex Optimization: Convex optimization problems, Equivalent problems*. New York, NY, USA: Cambridge University Press, 2004.

[30] R. Balan and J. Rosca, "Microphone array speech enhancement by Bayesian estimation of spectral amplitude and phase," in *Proc. of IEEE Sensor Array and Multichannel Signal Processing Workshop*, pp. 209–213, 2002.

[31] J. S. Garofolo, "DARPA TIMIT acoustic-phonetic speech database," *National Institute of Standards and Technology (NIST)*, 1988.

[32] E. A. P. Habets, "Room impulse response generator," tech. rep., Technische Universiteit Eindhoven, Eindhoven, 2010.