IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 22 May 2024, accepted 22 August 2024, date of publication 26 August 2024, date of current version 5 September 2024.

Digital Object Identifier 10.1109/ACCESS.2024.3450268

RESEARCH ARTICLE

Joint Far- and Near-End Speech and Listening Enhancement With Minimum Processing

ANDREAS JONAS FUGLSIG[1], (Member, IEEE), ZHENG-HUA TAN[1], (Senior Member, IEEE), LARS SØNDERGAARD BERTELSEN[2], JESPER JENSEN[1], JENS CHRISTIAN LINDOF[2], (Member, IEEE), AND JAN ØSTERGAARD[1], (Senior Member, IEEE)
[1]Department of Electronic Systems, Aalborg University, 9220 Aalborg, Denmark
[2]RTX A/S, 9400 Nørresundby, Denmark

Corresponding author: Andreas Jonas Fuglsig (ajf@es.aau.dk)

This work was supported in part by the Innovation Fund Denmark under Grant 9065-00204B.

ABSTRACT This paper considers speech and listening enhancement for signals captured in one noisy environment that must be played back to a listener in another noisy environment. In both far-end speech enhancement and near-end listening enhancement, overly prioritizing noise suppression or maximizing intelligibility can result in undue speech distortions and reduced quality, especially when intelligibility is already high in favorable noise conditions. To address this, the use of a minimum processing framework has been proposed with the aim of reducing noise or enhancing listening to a minimum degree while ensuring that a specified intelligibility level is maintained. Furthermore, results have shown that jointly considering both environments improves performance compared to blindly concatenating far- and near-end methods. In blind processing, near-end listening enhancement typically assumes that the far-end signal is devoid of noise, potentially leading to erroneously interpreting noise as speech. Additionally, if the transmitter and receiver are blind to each other's presence, multiple instances of far- and near-end enhancement may occur and possibly work opposite directions, thus leading to degradations in the enhancement performance. In this paper, we perform a comprehensive exploration of our previously proposed joint far- and near-end minimum processing framework with systematic analysis and discussion. We derive a closed-form solution to the joint far- and near-end minimum processing optimization problem, with mean-square error processing penalty, a speech intelligibility constraint based on the approximated speech intelligibility index, and a noise power constraint. Performance was systematically studied using objective measures and listening tests for intelligibility, listening effort, and quality. We compared against relevant joint and blind methods with minimum and maximum processing. The results suggest that minimum processing achieves intelligibility comparable to maximum processing while preserving quality in higher signal-to-noise ratios, indicating its benefits in end-to-end communication. Joint processing provides advantages in objective estimated speech intelligibility for the minimum processing case, but not for maximum processing. However, no significant differences were observed in listening test results. This suggests that in certain speech and listening scenarios, it is feasible to optimize near- and far-end aspects separately, offering a more practical and convenient approach compared to joint optimization.

INDEX TERMS Minimum processing, joint processing, far-end, near-end, speech enhancement, listening enhancement.

I. INTRODUCTION

Speech communication systems are used in various contexts, such as mobile phones, hearing aids, intercoms, and public address systems. Hence, they need to work in diverse situations, where background noise can significantly impact both Speech Intelligibility (SI) and Speech Quality (SQ).

In speech communication systems, we can distinguish between two separate environments, cf. Fig.1: The Far-End (FE) environment (the target talker's location) and the

The associate editor coordinating the review of this manuscript and approving it for publication was Manuel Rosa-Zurera.

Near-End (NE) environment (the listener's location). Typically, both the FE and NE environments are subject to interfering acoustic noise sources, resulting in degradation of both SQ and SI for the listener. To mitigate this, noise reduction and speech enhancement algorithms can be applied in both FE and NE environments.

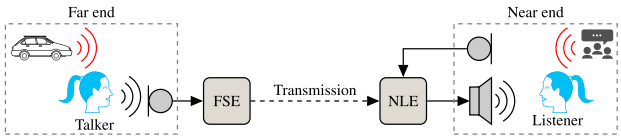


FIGURE 1. Speech communication system with Far-end Speech Enhancement (FSE) and Near-end Listening Enhancement(NLE).

Based on the number of available microphones, Far-end Speech Enhancement (FSE) methods may employ either single- or multi-microphone noise reduction algorithms to remove noise from recorded signals [1], [2], [3], [4], [5], [6], [7], [8], [9], [10], [11] before transmission. While FSE methods can remove noise after it has been mixed with the target speech, Near-end Listening Enhancement (NLE) methods must process the signal received from the FE environment prior to playback in the noisy NE environment. Thus, FSE techniques cannot typically be used at the NE. Instead, NLE may utilize knowledge of the NE noise to pre-process the FSE signal coming from the FE to increase the SI and SQ in the NE background noise [12], [13], [14], [15], [16], [17], [18], [19], [20], [21], [22], [23], [24]. A particular type of NLE technique is Active Noise Control (ANC) [23], [25], [26], where the aim is to cancel the noise by adding an anti-phase noise component to the speech signal before playout. However, ANC performance is insufficient outside headsets and handheld mobile phone scenarios [23]. Therefore, because we are not only concerned with these scenarios, we only consider NLE based on speech modifications in this work. Recently Deep Neural Networks (DNNs) have shown good results when optimizing for advanced SQ and SI predictors in both FSE [7], [8], [9] and NLE [20], [21], [22], [23]. However, increases in such predictions do not always translate to subjective performance gains [10]. Furthermore, DNNs have large memory usage, are difficult to interpret, and can have difficulties with generalizing to new acoustic scenarios. Hence, in this paper we take a more classic signal processing approach, providing closed-form and interpretable solutions.

Both SI and SQ are important in shaping the listening experience, and their significance varies depending on the acoustic setting [27], [28], [29], [30]. However, traditionally, the aim of NLE methods has been to exclusively maximize SI. Similarly, FSE methods have been designed according to the inherent undesirability of noise, and thus, with the purpose of maximizing noise reduction, leaving only the clean speech signal. In noisy conditions, enhanced intelligibility may positively impact the perceived SQ [28], [29], thus making SI a crucial contributor to SQ in noisy conditions [29]. To increase SI, NLE techniques potentially introduce speech distortions, which could diminish SQ; however, these distortions may be masked by more severe environmental NE noise [29], [30]. However, if the environmental noise subsides, speech becomes more intelligible and further SI enhancement is unattainable, and distortions become more disturbing [11], [19], [27], [28], [29], [30]. Furthermore, aggressive noise reduction by FSE also leads to significant speech distortions and possibly to the loss of contextual noise from the FE environment [11], [31].

To remedy the effects of excessive FE processing, it was proposed in [11] to apply a *minimum processing principle* to multi-microphone FSE, such that the beamformer output is minimally processed with respect to a certain reference signal, provided that a given performance criterion is fulfilled [11]. In particular, two cases wee investigated in [11]: in the first case, the processing of the noisy signal is limited to the minimal amount necessary for fulfilling an SI requirement, and in the second case, noise is completely eliminated with an aggressive beamformer unless the resulting distortion of the clean speech violates the SI requirement. To remedy the effects of excessive processing at the NE, [19] applied the minimum processing principle to NLE. This provided an adaptive NLE that limits the processing of the signal received from the FE to the minimum required to achieve a target SI, thereby minimizing speech distortions in relation to the received signal.

Conventionally, the FE and NE scenarios have been considered separately [1], [16], [32]. However, several recent studies have proposed a joint approach to FE and NE speech enhancement in both single- [33], [34], [35], [36], [37] and multi-channel cases [38], [39], [40]. Particularly, [33] proposed a new training strategy for DNNs in cases where FSE is performed more than once. In [34], both FE noise reduction and modifications to limit speech distortions were added to an existing state-of-the-art NLE technique [24]. In [35], a DNN was trained to jointly remove FE noise and enhance SI using the method [24] as a "teacher". They showed improvements compared to the joint method in [34], but no comparison was made against blind approaches. In [37], a DNN was used to jointly enhance for multiple SQ and SI estimators, and showed improvements against versions of the proposed joint DNN method and blind concatenation of DNNs and classic signal processing methods, but no comparison was made against other joint approaches. The work of [36] proposed a classic signal processing approach to jointly control noise reduction and an NLE post-filter gain to increase SI, and improvements were reported against a blind signal processing approach in an informal preference test. For the multi-microphone case, [38] proposed a maximization of SI by closed-form optimization based on approximated mutual information, providing some small improvements in objective SI measures and an informal listening test against methods that were unaware of the remaining noise after FSE. It was then shown in [39] that similar performance could

be obtained by a simpler closed-form optimization of the Approximated Speech Intelligibility Index (ASII) [18], [41].

Apart from [34], [36], and [37], the main goal of joint approaches has been to maximize SI. However, recently we proposed a *joint minimum processing beamforming and NLE* framework [40]. In contrast to existing joint processing works, this framework processes the signal the minimum amount required to achieve a desired target SI while preserving SQ in favorable noise condition [40]. Additionally, it extends the existing single-ended minimum processing frameworks [11], [19] to jointly consider all FSE, NLE, and environmental noises simultaneously. However, [40] only solved the optimization problem numerically, and a comparison was only made against the blind concatenation of the single-ended minimum processing frameworks of [11] and [19]. Thus, the joint minimum processing problem warrants further theoretical and experimental investigation.

In this paper, we conduct a comprehensive exploration of the joint FE and NE minimum processing framework initially introduced in [40]. The core contribution involves deriving a closed-form analytical solution for the joint FE and NE minimum processing optimization problem with a Mean-Square Error (MSE) processing penalty, an estimated SI constraint represented by the ASII, and a noise power SQ constraint. A systematic performance study encompassing objective measures and listening tests for SI, listening effort, and SQ, is conducted. We evaluate the effects of two aspects: minimum processing versus SI maximization and joint processing versus blind processing. Therefore, we compare against several methods: The joint ASII maximization of [39]; the blind concatenation of minimum processing FSE [11] and minimum processing NLE [19]; and blind concatenation classic "maximum" processing, i.e., a Minimum Variance Distortionless Response (MVDR) beamformer [1] and NLE ASII maximization [18]. The results suggest that minimum processing achieves a comparable SI to maximum processing while preserving good SQ in higher Signal-to-Noise Ratio (SNR) settings, emphasizing the benefits of applying the minimum processing principle in end-to-end communication scenarios. However, joint processing was only advantageous for estimated SI in the minimum processing case, but not in the maximum processing case. Finally, the subjective listening tests showed no significant differences between any of the tested methods. This leads to overall inconclusive but interesting results suggesting that in certain speech and listening scenarios, it is feasible to optimize the near- and far-end aspects separately, offering a more practical and convenient approach compared to joint optimization.

The remainder of this paper is organized as follows. In Section II we present our signal model. Section III introduces the minimum processing concept and details the differences between joint and blind approaches. In Section IV, we derive the case study optimization problem and its solution. Section V presents the experimental evaluation. Sections VI and VII present objective and subjective performance results, respectively. Finally, we discuss and conclude the paper in Sections VIII and IX.

A. ABBREVIATIONS

For convenience Table 1 lists the abbreviations used in this paper.

TABLE 1. Abbreviations used in this paper.

Abbreviation	Description
ANC	Active Noise Control
ASII	Approximated Speech Intelligibility Index
DFT	Discrete Fourier Transform
DNN	Deep Neural Network
ESTOI	Extended Short-Time Objective Intelligibility
FE	Far-End
FSE	Far-End Speech Enhancement
KKT	Karush-Kuhn-Tucker
NE	Near-End
NLE	Near-End Listening Enhancement
SI	Speech Intelligibility
SII	Speech Intelligibility Index
SQ	Speech Quality
MSE	Mean-Square Error
MUSHRA	Multiple Stimuli with Hidden Reference and Anchor
MVDR	Minimum Variance Distortionless Response
MWF	Multichannel Wiener Filter
SNR	Signal-to-Noise Ratio
STFT	Short-Time Fourier Transform
PESQ	Perceptual Evaluation of Speech Quality

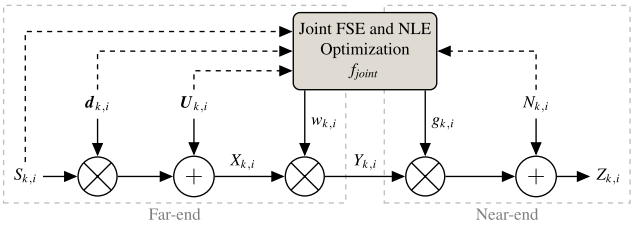


FIGURE 2. Signal model for joint FE and NE optimization.

II. SIGNAL MODEL

We consider the following time-frequency domain signal model with frequency index k and time index i, cf. Fig. 2,

$$Z_{k,i} = g_{k,i}\boldsymbol{w}_{k,i}^H \boldsymbol{d}_{k,i} S_{k,i} + g_{k,i}\boldsymbol{w}_{k,i}^H \boldsymbol{U}_{k,i} + N_{k,i} \quad (1)$$

$$= g_{k,i}\boldsymbol{w}_{k,i}^H \boldsymbol{X}_{k,i} + N_{k,i} \quad (2)$$

$$= g_{k,i} Y_{k,i} + N_{k,i}. \quad (3)$$

Here, $S_{k,i}$ is the clean speech signal at the source location, $\boldsymbol{U}_{k,i} \in \mathbb{C}^M$ is the additive FE noise, and $\boldsymbol{X}_{k,i} \in \mathbb{C}^M$ is the noisy multi-microphone signal picked up by M microphones, where $\boldsymbol{d}_{k,i} \in \mathbb{C}^M$ are acoustic transfer functions from the source to the microphones. First, to improve the SI and SQ of the noisy signal, $\boldsymbol{X}_{k,i}$, we employ a noise reduction beamformer, $\boldsymbol{w}_{k,i} \in \mathbb{C}^M$, as FSE with output signal $Y_{k,i}$. To further enhance SI and SQ, we apply an NLE gain, $g_{k,i} \in \mathbb{R}_+$, before the enhanced speech is played out in the noisy environment. The purpose of the beamformer $\boldsymbol{w}_{k,i}$ is to represent all processing focussed on the FE noise, whereas

the purpose of the NLE gain is to amplify relevant speech regions over the NE noise. Finally, the signal received by the NE listener, $Z_{k,i}$, is contaminated by ambient noise, $N_{k,i}$ in the NE environment.

We model the speech and noise as complex random vector processes comprised of Short-Time Fourier Transform (STFT) coefficients. As is common in the literature, we assume that the speech and noise processes are uncorrelated and zero-mean random processes with independence across frequencies [32]. From these assumptions, we can obtain the speech distortion weighted covariance matrix $C^{(\mu)}_{X_{k,i}}$ of $\boldsymbol{X}_{k,i}$ [11],

$$C^{(\mu)}_{X_{k,i}} \triangleq C_{S_{k,i}} + \mu C_{U_{k,i}} = \sigma^2_{S_{k,i}} \boldsymbol{d}_{k,i} \boldsymbol{d}^H_{k,i} + \mu C_{U_{k,i}}, \quad (4)$$

where $\sigma^2_{S_{k,i}}$ is the clean speech power spectrum level, and $C_{U_{k,i}} \triangleq \mathrm{E}\left[\boldsymbol{U}_{k,i}\boldsymbol{U}^H_{k,i}\right]$ is the FE noise covariance matrix and $\mu \in \mathbb{R}_+$ is the speech distortion weight [2], [3], and $\mu = 1$ leads to the standard covariance matrix.

Similar to the existing minimum processing frameworks [11], [19], [40], we focus on signal processing within perceptually relevant subbands. That is, signals are analyzed and processed in, e.g., octave bands, fractional octave bands, or critical bands that all mimic aspects of human auditory perception. We define perceptually motivated subbands such that multiple frequency bins may be included in the same and/or more subbands, and denote subbands with index j and frequencies with index k. Hence, we can encompass the effect of non-rectangular auditory filters. Therefore, each frequency bin-subband pairing is assigned a weight. For the j'th subband, we denote the non-negative filter weights as $\omega_{j,k}$, and let $\mathbb{B}_j$ be the set of frequency bins that contribute to the j'th subband, where $j \in \{1, \ldots, J\}$ and J is the total number of subbands. Thus, the NE noise spectrum level within one subband, j, and time-frame, i, is given as

$$\sigma^2_{N_{j,i}} \triangleq \sum_{k \in \mathbb{B}_j} \omega_{j,k} \sigma^2_{N_{k,i}}. \quad (5)$$

Any normalization for the filtering operation is already included within the subband filter weights, $\omega_{j,k}$. We provide additional details on the definition of $\omega_{j,k}$ and the connection between subbands and frequency bins in Appendix A.

In practical settings, joint FE and NE optimization requires sufficiently fast updating of signal properties and synchronization between the FE and NE for efficient processing. Similarly, the relevant signals must be both encoded, which leads to quantization noise, and transmitted across a possibly lossy channel. In this work, we are not concerned with these practical challenges and investigate performance under the assumption that they can be handled. Furthermore, with a similar motivation, we assume that no coupling issues exist between the microphones and loudspeakers at the FE and NE. Finally, in practice, the statistics of speech and noise processes must be estimated online, and we can apply the mathematical framework on a per time-frame basis. Hence, for brevity of notation, we disregard the time-index, i, and assume that we are working within a certain time frame, i, unless otherwise stated.

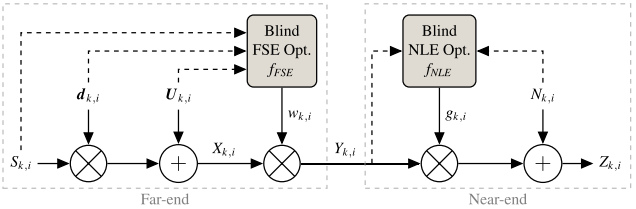


FIGURE 3. Block diagram of blind concatenation of FSE and NLE.

III. CONCEPTS

A. BLIND VERSUS JOINT PROCESSING

With joint processing, cf. Fig. 2, the beamformer, $\boldsymbol{w}_k$, and NLE gain, g_k, are determined as a function of all relevant signals, i.e.,

$$\left(\boldsymbol{w}^{\text{joint}},\ g^{\text{joint}}\right) = f_{\text{joint}}(S, U, N). \quad (6)$$

Hence, in joint processing all processing is optimized and derived jointly.

On the other hand, with blind processing, cf. Fig. 3, the beamformer, $\boldsymbol{w}_k$, is first determined as a function of only the speech and FE noise. Then, the NLE gain, g_k, is a function of the beamformer output signal and the NE noise, i.e,

$$\boldsymbol{w}^{\text{blind}} = f_{\text{FSE}}(S, U), \quad \text{and} \quad g^{\text{blind}} = f_{\text{NLE}}(Y, N) \quad (7)$$

Thus, the NLE is blind towards the effects of the FSE and cannot distinguish between speech and noise power in different parts of the spectrum. In particular, unlike joint processing, the signal after NLE is the output of a composite function of FSE and NLE.

B. THE MINIMUM PROCESSING CONCEPT

As in previous works [11], [19], [40], we assume that a designated target reference signal, S^R_k, is available, which could be the output signal from a beamformer with some specific desired characteristics; see [11] for more details. Focusing on a specific subband, j, we denote the number of frequency bins in this subband by $|\mathbb{B}_j|$. We then create the vectors $\boldsymbol{S}^R_j \in \mathbb{C}^{|\mathbb{B}_j|}$, $\boldsymbol{S}_j \in \mathbb{C}^{|\mathbb{B}_j|}$ and $\boldsymbol{Z}_j \in \mathbb{C}^{|\mathbb{B}_j|}$ by gathering all S^R_k, S_k and Z_k for $k \in \mathbb{B}_j$. Furthermore, we let $\mathcal{D}_j(\cdot,\cdot)$ and $\mathcal{I}_j(\cdot,\cdot)$ be two finite non-negative functionals indicating processing performance. Here, $\mathcal{D}_j\left(\boldsymbol{S}^R_j, \boldsymbol{Z}_j\right)$ measures the distortion (processing penalty) between the target reference signal, S^R_k, and the signal perceived by the NE listener, Z_k, while $\mathcal{I}_j\left(\boldsymbol{S}^R_j, \boldsymbol{Z}_j\right)$ is an intelligibility or performance estimator for the speech and listening enhancement in subband j. Consequently, the joint FE and NE minimum processing beamformer, $\boldsymbol{w}^{MP}_k$, and NLE gain, g^{MP}_k, for subband j, is defined as the solution to the optimization

problem [40],

$$\underset{\{\boldsymbol{w}_k\},\{g_k\},k\in\mathbb{B}_j}{\arg\min}\ \mathcal{D}_j\left(\boldsymbol{S}_j^R,\boldsymbol{Z}_j\right)\qquad \text{s.t.}\quad \mathcal{I}_j\left(\boldsymbol{S}_j,\boldsymbol{Z}_j\right)\geq I_j', \tag{8}$$

where $I_j' \triangleq \min\left(I_j, I_j^{\max}\right)$ with I_j the desired minimum requirement on the SI performance $\mathcal{I}_j\left(\boldsymbol{S}_j,\boldsymbol{Z}_j\right)$, and $I_j^{\max}$ the maximum achievable performance, when the performance $\mathcal{I}_j\left(\boldsymbol{S}_j,\boldsymbol{Z}_j\right)$ is maximized in an unconstrained manner.

In contrast to the existing minimum processing methods [11], [19], the joint approach considers the combined effects of all noise sources with FSE and NLE simultaneously.

Finally, we note that a straightforward approach to enhancing the NE output SNR involves increasing g_k indefinitely. However, this results in an infinite playback volume, and a substantial gap from the reference signal, thereby violating the fundamental principle of minimum processing.

IV. JOINT FAR- AND NEAR-END MINIMUM PROCESSING

In this section, we introduce our previously proposed joint FE and NE minimum processing case study reported in [40] with preliminary results. We consider an MSE processing penalty, $\mathcal{D}_j$, and two performance criteria; an SI estimator, $\mathcal{I}_j^{\text{SI}}$, based on ASII [18], [19] and a noise power constraint, $\mathcal{I}_j^{\text{NP}}$, for SQ [40].

Inspired by the solutions in [11] and [36], we define a multichannel noise reduction vector (beamformer):

$$\boldsymbol{w}_k \triangleq \alpha_k \boldsymbol{v}_k^R + (1-\alpha_k)\boldsymbol{w}_k^{\mu\text{MWF}}, \tag{9}$$

where $\alpha_k \in [0,1]$, $\boldsymbol{v}_k^R$ is a pre-selected reference beamformer with a desired property, for example, low speech distortion (MVDR) or an ambient noise preserving beamformer [11]. Further, $\boldsymbol{w}_k^{\mu\text{MWF}}$ is the speech distortion weighted Multichannel Wiener Filter (MWF) MWF [2],

$$\boldsymbol{w}_k^{\mu\text{MWF}} \triangleq \left(C_{X_k}^{(\mu)}\right)^{-1}\sigma_{S_k}^2\boldsymbol{d}_k \tag{10}$$

with pre-selected speech distortion weight, μ, e.g., $\mu \gg 1$ leading to high noise reduction and high speech distortion [11]. The parameter α_k provides a way to control the trade-off between the two beamformers, $\boldsymbol{v}_k^R$ and $\boldsymbol{w}_k^{\mu\text{MWF}}$, and their processing.

Early experiments show that solving (8) for our choice of $\mathcal{D}_j$ and $\mathcal{I}_j$ leads to solutions where only a single frequency within a subband is processed, leading to unpleasant artifacts. To avoid such solutions and ensure that we obtain more uniform processing across a subband, we assume that the NLE gains g_k and the combination weights α_k are fixed across an entire subband. That is,

$$\alpha_k = \alpha_i \quad \forall k, i \in \mathbb{B}_j \tag{11}$$

$$g_k = g_i \quad \forall k, i \in \mathbb{B}_j. \tag{12}$$

This also aligns with the results achieved in [11] and existing NLE studies [16], [19], [36], [38], [39], where the combination weights and NLE gains were derived to be fixed across subbands. Thus, we may write

$$\boldsymbol{w}_{j,k} \triangleq \alpha_j \boldsymbol{v}_k^R + (1-\alpha_j)\boldsymbol{w}_k^{\mu\text{MWF}}. \tag{13}$$

A. PROCESSING PENALTY

As suggested in [11], [19], and [40], we consider a processing penalty, $\mathcal{D}_j(\cdot)$ based on an MSE criterion. We want to minimize the processing in relation to the reference signal, S_k^R. The processing consists of two parts: the beamformer, $\boldsymbol{w}_{j,k}$, and the NLE gain, g_k. Therefore, we consider a processing penalty with two penalty terms: one that penalizes the processing caused by the beamformer, i.e., the distance between $\boldsymbol{S}_j^R$ and $\boldsymbol{Y}_j$, $\mathcal{D}_j\left(\boldsymbol{S}_j^R,\boldsymbol{Y}_j\right)$, and another term that punishes the processing caused by the NLE gain, i.e., the distance between $\boldsymbol{Y}_j$ and $\boldsymbol{Z}_j$, $\mathcal{D}_j\left(\boldsymbol{Y}_j,\boldsymbol{Z}_j\right)$. That is,

$$\mathcal{D}_j\left(\boldsymbol{S}_j^R,\boldsymbol{Z}_j\right) = \mathcal{D}_j\left(\boldsymbol{S}_j^R,\boldsymbol{Y}_j\right) + \mathcal{D}_j\left(\boldsymbol{Y}_j,\boldsymbol{Z}_j\right). \tag{14}$$

Since the reference signal, S_k^R, is the output of the reference beamformer, $\boldsymbol{v}_k^R$, the minimum processing solution to (8), i.e., $\boldsymbol{w}_{j,k}$ and g_k, should minimize the distance to $\boldsymbol{v}_k^R$ [40]. Therefore, we have that the processing penalty for j'th subband is,

$$\mathcal{D}_j(\boldsymbol{S}_j^R,\boldsymbol{Z}_j) = \left(1-\alpha_j\right)^2 + \left(1-g_j\right)^2, \tag{15}$$

where the details of the derivation are shown in Appendix B. The first term represents the processing penalty imposed on to the beamformer, urging $\boldsymbol{w}_{j,k}$ towards $\boldsymbol{v}_k^R$. The subsequent component signifies the penalty associated with the NLE gain, urging $g_k\boldsymbol{w}_{j,k}$ to approach $\boldsymbol{v}_k^R$ and mitigating potential speech distortions and excessive playback volume induced by the NLE gain [40].

B. PERFORMANCE CRITERIA

For the performance criteria we consider both an intelligibility performance criterion, and a noise power criterion to increase quality performance [40].

1) AUDIBILITY CONSTRAINT

The power of the processed speech within a subband for a given α_j is defined as

$$\delta_{S_j}(\alpha_j) \triangleq \sum_{k\in\mathbb{B}_j}\omega_{j,k}\boldsymbol{w}_k^H C_{S_k}\boldsymbol{w}_k \tag{16}$$

$$= \alpha_j^2\delta_{S_j}^R + (1-\alpha_j)^2\delta_{S_j}^{\mu\text{MWF}} + \alpha_j(1-\alpha_j)\delta_{S_j}^{\text{cross}} \tag{17}$$

where

$$\delta_{S_j}^R \triangleq \sum_{k\in\mathbb{B}_j}\omega_{j,k}\left(\boldsymbol{v}_k^R\right)^H C_{S_k}\boldsymbol{v}_k^R \tag{18}$$

$$\delta_{S_j}^{\mu\text{MWF}} \triangleq \sum_{k\in\mathbb{B}_j}\omega_{j,k}\left(\boldsymbol{w}_k^{\mu\text{MWF}}\right)^H C_{S_k}\boldsymbol{w}_k^{\mu\text{MWF}} \tag{19}$$

$$\delta_{S_j}^{\text{cross}} \triangleq \sum_{k\in\mathbb{B}_j}\omega_{j,k}2\Re\left\{\left(\boldsymbol{w}_k^{\mu\text{MWF}}\right)^H C_{S_k}\boldsymbol{v}_k^R\right\}. \tag{20}$$

A similar definition applies to the processed FE noise subband power, $\delta_{U_j}(\alpha_j)$. We now define the processed NE and FE subband SNRs as

$$\xi_j^N \triangleq \frac{g_j^2\delta_{S_j}(\alpha_j)}{g_j^2\delta_{U_j}(\alpha_j)+\sigma_{N_j}^2}, \qquad \xi_j^F \triangleq \frac{\delta_{S_j}(\alpha_j)}{\delta_{U_j}(\alpha_j)}. \tag{21}$$

Thus, we define (processed) SNR as the ratio between the (processed) speech power and the (processed) noise power. This is in contrast to [11], where the signal-to-distortion ratio is used, i.e., the noise term is defined as the MSE between the clean speech and processed noise signal; hence all noise and speech distortions are considered as noise.

Remark 1: We note that the processed NE SNR is upper bounded by the FE SNR, that is,

$$\lim_{g_j\to\infty}\xi_j^N = \lim_{g_j\to\infty}\frac{g_j^2\delta_{S_j}(\alpha_j)}{g_j^2\delta_{U_j}(\alpha_j)+\sigma_{N_j}^2} = \frac{\delta_{S_j}(\alpha_j)}{\delta_{U_j}(\alpha_j)} = \xi_j^F. \tag{22}$$

Hence, the SI at the NE, as determined by the SNR, is upper bounded by the SI of the signal coming from the FE, since the NE noise can only lower the SI and we must compensate in the best possible way for this using the NLE gain.

Similar to [19] and [40], we derive optimal processing in relation to a performance criterion based on the ASII [18]. The original FSE minimum processing method [11] used SII. Both ASII [18] and SII [41] define SI as a weighted sum of intermediate subband audibility measures, where specifically for the ASII the subband audibility is given as a sigmoidal function of the NE subband SNR, ξ_j^N. We let I_j be a given minimum requirement on the ASII subband audibility performance [18] in subband, j. Then, by definition of the ASII audibility measures [18], it was shown in [19, App. C], that the SI constraint for the j'th subband in terms of the NE subband SNR, ξ_j^N, is [40],

$$\frac{g_j^2\delta_{S_j}(\alpha_j)}{g_j^2\delta_{U_j}(\alpha_j)+\sigma_{N_j}^2} \geq I_j^\xi \tag{23}$$

$$\mathcal{I}_j^{\text{SI}} = g_j^2\left(\delta_{S_j}(\alpha_j) - \delta_{U_j}(\alpha_j)I_j^\xi\right) \geq \sigma_{N_j}^2 I_j^\xi, \tag{24}$$

where $I_j^\xi \triangleq \frac{I_j}{1-I_j}$. As stated above, the NE subband SI is upper bounded by the FE subband SI. Considering the terms inside the parenthesis in (24), we see how the subband SI constraint is only feasible if an α_j exists such that $\xi_j^F > I_j^\xi$, i.e., if the parameterized beamformer can provide a feasible FE SNR. Therefore, as proposed in [40], we define the following parameter,

$$D_j^R \triangleq \delta_{S_j}^R - \delta_{U_j}^R I_j^\xi, \tag{25}$$

which indicates the ability of the reference beamformer, $\boldsymbol{v}_k^R$, to provide a feasible FE SNR. That is, D_j^R is positive only if the processed FE SNR resulting from the reference beamformer is above the desired audibility limit. Similarly, we can define the parameters $D_j^{\mu\text{MWF}}$ and D_j^{cross} which indicate the ability of $\boldsymbol{w}_k^{\mu\text{MWF}}$ and the cross combination of beamformers to provide a feasible FE SNR. Expanding the terms inside the parenthesis in (24) and using the above defined parameters, we can define the polynomial,

$$\begin{aligned} p_{\text{FSE}}(\alpha_j) \triangleq{}& \alpha_j^2 D_j^R + (1-\alpha_j)^2 D_j^{\mu\text{MWF}} \\ &+ \alpha_j(1-\alpha_j)D_j^{\text{cross}} \end{aligned} \tag{26}$$

which represents the ability of the parameterized beamformer to remove sufficient FE noise for various values of α_j. For example, for $\alpha_j = 1$ the polynomial is equal to D_j^R and is positive only if the reference beamformer can remove sufficient FE noise. Using this polynomial, we can write the audibility constraint as

$$g_j^2 p_{\text{FSE}}(\alpha_j) \geq \sigma_{N_j}^2 I_j^\xi. \tag{27}$$

2) NOISE POWER CRITERION

The joint approach has the advantage of having knowledge about the noise situation at both the FE and NE [40]. Thus, in contrast to the blind minimum processing of [11] and [19], we have the ability to try to control the processing in relation to noise in both environments [40].

From (23), it becomes apparent that, to enhance the SNR and meet the audibility requirement, it may be necessary for the processed FE noise to surpass the NE noise. However, the improvement in SI could lead to an undesirable decline in SQ due to elevated overall noise levels, depending on the noise powers. Consequently, to mitigate distortions resulting from excessive noise levels, the joint minimum processing approach introduces a constraint, $\mathcal{I}_j^{\text{NP}}$, on the power of the processed FE noise [40],

$$10\log_{10}\left(g_j^2\delta_{U_j}(\alpha_j)\right) \leq 10\log_{10}\sigma_{N_j}^2 + \Delta_{U_j}. \tag{28}$$

Here, the parameter Δ_{U_j} is used to regulate the amount of dB the processed FE noise power can overpower or must stay below the NE noise power in subband j [40].

C. OPTIMIZATION PROBLEM AND SOLUTION

Combining the cost function and performance constraints, we have that the joint FE and NE minimum processing speech enhancement problem (8) with MSE processing penalty (15), ASII performance constraint (27) and noise power constraint (28) is [40],

$$\begin{aligned} &\underset{\alpha_j, g_j\in\mathbb{R}_+}{\arg\min} \quad \left(1-\alpha_j\right)^2 + \left(1-g_j\right)^2 \qquad (P_0)\\ &\text{s.t. } \mathcal{C}_1: g_j^2 p_{\text{FSE}}(\alpha_j) \geq \sigma_{N_j}^2 I_j^\xi, \quad \mathcal{C}_3: 0 \leq \alpha_j \leq 1, \\ &\qquad \mathcal{C}_2: g_j^2\delta_{U_j}(\alpha_j) \leq \sigma_{N_j}^2 c_{U_j}, \mathcal{C}_4: 1 \leq g_j, \end{aligned}$$

where $c_{U_j} = 10^{\Delta_{U_j}/10}$.

Remark 2: The interaction between the two performance constraints $\mathcal{C}_1$ and $\mathcal{C}_2$ and specifically the parameters I_j^ξ and Δ_{U_j} determine the feasibility of the optimization problem. For example, as the SI target increases, the audibility target, I_j^ξ, also increases, and it becomes more difficult to satisfy constraint $\mathcal{C}_1$. Similarly, as we lower how much the FE noise

is allowed to overpower the NE noise by lowering Δ_{U_j} it becomes increasingly difficult to satisfy constraint $\mathcal{C}_2$.

Thus, if the beamformers are sufficiently good, such that there exists an α_j where $p_{FSE}(\alpha_j) > 0$, then as the audibility constraint increases, it requires a larger NLE gain, g_j to satisfy $\mathcal{C}_1$. The larger NLE gain then leads to increased boosting of the processed FE noise power, $\delta_{U_j}(\alpha_j)$. However, if Δ_{U_j} is chosen such that the FE noise is not allowed to sufficiently overpower the NE noise, the optimization may not have a feasible solution even though the intelligibility constraint is satisfied.

We show that the solution to the optimization problem is found at the boundary of the feasible set or at stationary points. Therefore, we identify the following sets from the constraints,

$$\mathcal{F}^{\mathcal{C}_1} \triangleq \{\alpha \in [0,1] : p_{\text{FSE}}(\alpha_j) \geq \sigma_{N_j}^2 I_j^{\xi}\} \tag{29a}$$

$$\mathcal{F}^{\mathcal{C}_2} \triangleq \{\alpha \in [0,1] : \delta_{U_j}(\alpha_j) \leq \sigma_{N_j}^2 c_{U_j}\} \tag{29b}$$

$$\mathcal{F}^{\text{SI}} \triangleq \left\{\alpha \in [0,1] : \sigma_{N_j}^2 I_j^{\xi} > p_{\text{FSE}}(\alpha_j) > 0\right\} \tag{29c}$$

$$\mathcal{F}^{\text{NP}} \triangleq \left\{\alpha \in [0,1] : I_j^{\xi} \delta_{U_j}(\alpha) - c_{U_j} p_{\text{FSE}}(\alpha) \leq 0\right\}. \tag{29d}$$

Finally, we have the set,

$$\mathcal{F}_S \triangleq \left\{\alpha \in [0,1] : h'(\alpha) = 0 \text{ and } h''(\alpha) \geq 0\right\}, \tag{30}$$

containing the stationary points of the convex regions of the following helper function,

$$h(\alpha_j) \triangleq \sqrt{\frac{\sigma_{N_j}^2 I_j^{\xi}}{p_{\text{FSE}}(\alpha_j)}} - \alpha_j. \tag{31}$$

Details about the first and second derivatives of $h(\alpha_j)$ are provided in Appendix C-A4. Denoting the boundary of a set $\mathcal{F}$ by $\partial\mathcal{F}$, we have the following theorem, stating the solution to the optimization problem.

Theorem 1: The optimal minimum processing beamformer weight, α_j^*, and NLE gain, g_j^*, solution to the optimization problem (p_0) are:

If $\mathcal{F}^{\mathcal{C}_1} \cap \mathcal{F}^{\mathcal{C}_2} \neq \varnothing$ or $\mathcal{F}^{SI} \cap \mathcal{F}^{NP} \neq \varnothing$: α_j^* is the minimum of the stationary and boundary solutions, i.e.,

$$\begin{aligned}\alpha_j^* = &\arg\min_{\alpha} \pi(\alpha), \\ &s.t.\ \alpha \in \left(\mathcal{F}_S \cap \mathcal{F}^{SI} \cap \mathcal{F}^{NP}\right) \\ &\quad \cup \partial\left(\mathcal{F}^{SI} \cap \mathcal{F}^{NP}\right) \cup \partial\left(\mathcal{F}^{\mathcal{C}_1} \cap \mathcal{F}^{\mathcal{C}_2}\right),\end{aligned} \tag{32}$$

where

$$\pi(\alpha) = \begin{cases} 1-\alpha & \text{if } \alpha \in \partial\left(\mathcal{F}^{\mathcal{C}_1} \cap \mathcal{F}^{\mathcal{C}_2}\right) \\ h(\alpha) & \text{if } \alpha \in \left(\mathcal{F}_S \cap \mathcal{F}^{SI} \cap \mathcal{F}^{NP}\right) \\ & \text{or if } \alpha \in \partial\left(\mathcal{F}^{SI} \cap \mathcal{F}^{NP}\right). \end{cases} \tag{33}$$

and

$$g_j^* = \begin{cases} 1 & \text{if } \alpha_j^* \in \partial\left(\mathcal{F}^{\mathcal{C}_1} \cap \mathcal{F}^{\mathcal{C}_2}\right) \\ \sqrt{\dfrac{\sigma_{N_j}^2 I_j^{\xi}}{p_{FSE}(\alpha_j^*)}} & \text{if } \alpha_j^* \in \left(\mathcal{F}_S \cap \mathcal{F}^{SI} \cap \mathcal{F}^{NP}\right) \\ & \text{or if } \alpha_j^* \in \partial\left(\mathcal{F}^{SI} \cap \mathcal{F}^{NP}\right). \end{cases} \tag{34}$$

If $\mathcal{F}^{\mathcal{C}_1} \cap \mathcal{F}^{\mathcal{C}_2} = \varnothing$ and $\mathcal{F}^{SI} \cap \mathcal{F}^{NP} = \varnothing$: No feasible solution exists.

The proof of Theorem 1 is found in Appendix C.

Remark 3: We note that the set of stationary points, $\mathcal{F}_S$ is a finite discrete set. Similarly, the boundary of an interval in $\mathbb{R}$ is a finite discrete set, and hence, the boundaries of the intersections $\partial\left(\mathcal{F}^{\mathcal{C}_1} \cap \mathcal{F}^{\mathcal{C}_2}\right)$ and $\partial\left(\mathcal{F}^{SI} \cap \mathcal{F}^{NP}\right)$ are finite discrete sets. Thus, to find the optimal solution, we avoid searching over a continuum of points and instead only need to compare a small finite number of points. In fact, from the proof in Appendix C it can be seen that, $\|\mathcal{F}_S\|_0 \leq 2$ and $\|\mathcal{F}^i\|_0 \leq 4$, for $i \in \{\mathcal{C}_1, \mathcal{C}_2, SI, NP\}$.

With the optimal solution (α_j^*, g_j^*) the optimal minimum processing beamformer is then expressed as

$$\boldsymbol{w}_{j,k}^* = \alpha_j^* \boldsymbol{v}_k^R + (1-\alpha_j^*)\boldsymbol{w}_k^{\mu\text{MWF}}. \tag{35}$$

Based on the subband definition, various frequencies may contribute to multiple subbands indexed by $\mathbb{F}_k$. Consequently, the optimal beamformer $\boldsymbol{wj}, k^*$ and NLE gain g_j^* can influence multiple subbands. Denoting the weight that reflects the influence of this contribution as $\eta_{j,k}$, the optimal beamformer and NLE gain for each frequency are,

$$\boldsymbol{w}_k^* = \sum_{j \in \mathbb{F}_k} \eta_{j,k} \boldsymbol{w}_{j,k}^* \quad \text{and} \quad g_k^* = \sum_{j \in \mathbb{F}_k} \eta_{j,k} g_j^*. \tag{36}$$

See Appendix A for definition of the weights $\eta_{j,k}$.

1) INFEASIBLE CASES

For the sets in (29), $\mathcal{F}^{\mathcal{C}_1}$ represents the feasible α for which $g = 1$ satisfies $\mathcal{C}_1$, $\mathcal{F}^{\mathcal{C}_2}$ represents the feasible α for which $g = 1$ satisfies $\mathcal{C}_2$, $\mathcal{F}^{\text{SI}}$ represents the feasible α for which we must have $g > 1$ to satisfy $\mathcal{C}_1$, and $\mathcal{F}^{\text{NP}}$ represents the feasible α for which $g^2 = \sigma_{N_j}^2 I_j^{\xi} / p_{\text{FSE}}(\alpha_j) > 1$ satisfies $\mathcal{C}_2$. Particularly, this means, $\mathcal{F}^{\mathcal{C}_1} = \varnothing$ implies no α exists for which the beamformer can provide a feasible FE SNR for the optimal $g = 1$; $\mathcal{F}^{\mathcal{C}_2} = \varnothing$ implies no α exists for which the beamformer can provide a feasible FE noise power relative to NE noise for the optimal $g = 1$; $\mathcal{F}^{\text{SI}} = \varnothing$ implies no α exists for which the beamformer provides a feasible FE SNR where the optimal NLE gain is $g > 1$; and finally $\mathcal{F}^{\text{NP}} = \varnothing$ implies no α exists for which the beamformer provides a feasible FE noise power relative to the NE noise for the optimal $g^2 = \sigma_{N_j}^2 I_j^{\xi} / p_{\text{FSE}}(\alpha_j) > 1$.

For a feasible solution, the constraints $\mathcal{C}_1$ and $\mathcal{C}_2$ must both be satisfied simultaneously. Thus, we have that $\mathcal{F}^{\mathcal{C}_1} \cap \mathcal{F}^{\mathcal{C}_2} = \varnothing$ indicates that no α exists for which the optimal NLE gain is $g = 1$, and $\mathcal{F}^{\text{SI}} \cap \mathcal{F}^{\text{NP}} = \varnothing$

indicates that no α exists for which the optimal NLE gain is $g^2 = \sigma_{N_j}^2 I_j^{\xi} / p_{\text{FSE}}(\alpha_j) > 1$. If either constraint cannot be satisfied individually, then the constraints cannot be jointly feasible. Therefore, and because the intersection between an empty set and another set is always empty, we have a total of 16 different combinations of the sets in (29) that can result in an infeasible solution, cf. Table 2. For simplicity, we combine these configurations into four categories, as indicated in Table 2: (1) infeasibility due to empty intersections, i.e., the beamformer can provide a feasible FE SI and a feasible FE noise power but not for the same α; (2) infeasibility because the FE beamformer cannot provide a sufficient SI for any choice of g; (3) infeasibility because the processed FE noise power is too high in relation to the NE noise power for any choice of g; and (4) infeasibility because the beamformer cannot provide either a feasible FE SI or noise power for any choice of g. In Section VI-A, we further investigate how the number of feasible and infeasible subbands changes with the FE and NE SNRs.

V. EXPERIMENTAL EVALUATION

We investigate performance in two noise scenarios: (Babble-Car) where the target talker is in a babble noise setting and the NE listener is inside a car, and (Car-Babble) the reverse situation where the target talker is in car noise and the NE listener is in babble noise. Additional investigations, not reported here, have shown that the results generalize to other noise types.

TABLE 2. Categorization of infeasible cases for the optimal joint minimum processing optimization problem.

Category	$\mathcal{F}^{\mathcal{C}_1}$	$\mathcal{F}^{\mathcal{C}_2}$	$\mathcal{F}^{\text{SI}}$	$\mathcal{F}^{\text{NP}}$	$\mathcal{F}^{\mathcal{C}_1} \cap \mathcal{F}^{\mathcal{C}_2}$	$\mathcal{F}^{\text{SI}} \cap \mathcal{F}^{\text{NP}}$
Intersection	$\neq \varnothing$	$\neq \varnothing$	$\neq \varnothing$	$\neq \varnothing$	$\varnothing$	$\varnothing$
Intersection	$\neq \varnothing$	$\neq \varnothing$	$\neq \varnothing$	$\varnothing$	$\varnothing$	$\varnothing$
Intersection	$\neq \varnothing$	$\neq \varnothing$	$\varnothing$	$\neq \varnothing$	$\varnothing$	$\varnothing$
Intersection	$\neq \varnothing$	$\neq \varnothing$	$\varnothing$	$\varnothing$	$\varnothing$	$\varnothing$
Intersection	$\neq \varnothing$	$\varnothing$	$\neq \varnothing$	$\neq \varnothing$	$\varnothing$	$\varnothing$
Intersection	$\varnothing$	$\neq \varnothing$	$\neq \varnothing$	$\neq \varnothing$	$\varnothing$	$\varnothing$
Intersection	$\varnothing$	$\varnothing$	$\neq \varnothing$	$\neq \varnothing$	$\varnothing$	$\varnothing$
FE SI	$\varnothing$	$\neq \varnothing$	$\varnothing$	$\neq \varnothing$	$\varnothing$	$\varnothing$
FE SI	$\varnothing$	$\neq \varnothing$	$\varnothing$	$\varnothing$	$\varnothing$	$\varnothing$
FE SI	$\varnothing$	$\varnothing$	$\varnothing$	$\neq \varnothing$	$\varnothing$	$\varnothing$
FE Noise power	$\neq \varnothing$	$\varnothing$	$\neq \varnothing$	$\varnothing$	$\varnothing$	$\varnothing$
FE Noise power	$\neq \varnothing$	$\varnothing$	$\varnothing$	$\varnothing$	$\varnothing$	$\varnothing$
FE Noise power	$\varnothing$	$\varnothing$	$\neq \varnothing$	$\varnothing$	$\varnothing$	$\varnothing$
SI+NP	$\neq \varnothing$	$\varnothing$	$\varnothing$	$\neq \varnothing$	$\varnothing$	$\varnothing$
SI+NP	$\varnothing$	$\neq \varnothing$	$\neq \varnothing$	$\varnothing$	$\varnothing$	$\varnothing$
SI+NP	$\varnothing$	$\varnothing$	$\varnothing$	$\varnothing$	$\varnothing$	$\varnothing$

A. REFERENCE METHODS

We compare performance against three relevant reference methods: (1) **Blind Min**; The blind concatenation of minimum processing FSE [11] and minimum processing NLE [19]. The FSE beamformer of [11] is parameterized according to a reference beamformer and a μ-MWF. Therefore, these beamformers are selected to be the same as those used in the proposed joint method. This allows us to investigate the effect of combining joint processing with minimum processing in the end-to-end speech enhancement setting. (2) **Joint Max**: The joint FE and NE optimization method based on maximizing ASII [39], since this is also a joint approach based on ASII but without minimum processing. Thus, we investigate effects of adding minimum processing to the joint setting. (3) **Blind Max**: The blind concatenation of an MVDR beamformer at the FE (as this is the beamformer used in [39]) and optimal ASII maximization at the NE [18]. Thus, we investigate the performance of the simple classic maximum processing, and the effect of adding minimum processing to the classic blind end-to-end approach. Furthermore, we investigate the effect of using a joint approach [39] against its classic blind counter part without any relation to minimum processing.

The reference methods [18] and [39] are based on a target speech power equality constraint. That is, the output power should match the power of the input signal, P_{ref}. However, the proposed method is allowed to gain the input signal to overpower the NE noise, resulting in a total processed power level of P_{prop}. Therefore, for a fair comparison, the power constrained methods [18] and [39] are implemented with the same power gain to their input signals such that $P_{ref} = P_{prop}$. The blind NLE reference of [19] was, similar to the proposed method, designed to gain the input signal a sufficient amount to overpower the NE noise. However, we cannot control the amount of total applied gain, because the method was designed specifically without this option.

Finally, the blind concatenation methods are implemented such that the NLE parts [18] and [19] interpret the signal coming from the FE, Y_k, as clean speech.

B. HANDLING INFEASIBLE SUBBANDS

The joint minimum processing optimization problem (p_0) is solved per subband. However, as shown in Theorem 1, the existence of a feasible solution is determined by the subband noise and speech powers. Therefore, to investigate performance using real speech and noise signals, infeasible subbands must be handled. In [40], various heuristic approaches were proposed to find good solutions in infeasible subbands. However, to better investigate the performance differences between the proposed joint minimum processing and the blind concatenated minimum processing, in this paper we instead default to using the blind minimum processing concatenation within these infeasible subbands. Thus, all differences between the proposed and blind minimum processing method are due to our joint solution in the feasible subbands.

C. ESTIMATING STATISTICS

The statistics of speech and non-stationary noisy signals change over time, and must therefore be estimated and updated in time. However, updating too fast may lead to abrupt changes in the processing between time frames, leading to audible distortions. Therefore, slow time-varying processing is commonly employed. Hence, in this paper, for simplicity, we estimate the average speech and noise power

per Discrete Four Transform (DFT) bin using a long-term average over several short-time frames,

$$\sigma^2_{S_k} \triangleq \frac{1}{I}\sum_i |S_{k,i}|^2, \tag{37}$$

$$C_{U_k} \triangleq \frac{1}{I}\sum_i \boldsymbol{U}_{k,i}\boldsymbol{U}^H_{k,i}, \tag{38}$$

where I denotes the total number of frames. An expression similar to (37) holds for the NE noise signal, $N_{k,i}$. Thus, the estimated statistics do not change with time, and we process signals in a time-invariant manner. For time-varying processing in practice, the statistics must be updated with, for example, a recursive average. Furthermore, in the simulations, we assume that the speech and noise spectra are known. However, in practical scenarios, the speech and noise spectra must be estimated from noisy microphone recordings [32].

D. EXPERIMENTAL SETUP

Unless otherwise stated, the target speech material used for the evaluations are English sentences from the TIMIT-database [42] test set sampled at 16 kHz. We pad each target speech excerpt with 1.5 s of silence at the beginning and end to allow ramping up the noise level before the speech segment starts and down after it ends, ensuring a more pleasant listening experience in subjective listening tests.

The babble noise is created by mixing talkers from the TIMIT training set. In the (Babble-Car) scenario, we use six talkers per FE noise source position, and in the (Car-Babble) setting, we use six competing talkers at the NE. The car noise is generated by taking a random excerpt of an appropriate length from noise recorded inside a car traveling at 130 km/h.

We consider an FE room with dimensions $3 \times 4 \times 3\,\mathrm{m}^3$, four noise sources at $[0.50, 1.00, 1]$ m, $[0.75, 3.00, 1]$ m, $[3.00, 2.40, 1]$ m, and $[2.70, 1.30, 1]$ m, and a target talker at $[1.50, 3.00, 1]$ m. The FE beamformer has three microphones at $[1.50, 2.00, 1]$ m, $[1.50, 2.02, 1]$ m and $[1.50, 1.98, 1]$ m, where each microphone is subject to a 60 dB SNR white noise. We assume that the room transfer functions are known and generated without reverberation using [43]. The speech and noise signals are converted to the time-frequency domain using an STFT with 32 ms Hann windows with 50% overlap and a sampling rate of 16 kHz. We consider a total of $J = 30$ auditory subband filters with center frequencies linearly spaced on the equivalent rectangular bandwidth scale from 150 Hz to 8000 Hz [44].

We investigate performance with a focus on achieving high SI with minimal speech distortion. Therefore, we select the reference beamformer, v^R, as the MVDR beamformer, and $\mu = 5$ for the μ-MWF. The per-band audibility targets, I_j, and noise power constraints, Δ_{U_j}, are derived from the overall parameters A^* and Δ_U, using the band importance functions of the SII as weights, cf. [19, Sec. IV.B]. Through informal listening and objective scoring we tuned the parameters of the proposed method, such that the target total ASII is $A^* = 0.9$ for all scenarios, while $\Delta_U = 0$ dB in the (Babble-Car) scenario and $\Delta_U = -5$ dB in the (Car-Babble) scenario. We note that these are not generally applicable values, and a new tuning should be made when working with other scenarios.

Finally, to further control the gains and limit excessive sound levels, we set an additional maximum gain limit on the NLE gain, g_j^*, of 60 dB.

VI. OBJECTIVE PERFORMANCE

Performance is evaluated and averaged across 10 trials[1] from the TIMIT dataset. We consider performance in terms of estimated SI with the ESTOI metric and estimated SQ using the PESQ metric. Furthermore, we consider how the number of feasible bands changes in (p_0) with varying FE and NE SNRs.

A. FEASIBLE AND INFEASIBLE SUBBANDS

We consider 30 auditory subband filters and solve the joint minimum processing optimization problem (p_0) for each subband. However, as shown in Theorem 1, a feasible solution may not exist for all subbands depending on the noise situation at the FE and NE. As, these infeasible situations must be handled, it is of great interest to investigate how often we are in an infeasible case (and which one) or in a feasible case. Table 3 shows the average number of feasible and infeasible subbands for each infeasibility category (Table 2) for various SNR and noise combinations, where the average is taken across the 10 TIMIT trials.

The results show that, for fixed NE SNRs, as the FE SNR increases the number of feasible subbands increases, as the number of infeasible subbands due to insufficient FE SI decreases. For higher fixed NE SNRs, the number of infeasible subbands that satisfy neither the SI nor the noise power constraint first increases and then decreases. In particular, in (Babble-car) scenario (Table 4(a)) the number of infeasible subbands owing to empty intersections first increases and then decreases. Thus, as the FE SNR increases, it generally becomes easier for the beamformer to remove sufficient FE noise to satisfy the subband SI and noise power constraints. However, for higher NE SNRs, the remaining FE noise power is still so high that the beamformer cannot maintain the processed FE noise below the NE noise until the FE SNR is sufficiently high and the FE noise power subsides.

For fixed FE SNRs, as the NE SNR increases, the number of feasible subbands decreases. This seems counterintuitive at first, because the number of infeasible subbands due to insufficient FE SI decreases. However, this is caused by an increase in the number of infeasible subbands satisfying neither the SI nor the noise power constraint. Furthermore, for the (Babble-Car) scenario, the number of infeasible subbands owing to empty intersections first increases and then decreases. Thus, as the NE SNR increases, satisfying the

[1] Audio samples available: https://afugls.github.io/Joint_MinProc_FSE_and_NLE/

TABLE 3. Out of a total of 30 subbands the tables show the average number of subbands that were either feasible or infeasible according to the four categories in Table 2, for (a) FE babble and NE car noise and (b) FE car and NE babble noise. Background color scales with the score, brighter colors correspond to higher values.

SNR FE	NE Band Category	−40.0	−32.5	−25.0	−17.5	−10.0	−2.5	5.0	12.5	20.0	27.5	35.0
−12.5	Feasible	2.3	1.9	1.1	1.0	1.0	1.0	0.9	0.8	0.5	0.1	0.0
	Infeas. NP	0.0	0.0	0.0	0.1	0.6	0.7	0.7	0.7	0.7	0.7	0.7
	Infeas. SI	27.0	27.0	26.9	24.7	17.8	8.9	4.9	3.0	1.8	0.8	0.4
	Infeas. SI+NP	0.5	0.5	1.0	3.8	10.3	19.1	23.1	25.1	26.4	27.7	28.3
	Infeas. intersec.	0.2	0.6	1.0	0.4	0.3	0.3	0.4	0.4	0.6	0.7	0.6
−5.0	Feasible	6.1	6.1	4.8	3.1	3.0	2.8	2.4	1.7	0.6	0.2	0.0
	Infeas. NP	0.0	0.0	0.0	0.1	0.7	1.5	1.7	1.9	2.0	2.1	2.1
	Infeas. SI	21.6	21.6	21.6	20.0	11.8	3.2	1.2	0.3	0.0	0.0	0.0
	Infeas. SI+NP	1.2	1.2	1.3	4.4	13.3	21.3	23.2	24.6	25.6	26.6	27.0
	Infeas. intersec.	1.1	1.1	2.3	2.4	1.2	1.2	1.5	1.5	1.8	1.1	0.9
0.0	Feasible	9.8	9.8	9.6	6.0	4.5	4.0	3.1	1.9	0.9	0.2	0.0
	Infeas. NP	0.0	0.0	0.0	0.2	1.2	1.7	1.9	2.1	2.1	2.1	2.1
	Infeas. SI	17.7	17.7	17.7	17.6	11.5	2.1	0.4	0.0	0.0	0.0	0.0
	Infeas. SI+NP	1.7	1.7	1.7	2.2	11.7	21.3	23.2	24.7	26.0	27.0	27.7
	Infeas. intersec.	0.8	0.8	1.0	4.0	1.1	0.9	1.4	1.3	1.0	0.7	0.2
15.0	Feasible	25.6	25.6	25.6	25.6	24.2	9.9	4.7	3.6	2.4	1.0	0.3
	Infeas. NP	0.0	0.0	0.0	0.0	0.0	0.0	0.8	1.4	1.4	1.4	1.4
	Infeas. SI	0.7	0.7	0.7	0.7	0.7	0.6	0.5	0.0	0.0	0.0	0.0
	Infeas. SI+NP	1.3	1.3	1.3	1.3	1.3	7.0	18.7	23.7	25.0	26.5	27.6
	Infeas. intersec.	2.4	2.4	2.4	2.4	3.8	12.5	5.3	1.3	1.2	1.1	0.7

SNR NE	FE Band Category	−40.0	−32.5	−25.0	−17.5	−10.0	−2.5	5.0	12.5	20.0	27.5	35.0
−40.0	Feasible	0.0	0.0	0.0	1.1	3.6	7.9	12.8	20.8	29.4	30.0	30.0
	Infeas. NP	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
	Infeas. SI	30.0	30.0	30.0	28.7	25.6	19.7	14.6	3.1	0.5	0.0	0.0
	Infeas. SI+NP	0.0	0.0	0.0	0.1	0.4	1.1	1.6	3.5	0.1	0.0	0.0
	Infeas. intersec.	0.0	0.0	0.0	0.1	0.4	1.3	1.0	2.6	0.0	0.0	0.0
−30.0	Feasible	0.0	0.0	0.0	0.4	2.9	7.9	12.8	20.8	29.4	30.0	30.0
	Infeas. NP	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
	Infeas. SI	30.0	30.0	30.0	28.7	25.6	19.7	14.6	3.1	0.5	0.0	0.0
	Infeas. SI+NP	0.0	0.0	0.0	0.1	0.4	1.1	1.6	3.5	0.1	0.0	0.0
	Infeas. intersec.	0.0	0.0	0.0	0.8	1.1	1.3	1.0	2.6	0.0	0.0	0.0
−10.0	Feasible	0.0	0.0	0.0	0.3	1.7	3.9	5.7	16.5	29.4	30.0	30.0
	Infeas. NP	0.0	0.0	0.0	0.1	0.2	1.0	0.4	0.0	0.0	0.0	0.0
	Infeas. SI	29.7	28.3	26.1	21.5	14.8	11.3	12.7	3.1	0.5	0.0	0.0
	Infeas. SI+NP	0.3	1.7	3.9	8.1	13.0	12.5	7.5	4.0	0.1	0.0	0.0
	Infeas. intersec.	0.0	0.0	0.0	0.0	0.3	1.3	3.7	6.4	0.0	0.0	0.0
5.0	Feasible	0.0	0.0	0.0	0.3	1.6	3.0	3.7	4.3	6.8	22.8	29.9
	Infeas. NP	0.0	0.0	0.0	0.2	0.7	1.4	2.4	3.7	0.0	0.0	0.0
	Infeas. SI	25.9	21.1	14.5	7.3	3.3	0.7	0.3	0.7	0.5	0.0	0.0
	Infeas. SI+NP	4.1	8.9	15.5	22.1	23.9	23.7	22.7	19.0	9.8	0.9	0.0
	Infeas. intersec.	0.0	0.0	0.0	0.1	0.5	1.2	0.9	2.3	12.9	6.3	0.1

(a) Band counts for far-end Babble and near-end Car noise

SNR FE	NE Band Category	−40.0	−32.5	−25.0	−17.5	−10.0	−2.5	5.0	12.5	20.0	27.5	35.0
−40.0	Feasible	3.8	3.8	3.8	3.8	3.8	3.8	3.8	2.6	0.0	0.0	0.0
	Infeas. NP	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.7	1.5	1.5	1.5
	Infeas. SI	24.7	24.7	24.7	24.7	24.7	24.7	24.4	19.7	10.1	4.6	3.2
	Infeas. SI+NP	1.5	1.5	1.5	1.5	1.5	1.5	1.8	6.8	18.3	23.9	25.3
	Infeas. intersec.	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.2	0.1	0.0	0.0
−10.0	Feasible	26.7	26.7	26.7	26.7	26.7	26.7	26.7	25.3	22.1	13.7	8.5
	Infeas. NP	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	2.3	2.5
	Infeas. SI	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.0
	Infeas. SI+NP	2.5	2.5	2.5	2.5	2.5	2.5	2.5	2.7	5.2	12.9	18.5
	Infeas. intersec.	0.3	0.3	0.3	0.3	0.3	0.3	0.3	1.5	2.2	0.6	0.5
−5.0	Feasible	29.5	29.5	29.5	29.5	29.5	29.5	29.5	28.9	25.7	21.3	10.9
	Infeas. NP	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.5
	Infeas. SI	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
	Infeas. SI+NP	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.7	2.9	8.4	18.6
	Infeas. intersec.	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.4	1.4	0.3	0.0
5.0	Feasible	30.0	30.0	30.0	30.0	30.0	30.0	30.0	30.0	28.7	26.1	24.5
	Infeas. NP	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
	Infeas. SI	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
	Infeas. SI+NP	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.8	3.8	5.5
	Infeas. intersec.	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.5	0.1	0.0

SNR NE	FE Band Category	−40.0	−32.5	−25.0	−17.5	−10.0	−2.5	5.0	12.5	20.0	27.5	35.0
−20.0	Feasible	3.8	5.6	7.8	15.2	26.7	29.5	30.0	30.0	30.0	30.0	30.0
	Infeas. NP	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
	Infeas. SI	24.7	22.3	16.5	4.9	0.5	0.0	0.0	0.0	0.0	0.0	0.0
	Infeas. SI+NP	1.5	2.1	5.6	9.2	2.5	0.5	0.0	0.0	0.0	0.0	0.0
	Infeas. intersec.	0.0	0.0	0.1	0.7	0.3	0.0	0.0	0.0	0.0	0.0	0.0
−10.0	Feasible	3.8	5.6	7.8	15.2	26.7	29.5	30.0	30.0	30.0	30.0	30.0
	Infeas. NP	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
	Infeas. SI	24.7	22.3	16.5	4.9	0.5	0.0	0.0	0.0	0.0	0.0	0.0
	Infeas. SI+NP	1.5	2.1	5.6	9.2	2.5	0.5	0.0	0.0	0.0	0.0	0.0
	Infeas. intersec.	0.0	0.0	0.1	0.7	0.3	0.0	0.0	0.0	0.0	0.0	0.0
−5.0	Feasible	3.8	5.6	7.8	15.2	26.7	29.5	30.0	30.0	30.0	30.0	30.0
	Infeas. NP	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
	Infeas. SI	24.7	22.3	16.5	4.9	0.5	0.0	0.0	0.0	0.0	0.0	0.0
	Infeas. SI+NP	1.5	2.1	5.6	9.2	2.5	0.5	0.0	0.0	0.0	0.0	0.0
	Infeas. intersec.	0.0	0.0	0.1	0.7	0.3	0.0	0.0	0.0	0.0	0.0	0.0
15.0	Feasible	1.5	5.5	7.4	13.5	24.6	28.4	29.8	30.0	30.0	30.0	30.0
	Infeas. NP	1.0	0.9	2.0	0.4	0.0	0.0	0.0	0.0	0.0	0.0	0.0
	Infeas. SI	16.4	13.7	12.8	4.9	0.5	0.0	0.0	0.0	0.0	0.0	0.0
	Infeas. SI+NP	10.8	9.9	7.8	9.2	3.3	1.0	0.1	0.0	0.0	0.0	0.0
	Infeas. intersec.	0.3	0.0	0.0	2.0	1.6	0.6	0.1	0.0	0.0	0.0	0.0

(b) Band counts for far-end Car and near-end Babble noise

SI constraint becomes easier. However, too much FE noise still remains and causes infeasibility because the NE noise is so low that the FE noise still overpowers it.

B. ESTIMATED INTELLIGIBILITY

Tables 5(a) and 5(b) show the ESTOI performance of the proposed and reference methods in the (Babble-Car) and (Car-Babble) scenario, respectively. At low SNRs, all methods demonstrate an improvement in performance compared to the unprocessed scenario, except for instances of FE babble noise with an SNR less than −25 dB for all NE car noise SNRs. In FE car noise scenarios, performance improvement is more substantial than in babble noise situations because car noise is less detrimental to SI and easier to remove, thus providing better conditions for the NLE component to operate effectively. As the SNRs increase, the unprocessed performance naturally improves, making it challenging for the various processing methods to exhibit significant enhancements. Additionally, at higher SNRs, objective scores may penalize processing artifacts, even though the speech is highly intelligible. Thus, the differences in scores at this level may not accurately reflect real-world perceptual improvements.

Comparing processing methods, the maximum processing techniques generally exhibit slightly higher performance than minimum processing methods, with the difference becoming more pronounced at higher SNRs. This is expected because the maximum processing methods are designed to maximize SI. However, because we provided a high SI target the minimum processing methods also yield high ESTOI scores. The high SI target also leads to a larger power increase, which the maximum processing methods utilize to achieve a slightly higher ESTOI score.

When comparing the joint minimum processing and blind minimum processing methods, their performances are relatively similar. However, joint minimum processing tends to have slightly better ESTOI performance than blind minimum processing when the FE SNR is at or above −20 dB for all noise scenarios and NE SNRs. Looking at Table 3, we see that this corresponds to the number of feasible bands starting to rise (if not before depending on NE noise scenario). Because we chose to let the joint minimum processing default to the blind approach in the infeasible subbands, we expect joint minimum processing to behave similar to blind processing if there are no feasible bands. Thus, the differences in performance are caused by the joint solution in the feasible subbands, and joint processing can, for certain SNRs, outperform blind processing in the minimum processing setting. However, for the maximum processing methods, their performances remain largely similar across various noise and SNR combinations, with only marginal differences. In cases where differences exist, it is inconsistent whether joint or blind maximum processing is superior, with blind processing seemingly outperforming more frequently.

TABLE 4. Average ESTOI and PESQ scores in the (Babble-Car) scenario in (a)+(c) and (Car-Babble) scenario in (b)+(d). Best performance is highlighted in bold for each SNR and noise pair. Background color scales with the score. Brighter colors correspond to higher values (normalized per measure). Blind Min: [11] + [19], Joint Max: [39], Blind Max: MVDR + [18].

SNR FE	NE Method	−40.0	−32.5	−25.0	−17.5	−10.0	−2.5	5.0	12.5	20.0
	Unproc.	.123	.178	.214	.241	**.267**	**.285**	**.292**	**.294**	**.295**
	Joint Min	**.259**	**.260**	**.260**	**.260**	.261	.263	.267	.273	.279
−12.5	Blind Min	.250	.249	.250	.250	.250	.253	.259	.269	.278
	Joint Max	.251	.251	.251	.251	.252	.255	.260	.262	.262
	Blind Max	.252	.253	.253	.253	.254	.257	.262	.264	.264
	Unproc.	.155	.256	.330	.386	.444	**.497**	**.529**	**.542**	**.545**
	Joint Min	.442	.442	.442	.443	.446	.458	.478	.498	.511
−5.0	Blind Min	.425	.425	.425	.426	.429	.440	.466	.494	.511
	Joint Max	**.448**	**.448**	**.448**	**.449**	**.453**	.470	.486	.490	.491
	Blind Max	.446	.446	.446	.447	.451	.471	.488	.493	.493
	Unproc.	.164	.282	.382	.458	.541	**.621**	**.677**	**.703**	**.710**
	Joint Min	.548	.548	.548	.549	.554	.583	.626	.655	.670
0.0	Blind Min	.530	.530	.531	.532	.538	.561	.611	.650	.669
	Joint Max	**.571**	**.571**	**.571**	**.571**	**.581**	.618	.643	.650	.651
	Blind Max	.566	.566	.566	.566	.577	.618	.645	.652	.653
	Unproc.	.170	.301	.429	.537	.658	.790	**.895**	**.945**	**.963**
	Joint Min	.670	.670	.670	.672	.687	.746	.836	.897	.926
15.0	Blind Min	.667	.667	.667	.669	.685	.743	.833	.896	.926
	Joint Max	**.736**	**.736**	**.736**	**.737**	**.760**	**.840**	.895	.913	.917
	Blind Max	.731	.731	.731	.733	.757	.840	.895	.913	.917

SNR NE	FE Method	−40.0	−32.5	−25.0	−17.5	−10.0	−2.5	5.0	12.5	20.0
	Unproc.	.006	.014	.041	.089	.137	.161	.168	.170	.171
	Joint Min	.007	.015	**.057**	**.158**	**.318**	.500	.615	.663	.676
−40.0	Blind Min	.007	.015	**.057**	.153	.306	.481	.605	.659	.675
	Joint Max	.011	.019	.053	.148	.312	**.513**	**.654**	**.721**	**.737**
	Blind Max	**.011**	**.020**	.054	.150	.313	.509	.648	.716	.732
	Unproc.	.004	.015	.046	.124	.226	.305	.336	.345	.347
	Joint Min	.004	.017	**.058**	**.157**	**.318**	.500	.615	.663	.676
−30.0	Blind Min	.004	.017	**.058**	.153	.306	.481	.605	.659	.675
	Joint Max	.012	.020	.054	.148	.313	**.513**	**.654**	**.721**	**.737**
	Blind Max	**.013**	**.021**	.055	.150	.314	.509	.648	.716	.732
	Unproc.	.003	.012	.053	**.160**	**.326**	.496	.605	.651	.664
	Joint Min	.006	.017	**.057**	.158	.319	.505	.627	.679	.694
−10.0	Blind Min	.006	.017	**.057**	.154	.307	.486	.616	.676	.693
	Joint Max	.013	.021	.054	.149	.314	**.520**	**.674**	**.746**	**.766**
	Blind Max	**.013**	**.021**	.055	.150	.316	.517	.670	.742	.763
	Unproc.	.004	.013	.054	**.167**	**.367**	**.607**	**.787**	**.880**	.911
	Joint Min	.006	.018	**.058**	.160	.332	.556	.736	.822	.850
5.0	Blind Min	.006	.018	**.058**	.157	.322	.541	.724	.818	.850
	Joint Max	.013	.021	.054	.151	.329	.567	.769	.877	**.914**
	Blind Max	**.013**	**.021**	.055	.152	.331	.569	.770	.877	**.914**

(a) ESTOI scores for FE Babble and NE Car noise

SNR FE	NE Method	−40.0	−32.5	−25.0	−17.5	−10.0	−2.5	5.0	12.5	20.0
	Unproc.	.005	.009	.023	.063	.137	.212	.254	.268	.271
	Joint Min	.259	.271	.277	.277	.277	.277	.277	.277	.279
−40.0	Blind Min	.260	.272	.278	.278	.278	.278	.278	.278	.279
	Joint Max	.275	.281	.282	.283	.283	.283	.283	.283	.283
	Blind Max	**.284**	**.285**	**.285**	**.285**	**.285**	**.285**	**.285**	**.285**	**.285**
	Unproc.	.006	.009	.025	.081	.206	.391	.577	.702	.756
	Joint Min	.675	.685	.685	.685	.685	.688	.699	**.727**	**.760**
−10.0	Blind Min	.649	.657	.657	.657	.657	.659	.671	.718	.759
	Joint Max	.710	.722	.723	.723	.723	.723	.723	.724	.729
	Blind Max	**.713**	**.725**	**.726**	**.726**	**.726**	**.726**	**.726**	.727	.732
	Unproc.	.006	.009	.025	.082	.213	.411	.621	.769	**.837**
	Joint Min	.701	.711	.712	.712	.712	.717	.744	.795	.830
−5.0	Blind Min	.683	.693	.694	.694	.694	.697	.721	.786	.829
	Joint Max	.778	.792	.794	.794	.794	.794	.794	.795	.802
	Blind Max	**.780**	**.795**	**.796**	**.796**	**.796**	**.796**	**.796**	**.797**	.805
	Unproc.	.006	.009	.025	.082	.217	.429	.667	.845	**.934**
	Joint Min	.719	.731	.732	.732	.732	.740	.787	.868	.910
5.0	Blind Min	.715	.726	.727	.727	.727	.734	.777	.863	.909
	Joint Max	.865	.882	.884	.884	.884	.884	.884	.885	.896
	Blind Max	**.865**	**.882**	**.884**	**.884**	**.884**	**.884**	**.884**	**.886**	.897

SNR NE	FE Method	−40.0	−32.5	−25.0	−17.5	−10.0	−2.5	5.0	12.5	20.0
	Unproc.	.046	.049	.050	.052	.055	.056	.056	.056	.056
	Joint Min	.277	.392	.491	.587	.685	.721	.732	.734	.735
−20.0	Blind Min	.278	.393	.490	.579	.657	.706	.727	.733	.735
	Joint Max	.283	.401	.502	.607	.723	.824	.884	**.906**	**.911**
	Blind Max	**.285**	**.405**	**.506**	**.610**	**.726**	**.825**	**.884**	**.906**	**.911**
	Unproc.	.137	.164	.181	.194	.206	.215	.217	.218	.218
	Joint Min	.277	.392	.491	.587	.685	.721	.732	.734	.735
−10.0	Blind Min	.278	.393	.490	.579	.657	.706	.727	.733	.735
	Joint Max	.283	.401	.502	.607	.723	.824	.884	**.906**	**.911**
	Blind Max	**.285**	**.406**	**.506**	**.610**	**.726**	**.825**	**.884**	**.906**	**.911**
	Unproc.	.190	.239	.271	.300	.326	.345	.352	.353	.354
	Joint Min	.277	.392	.491	.587	.686	.724	.735	.738	.738
−5.0	Blind Min	.278	.393	.490	.579	.658	.708	.730	.736	.738
	Joint Max	.283	.401	.502	.607	.723	.824	.884	**.906**	**.911**
	Blind Max	**.285**	**.406**	**.506**	**.610**	**.726**	**.825**	**.884**	**.906**	**.911**
	Unproc.	.270	.395	.500	.610	.727	.829	.884	.901	.905
	Joint Min	.278	.394	.496	.610	**.739**	**.836**	.886	.901	.905
15.0	Blind Min	.279	.395	.497	.608	.735	.831	.882	.900	.904
	Joint Max	.283	.401	.503	.607	.724	.826	.887	.909	.915
	Blind Max	**.285**	**.406**	**.507**	**.610**	.727	.828	**.887**	**.909**	**.915**

(b) ESTOI scores for FE Car and NE Babble noise

SNR FE	NE Method	−17.5	−10.0	−2.5	5.0	12.5	20.0	27.5	35.0
	Unproc.	1.06	**1.11**	**1.15**	1.12	1.13	1.09	1.08	1.13
	Joint Min	1.09	1.09	1.10	1.13	1.27	1.13	**1.14**	1.15
−12.5	Blind Min	**1.10**	1.10	1.11	**1.14**	**1.27**	**1.13**	**1.14**	1.15
	Joint Max	1.05	1.05	1.05	1.13	1.03	1.03	1.05	1.05
	Blind Max	1.04	1.05	1.06	1.06	1.06	1.07	1.06	**1.19**
	Unproc.	1.04	**1.10**	**1.15**	**1.18**	1.20	1.20	1.20	1.20
	Joint Min	1.09	1.08	1.10	1.16	1.19	1.23	1.23	**1.23**
−5.0	Blind Min	**1.10**	1.09	1.11	1.17	**1.20**	**1.23**	**1.23**	**1.23**
	Joint Max	1.04	1.04	1.05	1.07	1.08	1.08	1.08	1.08
	Blind Max	1.03	1.04	1.05	1.07	1.07	1.08	1.08	1.08
	Unproc.	1.04	**1.12**	**1.24**	**1.31**	**1.35**	1.36	1.37	1.37
	Joint Min	1.11	1.12	1.18	1.27	1.33	1.37	1.38	**1.38**
0.0	Blind Min	**1.12**	1.12	1.18	1.28	1.34	**1.38**	**1.38**	**1.38**
	Joint Max	1.04	1.04	1.07	1.12	1.14	1.14	1.14	1.14
	Blind Max	1.04	1.04	1.07	1.12	1.14	1.14	1.14	1.14
	Unproc.	1.05	1.18	**1.54**	**2.03**	**2.45**	**2.64**	**2.70**	**2.72**
	Joint Min	**1.18**	**1.20**	1.43	1.86	2.22	2.46	2.53	2.56
15.0	Blind Min	1.15	1.18	1.42	1.86	2.22	2.46	2.54	2.56
	Joint Max	1.05	1.06	1.18	1.52	1.87	2.01	2.04	2.04
	Blind Max	1.05	1.07	1.18	1.52	1.87	2.01	2.04	2.05

SNR NE	FE Method	−17.5	−10.0	−2.5	5.0	12.5	20.0	27.5	35.0
	Unproc.	1.10	**1.13**	1.02	1.02	1.03	1.03	1.03	1.02
	Joint Min	1.16	1.07	1.10	**1.14**	**1.17**	**1.17**	**1.16**	**1.15**
−40.0	Blind Min	1.18	1.08	**1.11**	1.14	1.15	1.15	1.15	1.15
	Joint Max	1.06	1.04	1.04	1.04	1.05	1.05	1.05	1.05
	Blind Max	**1.22**	1.03	1.04	1.04	1.05	1.05	1.05	1.05
	Unproc.	1.08	1.02	1.02	1.02	1.02	1.02	1.02	1.02
	Joint Min	1.17	1.07	1.10	**1.14**	**1.17**	**1.17**	**1.16**	**1.16**
−30.0	Blind Min	**1.18**	**1.08**	**1.11**	1.14	1.15	1.16	1.16	1.16
	Joint Max	1.05	1.03	1.04	1.04	1.05	1.05	1.05	1.05
	Blind Max	1.11	1.03	1.04	1.04	1.05	1.05	1.05	1.05
	Unproc.	1.16	**1.07**	**1.11**	**1.15**	1.17	1.18	1.18	1.18
	Joint Min	**1.18**	1.06	1.10	1.15	**1.19**	**1.20**	**1.19**	**1.19**
−10.0	Blind Min	1.16	1.07	1.10	1.15	1.18	1.18	1.19	1.19
	Joint Max	1.04	1.04	1.04	1.05	1.06	1.06	1.06	1.06
	Blind Max	1.11	1.04	1.04	1.05	1.06	1.06	1.06	1.06
	Unproc.	1.11	1.10	**1.23**	**1.52**	**1.92**	**2.18**	**2.27**	**2.29**
	Joint Min	1.14	1.10	1.21	1.45	1.77	1.96	2.03	2.05
−5.0	Blind Min	1.14	**1.11**	1.22	1.46	1.77	1.96	2.03	2.05
	Joint Max	1.08	1.05	1.09	1.22	1.45	1.61	1.66	1.67
	Blind Max	**1.19**	1.05	1.09	1.22	1.45	1.61	1.66	1.67

(c) PESQ scores for FE Babble and NE Car noise

SNR FE	NE Method	−17.5	−10.0	−2.5	5.0	12.5	20.0	27.5	35.0
	Unproc.	1.03	1.03	1.02	1.02	1.02	1.02	1.02	1.02
	Joint Min	1.20	1.20	1.20	1.20	1.19	**1.17**	**1.14**	**1.14**
−40.0	Blind Min	**1.20**	**1.20**	**1.20**	**1.20**	**1.20**	**1.17**	**1.14**	**1.14**
	Joint Max	1.07	1.07	1.07	1.07	1.07	1.07	1.07	1.07
	Blind Max	1.07	1.07	1.07	1.07	1.07	1.07	1.07	1.07
	Unproc.	1.04	1.06	1.11	1.19	1.33	1.45	1.50	1.51
	Joint Min	2.04	2.04	2.06	2.07	**2.10**	**2.11**	2.31	2.41
−10.0	Blind Min	**2.06**	**2.06**	**2.07**	**2.09**	2.10	2.10	**2.31**	**2.42**
	Joint Max	1.50	1.50	1.50	1.50	1.50	1.53	1.65	1.71
	Blind Max	1.54	1.54	1.54	1.54	1.54	1.58	1.72	1.78
	Unproc.	1.04	1.06	1.12	1.23	1.49	1.77	1.91	1.95
	Joint Min	**2.12**	**2.12**	**2.13**	**2.15**	2.18	**2.23**	2.54	2.71
−5.0	Blind Min	2.11	2.11	2.13	2.15	**2.18**	2.22	**2.54**	**2.71**
	Joint Max	1.63	1.63	1.63	1.63	1.64	1.69	1.91	2.04
	Blind Max	1.65	1.65	1.65	1.65	1.66	1.72	1.96	2.09
	Unproc.	1.04	1.06	1.13	1.27	1.65	2.23	2.70	2.92
	Joint Min	**2.14**	**2.14**	**2.16**	**2.19**	2.25	2.39	2.93	3.34
5.0	Blind Min	2.14	2.14	2.16	2.19	**2.25**	**2.39**	**2.93**	**3.34**
	Joint Max	1.82	1.82	1.82	1.82	1.83	1.93	2.43	2.79
	Blind Max	1.82	1.82	1.82	1.82	1.84	1.94	2.43	2.80

SNR NE	FE Method	−17.5	−10.0	−2.5	5.0	12.5	20.0	27.5	35.0
	Unproc.	1.04	1.05	1.05	1.05	1.04	1.05	1.05	1.05
	Joint Min	1.84	2.04	**2.13**	**2.14**	**2.14**	**2.14**	**2.14**	**2.14**
−20.0	Blind Min	**1.85**	**2.06**	2.13	2.14	2.14	2.14	2.14	**2.14**
	Joint Max	1.31	1.50	1.69	1.82	1.87	1.88	1.89	1.89
	Blind Max	1.35	1.54	1.71	1.82	1.87	1.88	1.89	1.89
	Unproc.	1.05	1.06	1.06	1.06	1.06	1.06	1.06	1.06
	Joint Min	1.84	2.04	**2.13**	**2.14**	**2.14**	**2.14**	**2.14**	**2.14**
−10.0	Blind Min	**1.85**	**2.06**	2.13	2.14	2.14	2.14	**2.14**	**2.14**
	Joint Max	1.31	1.50	1.69	1.82	1.87	1.88	1.89	1.89
	Blind Max	1.35	1.54	1.71	1.82	1.87	1.88	1.89	1.89
	Unproc.	1.07	1.09	1.09	1.10	1.10	1.10	1.10	1.10
	Joint Min	1.84	2.05	**2.14**	**2.15**	**2.15**	**2.15**	**2.15**	**2.15**
−5.0	Blind Min	**1.86**	**2.06**	2.13	2.15	2.15	2.15	2.15	**2.15**
	Joint Max	1.31	1.50	1.69	1.82	1.87	1.88	1.89	1.89
	Blind Max	1.35	1.54	1.71	1.82	1.87	1.88	1.89	1.89
	Unproc.	1.12	1.38	1.67	1.84	1.94	1.96	1.96	1.96
	Joint Min	1.85	**2.11**	2.22	2.28	**2.31**	**2.32**	**2.32**	**2.32**
15.0	Blind Min	**1.85**	2.09	**2.22**	**2.28**	2.31	2.32	**2.32**	**2.32**
	Joint Max	1.31	1.50	1.71	1.85	1.90	1.91	1.92	1.92
	Blind Max	1.36	1.55	1.72	1.85	1.90	1.91	1.92	1.92

(d) PESQ scores for FE Car and NE Babble noise

This is surprising since the joint maximum processing method has no feasibility issues; thus, it should always be able to utilize joint knowledge to outperform blind maximum processing. Hence, based on objective SI estimation using ESTOI, it is inconclusive whether joint processing can consistently enhance performance over blind processing.

Finally, as either noise type vanishes (very high SNRs), we see, as expected, that there are no differences between the joint and blind methods. This is expected because if there is little to no noise at either the FE or NE, the end-to-end communication scenario tends towards a single-end scenario. Hence, both the joint and blind methods operate on the same terms, and it does not matter if the joint method can utilize joint knowledge if there is not much to be knowledgeable about. Thus, using the proposed joint minimum processing method, we automatically obtain the performance and behavior seen in the single-sided minimum processing [11], [19].

C. ESTIMATED QUALITY

Tables 5(c) and 5(d) show the PESQ performance of the proposed and reference methods in the (Babble-Car) and (Car-Babble) scenario, respectively. We note that PESQ is very sensitive to noise, and PESQ scores tend to be dominated by noise at lower SNRs, thus making it very difficult to discern any differences when noise is the dominant component [11]. As the SNRs increase, the environmental FE noise and environmental NE noise become less dominant, and the PESQ scores increase, and improvements become evident. Furthermore, PESQ compares clean speech to the (processed) noisy signal, thus the unprocessed scenarios may receive the highest score in higher SNRs where speech distortions exceed the noise level.

In the (Babble-Car) noise scenario, with FE SNRs below 0 dB and NE SNRs below -5 dB, PESQ scores remain consistently low, and differences are indiscernible between processed and unprocessed signals. For higher SNRs, the performance of the processing methods improves. However, the unprocessed performance remains the highest, with the minimum processing methods being able to better maintain a low distortion and higher speech quality than the maximum processing methods. For the (Car-Babble) noise scenario, all methods can improve performance over the unprocessed signal, even at low SNRs, where PESQ can be dominated by the noise signal [11], and increases in SI lead to increases in SQ as the speech signal becomes clearer within the noise.

Comparing processing methods, the minimum processing methods consistently outperform the maximum processing methods, particularly at high SNRs, where they exhibit improved speech quality compared to their maximum processing counterparts. At lower SNRs, the flexibility of the minimum processing methods to adopt a more aggressive MWF beamformer allows them to remove more noise than the MVDR beamformer employed in the maximum processing methods, resulting in a more substantial increase in PESQ. Thus, minimum processing methods achieve competitive estimated objective SI on par with maximum processing methods while concurrently achieving superior estimated objective SQ.

Comparing between joint and blind methods, there is no difference in performance in either the minimum or the maximum processing case. As mentioned earlier, at high FE SNRs, there are many feasible bands, cf. Table 3, but not much to gain using a joint approach method because there is not much noise to consider and the terms are the same for both joint and blind methods. However, surprisingly we see no difference at the medium SNRs where there are still many feasible bands. Again it is interesting, that there is no difference between joint and blind maximum processing, because there are no feasibility issues, so joint should always be able to utilize its joint knowledge to increase performance.

VII. SUBJECTIVE PERFORMANCE

In this section, we evaluate the performance of the proposed joint minimum processing along with the reference methods in subjective listening tests.

The objective results showed that the joint minimum processing method obtained a slightly better estimated SI than the blind minimum processing. However, the results were inconclusive regarding the effects of joint processing in the maximum processing setting. Although there was a small advantage to maximum processing over minimum processing in the estimated SI. Additionally, there were no differences between joint and blind processing in terms of estimated SQ, but there was a clearer difference between minimum and maximum processing in estimated SQ. However, subjective intelligibility and quality may not always be well represented by objective measures, and the differences observed in these methods may not reflect realistic performance [10]. Furthermore, we conducted informal listening tests, which indicated that it was difficult to distinguish between the intelligibility of the evaluated methods, even when there was a large difference in objective performance.

Therefore, to further evaluate the performance differences between the joint and blind methods, and the minimum processing and SI maximization methods, we conducted subjective listening tests. We perform a listening test for SI in combination with listening effort, and a separate listening test for perceptual SQ. We investigate self-reported listening effort, since SI also affects listening effort [27], and we only found small differences in estimated SI. Thus, it is interesting to see if joint processing has a clearer effect on listening effort than on SI.

The SNRs used in the listening tests were chosen such that there is noise present at both ends since we are interested in the joint setup with end-to-end noisy communication. If we consider very high SNRs at either end, we would repeat the studies of [11] and [19].

A. SHARED SETUP AND PROCEDURE

Both listening tests were performed in a silent room. A Lenovo T460s laptop connected to an external monitor

was used for reporting and displaying the user interface. The laptop was also equipped with a USB sound card (DragonFly Black) and a set of closed over-ear headphones (Beyerdynamic DT-770 Pro 32 ohm) for audio playback. All audio stimuli in both tests were normalized to a perceived loudness of −31 LUFS according to the EBU R128 [45] recommendation for loudness normalization as implemented in `ffmpeg-normalize`.[2] The test participants were allowed to adjust the overall volume to a comfortable level during a prior training session of each test. Both tests included a short training session to familiarize the participants with the test procedure, audio stimuli, and user interface and limit learning bias. The training scores were not included in the final test results. Finally, to limit listening fatigue, participants were not allowed to participate in both tests on the same day.

B. SPEECH INTELLIGIBILITY AND LISTENING EFFORT TEST

1) SETUP

A total of 22 (3 female, 19 male) native Danish speaking untrained listeners, with an age span of 25 to 65 years and an average age of 39.9 years, volunteered for participation. All participants had self-reported normal hearing. The average test time, including the training, was 52 minutes. The user interface was based on [46] and was modified to also enable self-reported listening effort.

2) PROCEDURE

We performed a closed-vocabulary matrix test combined with an additional rating of listening effort. The speech material used for this test was the Danish Dantale II corpus [47]. The utterances in the Dantale II corpus were recorded by a single female native Danish speaker in silent conditions. Each utterance has a syntactical structure of `name + verb + numeral + adjective + object` and was generated by randomly choosing a word from a set of 10 different candidate words for each word class [47].

For a series of trials, the user interface presented a matrix of all the candidate words for each of the five word classes. Participants initiated audio playback via a mouse click, and the audio was played *only once*. Subsequently, participants selected the words heard from the matrix of candidate words using the mouse. When participants were satisfied with their word selection, a second window automatically opened where participants were asked to rate, using a slider interface, how much listening effort they spent on understanding the words on a scale from 0 (no effort) to 10 (maximum effort). After rating the listening effort, the interface returned to the word matrix and the stimulus of the next trial was automatically played. This procedure was repeated until the end of the test. Intelligibility is measured as the percentage of correctly identified words.

Each test consisted of 2 noise scenarios × 2 SNR pairs × 5 processing types (including unprocessed) × 6 sentences = 120 trials. The target speech and the order of the trials were random for each participant. For the (Babble-Car) scenario, performance was evaluated at FE and NE SNR pairs (−12.5 dB, −40 dB) and (−7.5 dB, −30 dB). In the (Car-Babble) scenario, performance was evaluated at FE and NE SNR pairs (−40 dB, −20 dB) and (−10 dB, −10 dB). The SNR that showed the largest difference in ESTOI scores did not show differences in SI in informal listening, and indicated full intelligibility. Therefore, we performed the listening test at more severe SNRs where it might be possible to detect differences in SI.

The training session consisted of 20 trials (1 sentence per noise/SNR/processing pair).

C. SPEECH QUALITY TEST

1) SETUP

The speech quality listening test was conducted by 25 (4 female, 21 male) volunteer untrained listeners, with an age span of 24 to 65 years and an average age of 39.3 years. All participants had self-reported normal hearing. The average test time, including the training, was 46 minutes. The speech material used for the speech quality listening test was sentences from the English TIMIT test set. The test was conducted using a user interface that was slightly modified from [48].

2) PROCEDURE

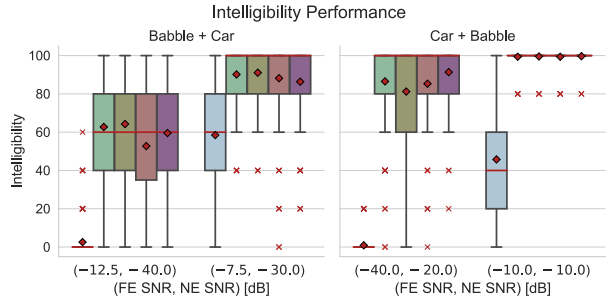
We carried out a listening test using the MUlti Stimulus with Hidden Reference and Anchor (MUSHRA) paradigm [49]. Participants evaluated audio quality on a scale ranging from 0 to 100, segmented into five equal intervals denoted as *bad, poor, fair, good*, and *excellent*. Participants were specifically directed to assess the *basic audio quality* in comparison to a known reference signal, with no additional specifications provided for the definition of audio quality. Each participant was presented with a sequence of 2 noise scenarios × 2 SNR pairs × 4 sentences = 16 trials. Both the reference sentences and the order of the trials were random for each participant. Each trial consisted of a clean reference signal and 7 other signals to be rated: 1 hidden reference, the 4 systems under test, 1 unprocessed signal, and 1 hidden anchor (unprocessed signal at lower FE and NE SNRs). For the (Babble-Car) scenario, performance was evaluated at the FE and NE SNR pairs (0 dB, −10 dB) and (15 dB, 5 dB), and the anchor SNR pair was (−10 dB, −20 dB). In the (Car-Babble) scenario, performance was evaluated at the FE and NE SNR pairs (−5 dB, −5 dB) and (5 dB, 20 dB), and the anchor SNR pair was (−25 dB, −10 dB).

The training session consisted of four trials (1 sentence per noise-SNR pair).

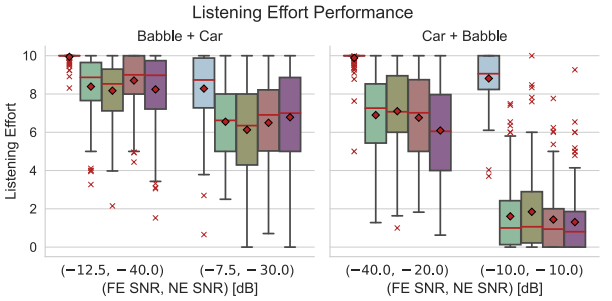
D. LISTENING TEST RESULTS

Figure 4 shows box plots of the results of the three listening tests for each noise, processing, and SNR condition. For statistical significance tests of the results, we consider the

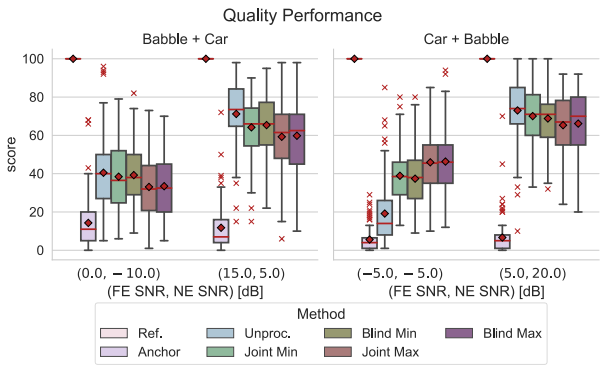
[2]https://github.com/slhck/ffmpeg-normalize



(a) Speech intelligibility test results. Higher is better.



(b) Listening effort test results. Lower is better.



(c) MUSHRA test results. Higher is better.

FIGURE 4. Boxplot of the (a) speech intelligibility, (b) listening effort, and (c) speech quality test results for FE babble and NE car noise (left) and FE car and NE babble noise (right). Medians and means are indicated by red horizontal lines and diamonds, respectively. Outliers are indicated by red crosses. Legend from bottom plot applies to all figures (no reference or anchor was used in the (a) intelligibility and (b) listening effort tests). Blind Min: [11] + [19], Joint Max: [39], Blind Max: MVDR + [18].

TABLE 5. *p*-values for (a) the intelligibility, (b) listening effort, and (c) MUSHRA test. *p*-values below a Bonferroni corrected significance level of 0.005 are marked in bold. Blind Min: [11] + [19], Joint Max: [39], Blind Max: MVDR + [18].

p-values	Babble + Car		Car + Babble	
Comparison	$(-12.5, -40)$	$(-7.5, -30)$	$(-40, -20)$	$(-10, -10)$
Joint Min - Blind Min	0.6657	0.6102	0.0426	0.7022
Joint Min - Joint Max	0.0075	0.9908	0.2195	1.0000
Joint Min - Blind Max	0.4083	0.1598	0.1328	0.4097
Joint Min - Unproc.	**0.0000**	**0.0000**	**0.0000**	**0.0000**
Blind Min - Joint Max	**0.0023**	0.6160	0.3242	0.7022
Blind Min - Blind Max	0.2230	0.0585	**0.0003**	0.6522
Blind Min - Unproc.	**0.0000**	**0.0000**	**0.0000**	**0.0000**
Joint Max - Blind Max	0.0600	0.1878	0.0051	0.4097
Joint Max - Unproc.	**0.0000**	**0.0000**	**0.0000**	**0.0000**
Blind Max - Unproc.	**0.0000**	**0.0000**	**0.0000**	**0.0000**

(a) *p*-values of the intelligibility test.

p-values	Babble + Car		Car + Babble	
Comparison	$(-12.5, -40)$	$(-7.5, -30)$	$(-40, -20)$	$(-10, -10)$
Joint Min - Blind Min	0.22380	0.30091	0.49752	0.43898
Joint Min - Joint Max	0.08729	0.80512	0.73846	0.44787
Joint Min - Blind Max	0.96382	0.25050	**0.00382**	0.12780
Joint Min - Unproc.	**0.00000**	**0.00000**	**0.00000**	**0.00000**
Blind Min - Joint Max	**0.00373**	0.24302	0.25136	0.12982
Blind Min - Blind Max	0.28745	0.03407	**0.00051**	0.02486
Blind Min - Unproc.	**0.00000**	**0.00000**	**0.00000**	**0.00000**
Joint Max - Blind Max	0.09525	0.29859	0.02168	0.41195
Joint Max - Unproc.	**0.00000**	**0.00000**	**0.00000**	**0.00000**
Blind Max - Unproc.	**0.00000**	**0.00000**	**0.00000**	**0.00000**

(b) *p*-values of the listening effort test.

p-values	Babble + Car		Car + Babble	
Comparison	$(0, -10)$	$(15, 5)$	$(-5, -5)$	$(5, 20)$
Joint Min - Blind Min	0.70117	0.62662	0.46110	0.51945
Joint Min - Joint Max	0.03357	0.03194	**0.00063**	0.04592
Joint Min - Blind Max	0.04324	0.06199	**0.00066**	0.11483
Joint Min - Unproc.	0.58908	**0.00079**	**0.00000**	0.13086
Blind Min - Joint Max	0.00937	0.00950	**0.00013**	0.16240
Blind Min - Blind Max	0.01275	0.02265	**0.00013**	0.34973
Blind Min - Unproc.	0.92308	**0.00371**	**0.00000**	0.02867
Joint Max - Blind Max	0.89596	0.89014	0.98147	0.67942
Joint Max - Unproc.	0.00659	**0.00000**	**0.00000**	**0.00055**
Blind Max - Unproc.	0.01126	**0.00000**	**0.00000**	**0.00253**

(c) *p*-values of the MUSHRA test.

nonparametric Kruskal-Wallis H test [50], since the assumption of normal distribution of the data is invalid according to the Kolmogorov-Smirnov test [51], and given the number of participants and their different interpretations of the scales [52]. The p-values for the comparisons considered in this paper are listed in Table 5. p-values are considered significant and are marked in bold if $p < 0.05/m = 0.005$ ($m = 10$), where we have corrected the significance level with the Bonferroni method [53], and m is the number of tested hypotheses.

1) SPEECH INTELLIGIBILITY

Considering Fig. 4(a) and Table 5(b) jointly, the results show across both noise scenarios and all SNR pairs that all processing methods significantly enhance SI compared to the unprocessed condition, a result consistent with the objective ESTOI scores. However, generally, there are large variations in the data, and it is difficult to determine a particular trend. Notably, no statistically significant differences were observed between the maximum and minimum processing methods. Similarly, no significant differences were found between joint and blind processing, with the exception of the (Car-Babble) scenario at very low SNRs. Surprisingly, in this case, blind minimum processing demonstrated better performance than joint and blind maximum processing. In the (Babble-Car) scenario, there appears to be a slight, insignificant advantage favoring minimum processing methods over maximum processing; this trend is also reflected in the ESTOI scores. Despite the expectation that maximum processing should outperform minimum processing and

joint processing outperform blind processing, the limited participant pool precluded the determination of statistically significant performance differences.

2) LISTENING EFFORT

Considering the listening effort results in Fig. 4(b) and Table 5(b), the results show that all methods significantly alleviate listening effort compared to the unprocessed condition across both noise scenarios and all SNR pairs. However, no discernible significant differences were observed between the maximum and minimum processing methods. Similarly, no clear distinctions were observed between joint and blind processing, with the exception of the low SNR case in the (Car-Babble) noise scenario, where blind maximum processing exhibited a significantly better outcome than the minimum processing methods. However, the data exhibits substantial variations, and no clear tendencies can be discerned.

3) SPEECH QUALITY

Considering Fig.4(c) and Table 5(c), we see no trends or significant differences in SQ between the blind and joint processing methods. However, some trend is seen aligning with PESQ scores at high SNRs, where the minimum processing methods outperform maximum processing. However, no significant differences could be determined between the minimum and maximum processing methods. Although, in the low SNR case of the (Car-Babble) scenario, the maximum processing methods significantly outperform minimum processing methods. This is in contrast with the PESQ results, where minimum processing was notably superior to maximum processing. However, the maximum processing methods had higher ESTOI scores, although minimum processing also achieved high ESTOI scores, cf. Table 5(b). In addition, for the high SNR case of (Car-Babble), maximum processing methods are significantly worse than the unprocessed performance, while no significant differences are observed between minimum processing methods and unprocessed or between minimum and maximum processing. Hence, the optimal balance between SI, speech distortions, and noise suppression remains unclear in scenarios with noise at both ends. In the high SNR case of the (Babble-Car) scenario, all methods were significantly worse than the unprocessed, aligning with the PESQ results. This indicates that the increase in estimated SI comes at the cost of speech distortion. Although not significant, the minimum processing methods appear to exhibit better quality than the maximum processing methods, which is also seen in the PESQ results.

4) JOINT VERSUS BLIND

The number of feasible bands is worth noting when comparing joint and blind minimum processing. There were only a limited number of feasible bands in both the SI and listening effort tests, and also in the (Babble-Car) scenario for the quality test. Hence, there is only slight variations between the processing of the joint and blind minimum processing methods. This can partially explain why we see only a slight performance difference between these methodologies. Notably, while higher SNRs presented more feasible bands, lower SNRs levels were specifically chosen for the listening tests because ESTOI and informal listening tests suggested that unprocessed SI was so high that all processing would lead to maximum SI and the absence of observable differences at higher SNRs. In contrast, in the (Car-Babble) scenario in the speech quality test, there were many feasible bands; however, no significant differences were observed between the joint and blind minimum processing. Furthermore, it is interesting that there were only very small variations between the joint and blind maximum processing despite the absence of feasibility constraints.

In summary, the subjective listening tests yielded inconclusive results regarding the superiority of joint processing over blind processing in both the minimum and maximum processing scenarios. Additionally, we did not establish statistically significant improvements in speech quality with minimum processing over maximum processing in high SNR cases, despite trends aligned with objective measures. The inherent challenge of small sample sizes underscores the need for future studies with larger participant pools to draw more definitive conclusions.

VIII. DISCUSSION

The results of this study revealed more distinct differences between minimum and maximum processing than between joint and blind processing, although the disparities remained statistically insignificant in subjective listening tests. Notably, this showed that the performance of minimum processing was not significantly inferior to that of maximum processing, since both the maximum and minimum processing methods significantly enhanced SI over unprocessed signals. This highlights the effectiveness of minimum processing approaches in SI enhancement.

The proposed joint minimum processing approach converges, by design, towards the performance of single-ended minimum processing methods in scenarios with high SNRs. Specifically, our method achieves comparable performance to the single-ended minimum processing methods [11] and [19] when there is minimal to no noise at either the NE or the FE. The adaptability of joint minimum processing is further emphasized by its ability to provide SI performance similar to joint maximum processing at low SNRs and when noise is present in both the FE and NE. In addition, for higher SNRs, where noise is absent at either end, our approach achieves a higher SQ. Consequently, the joint minimum processing method minimizes speech distortions while preserving a high SI. This indicates the beneficial application of the minimum processing principle in end-to-end communication scenarios.

Existing studies on joint FE and NE enhancement [33], [34], [35], [36], [37], [38], [39], [40] show varied results regarding the superiority of joint processing over blind methods. Although these studies compare a wide array of methods, they sometimes omit comparisons against blind methods

or other blind and joint approaches of the same nature, leading to a lack of clarity regarding their implementations under consistent conditions. In our approach, we compared similar methods to investigate the specific gains from using a joint approach versus a blind approach, with the aim of eliminating the confounding effects of vastly different processing types. Hence, this work provides a clearer indication of the dynamics between joint and blind processing, as well as between minimum and maximum processing. Therefore, it is interesting that our results were inconclusive regarding the superiority of joint processing over blind processing. Despite conducting experiments in oracle situations, the distinctions between the two approaches were very subtle. Although trends showed that joint processing outperformed blind processing in objective SI, the differences were statistically insignificant in the subjective listening tests.

The limitation in increasing the joint performance over blind minimum processing can, in certain cases, be attributed to the number of feasible bands in the optimization problem. Hence, the performance constraints must be carefully considered in practical implementations. However, the absence of feasibility issues in the maximum processing case indicates that the lack of performance difference between joint and blind maximum processing cannot be solely attributed to the optimization constraints.

We investigated the effect of joint processing when noise is present at both ends, because this is the only time that it makes sense to use a joint approach. However, conducting listening tests with noise at both the FE and NE is not very common. In [36] and [37], preference tests were used for both SI and SQ. Reference [38] conducted an informal closed-matrix SI test. References [34] and [35] conducted SI tests in which the words heard were typed in a computer interface. In [34] SQ was assessed only in quiet NE conditions. Hence, it is not clear what best practice is to judge especially SQ in the presence of both FE and NE noise. However, by using the MUSHRA test for SQ, we allowed the participants to freely judge what they deemed important in terms of SQ, under the assumption that they were familiar with the target speech, i.e., the reference signal. Hence, participants judged the importance of noise and speech distortions simultaneously. However, future studies may benefit from a clearer focus on which SQ degradations are caused by either (processed) environmental noise or speech distortions.

The results also show the impressive performance of the reference methods, suggesting limited potential for substantial improvement in terms of SI and SQ. Additionally, as SNRs improve at either end, the demand for joint FE and NE speech enhancement diminishes as the situation converges towards the single-ended cases. Generally, the better FSE methods are at removing noise in a distortionless manner by, for example, using DNNs [7], [8], [9] or an increased number of microphones [5], there is less need for joint SI and SQ enhancement.

IX. CONCLUSION

We extensively explored the joint FE and NE minimum processing framework introduced in [40]. The primary contribution lies in deriving a closed-form analytical solution for the optimization problem, with an MSE processing penalty, an estimated SI constraint represented by ASII and an SQ noise power constraint. We provided a thorough explanation of the key elements and conducted a systematic performance study, including objective measures and listening tests for SI, listening effort, and SQ. Performance was compared to joint ASII maximization, the blind concatenation of minimum processing FSE and NLE, and the blind concatenation of classic maximum processing, and revealed nuanced results.

For estimated SI measured by ESTOI, maximum processing methods generally exhibit slightly superior performance, with the proposed joint minimum processing framework showing a slight edge over blind minimum processing. However, the results were inconclusive regarding the consistent superiority of joint maximum processing over blind maximum processing in ESTOI. The PESQ results consistently show that minimum processing outperforms maximum processing, especially in high SNRs, but no significant differences were observed between the joint and blind methods. All subjective listening tests yielded inconclusive results. Subjective listening test results align with trends observed in objective measures, but fail to establish significant differences between maximum and minimum processing or between joint and blind processing. Hence, minimum processing performs on-par in SI with maximum processing while preserving a good SQ in higher SNR settings when noise is present at both ends. Additionally, because the joint minimum processing method has the single-ended solutions as special cases, the results and performance from the single-end minimum processing works extend to the joint case. This shows that it is also beneficial to apply the minimum processing principle in the context of end-to-end communication scenarios.

In essence, our work sheds light on the intricate relationship between SI, SQ, and the joint, blind, maximum and minimum processing methods. We provide in-depth insights into the optimization problem of joint minimum processing and underscore the importance of future investigations concerning optimization at both FE and NE in a joint context.

A. FUTURE WORK

Our results provide valuable insights into this field, emphasizing the need for more thorough investigations with controlled implementations of joint and blind methods under identical conditions, for a deeper understanding of the behavioral aspects of blind and joint processing and a combined review of the performance of existing results.

Future work, directly pertaining to our method, includes investigating the conditions for feasibility in the optimization

problem and how to derive an optimal performance in infeasible cases. Additionally, interesting extensions include other optimization targets, ANC, non-linear processing, and multiple FE environments with only one NE, together with a single FE broadcasting to multiple NE environments as in online meetings. It is also very interesting to investigate performance under more real-world conditions, with estimated speech and noise statistics, and in dimensions other than SI and SQ, such as complexity, synchronization requirements, and the need for bidirectional side-information transfer, are crucial. Particularly, when considering speech coding and how it can be incorporated into the joint end-to-end communication, where joint knowledge might be utilized in allocating limited bit rates to frequencies that are inaudible due to NE noise. Furthermore, it is interesting to investigate the impact of nonlinear processes between FE and NE, especially in the context of speech coding in real-life systems. Exploring the robustness of methods to coding added after the beamformer is also of interest. Similarly, the interplay between the dependence of the proposed solution on frequency-band energy and automatic gain control warrants further exploration. In addition, the potential increase in bit rates for carrying enhanced signals needs consideration, and adjustments to the solutions may be necessary.

APPENDIX A
SUBBAND FILTERING

Filtering frequency bins into subbands and determining the filter weights can be performed in various ways, cf. [11], [19, App. A]. In this paper, we consider auditory critical band filters [41] based on the gammatone filter bank model of [44].

As in [19], we let h_j be the impulse response of the j'th subband auditory filter. Now, for the j'th subband, the energy of the clean speech signal, $\mathcal{S}_{j,i}$, is given as the convolution between s and h_j, which in the time-subband domain is

$$\mathcal{S}_{j,i}^2 \triangleq \sum_{k\in\mathbb{B}_j} |S_{k,i}|^2 |H_j(k)|^2, \tag{39}$$

where $H_j(k)$ represents the DFT of h_j in frequency-bin k. We generate the frequency domain gammatone filters, $H_j(k)$, according to [44], and normalize them according to the mean total weight per frequency, that is.,

$$H_j'(k) = \frac{H_j(k)}{\frac{1}{K}\sum_m^K \sum_l^J H_l(m)}, \quad \forall j,k. \tag{40}$$

We then let the subband filter weights, $\omega_{j,k}$, be the normalized squared magnitude response of h_j, i.e., $\omega_{j,k} = |H_j'(k)|^2$.

When producing the final beamformers and NLE gains for each frequency bin, we need to apply a combination formula, such as that in (36). Because the weights, $\eta_{j,k}$, are applied to beamformer vectors and NLE gains, and not power spectra, we let the weights be given according to the normalized subband filter amplitudes, i.e., $\eta_{j,k} = |H_j'(k)|$.

APPENDIX B
PROCESSING PENALTY

A. BEAMFORMING COST

For far-end-only minimum processing speech enhancement, it was shown in [11, Sec. IV.A] that the minimum processing beamforming processing penalty is

$$\mathcal{D}_j(\boldsymbol{S}_j^R, \boldsymbol{Y}_j) = \sum_{k\in\mathbb{B}_j} \omega_{j,k} \left(\boldsymbol{v}_k^R - \boldsymbol{w}_k\right)^H C_{X_k}^{(\mu)} \left(\boldsymbol{v}_k^R - \boldsymbol{w}_k\right). \tag{41}$$

Considering the difference to the reference beamformer in the cost function we have

$$\boldsymbol{v}_k^R - \boldsymbol{w}_{j,k} = \boldsymbol{v}_k^R - \left(\alpha_j \boldsymbol{v}_k^R + (1-\alpha_j)\boldsymbol{w}_k^{\mu\text{MWF}}\right) \tag{42}$$

$$= (1-\alpha_j)\left(\boldsymbol{v}_k^R - \boldsymbol{w}_k^{\mu\text{MWF}}\right). \tag{43}$$

Inserting this into the minimum processing beamforming processing penalty gives

$$\mathcal{D}_j(\boldsymbol{S}_j^R, \boldsymbol{Y}_j) = \sum_{k\in\mathbb{B}_j} \Big[\omega_{j,k} (1-\alpha_j)^2 \cdot \left(\boldsymbol{v}_k^R - \boldsymbol{w}_k^{\mu\text{MWF}}\right)^H C_{X_k}^{(\mu)} \left(\boldsymbol{v}_k^R - \boldsymbol{w}_k^{\mu\text{MWF}}\right)\Big] \tag{44}$$

$$\propto (1-\alpha_j)^2. \tag{45}$$

B. LISTENING ENHANCEMENT COST

The minimum processing NLE processing penalty was shown in [19] to be

$$\mathcal{D}_j(\boldsymbol{Y}_j, \boldsymbol{Z}_j) = \sum_{k\in\mathbb{B}_j} \omega_{j,k} (1-g_k)^2 \sigma_{Y_k}^2. \tag{46}$$

Following the results of [19], where it was shown to be optimal to have fixed gains across the entire subband and to avoid comb filtering [40], we assume that the gains are equal across the subband. Therefore, we have that

$$\mathcal{D}_j(\boldsymbol{Y}_j, \boldsymbol{Z}_j) = (1-g_j)^2 \sum_{k\in\mathbb{B}_j} \omega_{j,k}\sigma_{Y_k}^2 \propto (1-g_j)^2. \tag{47}$$

APPENDIX C
PROOF OF THEOREM 1

We reformulate and solve the optimization problem as a maximization problem with a simpler cost function. That is,

$$\begin{aligned} &\underset{\alpha_j, g_j\in\mathbb{R}_+}{\arg\max} \quad \alpha_j - g_j \\ &\text{s.t. } \mathcal{C}_1 : g_j^2 p_{\text{FSE}}(\alpha_j) \geq \sigma_{N_j}^2 I_j^{\xi}, \quad \mathcal{C}_3 : 0 \leq \alpha_j \leq 1, \\ &\qquad \mathcal{C}_2 : g_j^2 \delta_{U_j}(\alpha_j) \leq \sigma_{N_j}^2 c_{U_j}, \mathcal{C}_4 : 1 \leq g_j. \end{aligned} \tag{48}$$

Looking at $\mathcal{C}_1$ and $\mathcal{C}_2$, we see that the optimization problem is only feasible if

$$\exists \alpha_j \in [0,1] \ : \ p_{\text{FSE}}(\alpha_j) > 0 \text{ and } \delta_{U_j}(\alpha_j) \leq \sigma_{N_j}^2 c_{U_j}. \tag{49}$$

That is, when there exists an α_j such that the FE SNR is above the desired audibility limit and the processed FE noise is below the upper noise limit. For convex optimization

problems, the Karush-Kuhn-Tucker (KKT) conditions are necessary and sufficient conditions that determine the global optimal solution, cf. [54]. However, we see from the constraints that we are optimizing over a non-convex set; hence, the optimization problem is non-convex. Therefore, the KKT conditions are not sufficient conditions for a global optimum to our optimization problem, but they are still necessary conditions [54]. Therefore, we investigate the KKT conditions to determine an optimum point.

Firstly, we formulate the Lagrangian,

$$\mathcal{L} = \alpha_j - g_j + \lambda_1\alpha_j + \lambda_2(1-\alpha_j) + \lambda_3(g_j - 1) + \lambda_4\left(g_j^2 p_{\text{FSE}}(\alpha_j) - \sigma_{N_j}^2 I_j^{\xi}\right) + \lambda_5\left(\sigma_{N_j}^2 c_{U_j} - g_j^2\delta_{U_j}(\alpha_j)\right). \tag{50}$$

Determining the gradient and writing up the KKT conditions we get

$$g_j^2 p_{\text{FSE}}(\alpha_j) \geq \sigma_{N_j}^2 I_j^{\xi}, \tag{51a}$$
$$g_j^2\delta_{U_j}(\alpha_j) \leq \sigma_{N_j}^2 c_{U_j}, \tag{51b}$$
$$\alpha_j \geq 0, \quad \alpha_j \leq 1, \quad g_j \geq 1, \tag{51c}$$
$$\lambda_1 \geq 0, \quad \lambda_2 \geq 0, \quad \lambda_3 \geq 0, \quad \lambda_4 \geq 0, \quad \lambda_5 \geq 0, \tag{51d}$$
$$\lambda_1\alpha_j = 0, \quad \lambda_2(\alpha_j - 1) = 0, \quad \lambda_3(g_j - 1) = 0, \tag{51e}$$
$$\lambda_4\left(g_j^2 p_{\text{FSE}}(\alpha_j) - \sigma_{N_j}^2 I_j^{\xi}\right) = 0, \tag{51f}$$
$$\lambda_5\left(\sigma_{N_j}^2 c_{U_j} - g_j^2\delta_{U_j}(\alpha_j)\right) = 0, \tag{51g}$$
$$\begin{bmatrix} 1 + \lambda_1 - \lambda_2 + \lambda_4 g_j^2 p'_{\text{FSE}}(\alpha_j) - \lambda_5 g_j^2\delta'_{U_j}(\alpha_j) \\ -1 + \lambda_3 + 2\lambda_4 g_j p_{\text{FSE}}(\alpha_j) - 2\lambda_5 g_j\delta_{U_j}(\alpha_j) \end{bmatrix} = \mathbf{0} \tag{51h}$$

To solve the optimization problem, we begin by determining the boundary and stationary solutions. Subsequently, we compare the different feasible solutions and select the one with the optimal cost function value as the optimal solution [54].

A. SOLVING KKT CONDITIONS

1) IF $\lambda_1 > 0,\ \lambda_2 = 0,\ \lambda_3 > 0,\ \lambda_4 \geq 0,\ \lambda_5 \geq 0$

Then we must have $\alpha_j^* = 0$ and $g_j^* = 1$ and the cost function value is $f_{\text{cost}} = -1$. This is only a feasible solution if $p_{\text{FSE}}(0) = D_j^{\mu\text{MWF}} \geq \sigma_{N_j}^2 I_j^{\xi}$ and $\delta_{U_j}(0) = \delta_{U_j}^{\mu\text{MWF}} \leq \sigma_{N_j}^2 c_{U_j}$, i.e., the μMWF beamformer, $\boldsymbol{w}_k^{\mu\text{MWF}}$, provides a feasible FE SNR with a sufficiently low FE noise power.

2) IF $\lambda_1 = 0,\ \lambda_2 > 0,\ \lambda_3 > 0,\ \lambda_4 \geq 0,\ \lambda_5 \geq 0$

Then we must have $\alpha_j^* = 1$ and $g_j^* = 1$ and the cost function value is globally maximized at $f_{\text{cost}} = 0$. This is only a feasible solution if $p_{\text{FSE}}(1) = D_j^R \geq \sigma_{N_j}^2 I_j^{\xi}$ and $\delta_{U_j}(1) = \delta_{U_j}^R \leq \sigma_{N_j}^2 c_{U_j}$, i.e., the reference beamformer $\boldsymbol{v}_k^R$ provides a feasible FE SNR with sufficiently low FE noise power.

3) IF $\lambda_1 = 0,\ \lambda_2 = 0,\ \lambda_3 > 0,\ \lambda_4 \geq 0,\ \lambda_5 \geq 0$

Then we must have $g_j^* = 1$, and the cost function is $f_{\text{cost}} = \alpha_j - 1$. To ensure feasibility, it is necessary that the optimal α_j satisfies $p_{\text{FSE}}(\alpha_j) \geq \sigma_{N_j}^2 I_j^{\xi}$ and $\delta_{U_j}(\alpha_j) \leq \sigma_{N_j}^2 c_{U_j}$.

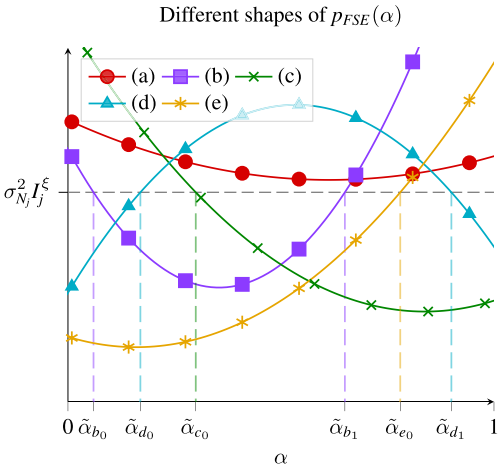


FIGURE 5. The figure illustrates the five different shapes and feasible regions that may be created by $p_{\text{FSE}}(\alpha)$.

The shape of the polynomial $p_{\text{FSE}}(\alpha_j)$ defines a set of points

$$\mathcal{F}^{\mathcal{C}_1} \triangleq \{\alpha \in [0,1] : p_{\text{FSE}}(\alpha_j) \geq \sigma_{N_j}^2 I_j^{\xi}\}, \tag{52}$$

that satisfy $\mathcal{C}_1$. The polynomial may define the set to have one of five possible shapes, (a)–(e), as illustrated in Fig. 5, which are

$$\begin{aligned} \mathcal{F}_{(a)}^{\mathcal{C}_1} &= [0,1], & \mathcal{F}_{(b)}^{\mathcal{C}_1} &= [0,\tilde{\alpha}_{b_0}] \cup [\tilde{\alpha}_{b_1},1], \\ \mathcal{F}_{(c)}^{\mathcal{C}_1} &= [0,\tilde{\alpha}_{c_0}], & \mathcal{F}_{(d)}^{\mathcal{C}_1} &= [\tilde{\alpha}_{d_0},\tilde{\alpha}_{d_1}], \\ \mathcal{F}_{(e)}^{\mathcal{C}_1} &= [\tilde{\alpha}_{e_0},1]. \end{aligned} \tag{53}$$

Similarly, the shape of the polynomial $\delta_{U_j}(\alpha_j)$ defines a set of points,

$$\mathcal{F}^{\mathcal{C}_2} \triangleq \{\alpha \in [0,1] : \delta_{U_j}(\alpha_j) \leq \sigma_{N_j}^2 c_{U_j}\}, \tag{54}$$

that satisfy $\mathcal{C}_2$. Again there are five possible shapes to the set,

$$\begin{aligned} &\mathcal{F}_{(1)}^{\mathcal{C}_2} = [0,1],\ \mathcal{F}_{(2)}^{\mathcal{C}_2} = [0,\hat{\alpha}_{b_0}] \cup [\hat{\alpha}_{b_1},1] \\ &\mathcal{F}_{(3)}^{\mathcal{C}_2} = [0,\hat{\alpha}_{c_0}],\ \mathcal{F}_{(4)}^{\mathcal{C}_2} = [\hat{\alpha}_{d_0},\hat{\alpha}_{d_1}], \\ &\mathcal{F}_{(5)}^{\mathcal{C}_2} = [\hat{\alpha}_{e_0},1]. \end{aligned} \tag{55}$$

Then,

$$\mathcal{F}^{g=1} = \mathcal{F}^{\mathcal{C}_1} \cap \mathcal{F}^{\mathcal{C}_2}, \tag{56}$$

is the jointly feasible region for both $\mathcal{C}_1$ and $\mathcal{C}_2$, and we have joint feasibility if $\mathcal{F}^{g=1} \neq \varnothing$, i.e, the feasible region is non-empty. Since $p_{\text{FSE}}(\alpha_j)$ and $\delta_{U_j}(\alpha_j)$ are second-order polynomials, we note that the feasible set, $\mathcal{F}^{g=1}$, is not necessarily a convex set. Now since the cost function, $f_{\text{cost}} = \alpha_j - 1$, is strictly monotonically increasing with α_j, the optimal α_j^* is the largest α in $\mathcal{F}^{g=1}$, i.e., the rightmost boundary of the feasible region, $\partial\mathcal{F}^{g=1}$. For example, if $\mathcal{F}^{g=1} = \mathcal{F}_{(d)}^{\mathcal{C}_1} \cap \mathcal{F}_{(4)}^{\mathcal{C}_2}$ and $\tilde{\alpha}_{d_0} \leq \hat{\alpha}_{d_0} \leq \tilde{\alpha}_{d_1} \leq \hat{\alpha}_{d_1}$, then $\mathcal{F}^{g=1} = [\hat{\alpha}_{d_0},\tilde{\alpha}_{d_1}]$ and the optimal value is $\alpha_j^* = \tilde{\alpha}_{d_1}$.

4) IF $\lambda_1 \geq 0,\ \lambda_2 \geq 0,\ \lambda_3 = 0,\ \lambda_4 \geq 0,\ \lambda_5 \geq 0$

Finally, we consider the interior stationary solution. Isolating g_j in the audibility constraint, $\mathcal{C}_1$, we have

$$g_j \geq \sqrt{\frac{\sigma_{N_j}^2 I_j^{\xi}}{p_{\mathrm{FSE}}(\alpha_j)}}, \tag{57}$$

where $p_{\mathrm{FSE}}(\alpha) > 0$. The only feasible root is the positive principal root since we must have $g_j \geq 1$. Combining the above limit with $g_j \geq 1$ we obtain

$$g_j = \max\left\{\sqrt{\frac{\sigma_{N_j}^2 I_j^{\xi}}{p_{\mathrm{FSE}}(\alpha_j)}}, 1\right\}. \tag{58}$$

We see that $g_j = 1$ if for the optimal α_j we have $p_{\mathrm{FSE}}(\alpha_j^*) \geq \sigma_{N_j}^2 I_j^{\xi}$, which we haved already solved in the previous case. Therefore, we focus on the case where $p_{\mathrm{FSE}}(\alpha_j) < \sigma_{N_j}^2 I_j^{\xi}\ \forall \alpha_j \in [0, 1]$. In this case, the optimal g_j is $g_j^* = \sqrt{\frac{\sigma_{N_j}^2 I_j^{\xi}}{p_{\mathrm{FSE}}(\alpha_j^*)}}$. Inserting this into (p_0) and rearranging the terms of $\mathcal{C}_2$, we find the optimal α_j^* by solving the optimization problem

$$\begin{aligned} &\underset{\alpha_j \in [0,1]}{\arg\max}\ \alpha_j - \sqrt{\frac{\sigma_{N_j}^2 I_j^{\xi}}{p_{\mathrm{FSE}}(\alpha_j)}} \\ &\text{s.t. } \mathcal{C}_{SI} : \sigma_{N_j}^2 I_j^{\xi} > p_{\mathrm{FSE}}(\alpha_j) > 0, \\ &\qquad \mathcal{C}_{NP} : I_j^{\xi} \delta_{U_j}(\alpha_j) - c_{U_j} p_{\mathrm{FSE}}(\alpha_j) \leq 0. \end{aligned} \tag{59}$$

First, we consider the conditions for feasibility. Similar to the above case, we can define

$$\mathcal{F}^{\mathrm{SI}} \triangleq \left\{\alpha \in [0, 1] : \sigma_{N_j}^2 I_j^{\xi} > p_{\mathrm{FSE}}(\alpha) > 0\right\} \tag{60}$$

as the set of points that satisfy $\mathcal{C}_{SI}$. Additionally, for the second-order polynomial in the noise power constraint, we define the set of points that satisfy $\mathcal{C}_{NP}$,

$$\mathcal{F}^{\mathrm{NP}} \triangleq \left\{\alpha \in [0, 1] : I_j^{\xi} \delta_{U_j}(\alpha) - c_{U_j} p_{\mathrm{FSE}}(\alpha) \leq 0\right\}. \tag{61}$$

Letting $\dot{\alpha}_l, \dot{\alpha}_r \in \mathbb{R}$ be the real roots of the polynomial, we have four possible regions for the noise power constraint:

$$\mathcal{F}_{(i)}^{\mathrm{NP}} = [0, 1],\ \mathcal{F}_{(ii)}^{\mathrm{NP}} = [\dot{\alpha}_l, \dot{\alpha}_r], \tag{62}$$

$$\mathcal{F}_{(iii)}^{\mathrm{NP}} = [0, \dot{\alpha}_l] \cup [\dot{\alpha}_r, 1],\ \mathcal{F}_{(iv)}^{\mathrm{NP}} = \varnothing. \tag{63}$$

The jointly feasible set is then $\mathcal{F}^{g(\alpha)} = \mathcal{F}^{\mathrm{SI}} \cap \mathcal{F}^{\mathrm{NP}}$. As we have seen previously, this is not necessarily a convex set.

Secondly, letting

$$h(\alpha_j) \triangleq \alpha_j - \sqrt{\frac{\sigma_{N_j}^2 I_j^{\xi}}{p_{\mathrm{FSE}}(\alpha_j)}} \tag{64}$$

be the cost function, we investigate the concavity/convexity of the cost function. It can be shown that

$$\frac{\mathrm{d}}{\mathrm{d}\alpha} h(\alpha) = 1 + \frac{\sqrt{\sigma_{N_j}^2 I_j^{\xi}}\, p'_{\mathrm{FSE}}(\alpha_j)}{2\left(p_{\mathrm{FSE}}(\alpha_j)\right)^{3/2}} \tag{65}$$

$$\begin{aligned} \frac{\mathrm{d}^2}{\mathrm{d}\alpha^2} h(\alpha) &= -\frac{3\sqrt{\sigma_{N_j}^2 I_j^{\xi}}\left(p'_{\mathrm{FSE}}(\alpha_j)\right)^2}{4\left(p_{\mathrm{FSE}}(\alpha_j)\right)^{5/2}} + \frac{\sqrt{\sigma_{N_j}^2 I_j^{\xi}}\, p''_{\mathrm{FSE}}(\alpha_j)}{2\left(p_{\mathrm{FSE}}(\alpha_j)\right)^{3/2}} \qquad (66) \\ &= -\frac{2\sqrt{\sigma_{N_j}^2 I_j^{\xi}}\, \phi(\alpha_j)}{4\left(p_{\mathrm{FSE}}(\alpha_j)\right)^{5/2}}, \qquad (67) \end{aligned}$$

where

$$\begin{aligned} &\phi(\alpha_j) \\ &\triangleq \left(D_j^{\mathrm{cross}} - D_j^R - D_j^{\mu\mathrm{MWF}}\right)^2 \alpha_j^2 \\ &\quad - \left(D_j^{\mathrm{cross}} - 2D_j^{\mu\mathrm{MWF}}\right)\left(D_j^{\mathrm{cross}} - D_j^R - D_j^{\mu\mathrm{MWF}}\right)\alpha_j \\ &\quad + \left(D_j^{\mu\mathrm{MWF}}\right)^2 - \left(D_j^{\mathrm{cross}} + \frac{D_j^R}{2}\right) D_j^{\mu\mathrm{MWF}} + \frac{3\left(D_j^{\mathrm{cross}}\right)^2}{8} \end{aligned} \tag{68}$$

is a second-order polynomial in α_j. Because the optimization problem is only feasible for $p_{\mathrm{FSE}}(\alpha_j) > 0$, the sign of the second derivative depends on the sign of $\phi(\alpha_j)$. Let $\overline{\alpha}_l \in \mathbb{R}$ and $\overline{\alpha}_r \in \mathbb{R}$ be the real roots of $\phi(\alpha_j)$, if they exist, where we assume $\overline{\alpha}_l \leq \overline{\alpha}_r$. Then, we may consider four different scenarios:

(*a*) If $\phi(\alpha) \geq 0\ \forall\ \alpha \in [0, 1]$ then $h(\alpha)$ is concave in the entire interval $[0, 1]$.

(*b*) If $\phi(\alpha) \geq 0\ \forall\ \alpha \in [0, \overline{\alpha}_l] \cup [\overline{\alpha}_r, 1]$, and $\phi(\alpha) < 0\ \forall\ \alpha \in (\overline{\alpha}_l, \overline{\alpha}_r)$, then $h(\alpha)$ is concave in the intervals $[0, \overline{\alpha}_l]$ and $[\overline{\alpha}_r, 1]$, and $h(\alpha)$ is convex in the interval $(\overline{\alpha}_l, \overline{\alpha}_r)$.

(*c*) If $\phi(\alpha) \geq 0\ \forall\ \alpha \in [\overline{\alpha}_l, \overline{\alpha}_r]$, and $\phi(\alpha) < 0\ \forall\ \alpha \in [0, \overline{\alpha}_l) \cup (\overline{\alpha}_r, 1]$, then $h(\alpha)$ is concave in the interval $[\overline{\alpha}_l, \overline{\alpha}_r]$, and $h(\alpha)$ is convex in the intervals$[0, \overline{\alpha}_l)$ and $(\overline{\alpha}_r, 1]$.

(*d*) If $\phi(\alpha) \leq 0\ \forall\ \alpha \in [0, 1]$, then $h(\alpha)$ is convex in the entire interval $[0, 1]$.

We note that the stationary points in the intervals where $h(\alpha)$ is concave are maxima, and the stationary points in the intervals where $h(\alpha)$ is convex are minima. Thus, the optimal points are either at the stationary points of the concave regions or at the boundary of the convex intervals. However, these optimal points may not necessarily be feasible.

Let α_s be a stationary point in a concave region of $h(\alpha)$. The stationary points can be determined by explicitly solving $h'(\alpha) = 0$ or via a simple bisection of $h'(\alpha)$ on the concave regions. If $\alpha_s \in \mathcal{F}^{g(\alpha)} \cap [0, 1]$, then α_s is an optimal and feasible point. On the other hand, if $\alpha_s \notin \mathcal{F}^{g(\alpha)}$, i.e., the stationary point is not feasible, then the optimal solution is at the boundary of the feasible set close to α_s. For example, if we are in the case of (*a*) and $\mathcal{F}_{(iii)}^{\mathrm{NP}}$, then we might have $\dot{\alpha}_l < \alpha_s < \dot{\alpha}_r$, and α_s is not feasible. Therefore, the optimal feasible value in this case is found to be either $\dot{\alpha}_l$ or $\dot{\alpha}_r$.

Finally, let

$$\mathcal{F}_S^{g(\alpha)} = \{\alpha : \alpha \in \mathcal{F}^{g(\alpha)},\ h'(\alpha) = 0,\ h''(\alpha) \leq 0\} \tag{69}$$

be the set of feasible stationary points in the concave regions of $h(\alpha)$, and denote the boundary of $\mathcal{F}^{g(\alpha)}$ by $\partial\mathcal{F}^{g(\alpha)}$. Then, the optimal α_j is given as

$$\alpha_j^* = \arg\max_{\alpha} \; h(\alpha), \quad \text{s.t. } \alpha \in \mathcal{F}_S^{g(\alpha)} \cup \partial\mathcal{F}^{g(\alpha)}, \tag{70}$$

which is a simple combinatorial problem.

B. COMPARING COST FUNCTIONS

We now combine all the above cases such that we find the optimal solution by comparing the cost function values for the various optimum points:

$$f_{\text{cost}}(1,1) = 0 \tag{71}$$

$$f_{\text{cost}}(0,1) = -1 \tag{72}$$

$$f_{\text{cost}}(\alpha_j,1) = \alpha_j - 1 \tag{73}$$

$$f_{\text{cost}}\left(\alpha_j, \sqrt{\frac{\sigma_{N_j}^2 I_j^{\xi}}{f_{\text{con}}(\alpha_j)}}\right) = \alpha_j - \sqrt{\frac{\sigma_{N_j}^2 I_j^{\xi}}{f_{\text{con}}(\alpha_j)}} \tag{74}$$

First, if $D_j^R \geq \sigma_{N_j}^2 I_j^{\xi}$ and $\delta_{U_j}^R \leq \sigma_{N_j}^2 c_{U_j}$ then the optimal feasible cost function is $f_{\text{cost}}(1,1) = 0$, which is clearly the global optimum value. Therefore, in this case $(\alpha_j^*, g_j^*) = (1,1)$ regardless of the value of any of the other parameters.

We can cover all other cases by first solving the following combinatorial problem to determine the optimal α_j,

$$\alpha_j^* = \arg\max_{\alpha} \; \pi(\alpha), \quad \text{s.t. } \alpha \in \mathcal{F}_S^{g(\alpha)} \cup \partial\mathcal{F}^{g(\alpha)} \cup \partial\mathcal{F}^{g=1}, \tag{75}$$

where

$$\pi(\alpha) = \begin{cases} \alpha - 1 & \text{if } \alpha \in \partial\mathcal{F}^{g=1} \\ h(\alpha) & \text{if } \alpha \in \mathcal{F}_S^{g(\alpha)} \text{ or } \alpha \in \partial\mathcal{F}^{g(\alpha)}. \end{cases} \tag{76}$$

Finally, the optimal g_j is

$$g_j^* = \begin{cases} 1 & \text{if } \alpha_j^* \in \partial\mathcal{F}^{g=1} \\ \sqrt{\dfrac{\sigma_{N_j}^2 I_j^{\xi}}{f_{\text{con}}(\alpha_j^*)}} & \text{if } \alpha_j^* \in \mathcal{F}_S^{g(\alpha)} \text{ or } \alpha_j^* \in \partial\mathcal{F}^{g(\alpha)}. \end{cases} \tag{77}$$

This completes the proof.

REFERENCES

[1] S. Gannot, E. Vincent, S. Markovich-Golan, and A. Ozerov, "A consolidated perspective on multimicrophone speech enhancement and source separation," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 25, no. 4, pp. 692–730, Apr. 2017.

[2] M. Brandstein and D. Ward, *Microphone Arrays: Signal Processing Techniques and Applications* (Digital Signal Processing). New York, NY, USA: Springer, 2001.

[3] S. Doclo, S. Gannot, M. Moonen, and A. Spriet, "Acoustic beamforming for hearing aid applications," in *Handbook on Array Processing and Sensor Networks*. Hoboken, NJ, USA: Wiley, 2010, pp. 269–302.

[4] K. Eneman, H. Luts, J. Wouters, M. Büchler, N. Dillier, W. Dreschler, M. Froehlich, G. Grimm, V. Hohmann, R. Houben, A. Leijon, A. Lombard, D. Mauler, M. Moonen, H. Puder, M. Schulte, A. Spriet, and M. Vormann, "Evaluation of signal enhancement algorithms for hearing instruments," in *Proc. 16th Eur. Signal Process. Conf.*, Aug. 2008, pp. 1–5.

[5] E. A. P. Habets, J. Benesty, I. Cohen, S. Gannot, and J. Dmochowski, "New insights into the MVDR beamformer in room acoustics," *IEEE Trans. Audio, Speech, Language Process.*, vol. 18, no. 1, pp. 158–170, Jan. 2010.

[6] Y. Huang, J. Benesty, and J. Chen, "Analysis and comparison of multichannel noise reduction methods in a common framework," *IEEE Trans. Audio, Speech, Language Process.*, vol. 16, no. 5, pp. 957–968, Jul. 2008.

[7] K. Tesch and T. Gerkmann, "Insights into deep non-linear filters for improved multi-channel speech enhancement," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 31, pp. 563–575, 2023.

[8] D. Wang and J. Chen, "Supervised speech separation based on deep learning: An overview," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 26, no. 10, pp. 1702–1726, Oct. 2018.

[9] M. Kolbæk, Z.-H. Tan, and J. Jensen, "Speech intelligibility potential of general and specialized deep neural network based speech enhancement systems," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 25, no. 1, pp. 153–167, Jan. 2017.

[10] I. López-Espejo, A. Edraki, W.-Y. Chan, Z.-H. Tan, and J. Jensen, "On the deficiency of intelligibility metrics as proxies for subjective intelligibility," *Speech Commun.*, vol. 150, pp. 9–22, May 2023.

[11] A. Zahedi, M. S. Pedersen, J. Østergaard, T. U. Christiansen, L. Bramsløw, and J. Jensen, "Minimum processing beamforming," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 29, pp. 2710–2724, 2021.

[12] C. Chermaz, C. Valentini-Botinhao, H. Schepker, and S. King, "Evaluating near end listening enhancement algorithms in realistic environments," in *Proc. 23rd Int. Congr. Acoust., Integrating 4th EAA Euroregio*, Sep. 2019, pp. 5731–5735.

[13] M. Cooke, C. Mayo, C. Valentini-Botinhao, Y. Stylianou, B. Sauert, and Y. Tang, "Evaluating the intelligibility benefit of speech modifications in known noise conditions," *Speech Commun.*, vol. 55, no. 4, pp. 572–585, May 2013.

[14] M. Cooke, S. King, M. Garnier, and V. Aubanel, "The listening talker: A review of human and algorithmic context-induced modifications of speech," *Comput. Speech Lang.*, vol. 28, no. 2, pp. 543–571, Mar. 2014.

[15] M. Cooke, C. Mayo, and C. Valentini-Botinhao, "Intelligibility-enhancing speech modifications: The hurricane challenge," in *Proc. INTERSPEECH*, Aug. 2013, pp. 3552–3556.

[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, Mar. 2015.

[17] B. Sauert, "Near-end listening enhancement: Theory and application," Ph.D. dissertation, Institut für Nachrichtengeräte und Datenverarbeitung, Wissenschaftsverlag Mainz, Aachen, Germany, 2014.

[18] 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.

[19] A. J. Fuglsig, J. Jensen, Z.-H. Tan, L. S. Bertelsen, J. C. Lindof, and J. Østergaard, "Minimum processing near-end listening enhancement," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 31, pp. 2233–2245, 2023.

[20] H. Li and J. Yamagishi, "Multi-metric optimization using generative adversarial networks for near-end speech intelligibility enhancement," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 29, pp. 3000–3011, 2021.

[21] G. Li, R. Hu, R. Zhang, and X. Wang, "A mapping model of spectral tilt in normal-to-Lombard speech conversion for intelligibility enhancement," *Multimedia Tools Appl.*, vol. 79, nos. 27–28, pp. 19471–19491, Jul. 2020.

[22] J. Rennies, H. Schepker, C. Valentini-Botinhao, and M. Cooke, "Intelligibility-enhancing speech modifications—The hurricane challenge 2.0," in *Proc. INTERSPEECH*, Oct. 2020, pp. 1341–1345.

[23] G. Li, R. Hu, X. Wang, and R. Zhang, "A near-end listening enhancement system by RNN-based noise cancellation and speech modification," *Multimedia Tools Appl.*, vol. 78, no. 11, pp. 15483–15505, Jun. 2019.

[24] T.-C. Zorila, V. Kandia, and Y. Stylianou, "Speech-in-noise intelligibility improvement based on spectral shaping and dynamic range compression," in *Proc. 20th Eur. Signal Process. Conf. (EUSIPCO)*, Portland, OR, USA, Sep. 2012, pp. 2075–2079.

[25] N. V. George and G. Panda, "Advances in active noise control: A survey, with emphasis on recent nonlinear techniques," *Signal Process.*, vol. 93, no. 2, pp. 363–377, Feb. 2013.

[26] S. M. Kuo and D. R. Morgan, "Active noise control: A tutorial review," *Proc. IEEE*, vol. 87, no. 6, pp. 943–975, Jun. 1999.

[27] J. Rennies, A. Pusch, H. Schepker, and S. Doclo, "Evaluation of a near-end listening enhancement algorithm by combined speech intelligibility and listening effort measurements," *J. Acoust. Soc. Amer.*, vol. 144, no. 4, pp. EL315–EL321, Oct. 2018.

[28] R. Pricken, M. Wältermann, E. Parotat, M. Soloducha, and A. Raake, "Quality aspects of near-end listening enhancement approaches in telecommunication applications," in *Proc. DAGA*. Kiel, Germany: German Acoustical Society (DEGA), 2017, pp. 872–875.

[29] Y. Tang, C. Arnold, and T. J. Cox, "A Study on the relationship between the intelligibility and quality of algorithmically-modified speech for normal hearing listeners," *J. Otorhinolaryngol., Hearing Balance Med.*, vol. 1, no. 1, p. 5, Jun. 2018.

[30] 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, no. 4, pp. 858–872, Jul. 2014.

[31] B. Hect, J. Teevan, and A. Sellen, *The 'Leaf Blower Problem' and the Importance of Common Ground*. Redmond, WA, USA: Microsoft Research, 2021.

[32] P. C. Loizou, *Speech Enhancement: Theory and Practice*, 2nd ed., Boca Raton, FL, USA: CRC Press, 2013.

[33] K. Tan and D. Wang, "Improving robustness of deep learning based monaural speech enhancement against processing artifacts," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Barcelona, Spain, May 2020, pp. 6914–6918.

[34] T.-C. Zorilă and Y. Stylianou, "On the quality and intelligibility of noisy speech processed for near-end listening enhancement," in *Proc. INTERSPEECH*. Singapore: ISCA, Aug. 2017, pp. 2023–2027.

[35] M. P. V. Shifas, C. Zorila, and Y. Stylianou, "End-to-end neural based modification of noisy speech for speech-in-noise intelligibility improvement," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 30, pp. 162–173, 2022.

[36] M. Niermann, P. Jax, and P. Vary, "Joint near-end listening enhancement and far-end noise reduction," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Mar. 2017, pp. 4970–4974.

[37] H. Li, Y. Liu, and J. Yamagishi, "Joint noise reduction and listening enhancement for full-end speech enhancement," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Jun. 2023, pp. 1–5.

[38] S. Khademi, R. C. Hendriks, and W. B. Kleijn, "Intelligibility enhancement based on mutual information," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 25, no. 8, pp. 1694–1708, Aug. 2017.

[39] A. J. Fuglsig, J. Østergaard, J. Jensen, L. S. Bertelsen, P. Mariager, and Z.-H. Tan, "Joint far- and near-end speech intelligibility enhancement based on the approximated speech intelligibility index," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Singapore, May 2022, pp. 7752–7756.

[40] A. J. Fuglsig, J. Jensen, Z.-H. Tan, L. S. Bertelsen, J. C. Lindof, and J. Østergaard, "Joint minimum processing beamforming and near-end listening enhancement," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. Workshops (ICASSPW)*, Apr. 2024, pp. 485–489.

[41] *Methods for Calculation of the Speech Intelligibility Index Ansi S.35-1997*, Acoust. Soc. Amer., Amer. Nat. Standards Inst., New York, NY, USA, 1997.

[42] J. S. Garofolo, L. F. Lamel, W. M. Fisher, J. G. Fiscus, D. S. Pallett, N. L. Dahlgren, and V. Zue, *TIMIT Acoustic-Phonetic Continuous Speech Corpus*. Philadelphia, PA, USA: Linguistic Data Consortium, 1993.

[43] J. B. Allen and D. A. Berkley, "Image method for efficiently simulating small-room acoustics," *J. Acoust. Soc. Amer.*, vol. 65, no. 4, pp. 943–950, Apr. 1979.

[44] S. van de Par, A. Kohlrausch, R. Heusdens, J. Jensen, and S. H. Jensen, "A perceptual model for sinusoidal audio coding based on spectral integration," *EURASIP J. Adv. Signal Process.*, vol. 2005, no. 9, pp. 1292–1304, Jun. 2005.

[45] *EBU Recommendation R128—Loudness Normalisation and Permitted Maximum Level of Audio Signals*, document R-128, Eur. Broadcast. Union, 2014.

[46] A. H. Andersen, "Speech intelligibility prediction for hearing aid systems," Ph.D. dissertation, Dept. Electron. Syst., Aalborg Universitetsforlag, Aalborg, Denmark, 2017.

[47] K. Wagener, J. L. Josvassen, and R. Ardenkjær, "Design, optimization and evaluation of a Danish sentence test in noise," *Int. J. Audiol.*, vol. 42, no. 1, pp. 10–17, Jan. 2003.

[48] E. Vincent. (2005). *MUSHRAM—A MATLAB Interface for MUSHRA Listening Tests*. Accessed: Dec. 12, 2023. [Online]. Available: https://c4dm.eecs.qmul.ac.uk/downloads/#mushram

[49] *Method for the Subjective Assessment of Intermediate Quality Level of Audio Systems*, document BS.1534-3 Recommendation BS.1534-3, Oct. 2015.

[50] W. Kruskal and W. Wallis, "Use of ranks in one-criterion variance analysis," *J. Amer. Stat. Assoc.*, vol. 47, no. 260, pp. 583–621, 1952.

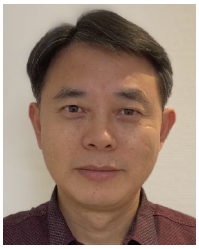
[51] F. J. Massey, "The Kolmogorov–Smirnov test for goodness of fit," *J. Amer. Stat. Assoc.*, vol. 46, no. 253, p. 68, Mar. 1951.

[52] C. Mendonça and S. Delikaris-Manias, *Statistical Tests With MUSHRA Data*. New York, NY, USA: Audio Engineering Society, May 2018.

[53] A. Field, *Discovering Statistics Using IBM SPSS Statistics*, 5th ed., New York, NY, USA: SAGE, 2018.

[54] S. P. Boyd and L. Vandenberghe, *Convex Optimization*. Cambridge, U.K.: Cambridge Univ. Press, 2004.

ANDREAS JONAS FUGLSIG (Member, IEEE) received the B.Sc. and M.Sc. degrees in mathematical engineering from Aalborg University, in 2017 and 2019, respectively, and the Ph.D. degree from the Centre on Acoustic Signal Processing Research (CASPR), Aalborg University, and RTX A/S, in 2024. He was with RTX A/S as a Research and Development Engineer, from 2019 to 2024. He joined the Audio Analysis Laboratory, Aalborg University, in 2024. His research interests include acoustic signal processing, speech enhancement, and information theory.

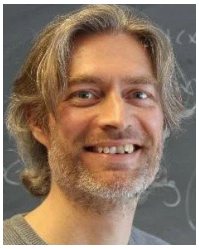


ZHENG-HUA TAN (Senior Member, IEEE) received the B.Sc. and M.Sc. degrees in electrical engineering from Hunan University, Changsha, China, in 1990 and 1996, respectively, and the Ph.D. degree in electronic engineering from Shanghai Jiao Tong University (SJTU), Shanghai, China, in 1999. He is a Professor with the Department of Electronic Systems and the Co-Head of the Centre for Acoustic Signal Processing Research, Aalborg University, Aalborg, Denmark. He is also a Co-Lead of the Pioneer Centre for AI, Denmark. He was a Visiting Scientist with the Computer Science and Artificial Intelligence Laboratory, MIT, Cambridge, USA; an Associate Professor with the Department of Electronic Engineering, SJTU; and a Postdoctoral Fellow with the AI Laboratory, KAIST, Daejeon, South Korea. He has (co)-authored over 250 refereed publications. His research interests include machine learning, deep learning, speech and speaker recognition, noise-robust speech processing, and multimodal signal processing. His works have been recognized by the prestigious IEEE Signal Processing Society 2022 Best Paper Award and International Speech Communication Association 2022 Best Research Paper Award. He is a member of the Speech and Language Processing Technical Committee (SLTC). He was the Elected Chair of the IEEE Signal Processing Society Machine Learning for Signal Processing TC (MLSP TC). He is the TPC Vice-Chair of ICASSP 2024 and was the General Chair of IEEE MLSP 2018. He is a Lead Guest Editor of IEEE JOURNAL OF SELECTED TOPICS IN SIGNAL PROCESSING, Inaugural Special Series on AI in Signal and Data Science–Toward Explainable, Reliable, and Sustainable Machine Learning. He served as an Associate Editor for IEEE/ACM TRANSACTIONS ON AUDIO, SPEECH AND LANGUAGE PROCESSING, *Computer Speech and Language*, and several other journals.

LARS SØNDERGAARD BERTELSEN received the M.S.E.E. degree from Aalborg University, in 2000. After that, he has worked in the telecommunication industry for several years doing both software, ASIC/FPGA, and system design. In recent years, he has worked with DSP audio processing in wireless audio products.

JESPER JENSEN received the M.Sc. degree in electrical engineering and the Ph.D. degree in signal processing from Aalborg University, Aalborg, Denmark, in 1996 and 2000, respectively. From 1996 to 2000, he was with the Center for Person Kommunikation (CPK), Aalborg University, as a Ph.D. student and an Assistant Research Professor. From 2000 to 2007, he was a Postdoctoral Researcher and an Assistant Professor with Delft University of Technology, Delft, The Netherlands; and an External Associate Professor with Aalborg University. Currently, he is a fellow with Oticon A/S, Denmark, where his main responsibility is scouting and development of new signal processing concepts for hearing aid applications. He is also a Professor with the Section for Signal and Information Processing (SIP), Department of Electronic Systems, Aalborg University. He is also a Co-Founder of the Centre for Acoustic Signal Processing Research (CASPR), Aalborg University. His main interests include acoustic signal processing, including signal retrieval from noisy observations, coding, speech and audio modification and synthesis, intelligibility enhancement of speech signals, signal processing for hearing aid applications, and perceptual aspects of signal processing.

JENS CHRISTIAN LINDOF (Member, IEEE) received the M.Sc.E.E. degree in telecommunication from Aalborg University, Denmark, in 1991. He is currently with RTX A/S, as a Chief Technology Officer, where he is heading the company's technology road-mapping activities. Previously, he was with Texas Instruments, Denmark, as a Senior Member of Technical Staff, the Site Manager, and responsible for advanced technology development. He was a company supervisor for several industrial Ph.D. students ranging from novel power amplifier designs to sound processing algorithms and to the importance of cooperative technical clusters. He is the co-author of several papers and patents within the area of RF, cellular, and audio. His keen intersts on entrepreneurship, he has involved with many start-ups, including a handful cofounded over the last 25 years.



JAN ØSTERGAARD (Senior Member, IEEE) received the M.Sc.E.E. degree from Aalborg University, Aalborg, Denmark, in 1999, and the Ph.D.E.E. degree (cum laude) from Delft University of Technology, Delft, The Netherlands, in 2007. From 1999 to 2002, he was a Research and Development Engineer with ETI A/S, Aalborg. From 2002 to 2003, he was a Research and Development Engineer with ETI Inc., VA, USA. From September 2007 to June 2008, he was a Postdoctoral Researcher with The University of Newcastle, NSW, Australia. He has been a Visiting Researcher with Tel Aviv University, Israel, and the Universidad Técnica Federico Santa María, Valparaíso, Chile. He is currently a Full Professor of information theory and signal processing and the Head of the Section on AI and Sound and of the Centre on Acoustic Signal Processing Research (CASPR), Aalborg University. His research interests include acoustic signal processing, statistical signal processing, information theory, joint source-channel coding, and networked control theory. He has received the Danish Independent Research Council's Young Researcher's Award, the Best Ph.D. Thesis Award by the European Association for Signal Processing (EURASIP), and the fellowships from the Danish Independent Research Council and the Villum Foundations Young Investigator Programme. He is an Associate Editor of IEEE TRANSACTIONS ON INFORMATION THEORY.