Intelligibility Enhancement Based on Mutual Information

Seyran Khademi, *Student Member, IEEE*, Richard C. Hendriks, *Member, IEEE*, and W. Bastiaan Kleijn, *Fellow, IEEE*

***Abstract*—Speech intelligibility enhancement is considered for multiple-microphone acquisition and single loudspeaker rendering. This is based on the mutual information measured between the message spoken at far-end environment and the message perceived by a listener at near-end. We prove that the joint optimal processing can be decomposed into far-end and near-end processing. The former is a minimum variance distortionless response beamformer that reduces the noise in the talker environment and the latter is a post-filter that redistributes the power over the frequency bands. Disjoint processing is optimal provided that the post-filtering operation is aware of the residual noise from the beamforming operation. Our results show that both processing steps are necessary for the effective conveyance of a message and, importantly, that the second step must be aware of the remaining noise from the beamforming operation in the first step. In addition, we study the use of the mutual information applied on the perceptually more relevant powers per critical band.**

***Index Terms*—Minimum variance distortionless response (MVDR) beamformer, mutual information, multi-microphone, speech intelligibility enhancement.**

I. Introduction

THE use of speech processing devices in our society has become widespread. Examples are public address systems, mobile telephony, hearing aids and internet telephony. The trend to make these devices more mobile also increased the expectancy that these systems can be used in all daily life situations. However, this also means that the user environment is typically also more noisy with increased intelligibility problems as a result.

In many speech communication applications we can distinguish two environments: the far-end and the near-end environment, see Fig. 1. In a typical mobile phone application or public address system, the far-end is the environment where the (noisy) target is recorded, and the near-end is the environment where the recording is played back to the listener. Note that in such applications the loudspeaker (at near-end) and microphone (at far-end) are physically separated, while with hearing assisting devices and conferencing applications the microphone and loudspeaker are typically in the same environment.

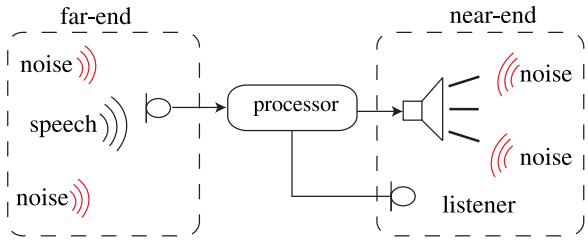

Fig. 1. A block diagram for a speech communication system.

Commonly, the target source and the listener are surrounded by other interfering acoustic sources. This can lead to a degradation of speech quality, causing an increased effort for the listener, as well as a decreased speech intelligibility. To increase the speech quality and intelligibility, it is common practice to apply noise reduction and speech enhancement algorithms to the signals recorded at the far end and played back at the near end, respectively.

The processing to overcome the effect of interfering signals present in the far-end and near-end environment have always been considered separately in the literature. To reduce far-end noise, it is common to apply single or multi-microphone noise reduction algorithms at the far-end (for example [2]–[5]), although methods operating at the near-end to remove far-end noise do exist [6]. Single-microphone methods are mostly effective for increasing the speech quality, while multi-microphone methods are able to improve the speech intelligibility as well [7].

At the near-end, the pre-processing is performed on the speech signal that is received from the far-end, to maintain its intelligibility when played out in the noisy near-end environment. One way to classify existing near-end intelligibility techniques, is with respect to the present interference in the noisy environment. A great amount of these methods focus primarily on additive sound sources that are uncorrelated with the target [6]–[14]. However, depending on the exact scenario (e.g., in the case of a public address system), reverberation might also be present at the near-end [8], [9]. In some more recent contributions the presence of both reverberation and additive noise was investigated, e.g., [13], [15], [16]. In this paper, we will neglect the presence of reverberation. However, the presented model can easily be extended to also take certain aspects of (late) reverberation into account, in a similar way as presented in [16].

Manuscript received November 7, 2016; revised February 27, 2017, April 25, 2017, and May 24, 2017; accepted May 25, 2017. Date of publication June 12, 2017; date of current version June 28, 2017. This work was supported in part by the Dutch Technology Foundation STW and Bosch Security Systems B.V. This paper was presented in part at IEEE International Conference on Acoustics, Speech, and Signal Processing, Shanghai, China, March 2016. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Simon Doclo. *(Corresponding author: Seyran Khademi.)*

The authors are with the Faculty of Electrical Engineering, Mathematics, and Computer Science, Delft University of Technology, Delft, 2628 CD, The Netherlands (e-mail: s.khademi@tudelft.nl; r.c.hendriks; w.b.kleijn@tudelft.nl).

Digital Object Identifier 10.1109/TASLP.2017.2714424

Another way to classify intelligibility enhancement methods is based on how intelligibility, perception and/or audibility is taken into account. The first class is based on somewhat more heuristic and empirical considerations. The near-end speech enhancement has been studied as early as in the late 1940s. Licklider and Pollack [17] showed that there is a negligible negative impact on intelligibility of speech by first differentiating the signal followed by infinite clipping and integration. Niederjohn and Groteluesch en proposed an effective technique [10], which is based on simple techniques, such as high-pass filtering, clipping and dynamic range compression, being motivated by the empirical finding that high-frequency and low energy components of speech (frequently belonging to consonants) promote intelligibility [18], [19]. See also e.g., [11], [12]. Even though these sort of heuristic techniques offer an acceptable complexity-performance trade-off, often extensive tuning is required and there is no optimality criteria for direct evaluation and comparison.

Therefore, the second class is based on optimization of more formal mathematical models of speech intelligibility, see e.g., [14] for an overview. To bound power usage, satisfy average loudspeaker power constraints or to overcome hearing discomfort due to loud sounds, typically, these models are optimized under an energy constraint.

Two historically important intelligibility predictors of speech in noise are the articulation index (AI) [20], [21] and the speech intelligibility index (SII) [22]. Within the near-end intelligibility enhancement context, the SII has been optimized in [23], [24]. More recently the SII has been approximated as proposed in [25] to make constrained optimization tractable. The approximated SII was used in [16] to increase the intelligibility of speech when degraded by both speech reverberation and noise. Instrumental speech intelligibility measures referred to as short-time objective intelligibility (STOI) measure [26] and the Glimpse proportion metric [27] are amongs the recently proposed measures. STOI is based on calculating normalized correlation coefficients between the temporal envelopes of the degraded and the clean speech per critical band. Although STOI was developed to predict the intelligibility of processed noisy speech, it was used in [28] to perform optimal channel selection in a cochlear implant setup.

The glimpse proportion metric [27] measures the proportion of spectro-temporal regions whose local signal-to-noise ratio (SNR) exceeds a pre-determined threshold. The glimpse proportion metric was used in [29]–[31] to optimize the intelligibility of speech in noise. A somewhat different metric is the spectro-temporal auditory model presented in [32]. This model was used in [33], [34] to optimally redistribute speech energy over frequency and time for the application where processed speech is exposed to background noise. In [15], [35] the model from [32] was adapted to the application of speech rendered in noisy reverberant environments.

Most of the aforementioned speech intelligibility metrics model certain stages of the human ear and determine speech intelligibility based SNR (e.g., [20]–[22], [27], [29]) or the correlation between the clean and noisy processed signal (e.g., [36]). As speech intelligibility expresses the amount of information that is transferred from the source, i.e., the intended message in the speakers brain, to the receiver, i.e., the listener's brain, it is natural to use information theory and express intelligibility in terms of mutual information (MI). This is an emerging theory that can be considered to unify all the qualitative and quantitative design methodologies for speech enhancement. Indeed, many of the experienced-based measures and historically developed techniques for speech enhancement fall as an special case within the overreaching mutual information framework.

A few examples of speech intelligibility predictors based on the concept of mutual information have been proposed over the last few years, [37]–[40]. The model proposed in [40] is an effective model of human communication based on MI. This model takes the noise inherent in the speech production process and the speech interpretation process into account. Unlike the environmental acoustical noise sources, the speech production noise is not a physical acoustical noise source, but can be interpreted as variations on the intended message that are introduced during the speech production process. Although not a physical noise source, it is possible to measure its variance using speech databases where the underlying message in the speech realizations are identical, see e.g., [41]. An important aspect of the speech production and the interpretation noise is the fact that they can be argued to scale with the signal, resulting in a fixed SNR. Such a constant production and/or interpretation SNR has a significant effect on a power constrained communication system. The usefulness of a particular communication channel saturates near the production SNR or the interpretation SNR, whichever is lower. Although originating from the concept of MI, the predictor from [40] resembles the heuristically derived classical measures of intelligibility such as the AI and the SII [1]. Further, as shown in [40], this measure is effective for near-end speech intelligibility enhancement.

A typical assumption of many near-end intelligibility enhancement algorithms (e.g., [16], [23], [25], [30], [31], [38], [42]) is to assume the input speech to be clean. However, in many daily life situations this assumption is invalid. Interfering sources present at the far-end and the subsequent processing to reduce them, influence the final intelligibility at the near-end. However, processing to overcome the effects of noise at the far-end and the near-end have always been considered as two separate problems.

In [1], we considered the presence of a microphone array at the far-end and the intelligibility optimization is performed with respect to the far-end and the near-end in a joint manner by taking both the disturbances at the near-end and the far-end into account. By doing so, we extended the model from [40] to also include noise sources present at the far-end. Similar to the fixed production SNR, a finite far-end SNR influences the effectiveness with which the intelligibility can be improved at the near-end. More specifically, depending on the 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 increase the power of channels with a high far-end SNR.

In addition to [1], in this work, we consider the critical band model. A thorough analysis is offered to derive a proper statistical model to evaluate the mutual information in this model. Moreover, we approach the intelligibility enhancement problem from a different point of view that provides better understanding of the problem, i.e., first by having no statistical assumption about the signal in critical bands and proving the optimality analysis regardless of the distribution of the signal and then considering the stochastic model for completion. At the end, more extensive experimental results are added and listening tests are performed to validate the proposed model and algorithm. Our analysis suggests that the Gaussian assumption of critical band powers in [40] offers a valid and accountable model for intelligibility studies. We prove that, given the transparency between the processors at the far-end and the near-end, the joint intelligibility optimization problem can be decomposed into the well-known minimum variance distortionless response (MVDR) beamformer (at far-end), followed by a linear near-end processor which redistributes the power over the frequency bands.

II. Signal Model

In this section we present the communication model, which is partially adopted from [40]. We briefly summarize the main aspects of the model presented in [40] and extend this to also take the far-end noise into account. In Section IV we further extend the model to multiple microphones.

We use standard bold upper case, bold lower case and normal symbols to indicate matrices, vectors and scalars, respectively. Scalar stochastic variables are distinguished by uppercase letters whereas stochastic vector and matrix variables are left to be recognized from the context. The conjugate, conjugate transpose and inverse of a matrix $\mathbf{X}$ are denoted as $\mathbf{X}^*$, $\mathbf{X}^H$ and $\mathbf{X}^{-1}$, respectively. The cardinality of a set A is denoted by its zero-norm as $||A||_0$.

The speech process S is assumed to be a sequence of complex random vectors, with each coefficient $S_{k,i}$ describing a complex DFT at frequency-bin index k and time-frame index i, so the signal model is represented in DFT domain. We take the natural variation of speech, referred to as the production noise $Q_{k,i}$, into account. The produced speech is the convolution of the human vocal tract filter and the excitation signal in time domain[1]. The excitation signal can be considered as the carrier, while the vocal tract can be considered as the information. To take the natural variations of speech into account, i.e., the production noise, we introduce the production noise by means of the variable $Q_{k,i}$ that is thus assumed here to be additive on the intended speech message S like [40]. As a result, the acoustic signal produced at time-frequency point (k, i) is given by

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

We make the plausible assumption that the SNR due to the noise Q in the produced speech signal is independent of the presentation level. This can be explained by the fact that the inter and intra person variations of a specific speech sound scale with the representation level. As an important consequence, the speech production SNR, remains constant.

[1]This is equivalent to multiplication in frequency domain, which can be modeled as a multiplicative noise [41], however we consider the production noise to be additive for the sake of simplicity.

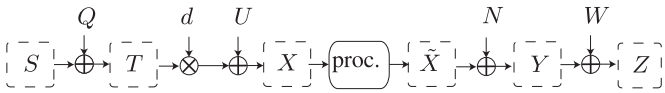

Fig. 2. System model schematic.

The noise present at the far-end is denoted by $U_{k,i}$. The spectral coefficients of the recorded signal are therefore

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

where $d_{k,i}$ is the room transfer function from source to the microphone.

To increase the intelligibility of the interpreted message, the recorded signal is being processed prior to rendering in the noisy near-end environment. The modified coefficients are denoted by $\tilde{}$. The signal as received by the observer is contaminated by the noise in the near-end environment $N_{k,i}$.

As proposed in [40], similar to the production noise, natural variation of the human auditory system including internal noise, the absolute hearing threshold, variations in the message interpretation process and an increased hearing threshold can be modeled by an additional noise source, which we refer to as *interpretation noise* and is denoted by $W_{k,i}$. Part of $W_{k,i}$ scales with the signal. This is consistent with the notion of instantaneous masking (e.g., [43]). As a consequence, we assume that the interpretation SNR ($\sigma^2_{Y_k}/\sigma^2_{W_k}$), remains constant.

The schematic overview in Fig. 2 illustrates the considered signal model. In the absence of the processor, this is summarized as

$$Z_{k,i} = d_{k,i}S_{k,i} + d_{k,i}Q_{k,i} + U_{k,i} + N_{k,i} + W_{k,i}, \tag{3}$$

where $d_{k,i}, U_{k,i}$ and $N_{k,i}$ are environmental distortions while $Q_{k,i}$ and $W_{k,i}$ are natural variations.

The considered signal model constitutes a Markov chain, i.e., $S_{k,i} \rightarrow T_{k,i} \rightarrow X_{k,i} \rightarrow \tilde{X}_{k,i} \rightarrow Y_{k,i} \rightarrow Z_{k,i}$. The relation at each Markov step is described by the correlation coefficient between the corresponding variables, which is a measure of dependency.

We need to find an effective objective measure for designing the processor that maximizes the correlation between the original signal S and the perceived signal Z. This is studied in the next section and the optimal linear processor is discussed and developed in the subsequent sections.

III. Mutual Information Measure

Mutual information is a unique measure of dependence between two arbitrary random variables i.e., the information one can obtain about random variable S by observing Z. This is by definition the difference between the differential entropy in variable S ($h(S)$) and the conditional differential entropy ($h(S|Z)$). The functions $p(s)$, $p(z)$ and $p(s, z)$ are the probability density function (pdf) of the random variables S and Z, and the joint pdf of S and Z, respectively. Mutual information is quantified

by the unit of bit or nat per communication depending on the base of the logarithm [44], which is given by

$$I(S;Z) := h(S) + h(Z) - h(S,Z) = \int_{-\infty}^{+\infty} p(s,z) \log \frac{p(s,z)}{p(s)p(z)} dsdz. \quad (4)$$

The mutual information for two random variables does not change by scaling the variables. Note that, this holds for any invertible and differentiable mapping of the random variables.

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 $\mathbf{z}_i$, denoted by $I(\mathbf{s}_i;\mathbf{z}_i)$, describes the effectiveness of the communication process in this context. This is dependent on the SNR of the system (ξ_k) at each frequency bin, which varies with the processor, and consequently is related to $\rho_{T_{k,i}\tilde{X}_{k,i}}\rho_{\tilde{X}_{k,i}Y_{k,i}}$. These are product moment correlation coefficients between two arbitrary random variables, which is defined as $\rho_{XY} = \frac{\mathrm{cov}(X,Y)}{\sigma_X \sigma_Y}$, where the covariance for the zero-mean complex-valued random variables is defined by $E\{XY^*\}$ and the denominator is the product of the standard variation of the two variables.

We summarize and discuss the assumptions that we make to further introduce the enhancement operator based on the mutual information measure:

1) The processing is performed by a linear time-invariant operator (gain), which implies that $\rho_{T_{k,i}\tilde{X}_{k,i}} = \rho_{T_{k,i}X_{k,i}}$.
2) All processes are stationary, and memoryless so we omit the time-frame index i for notational convenience, i.e., $\rho_{S_{k,i}Z_{k,i}} = \rho_{S_kZ_k}$.
3) The speech signal and the environmental noise processes are uncorrelated, i.e., $E\{T_kU_k\} = \mathbb{E}\{T_k\}E\{U_k\}$: This assumption is critical for the mathematical tractability of the problem, i.e., valid to a good extent unless the environment is reverberant.
4) The signal model follows the Markov chain model as $S_k \to T_k \to X_k \to \tilde{X}_k \to Y_k \to Z_k$, and the correlation coefficient is given by $\rho_{S_kZ_k} = \rho_{S_kT_k}\rho_{T_k\tilde{X}_k}\rho_{\tilde{X}_kY_k}\rho_{Y_kZ_k}$: Under the Markov chain property, each random variable in the chain is conditionally dependent of only the previous one. This requires the individual sources to be independent from each other which is already assumed in our model.
5) Individual component signals in the time-frequency representation are independent so we can then write

$$I(\mathbf{s}_i;\mathbf{z}_i) := I(S;Z) = \sum_k I(S_k;Z_k). \quad (5)$$

Considering the independence of DFT coefficients across neighboring frequency bands is a crucial simplification, which is however commonly made within the context of speech enhancement, e.g., [45]. We emphasize that obviously, neighboring DFT bins in short time frames are dependent as naturally the speech modulation ranges across critical bands. Thus, this assumption is merely made for mathematical tractability.

6) The mutual information at the kth frequency bin is an increasing function of the SNR at that band (ξ_k), regardless of the distribution of S and Z: The mutual information between S and Z is a nonlinear function of the overall correlation coefficient and hence the SNR, which itself is a function of both far-end and near-end SNR, that we aim to maximize by tuning the processor. In fact, the mutual information between S and Z is increased via pre-processing of the speech prior to play back in the noisy near-end environment. This assumption complies with data processing inequality and we show later that it is correct for the two proposed distribution of the critical bands.
7) The production and interpretation noise are independent of the presentation level and they are represented by a fixed gain at each frequency band denoted by $\rho_{0_k} := \rho_{S_{k,i}T_{k,i}}\rho_{Y_{k,i}Z_{k,i}}$ which is a fixed number between $[0,1]$: For production noise, we would expect some variation in the production SNR if a person attempts to talk at a "not normal level" or strains to talk in loud noise, this is not considered in our model.

   The fixed-SNR model for interpretation noise is not valid for very high and very low listening levels. However, this is entirely consistent with the notion of having a masking curve, which has a level relative to the masker. Thus, we expect it to be an accurate representation when conventional masking models are accurate. Notice that although we did not take masking explicitly into account, the concept of interpretation noise that is introduced is to a high degree related to the concept of masking.

The maximum mutual information is obtained by giving $\tilde{X}_{k,i}$ the maximum amount of power prior to degradation by the environmental noise $N_{k,i}$. Setting the power to infinity, the mutual information will be infinite, i.e., infeasible in physical systems. Limiting the power in a band thus decreases the mutual information in that band. In this paper we consider the situation where the total power summed across all frequency bands is constrained. The optimal power distribution across frequency bands for mutual information maximization is then expected to depend on the power of the noise sources present at the different stages of the communication chain in the different frequency bands.

So far, we have not made any assumptions on the distribution of the random variables, nevertheless, we assume that the mutual information is an increasing function of SNR. Later, we verify this assumption for our proposed statistical model. In the following, we extend our signal model to a multi-microphone system in order to partly suppress the far-end noise that affects the speech intelligibility.

IV. Optimal Linear Processor

The correlation between the intended and perceived message is a function of SNR. In this work we assume linear time invariant processing and knowledge of the second order statistical information about the signal and present noise sources. Below we will extend the signal model from Section II to multiple microphones.

A. Multi-Microphone Signal Model

Let the linear processor for multi-channel processing be denoted by $\mathbf{v}_k$. In the multi-microphone setting, we aim to find, for each time-frequency component, a linear processor $\mathbf{v}_k$ that maximizes the mutual information between S and Z subject to a power constraint. We denote the acoustic transfer function from source to microphone m by $d_{k,m}$ and we use the following vector notation i.e., $\mathbf{d}_k = [d_{k,1}, ..., d_{k,M}]^T$. Let the far-end noise recorded by the microphones be denoted by the vector $\mathbf{u}_k$. The processed noisy microphone data is then given by

$$\tilde{X}_k = \mathbf{v}_k^H \mathbf{d}_k T_k + \mathbf{v}_k^H \mathbf{u}_k, \tag{6}$$

with T_k being the clean speech available at the source location for the kth frequency bin and $\mathbf{v}_k$ the multi-microphone processor. Given the communication model in Fig. 2, we can define a SNR at several positions in this channel. At first we define the acoustical SNR ξ_k, being the ratio between the variance of the speech process T (i.e., including production noise), and the variance of the environmental noise processes, that is,

$$\xi_k = \frac{\mathbf{v}_k^H \mathbf{d}_k \mathbf{d}_k^H \mathbf{v}_k \sigma_{T_k}^2}{\mathbf{v}_k^H \mathbf{R}_{U_k} \mathbf{v}_k + \sigma_{N_k}^2}, \tag{7}$$

where the far-end noise covariance matrix at the microphones is given by $\mathbf{R}_{U_k} = E\{\mathbf{u}_k \mathbf{u}_k{}^H\}$. Note that (7) describes the SNR of the speech including production noise, i.e., process T_k, with respect to the environmental noise sources only. The SNR value, ξ_k, is bounded by the remaining far-end noise $\mathbf{v}_k^H \mathbf{R}_{U_k} \mathbf{v}_k$, consequently, reducing the far-end noise makes the bound on the final SNR (ξ_k) less tight. Secondly, we can also define the SNR between the message S and the production noise Q. Note that the speech production noise Q, present in process T, already limits the SNR of the speech process S (the intended message), where no linear processing of process T can increase this SNR. This effectively limits the SNR of the overall communication channel.

The linear processor $\mathbf{v}_k$ is designed to maximize the mutual information with respect to the average power constraint. For now we do not assume any statistical model for the signal. The mutual information is considered to be a generic increasing function of the SNR denoted by $f(\xi_k(\mathbf{v}_k)) : \mathbb{C}^M \rightarrow \mathbb{R}_+$, where $\mathbb{R}_+$ is a set of real and positive scalars. This assumption facilitates the proof of the optimal linear processor regardless of the distribution of the signal.

So far, the signal model is built up based on the complex DFT domain. In Section V, we consider the critical band model for human auditory systems where the linear processor is determined for each critical band rather than every single complex DFT bin.

B. Spatial Processing

In this subsection, we derive the optimal multi-channel linear processor. The optimization problem for intelligibility improvement subject to the total power constraint is given by

$$\begin{aligned} \sup_{\{\mathbf{v}_k\} \in \mathbb{C}^M} \quad & \textstyle\sum_k f\left(\frac{\mathbf{v}_k^H \mathbf{d}_k \mathbf{d}_k^H \mathbf{v}_k \sigma_{T_k}^2}{\mathbf{v}_k^H \mathbf{R}_{U_k} \mathbf{v}_k + \sigma_{N_k}^2}\right) \\ \text{subject to} \quad & \textstyle\sum_k \mathbf{v}_k^H \mathbf{d}_k \mathbf{d}_k^H \mathbf{v}_k \sigma_{T_k}^2 = \sum_k \sigma_{T_k}^2, \end{aligned} \tag{8}$$

where the objective denotes the mutual information between S_k and Z_k which is given in terms of the processor $\mathbf{v}_k$.

The mutual information measure in (8) includes a sum of non-linear fractional terms which, in general, can not be transformed into standard convex programming framework [46]. To find an optimizer for (8), we introduce a variable transformation $\mathbf{v}_k = \alpha_k^{1/2} \mathbf{w}_k$ together with appending an additional constraint $\mathbf{v}_k^H \mathbf{d}_k \mathbf{d}_k^H \mathbf{v}_k = \alpha_k$. The variable α_k is a real and positive scalar. Then (8) can be rewritten as

$$\begin{aligned} \sup_{\mathbf{v}_k \in \mathbb{C}^M, \alpha_k \in \mathbb{R}_+} \quad & \textstyle\sum_k f\left(\frac{\mathbf{v}_k^H \mathbf{d}_k \mathbf{d}_k^H \mathbf{v}_k \sigma_{T_k}^2}{\mathbf{v}_k^H \mathbf{R}_{U_k} \mathbf{v}_k + \sigma_{N_k}^2}\right) \\ \text{subject to} \quad & \mathcal{C}_1 : \textstyle\sum_k \alpha_k \sigma_{T_k}^2 = \sum_k \sigma_{T_k}^2 \\ & \mathcal{C}_2 : \mathbf{v}_k^H \mathbf{d}_k \mathbf{d}_k^H \mathbf{v}_k = \alpha_k, \ \forall k. \end{aligned} \tag{9}$$

This is an equivalent problem to (8), where the objective function can be rewritten in terms of $\mathbf{w}_k$ and α_k:

$$I(S_k; Z_k) = f(\xi_k(\alpha_k, \mathbf{w}_k)) = f\left(\frac{\alpha_k \sigma_{T_k}^2}{\alpha_k \mathbf{w}_k^H \mathbf{R}_{U_k} \mathbf{w}_k + \sigma_{N_k}^2}\right).$$

The constraints are independently separated into $\mathcal{C}_1$ and $\mathcal{C}_2$. This step, although simple, is important for the final decomposition of the problem, using the fact that in general $\sup_{x,y} f(x,y) = \sup_x \sup_y f(x,y)$ (see [46, p. 133]), (9) can be rephrased as

$$\sup_{\alpha_k \in \mathbb{R}_+, \mathcal{C}_1} \ \sup_{\mathbf{w}_k \in \mathbb{C}^M, \mathcal{C}_2} \ \textstyle\sum_k f(\xi_k(\alpha_k, \mathbf{w}_k)). \tag{10}$$

The benefit appears from the independency of the constraints, which makes it amenable for standard solutions. The inner maximization problem in (10) over $\mathbf{w}_k$ leads to the standard MVDR beamforming problem where the solution is given by [4] as

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

Using $\mathbf{w}_k^*$, the outer maximization in (10) is only over the α_k variables, hence (9) simplifies to

$$\begin{aligned} \sup_{\alpha_k \in \mathbb{R}_+} \quad & \textstyle\sum_k f\left(\frac{\alpha_k \sigma_{T_k}^2}{\alpha_k \sigma_{M_k}^2 + \sigma_{N_k}^2}\right) \\ \text{subject to} \quad & \textstyle\sum_k \alpha_k \sigma_{T_k}^2 = \sum_k \sigma_{T_k}^2 \end{aligned} \tag{11}$$

where $\sigma_{M_k}^2 = \mathbf{w}_k^{*H} \mathbf{R}_{U_k} \mathbf{w}_k^*$ is the far-end noise that remains after processing by the MVDR beamformer. Thus, to solve (11) it is required to know at least the output noise of the MVDR processing from the far-end side. This is a crucial step in the decomposition of the processor. The problem in (11) is a convex problem provided that $f(\xi_k(\alpha_k))$ is a convex function of α_k since the sum of convex functions is convex.

The MVDR beamformer is the optimal linear pre-processor ($\mathbf{w}_k$) with respect to any intelligibility measure that is an increasing function of SNR, including the mutual information measure, and is operating in the complex DFT domain. This can be easily shown, e.g., for the ASII measure [25] as well. In fact, the MVDR power is the sufficient statistics for the near-end proceeding operation [47], [48].

V. Mutual Information for Critical Bands

As the ultimate recipient of the final signal is a human, we investigate in this section the post-processor (α_k) based on the perceptually motivated critical band model with real positive (power) values. The main challenge of taking the critical band model into account is to properly model the distribution of the signal within the critical bands in order to calculate the explicit mutual information expression.

A. Critical Band Signal Model

The critical band representation assumes an auditory filter bank that models the processing at the level of the basilar membrane in the cochlea. The effect of this filter bank is to transform the spectral information to real positive values that represent the envelope of the acoustic signal in the frequency domain. Then, adjacent DFT bins are grouped and their powers are summed to form the corresponding spectral power in the critical band i.e., the output of the human ear to be processed by the brain [49]. Instrumental intelligibility measures commonly work on critical band amplitudes, e.g., calculated as

$$\bar{S}'_{m,i} = \sqrt{\sum_{k\in B_m} |S_{k,i}|^2}\,, \tag{12}$$

where B_m denotes a set of spectral coefficient indices that belong to the mth critical band. The square root used in calculating critical band amplitudes often complicates analytical derivations. Using critical band powers (instead of magnitudes) simplifies the expression. In this section we derive an exact expression for the mutual information under the model introduced in Section II, using critical band powers. We define the speech critical band power as

$$\bar{S}_{m,i} = \sum_{k\in B_m} |S_{k,i}|^2, \tag{13}$$

and the processed noisy speech critical band power as

$$\bar{Z}_{m,i} = \sum_{k\in B_m} |Z_{k,i}|^2. \tag{14}$$

In a similar way as $\bar{S}_{m,i}$ and $\bar{Z}_{m,i}$ we define $\bar{N}_{m,i}$, $\bar{M}_{m,i}$ and $\bar{T}_{m,i}$, i.e., $\bar{T}_{m,i} = \sum_{k\in B_m} |\mathbf{w}_k^H \mathbf{d}_k T_{k,i}|^2$, and the processed noise from the far-end is denoted by $\bar{M}_{m,i} = \sum_{k\in B_m} |\mathbf{w}_k^H \mathbf{u}_{k,i}|^2$. In turn, $\bar{X}_{m,i}$ and $\bar{\tilde{X}}_{m,i}$ are related by

$$\bar{\tilde{X}}_{m,i} = \alpha_m \bar{X}_{m,i}. \tag{15}$$

A critical issue is the form that the distribution of $\bar{S}_m$ and $\bar{Z}_m$ take, as it is the key to finding the mutual information between $\bar{S}_m$ and $\bar{Z}_m$ (assuming stationary speech and dropping the time index i). Further, we need to derive $f(\xi_m(\alpha_m))$ to reformulate (11) based on the critical band model.

Random variables $\bar{S}_m$ and $\bar{Z}_m$ can be argued to be Chi-squared distributed with $l_m = 2||B_m||_0$ degrees of freedom [50], being twice the number of frequency coefficients in mth critical band. This is based on the assumption that the individual complex DFT coefficients are independent and identically Gaussian distributed, i.e., assuming within a critical band the power spectral density is fixed. Note that the latter assumption, on Gaussianity of DFT coefficients, does not change the MVDR optimality results from Section IV. Note that, in reality DFT coefficients are not independent because they are caused by physical modulation.

The factor two in the definition of l_m is due to the fact that the DFT bins are complex Gaussian variables. This means that l_m is at least two when the cardinality of B_m equals one. In this case, the Chi-squared variable is the power of a single complex coefficient of a DFT bin. The differential entropy for Chi-squared random variables with l_m degrees of freedom is a known expression derived in [51] so the mutual information can be derived accordingly.

B. Mutual Information for Chi-Squared Random Variables

The mutual information expression for two arbitrary Chi-squared random variables $\bar{S}$ and $\bar{Z}$ can be derived using the results from [51] and is given by

$$\begin{aligned} I(\bar{S};\bar{Z}) &= l\frac{\rho^2}{\rho^2-1} - \frac{l}{2}\log(1-\rho^2) \\ &\quad + E\left\{ \log\left(\Gamma\left(\frac{l}{2}\right)\left(\frac{\sqrt{\rho^2\bar{S}\bar{Z}}}{2(1-\rho^2)}\right)^{1-\frac{l}{2}} \mathrm{I}_{\frac{l}{2}-1}\left(\frac{\sqrt{\rho^2\bar{S}\bar{Z}}}{1-\rho^2}\right)\right)\right\} \\ &= l\frac{\rho^2}{\rho^2-1} - \frac{l}{2}\log(1-\rho^2) + \log\left(\Gamma\left(\frac{l}{2}\right)\right) \\ &\quad + \left(1-\frac{l}{2}\right)\left(\log\left(\frac{\sqrt{\rho^2}}{2(1-\rho^2)}\right) + \frac{1}{2}\underbrace{E\{\log(\bar{S}\bar{Z})\}}_{2\psi(\frac{l}{2})+2\log(2)}\right) \\ &\quad + E\left\{\log\left(\mathrm{I}_{\frac{l}{2}-1}\left(\frac{\sqrt{\rho^2\bar{S}\bar{Z}}}{1-\rho^2}\right)\right)\right\}. \end{aligned} \tag{16}$$

where ρ is the correlation coefficient between arbitrary Gaussian variables, that are summed to form the Chi-squared variables as defined in [51]. Note that if the Gaussian variables are complex-valued, as in our setting, then the covariance and hence the correlation coefficient, in general, are complex [52, Chapter 7], so we define $\rho^2 := \rho\rho^*$ here. In turn, l is the number of Gaussian variables in the summation. The extended derivation is given in Appendix A. The last expectation term in (16) does not have a known closed form expression. This complicates analytical evaluation of the mutual information for the Chi-squared distributions.

The Chi-squared distribution is the sum of l independent random variables with finite mean and variance, so it converges to a normal distribution for large l, according to the central

limit theorem. It is reported in [53] that for $l > 50$, the Chi-squared distribution is very close to a normal distribution and the difference can be ignored. Accordingly for large l we expect that the distribution of $\bar{S}$ and $\bar{Z}$ converges to a Gaussian distribution with a known mutual information [44]. Assuming, that based on the central limit theorem, $\bar{S}$ and $\bar{Z}$ are (non-zero mean) Gaussian distributed, an approximation of their mutual information is given by

$$I(\bar{S};\bar{Z}) \approx -\frac{1}{2}\log(1-\bar{\rho}^2), \tag{17}$$

where $\bar{\rho}$ is the correlation coefficient of two Chi-squared random variables. The correlation coefficient for $\bar{S}$ and $\bar{Z}$ can be easily derived since we know the mean and variance of Chi-squared random variables as $E\{\bar{S}\} = E\{\bar{Z}\} = l$,[2] in turn

$$E\{(\bar{S} - E\{\bar{S}\})^2\} = E\{(\bar{Z} - E\{\bar{Z}\})^2\} = 2l, \tag{18}$$

and from [51] we have

$$E\{\bar{S}\bar{Z}\} = l^2 + 2l\rho^2. \tag{19}$$

Therefore, the correlation coefficient for two correlated Chi-squared distributions is given by

$$\bar{\rho} = \frac{\mathrm{cov}\{\bar{S}\bar{Z}\}}{\sqrt{\mathrm{var}\{\bar{S}\}\mathrm{var}\{\bar{Z}\}}} = \frac{l^2 + 2l\rho^2 - l^2}{2l} = \rho^2\,. \tag{20}$$

Hence, the approximated mutual information for large l is given by

$$I(\bar{S};\bar{Z}) \approx -\frac{1}{2}\log(1-\rho^4). \tag{21}$$

where again ρ is the correlation coefficient of Gaussian random variables that are summed to form $\bar{S}$ and $\bar{Z}$, assuming they have equal correlation coefficients and $\rho^4 = (\rho\rho^*)^2$.

There are empirical ways to verify the mutual information expression in (16). To evaluate the mutual information for a pair of arbitrary Chi-squared random variables $\bar{S}$ and $\bar{Z}$, and compare it to the Gaussian approximation in (21) in particular for smaller values of l, we evaluate the expectation sentence in (16) empirically using a large number of generated Chi-squared random variables. We evaluated the mutual information as a function of the correlation coefficient. The results are discussed in Appendix B.

So far we discussed the mutual information for arbitrary Chi-squared random variables. In the following we extend the discussion to our speech signal model. The mutual information maximization problem in (11) is reformulated based on the critical band model. However, even though the Chi-squared model might be more accurate with respect to auditory modeling, the expression for the MI in (16) is unknown in a closed form. We will show later that, even after the approximation of (16) to (21), the objective function for near-end processing turns out to be non-convex.

[2]This follows from the normalization assumption that $E\{|S_k|^2\} = E\{|Z_k|^2\} = 2$ as introduced in [51]. Note that neither ρ nor the mutual information is affected by the unit variance assumption on the real and imaginary elements of the complex Gaussian random variables.

VI. Near-End Processing

In this section, we use plain symbols to denote the critical band variables, e.g., S_m and Z_m. We also use the Gaussian approximation of the Chi-squared distribution given in (21). For this, the mutual information measure for the mth critical band is given by

$$f(\xi_m(\alpha_m)) = -\frac{1}{2}\log(1-\rho^4_{S_m Z_m}), \tag{22}$$

where $\rho_{S_m Z_m}$ refers to the equivalent correlation coefficient of the Gaussian DFT bins within a critical band. As we assume the variances of DFT coefficients within a critical band to be equal, we calculate the variance within a critical band by

$$\sigma^2_{T_m} = E\Big\{\sum_{k\in B_m}|T_{k,i}|^2\Big\} = \sum_{k\in B_m}E\{|T_{k,i}|^2\} = \sum_{k\in B_m}\sigma^2_{T_k}.$$

The overall correlation coefficient for the mth critical band is then given by $\rho^2_{S_m Z_m} = \rho^2_{0_m}\rho^2_{T_m Y_m}$ with $\rho^2_{T_m Y_m}$ capturing the environmental noise effect and is given by

$$\rho^2_{T_m Y_m} = \frac{\alpha_m\sigma^2_{T_m}}{\alpha_m\sigma^2_{T_m} + \alpha_m\sigma^2_{M_m} + \sigma^2_{N_m}}. \tag{23}$$

Thus, the optimization problem in (11), under the Chi-squared model, is given by

$$\begin{aligned} &\sup \quad -\tfrac{1}{2}\textstyle\sum_m \log\Big(1-\big(\frac{\rho^2_{0_m}\alpha_m\sigma^2_{T_m}}{\alpha_m\sigma^2_{T_m}+\alpha_m\sigma^2_{M_m}+\sigma^2_{N_m}}\big)^2\Big)\\ &\alpha_m \in \mathbb{R}_+\\ &\text{subject to } \textstyle\sum_m \alpha_m\sigma^2_{T_m} = \sum_m \sigma^2_{T_m}\ . \end{aligned} \tag{24}$$

Unfortunately, (24) is not a convex problem with respect to α_m, and taking the Lagrangian and solving the KKT conditions lead to a quartic equation on α_m. Quartic equations can be solved with numerical algorithms, however, this may lead to complex roots or double repeated roots due to the non-convexity of the problem. To overcome this problem we take a step further to replace the mutual information measure in (22) with a proper convex function that can be used for the purpose of intelligibility enhancement.

A. Proposed Processor

In order to have an effective and reliable intelligibility measure that can be optimized using efficient optimization algorithms, at least two properties must be satisfied:

1) $f(\xi_m(\alpha_m))$ is a concave function of α_m,
2) $f(\xi_m(\alpha_m)) \geq 0 \quad \forall\rho_m$.

A combination of these two criteria assures that the intelligibility measure gives a valid communication rate while the optimal power allocation weights can be found using the available convex optimization techniques. The calculated mutual information in (23) for the Chi-squared distributed critical power bands does not meet the first criterion. Thus, we use the

following factorization

$$-\frac{1}{2}\log(1-\rho^4_{S_m Z_m}) = -\frac{1}{2}\log(1-\rho^2_{S_m Z_m}) - \frac{1}{2}\log(1+\rho^2_{S_m Z_m}). \quad (25)$$

The first term in (25) is readily recognizable as the mutual information assuming S_m and Z_m are Gaussian distributed. The second term is always negative which means that it reduces the mutual information with respect to the mutual information between S_m and Z_m ($I(S_m, Z_m)$), assuming they have a Gaussian distribution. Hence, a simple projection to a proper measure is to discard the second term in (25), which is not a valid term with respect to the second criterion. Therefore, we stick to the first term in (25) for the speech enhancement problem. It is worth mentioning that this approximation is particularly tight when $\rho_{S_m Z_m}$ is close to one (high SNRs) since the second term in (25) approaches $\log(2)$ which is negligible compared to the first term.

By replacing the fourth-order power of ρ_{S_m, Z_m} in (22) by a quadratic power, the proposed intelligibility measure is

$$f(\xi_m(\alpha_m)) = -\frac{1}{2}\log(1-\rho^2_{S_m, Z_m}). \quad (26)$$

This means that we effectively assume the critical band powers to be zero-mean Gaussian random variables. Even though this may overestimate the actual mutual information, it provides a convex function in terms of α_m. Now, we can solve for the amplification factors α_m assuming the above proposed model. It is worth mentioning that even though the objective in (24) is not always a convex function of α_m (depending on the near-end SNR), it may predict the mutual information rate in the speech more accurately. This needs more through investigation which is out of the scope of this work.

B. Problem Formulation

Applying the above argued approximation to (24), we obtain the problem formulation

$$\begin{aligned} &\sup \quad -\tfrac{1}{2}\textstyle\sum_m \log\left(1-\frac{\rho^2_{0_m}\alpha_m\sigma^2_{T_m}}{\alpha_m\sigma^2_{T_m}+\alpha_m\sigma^2_{M_m}+\sigma^2_{N_m}}\right) \\ &\alpha_m \in \mathbb{R}_+ \\ &\text{subject to } \textstyle\sum_m \alpha_m\sigma^2_{T_m} = \sum_m \sigma^2_{T_m}. \end{aligned} \quad (27)$$

The Lagrangian for this problem is given by

$$\begin{aligned} &\mathcal{L}(\{\alpha_m\},\lambda,\{\mu_m\}) = \\ &-\frac{1}{2}\sum_m \log\left(1-\frac{\rho^2_{0,m}\alpha_m\sigma^2_{T_m}}{\alpha_m\sigma^2_{T_m}+\alpha_m\sigma^2_{M_m}+\sigma^2_{N_m}}\right) \\ &+\lambda\left(\sum_m \alpha_m\sigma^2_{T_m} - \sum_m \sigma^2_{T_m}\right) + \mu_m\alpha_m, \end{aligned} \quad (28)$$

where λ and μ_m are nonnegative. Taking the partial derivative of (28) with respect to α_m and putting it to zero leads to a stationary condition which gives the μ_m variables in terms of the other variables.

$$\begin{aligned} &\frac{\partial\mathcal{L}(\{\alpha_m\},\lambda,\{\mu_m\})}{\partial\alpha_m} = \frac{1}{2}\frac{\sigma^2_{T_m}+\sigma^2_{M_m}}{\alpha_m(\sigma^2_{T_m}+\sigma^2_{M_m})+\sigma^2_{N_k}} \\ &-\frac{1}{2}\frac{(1-\rho^2_{0_m})\sigma^2_{T_m}+\sigma^2_{M_m}}{\alpha_m(1-\rho^2_{0_m})\sigma^2_{T_m}+\alpha_m\sigma^2_{M_m}+\sigma^2_{N_m}} \\ &+\lambda\sigma^2_{T_m}-\mu_m = 0. \end{aligned} \quad (29)$$

Then going to the complementary slackness condition we have $\mu_m\alpha_m = 0, \forall k$, i.e., for nonzero α_ms, μ_m needs to be zero so α_m is given by putting $\mu_m = 0$ which itself is a quadratic equation in α_m and can lead to negative solutions, however, only a positive α_m is valid and the negative ones are put to zero. The quadratic function is given by

$$a_m\alpha_m^2 + b_m\alpha_m + c_m = 0, \quad (30)$$

where $\alpha_m = \frac{-b_m \pm\sqrt{{b_m}^2-4a_m c_m}}{2a_m}$ and

$$a_m = -(\sigma^2_{T_m}+\sigma^2_{M_m})((1-\rho^2_{0_m})\sigma^2_{T_m}+\sigma^2_{M_m})\lambda \quad (31)$$

$$b_m = -((2-\rho^2_{0_m})\sigma^2_{T_m}+2\sigma^2_{M_m})\sigma^2_{N_m}\lambda \quad (32)$$

$$c_m = \frac{1}{2}\rho^2_{0_m}\sigma^2_{N_m} - \sigma^4_{N_m}\lambda. \quad (33)$$

The variables $\{a_m\}$ and $\{b_m\}$ are always negative and $\{\alpha_m\}$ need to be nonnegative. We show in Appendix C that $b_m^2 - 4a_m c_m \geq 0$, therefore (30) always possesses real roots. Nevertheless, we need another equation to eliminate λ, therefore, we invoke the power equality constraint to overcome this ambiguity. This problem possesses a unique solution although it is not in closed form. The solution is reached based on a bisection algorithm to alternate between the (30) and the power constraint within a range for λ. This can be summarized as follows, where here the superscript (κ) is the iteration counter.

1) select $\lambda^{(\kappa)} \in [\lambda_{\min}, \lambda_{\max}]$
2) solve (30) for $\lambda^{(\kappa)}$
3) set any negative $\alpha_m^{(\kappa)}$ to zero
4) check if the total power $\sum_m \alpha_m^{(\kappa)}\sigma^2_{T_m}$ is sufficiently close to the desired power, if yes stop and take $\{\alpha\}^{(\kappa)}$ as the optimal weights.
5) If not, then adjust $\lambda^{(\kappa+1)}$ to be more negative if the power is higher and more positive if the power is lower.

The algorithm converges very fast in general and the values for $\lambda_{\min}$ and $\lambda_{\max}$ are chosen as a very small (close to zero) and very large negative numbers, respectively. Once $\{\alpha_m\}$ are found they are applied to the critical bands, hence the same weights are used for all the frequency bins inside each critical band.

VII. Experimental Results

The goal of this section is to show that, as our theory indicated, the optimal linear processor can be implemented by two consecutive processors, provided that the second processor is aware of both the remaining far-end noise and the processing operation applied at the first processor. We first discuss our reference methods and then the experimental setup, followed

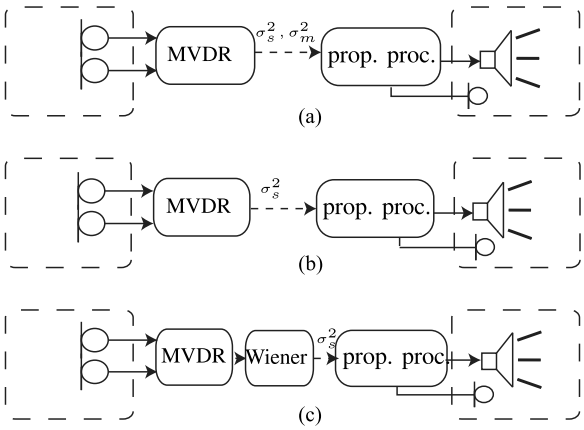

Fig. 3. Speech communication system with two processors. (a) Proposed transparent processors (MVDR+ Eq. (27)). (b) Blind processing (MVDR + [40]. (c) Blind processing (MVDR + Wiener + [40]).

by the results from instrumental measures and finally results originated from listening tests.

A. Reference Methods Based on Disjoint Processing

In Sections IV-B and VI we have derived the jointly optimal processor with respect to the noise at the near-end and the far-end. In this section we will investigate the consequence of having two processors, one spatial processor at the far-end and the proposed processor at the near-end. Given two processors, the initial model in Fig. 1 is changed into the one depicted in Fig. 3. In the case of two processors, the speech processing in a communication system can be categorized as a *transparent* or a *blind* approach. In the *transparent* mode, the processor at the near-end is aware of the remaining noise and speech power after far-end processing. This information is then assumed to be transmitted to the near-end as side information for the processor design. In such a transparent system, it can be shown that *any* linear (non-zero) processing at the far-end, will be completely compensated for by the proposed near-end processor. Hence the enhancement processor consists of a linear processing (scaling) by β_m. The problem formulation in (27) is then given by

$$\begin{aligned} &\sup \quad -\tfrac{1}{2}\textstyle\sum_m \log\left(1-\frac{\rho_{0m}^2\,\alpha_m\,\beta_m\,\sigma_{T_m}^2}{\alpha_m\,\beta_m\,\sigma_{T_m}^2+\alpha_m\,\beta_m\,\sigma_{M_m}^2+\sigma_{N_m}^2}\right) \\ &\alpha_m \in \mathbb{R}_+ \\ &\text{subject to } \textstyle\sum_m \alpha_m\beta_m\,\sigma_{T_m}^2 = \sum_m \sigma_{T_m}^2\,. \end{aligned} \tag{34}$$

The problem in (34) can be rephrased by substituting $G = \alpha_m\beta_m$, after which α_m is given by $\alpha_m = \frac{G}{\beta_m}$, which shows that the processing by β_m is completely compensated.

In the case of blind processing, the proposed near-end processor is blind to the far-end processing. We consider two different blind methods. At first one where the signal after MVDR processing is (erroneously) assumed to be noise free. The proposed near-end processor is thus blind to the remaining far-end noise. This reference method is depicted in Fig. 3(b) and denoted by (MVDR+[40]).

The second blind reference method is one where the MVDR output is processed prior to the processing at the near-end. A valid question posed here is on the best blind processing that minimizes the error between the disjoint and joint processing approaches. An evident candidate is a far-end processor that minimizes the noise remaining from the MVDR process, i.e., proved to be the Wiener processor. The Wiener gain is applied after MVDR beamformer at far-end, where the near-end processor calculates the proposed processor coefficients (α_m) assuming $\sigma_m^2 = 0$, i.e., assuming that the Wiener filter has perfectly removed all the noise at the far-end. This blind near-end processor was introduced in [40] and the combination with MVDR and a standard time-invariant multi-channel Wiener filter is considered as a reference blind method in this work, refered to as MWF+ [40]. We assume that the original overall clean speech power is known by the near-end processor. Moreover, we introduce MWF+ [25] and MVDR+ [25] where the ASII measure from [25] is used at the near-end instead of the mutual information measure to optimize for the α_m parameters. We show via instrumental measures that MWF+ [40] and MWF+ [25] performance coincide, which admits the generality of the mutual information framework. In the next section, we prove that ASII can be derived as an special case for the mutual information measure. The same experimental results are shown for MVDR+[25].

B. 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. The microphones are positioned at the coordinates $(1.5, 2, 1)$ and $(1.5, 2.2, 1)$. At far-end, the target source and the noise sources are located at $(1.5, 3, 1)$ $(0.5, 1, 1)$ $(0.75, 3, 1)$ $(3, 1.6, 1)$ coordinates, respectively. We used 36 seconds of speech material originating from the TIMIT-database [54], sampled at 16 kHz. The impulse responses were generated using [55] (without reverberation).

We examined two different noise types: 1) natural noise sources including babble and inside-car recorded noise sampled at 16 kHz. 2) synthesized noise sources, where the far-end and near-end noise source consisted of spectrally generated Gaussian noise, with an overlapping spectral region from 1.5 kHz to 3 kHz as shown in [1][3]. The latter is specifically used to study the behavior of the proposed algorithm, once the far-end and near-end noise sources are partially overlapping. Signals were processed on a block-by-block basis by applying a 32 ms square-root-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., $\sigma_{T_m}^2$, $\sigma_{M_m}^2$ and $\sigma_{N_m}^2$ (m = 64) were obtained by taking the sample-mean of the critical band energy over the entire utterance, leading to a time-invariant filter.

In summary, from the far-end environment, the knowledge of the clean speech and noise power spectrum is assumed and the

[3]The noise spectral power is high in this band for both near-end and far-end noise processes.

room transfer function is known. From the near-end side, the knowledge of the near-end noise power spectrum is assumed. These are not the per-frame clean speech and noise power spectral density (PSD)s, but the long-term averages of the speech and noise PSDs.

C. Instrumental Measures

For the instrumental evaluations, we compare seven approaches using three independent intelligibility measures including STOI [56], ASII [25] and SII [22]. The methods that we compare are

1) The signal after MVDR beamformer (MVDR output).
2) The single channel signal with only the proposed near-end processor based on our solution of (27) (Single mic + Eq.(27)).
3) The proposed method as outlined in Fig. 3(a) (Prop. MVDR + Eq.(27)).
4) Blind processing based on Fig. 3(b) (MVDR+ [40]) and (MVDR+ [25]).
5) Blind processing based on Fig. 3(c) (MWF+ [40]) and (MWF+ [25]).

Fig. 4 shows the improvement of all the above discussed methods with respect to the intelligibility of the unprocessed signal at the reference far-end microphone versus different intelligibility measures. The left and right-hand side plots show the same information by fixing the SNR at one side and plotting the variation in the intelligibility improvement versus the other side SNR. Two different noise types are used in the simulations: Fig. 4(a) shows the performance using natural noise sources where Fig. 4(b) makes use of synthesized noise sources. The MVDR beamformer followed by the transparent proposed processor in (27) dominantly outperforms the other approaches in all the presented measures with different noise sources.

In Fig. 4(a), we used babble noise sources at far-end and inside car noise at near-end. The near-end SNR is fixed at -5 dB and the improvements in intelligibility are shown versus far-end. The results in the left plots indicate that the improvement is maximal in the low-SNR range of the far-end SNRs. As expected, the improvement of all the presented algorithms (except for the plain MVDR output) with respect to the intelligibility of the unprocessed signal vanishes once the far-end SNR reaches 20 dB. Particularly, the performance of the single channel processing reaches other approaches for increasing far-end SNR. At the right-hand side, the far-end SNR is fixed at -15 dB and the variations are plotted against near-end SNR. In contrast to the left plots, the performance of MVDR is approaching to the other algorithms once the near-end SNR is increasing, while the single channel processing behaves oppositely.

In Fig. 4(b) the synthesized noise sources as explained in Section VII-B, are used at both near-end and far-end environments. The results are consistent with Fig. 4(a) even though the improvement with respect to the increasing near-end SNR is decreased relatively in Fig. 4(b) . We use the instrumental measures as a prediction for the real performance of the algorithm, which is evaluated using listening tests. This is discussed next for the synthesized noise sources.

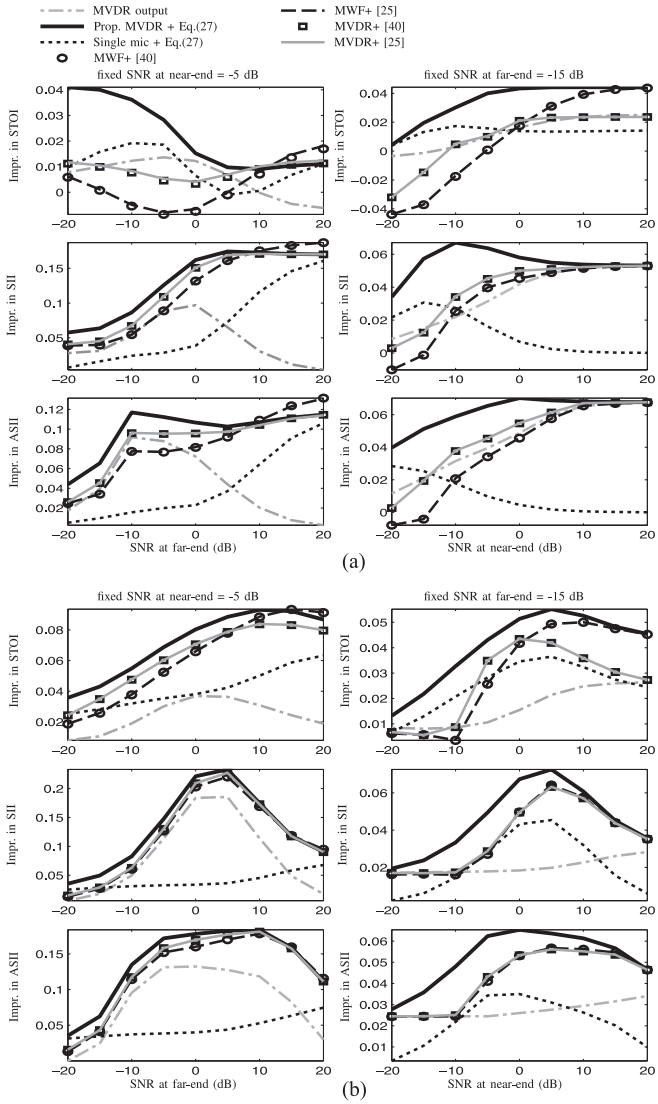

Fig. 4. Intelligibility improvement compared to the unprocessed signal at reference far-end microphone in terms of STOI, ASII and SII with different noise sources. (a) babble noise for 40 talkers at far-end and inside-car-noise at near-end. (b) partially overlapped synthesized noise sources from [1].

D. Listening Test

We conducted an informal listening test where seven native Dutch speakers (excluding the coauthors) listened to five-word sentences created from a closed set of words and had to select each word from a set of 10 [57]. In total, four algorithms (the 3 approaches in Fig. 3 and the MVDR without near-end processing) have been chosen to be tested in three different near-end SNRs as $[-7.5, 0, 5]$ dB and two far-end SNRs as $[-10, 2.5]$ dB. Note that each scenario is tested with three different sentences. Therefore, for each listener a data base including 72 signals is provided.

The listening test results are shown in Fig. 5. This confirms the optimality of the transparent (joint) processing over the blind (disjoint) approaches i.e., the proposed model outperforms the MWF+[40] and MVDR+[40]. The results illustrate intelligibility in the scale of 0 to 100 percent versus far-end SNR for

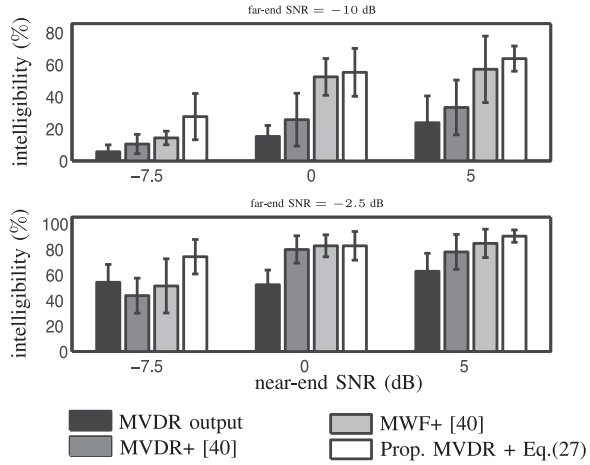

Fig. 5. Listening test results.

TABLE I
THE P-VALUE MEASURES (IN PERCENT) FOR THE SIGNIFICANCE LEVEL OF $\zeta_0 = 5\%$ WITH RESPECT TO DIFFERENT NEAR-END (HORIZONTAL) AND FAR-END (VERTICAL) SNRS BASED ON THE LISTENING TEST RESULTS

		-7.5 dB	0 dB	5 dB
Joint vs. MVDR	-10 dB	0.18	0	0.11
	-2.5 dB	1.9	0.28	0.28
Joint vs. MVDR+ [40]	-10 dB	0.9	0.46	0.49
	-2.5 dB	0.16	30	6.4
Joint vs. MWF+ [40]	-10 dB	3.7	37	20
	-2.5 dB	4.9	50	16

three near-end SNRs. The listening test results suggest that the proposed joint approach in Fig. 3(a) is significantly better than the reference blind method (MWF+ [40]) (Fig. 3(c)) when the SNR is low, both in near-end and far-end environment. This significance drops as the channel condition improves at both sides.

To evaluate the effectiveness of the results we calculated the statistical significance for the different algorithms. The Matlab function *ttest* is used here to obtain the p-value for the proposed Gaussian models over MVDR output and the reference blind methods are collected in Table I as "Joint vs. MVDR", "Joint vs. MVDR+ [40]" and "Joint vs. MWF+ [40]", respectively.

These results show that the proposed algorithm based on MVDR beamforming followed by near-end processing based on (27) is always significantly better that the MVDR output without near-end processing. In turn, the proposed approach is significantly better than (MVDR+ [40]), except for high near-end and far-end SNRs. Moreover, the p-value analysis suggests that the joint processing is significantly better than optimal blind processing (MWF+ [40]) for both tested far-end SNRs as long as the near-end SNR is relatively low (-7.5 dB). The intelligibilities of all the considered algorithms converges for high near-end SNR, that was predicted in the instrumental measures.

VIII. DISCUSSION AND CONCLUDING REMARKS

In this section, we provide more insight to the problem, by analyzing the solutions and studying the extreme and special cases. We look at the relation between the mutual information and well known classical measures of intelligibility.

A. Relation to Classical Intelligibility Measures

It is interesting to see the proposed intelligibility measure in (26) in the context of classical measures of intelligibility. Indeed, a reliable intelligibility measure should incorporate valid upper and lower bounds for extreme cases of the environmental noise, i.e., to reflect that the intelligibility can not go beyond certain levels. Moreover, it should accurately represent the effect of the inherent noise sources (production and interpretation noise) on the intelligibility of the speech. In the following we investigate the aforementioned properties for our proposed intelligibility measure.

For a single microphone scenario with no enhancement, the overall channel SNR due to environmental noise sources is given by

$$\xi_m = \frac{\sigma^2_{T_m}}{\sigma^2_{N_m} + \sigma^2_{U_m}}.$$

This yields the mutual information measure that is optimized in (27) as

$$I(\bar{S};\bar{Z}) \approx -\sum_m \frac{1}{2}\log\left(\frac{(1-\rho^2_{0_m})\xi_m + 1}{\xi_m + 1}\right). \tag{35}$$

In fact (35) is closely related to the classical intelligibility measures AI [20], [21], SII [22] and ASII [25] once we write the mutual information as

$$I(\bar{S};\bar{Z}) = \sum_m I_m A_m(\xi_m), \tag{36}$$

where

$$A_m(\xi_m) = \log\frac{(1-\rho^2_{0_m})\xi_m + 1}{\xi_m + 1} / \log(1-\rho^2_{0_m}), \tag{37}$$

$$I_m = -\frac{1}{2}\log(1-\rho^2_{0_m}). \tag{38}$$

One can readily identify I_m as the *unnormalized band-importance function* and $A_m(\xi_m)$ as the weighting function, comparing to the classical intelligibility measures. The former is simply the information rate transmitted in a band when no environmental noise is present, i.e., $\rho^2_{T_m,\tilde{X}_m}\rho^2_{\tilde{X}_m,Y_m} = 1$, hence, ξ_m is infinite. The latter represents the conventional, normalized band-importance function by dividing I_m by the overall mutual information rate $I(\bar{S};\bar{Z})$. The maximum information rate defines the importance of a band for speech intelligibility which is a decreasing function of the interpretation and production noise. This is naturally limited by the band importance value I_m and by zero (as mutual information is nonnegative).

In this work, the values for ρ_{0_m} are derived based on the band importance coefficients from [22], in (38), therefore $\rho_{0_m} = \sqrt{1-2^{(-2I_m)}}$. However, there is an ongoing research to extract more accurate values for production noise by statistically evaluating the variation in the human speech production process [41].

An interesting observation is the behavior of $A_m(\xi_m)$ once $\rho_{0_m}^2$ approaches zero, that is matching the approximated SII (ASII), which itself is an approximation of SII. Also, note that the proposed $A_m(\xi_m)$ are differentiable and concave functions of the ξ_m and α_m, whereas $A_m^{AI}(\xi_m)$ and $A_m^{SII}(\xi_m)$ are piecewise linear functions and not differentiable and convex, in general. In fact, ASII can be derived from the first order Taylor approximation of the mutual information (MI), accordingly we can write

$$\begin{aligned} A_m(\xi_m) &= \frac{\log\frac{(1-\rho_{0_m}^2)\xi_m+1}{\xi_m+1}}{\log(1-\rho_{0_m}^2)} = \frac{\sum_{n=1}^{\infty}(\frac{\rho_{0_m}^2\xi_m}{\xi_m+1})^n/n}{\sum_{n=1}^{\infty}(\rho_{0_m}^2)^n/n} \\ &\approx \frac{\frac{\rho_{0_m}^2\xi_m}{\xi_m+1}}{\rho_{0_m}^2} = \frac{\xi_m}{\xi_m+1}. \end{aligned} \tag{39}$$

This gives a lower bound on the MI, which explains why the proposed $A_m(\xi_m)$ for small $\rho_{0_m}^2$ approaches the weighting function of ASII, as presented in [1]. Note that the information rate approaches zero when $\rho_{0_m}^2 \downarrow 0$, since there is infinite production or interpretation noise and the effect of the weighting function is minimal since the unnormalized band importance function is already zero. The saturation of $A_m(\xi_m)$, indicates that increasing SNR beyond the interpretation and production SNR will not increase the MI.

B. Environmental Noise Only

The enhancement process, in the case of only environmental noise (no interpretation and production noise ($\rho_{0_m}=1$)), leads to the post-filter amplification factors

$$\alpha_m = \frac{-(\sigma_{T_m}^2+2\sigma_{M_m}^2)\sigma_{N_m}^2+2\sqrt{r}}{2(\sigma_{T_m}^2+\sigma_{M_m}^2)\sigma_{M_m}^2}, \tag{40}$$

where r is a function of λ

$$\begin{aligned} r &= \sigma_{N_m}^4(\sigma_{T_m}^2+2\sigma_{M_m}^2)^2 \\ &\quad -\sigma_{M_m}^2(\sigma_{T_m}^2+\sigma_{M_m}^2)\left(\sigma_{M_m}^4-\frac{\sigma_{N_m}^2}{2\lambda}\right). \end{aligned}$$

Note that $2\sqrt{r}$ is required to be large enough to force the nominator to a positive value.

An interesting scenario is when there is no far-end noise ($\sigma_{M_k}^2=0, \rho_{0_m}=1$), then a_m' is zero and (30) turns to a linear equation. The solution to this problem is given by

$$\alpha_m\sigma_{T_m}^2 = \begin{cases} \frac{1}{2}\left(\frac{1}{\lambda}-\frac{1}{\lambda_0}\right), & \lambda\geq\lambda_0 \\ 0, & \lambda\leq\lambda_0 \end{cases} \tag{41}$$

where $\lambda_0=\frac{1}{2\sigma_{N_m}^2}$, and together with the equality constraint; $\sum_m \alpha_m\sigma_{T_m}^2=\sum_m \sigma_{T_m}^2$, the solution is the well-known waterfilling power adaptation over frequency [58]. This indicates that more energy is allocated for the "strong channels" with less noise, and less energy to the "weak channels" with larger noise level. This is not the case when the far-end noise and/or production noise are considered, since amplifying the channels that have less near-end noise but stronger remaining noise from the far-end, will degrade the performance.

C. Conclusion

The main conclusions are summarized as

1) The optimal intelligibility enhancement of speech in a communication system consisting of separate near-end and far-end environments (Fig. 1), which can be decomposed into MVDR beamforming which reduces the noise at far-end, followed by a near-end processor that redistributes power over spectrum, given the noise statistics at each frequency band.
2) We proved theoretically that the disjoint processing is optimal only if the near-end processing is aware of the noise variance remaining from the far-end processing, which has been ignored in existing speech reinforcement techniques.
3) Taking the mutual information expression for two Gaussian random variables is a simple yet valid model for calculating the overall mutual information in the speech signal compared to the more complicated Chi-squared model resulting from the critical bands distribution.
4) The noise at the far-end, (which is recorded with the microphones) can be estimated using existing techniques, that is, using [59], [60] or the overview in [3]. Using a measurement microphone at the near-end, a similar strategy can be followed. However, here the potential presence of reverberation of the noise sources on the measurement microphone is a problem that is of great interest for future research.

Appendix A
Derivation of the MI for Chi-Squared Distribution

From [44], we know the mutual information for two arbitrary random variables of S and Z is given by

$$\begin{aligned} I(S;Z) &= h(S)-h(S|Z) \\ &= h(S)+h(Z)-h(S,Z) \end{aligned} \tag{42}$$

where from [51] we know for two Chi-squared distributed random variables of $\bar{S}$ and $\bar{Z}$

$$\begin{aligned} h(\bar{S}) &= h(\bar{Z}) = \log\left(2\Gamma\left(\frac{l}{2}\right)\right) \\ &\quad +\left(\frac{l}{2}-\left(\frac{l}{2}-1\right)\psi\left(\frac{l}{2}\right)\right) \end{aligned} \tag{43}$$

and the mutual entropy is given by

$$\begin{aligned} -h(\bar{S},\bar{Z}) &= (l-2)\left(\log(2)+\psi\left(\frac{l}{2}\right)\right)-\frac{l}{1-\rho^2} \\ &\quad -\log\left(2^{(l-2)}\left(2\Gamma\left(\frac{l}{2}\right)\right)^2(1-\rho^2)^{\frac{l}{2}}\right) \\ &\quad +E\left\{\log\left({}_0F_1\left(\frac{l}{2};\frac{\rho^2\bar{S}\bar{Z}}{4(1-\rho^2)^2}\right)\right)\right\}. \end{aligned} \tag{44}$$

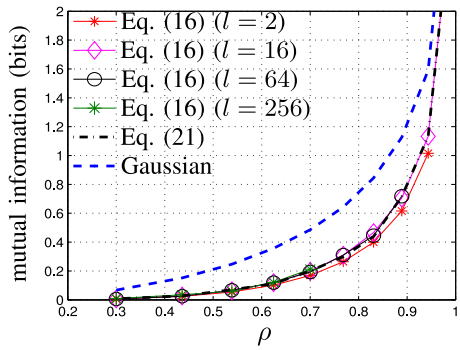

Fig. 6. Mutual information (bits per transmission) versus correlation coefficients.

Substituting the entropy expressions (43) and (44) in (42) gives

$$I(\bar{S};\bar{Z}) = l - \frac{l}{1-\rho^2} - \frac{l}{2}\log(1-\rho^2) + E\left\{\log\left({}_0F_1\left(\frac{l}{2};\frac{\rho^2\bar{S}\bar{Z}}{4(1-\rho^2)^2}\right)\right)\right\}.$$

Considering ${}_0F_1(a;x) := \Gamma(a)\sqrt{x}^{(1-a)}\mathrm{I}_{a-1}(2\sqrt{x})$, for arbitrary x and a, and $E\{\log(\bar{S}\bar{Z})\} = E\{\log(\bar{S})\} + E\{\log(\bar{Z})\} = 2(\psi(\frac{l}{2}) + \log(2))$, simplifies further the mutual information expression as given in (16). When ρ is zero, this means that the two variables $\bar{S}, \bar{Z}$ are independent, for the case that $h(\bar{S},\bar{Z}) = h(\bar{S}) + h(\bar{Z})$ or equivalently $h(\bar{S}|\bar{Z}) = h(\bar{S})$.

Appendix B
Empirical Evaluation of (21)

In Fig. 6, the mutual information for the four cases of $l = 2$, $l = 16$, $l = 64$ and $l = 256$ under the Chi-squared distribution is calculated according to (16) and the results are illustrated versus ρ which is the correlation coefficient between Gaussian variables. We emphasize that the expected value $E\{\log\left(\mathrm{I}_{\frac{l}{2}-1}(\frac{\sqrt{\rho^2\bar{S}\bar{Z}}}{1-\rho^2})\right)$ in (16) is calculated empirically for these plots. We used randomly generated complex Gaussian variables with uncorrelated real and imaginary variables and equal correlation coefficients ρ for real and imaginary parts. Moreover, the Gaussian approximated mutual information for the Chi-squared distributed variables according to (21) is plotted as well as for arbitrary two Gaussian variables, i.e., $I(S,Z) = -\frac{1}{2}\log(1-\rho^2)$.

As one can see in Fig. 6, the empirical results coincides with the calculated mutual information for correlated Chi-squared variables in (21), when $l \geq 16$. In fact, the simulations show that the amount of mutual information is changing slightly by having different values for l. It follows that the mutual information is considerably lower when measured on the power per critical band, compared to the case when the mutual information is directly calculated for two Gaussian random variables as in [40] with correlation coefficient of ρ (comparison between the Gaussian and Eq. (21) curves in Fig. 6). The former is referred to as the Chi-squared model and the latter is the Gaussian model.

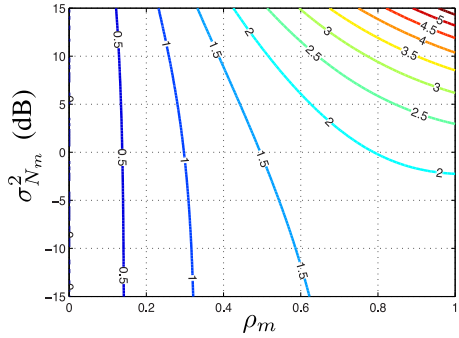

Fig. 7. Δ versus ρ_m and $\sigma^2_{N_m}$ for $\sigma^2_{M_m} = 0$ and $\sigma^2_{T_m} = 5$ dB.

Appendix C
Proof of the Existence of Real Roots for (30)

Using (31) to (33), $b_m^2 - 4a_m c_m \geq 0$ can be written as

$$\Delta := \left((1-0.5\rho_{0_m}^2)\sigma_{N\,m}\sigma_{T_m}^2 + \sigma_{N\,m}\sigma_{M_m}^2\right)^2 - \left((\sigma_{N_m}^2 - 0.5\rho_{0_m}^2)(\sigma_{T_m}^2 + \sigma_{M_m}^2)\right)\left((1-\rho_{0_m}^2)\sigma_{T_m}^2 + \sigma_{M_m}^2\right) \geq 0, \quad (45)$$

which is the difference between two terms. To have real roots for (30), we need to show that Δ is always non-negative.

By careful observation, the first term in (45) is the multiplication of two identical terms (squared) and the second term can be decomposed as the multiplication of somehow similar terms as $\left(\sigma_{N_m}^2 - 0.5\rho_{0_m}^2\right)\sigma_{T_m}^2 + (\sigma_{N_m}^2 - 0.5\rho_{0_m}^2)\sigma_{M_m}^2\big)$ times $((1-\rho_{0_m}^2)\sigma_{T_m}^2 + \sigma_{M_m}^2)$. To show that $\Delta \geq 0$, we plot Δ with respect to two variables ρ_m and $\sigma_{N_m}^2$ for a fixed $\sigma_{M_m}^2$ and $\sigma_{T_m}^2$. It is shown in Fig. 7 that Δ is always a non-negative value. The variation of $\sigma_{M_m}^2$ and $\sigma_{T_m}^2$ only scales the contours as they are common in both terms in (45).

Acknowledgment

MATLAB code for this paper is available at the authors' website http://cas.tudelft.nl.

References

[1] 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," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, Mar. 2016, pp. 654–658.

[2] P. Loizou, *Speech Enhancement Theory and Practice*. Boca Raton, FL, USA: CRC Press, 2007.

[3] 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*. Williston, VT, USA: Morgan & Claypool, 2013.

[4] J. Benesty, M. M. Sondhi, and Y. Huang, Eds. ,*Springer Handbook of Speech Processing*. New York, NY, USA: Springer, 2008.

[5] M. Brandstein and D. Ward, Eds. ,*Microphone Arrays: Signal Processing Techniques and Applications*. New York, NY, USA: Springer, 2001.

[6] V. Grancharov, J. Samuelsson, and W. B. Kleijn, "Noise-dependent post-filtering," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, Montreal, QC, Canada, 2004, pp. I457–I460.

[7] K. Eneman *et al.*, "Evaluation of signal enhancement algorithms for hearing instruments," in *Proc. 2008 16th Eur. Signal Process. Conf.*, Aug. 2008, pp. 1–5.

[8] A. Kusumoto, T. Arai, K. Kinoshita, N. Hodoshima, and N. Vaughan, "Modulation enhancement of speech by a pre-processing algorithm for improving intelligibility in reverberant environments," *ELSEVIER Speech Commun.*, vol. 45, no. 2, pp. 101–113, 2005.

[9] N. Hodoshima, T. Arai, A. Kusumoto, and K. Kinoshita, "Improving syllable identification by a preprocessing method reducing overlap-masking in reverberant environments," *J. Acoust. Soc. Amer.*, vol. 119, pp. 40–55, 2006.

[10] 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, Aug. 1976.

[11] 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.

[12] C. Tantibundhit *et al.*, "New signal decomposition method based speech enhancement," *Signal Process.*, vol. 87, pp. 2607–2628, 2007.

[13] J. Crespo and R. C. Hendriks, "Multizone speech reinforcement," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 22, no. 1, pp. 54–66, Jan 2014.

[14] 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, Mar. 2015.

[15] J. B. Crespo and R. C. Hendriks, "Speech reinforcement in noisy reverberant environments using a perceptual distortion measure," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, May 2014, pp. 910–914.

[16] R. C. Hendriks, J. Crespo, J. Jensen, and C. H. Taal, "Optimal near-end speech intelligibility improvement incorporating additive noise and late reverberation under an approximation of the short-time SII," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 23, no. 5, pp. 851–862, May 2015.

[17] J. C. R. Licklider and I. Pollack, "Effects of differentiation, integration, and infinite peak clipping upon the intelligibility of speech," *Acoust. Soc. Amer. J.*, vol. 20, 1948, Art. no. 42.

[18] G. A. Miller, *Language and Communication*. New York, NY, USA: McGraw-Hill, 1951.

[19] I. B. Thomas *The Significance of the Second Formant in Speech Intelligibility*. No. TR-10. Illinois Univ Urbana Biological Computer Lab, 1966.

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

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

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

[23] B. Sauert and P. Vary, "Near end listening enhancement optimized with respect to speech intelligibility index," in *Proc. EURASIP Eur. Signal Process. Conf.*, Aug. 2009, vol. 17, pp. 1844–1848.

[24] B. Sauert and P. Vary, "Near end listening enhancement optimized with respect to speech intelligibility index and audio power limitations," in *Proc. EURASIP Eur. Signal Process. Conf.*, Aug. 2010, pp. 1919–1923.

[25] 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, Mar. 2013.

[26] C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen, "An algorithm for intelligibility prediction of time-frequency weighted noisy speech," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 19, no. 7, pp. 2125–2136, Sep. 2011.

[27] M. Cooke, "A glimpsing model of speech perception in noise," *J. Acoust. Soc. Amer.*, vol. 119, pp. 1562–1573, Mar. 2006.

[28] C. H. Taal, R. C. Hendriks, and R. Heusdens, "Matching pursuit for channel selection in cochlear implants based on an intelligibility metric," in *Proc. EURASIP Eur. Signal Process. Conf.*, Aug. 2012, pp. 504–508.

[29] Y. Tang and M. Cooke, "Optimized spectral weightings for noise-dependent speech intelligibility enhancement," in *Proc. ISCA Interspeech*, Sep. 2012, pp. 955–958.

[30] V. Aubanel and M. Cooke, "Information-preserving temporal reallocation of speech in the presence of fluctuating maskers," in *Proc. ISCA Interspeech*, Aug. 2013, pp. 3592–3596.

[31] C. Valentini-Botinhao, J. Yamagishi, S. King, and R. Maia, "Intelligibility enhancement of HMM-generated speech in additive noise by modifying mel cepstral coefficients to increase the glimpse proportion," *Comput. Speech Lang.*, vol. 28, pp. 665–686, Jun. 2014.

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

[33] 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 *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, Mar. 2012, pp. 4061–4064.

[34] C. H. Taal, R. C. Hendriks, and R. Heusdens, "Speech energy redistribution for intelligibility improvement in noise based on a perceptual distortion measure," *Comput. Speech Lang.*, vol. 28, pp. 858–872, Dec. 2014.

[35] J. B. Crespo and R. C. Hendriks, "Speech reinforcement with a globally optimized perceptual distortion measure for noisy reverberant channels," in *Proc. Int. Workshop Acoust. Echo, Noise Control*, Sep. 2014, pp. 89–93.

[36] P. Stoica and Y. Selen, "Model-order selection: A review of information criterion rules," *IEEE Signal Process. Mag.*, vol. 21, pp. 36–47, Jul. 2004.

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

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

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

[40] W. B. Kleijn and R. C. Hendriks, "A simple model of speech communication and its application to intelligibility enhancement," *IEEE Signal Process. Lett.*, vol. 22, no. 3, pp. 303–307, Mar. 2015.

[41] S. van Kuyk, W. B. Kleijn, and R. C. Hendriks, "Intelligibility metric based on a simple model of speech communication," in *Proc. Int. Workshop Acoust. Signal Enhancement*, 2016, pp. 4266–4269.

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

[43] H. Fastl and E. Zwicker, *Psychoacoustics Facts and Models*. New York, NY, USA: Springer, 2006.

[44] T. M. Cover and J. A. Thomas, *Elements of Information Theory* (Wiley Series in Telecommunications and Signal Processing). Hoboken, NJ, USA: Wiley, 2006.

[45] P. Vary and R. Martin, *Digital Speech Transmission: Enhancement, Coding and Error Concealment*. Hoboken, NJ, USA: Wiley, 2006.

[46] S. Boyd and L. Vandenberghe, *Convex Optimization: Convex Optimization Problems, Equivalent Problems*. New York, NY, USA: Cambridge Univ. Press, 2004.

[47] H. L. Van Trees, *Detection, Estimation, and Modulation Theory. Part IV., Optimum Array Processing*. New York, NY, USA: Wiley, 2002.

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

[49] B. C. J. Moore, *An Introduction to the Psychology of Hearing*, 5th ed. San Diego, CA, USA: Academic, 2003.

[50] P. R. Krishnaiah, J. P. Hagis, and L. Steinberg, "A note on the bivariate chi distribution," *SIAM Rev.*, vol. 5, pp. 140–144, Apr. 1963.

[51] A. H. Joarder, A. Laradji, and M. H. Omara, "On some characteristics of bivariate chi-square distribution," *J. Theoretical Appl. Statist.*, vol. 46, no. 5, pp. 577–586, 2012.

[52] R. G. Gallager, *Principles of Digital Communication*. Cambridge, U.K.: Cambridge Univ. Press, 2008.

[53] W. G. H. George, E. P. Box, and J. S. Hunter, *Statistics for Experimenters*. Hoboken, NJ, USA: Wiley, 2005.

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

[55] E. A. P. Habets, "Room impulse response generator," Tech. Rep., Technische Universiteit Eindhoven, Eindhoven, 2010.

[56] C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen, "A short-time objective intelligibility measure for time-frequency weighted noisy speech," in *Proc. IEEE Int. Conf. Acoust, Speech, Signal Process.*, 2010, pp. 4214–4217.

[57] R. Houben *et al.*, "Development of a dutch matrix sentence test to assess speech intelligibility in noise," *Int. J. Audiology, Early Online*, vol. 127, pp. 1–4, 2014.

[58] J. G. Proakis, *Digital Communications* (McGraw-Hill series in electrical and computer engineering). New York, NY, USA: McGraw-Hill, 1995.

[59] R. Martin, "Noise power spectral density estimation based on optimal smoothing and minimum statistics," *IEEE Trans. Speech Audio Process.*, vol. 9, no. 5, pp. 504–512, Jul. 2001.

[60] R. C. Hendriks, J. Jensen, and R. Heusdens, "Noise tracking using DFT domain subspace decompositions," *IEEE Trans. Audio Speech Lang. Process.*, vol. 16, no. 3, pp. 541–553, Mar. 2008.

Seyran Khademi received the B.Sc. degree in electrical engineering, in 2005 from the University of Tabriz, Iran. She received the M.Sc. degree in communications engineering from Chalmers University of Technology in Gothenburg, Sweden, in 2010 and the Ph.D. from Circuits and Systems (CAS) Group, The Netherlands, in 2016. After her Bachelor's studies she held a position as an Application Engineer in Tehran for a telecommunication company. She was appointed as a Postdoctoral Researcher in CAS group at Delft University of Technology, from February 2015 to 2017 working on audio and speech processing for intelligibility enhancement. She is currently a Post Doctoral Researcher in Computer Vision Lab, Delft University of Technology, Delft, The Netherlands, working on image processing and machine learning algorithms.

Richard C. Hendriks was born in Schiedam, The Netherlands. He received the B.Sc., M.Sc. (*cum laude*), and Ph.D. (*cum laude*) degrees in electrical engineering from the Delft University of Technology, Delft, The Netherlands, in 2001, 2003, and 2008, respectively. He is currently an Assistant Professor in the Circuits and Systems (CAS) Group, Faculty of Electrical Engineering, Mathematics and Computer Science, Delft University of Technology. In March 2010, he received a VENI grant for his proposal " Intelligibility Enhancement for Speech Communication Systems."

W. Bastiaan Kleijn (F'99) received the M.Sc. degree in physics from the University of California, Riverside, CA, USA and the M.S.E.E. degree from Stanford University, Stanford, CA, USA. He received the Ph.D. degree in electrical engineering from Delft University of Technology, Delft, The Netherlands and the Ph.D. degree in soil science from the University of California, Riverside, CA, USA. He is a Professor at Victoria University of Wellington, New Zealand and a Professor (part-time) at Delft University of Technology. He was a Professor and the Head of the Sound and Image Processing Laboratory at KTH in Stockholm, 1996–2010. He was a founder of Global IP Solutions, a company that provided the enabling audio technology to Skype. It was acquired by Google in 2010. He has served on a number of editorial Boards including those of the IEEE TRANSACTIONS ON AUDIO, SPEECH, AND LANGUAGE PROCESSING, SIGNAL PROCESSING, IEEE SIGNAL PROCESSING LETTERS, and IEEE SIGNAL PROCESSING MAGAZINE. He was the Technical Chair of ICASSP 1999 and EUSIPCO 2010, and two IEEE workshops.