Journal of the Audio Engineering Society
Vol. 65, No. 1/2, January/February 2017 (© 2017)
DOI: https://doi.org/10.17743/jaes.2016.0047

Instrumental and Perceptual Evaluation of Dereverberation Techniques Based on Robust Acoustic Multichannel Equalization

INA KODRASI[1,*], **BENJAMIN CAUCHI**,[2] *AES Student Member*,
(ina.kodrasi@uni-oldenburg.de) (benjamin.cauchi@idmt.fraunhofer.de)

STEFAN GOETZE,[2] *AES Associate Member*, **AND SIMON DOCLO**,[1,2] *AES Associate Member*
(s.goetze@idmt.fraunhofer.de) (simon.doclo@uni-oldenburg.de)

[1] *Department of Medical Physics and Acoustics and Cluster of Excellence Hearing4All, University of Oldenburg*
[2] *Fraunhofer Institute for Digital Media Technology IDMT, Project Group Hearing, Speech and Audio Technology, Oldenburg, Germany*

Dereverberation techniques based on acoustic multichannel equalization, such as the relaxed multichannel least squares (RMCLS) technique and the partial multichannel equalization technique based on the multiple-input/output inverse theorem (PMINT), are known to be sensitive to room impulse response perturbations. In order to increase their robustness, several methods have been proposed, e.g., using a shorter reshaping filter length, incorporating regularization, or incorporating a sparsity-promoting penalty function. This paper focuses on evaluating the performance of these methods for single-source multi-microphone scenarios, both using instrumental performance measures as well as using subjective listening tests. While commonly used instrumental performance measures indicate that the regularized RMCLS technique yields the largest reverberant energy suppression, subjective listening tests show that the regularized and sparsity-promoting PMINT techniques yield the best perceptual speech quality. By analyzing the correlation between the instrumental and the perceptual results, it is shown that signal-based performance measures are more advantageous than channel-based performance measures to evaluate the perceptual speech quality of signals dereverberated by equalization techniques. Furthermore, this analysis also demonstrates the need to develop more reliable instrumental performance measures.

1 INTRODUCTION

Speech signals recorded in an enclosed space by microphones placed at a distance from the speaker are often corrupted by reverberation, which arises from the superposition of many delayed and attenuated copies of the clean signal. Reverberation causes signal degradation, typically leading to decreased speech quality and intelligibility [1–3] and performance deterioration in automatic speech recognition systems [4, 5]. With the continuously growing demand for high-quality hands-free communication in teleconferencing applications, voice-controlled systems, and hearing aids, speech enhancement techniques aiming at dereverberation have become indispensable.

In the last decades, many single and multichannel dereverberation techniques have been proposed [6], with multichannel techniques being generally preferred since they exploit both the spectro-temporal and the spatial characteristics of the received microphone signals. Existing multichannel dereverberation techniques can be broadly classified into spectral enhancement techniques [7, 8], probabilistic modeling-based techniques [9, 10], and acoustic multichannel equalization techniques [11–14]. Acoustic multichannel equalization techniques aim to reshape the available room impulse responses (RIRs) between the speaker and the microphone array. They can in theory achieve perfect dereverberation performance [11] and hence comprise an attractive approach to speech dereverberation.

A well-known multichannel equalization technique aiming at acoustic system inversion is the multiple-input/output inverse theorem (MINT) technique [11], which, however, suffers from drawbacks in practice. Since the available RIRs typically differ from the true RIRs due to, e.g., temperature

*This work was supported in part by the Cluster of Excellence 1077 "Hearing4All," funded by the German Research Foundation (DFG) and the Marie Curie Initial Training Network DREAMS (Grant No. 316969).

or position variations [15–17] or due to the sensitivity of blind and supervised system identification methods to near-common zeros or background noise [13, 18–21], MINT fails to invert the true RIRs [22]. This may lead to perceptually severe distortions in the output signal [13, 14]. In order to increase the robustness against RIR perturbations, partial multichannel equalization techniques such as relaxed multichannel least-squares (RMCLS) [13] and partial multichannel equalization based on MINT (PMINT) [14] have been proposed. Since early reflections tend to improve speech intelligibility [23] and late reflections are the major cause of speech intelligibility degradation, the objective of these techniques is to suppress only the late reflections.

Although partial equalization techniques are significantly more robust than MINT, their performance still remains susceptible to RIR perturbations [14]. Hence, several methods have been proposed to further increase the robustness of the RMCLS and PMINT techniques against RIR perturbations. In [24] it has been proposed to use a shorter reshaping filter length than conventionally used, resulting in a better-conditioned optimization criterion. In [14] it has been proposed to incorporate regularization in the filter design such that the distortion energy due to RIR perturbations is reduced. In [25, 26] it has been proposed to incorporate a signal-dependent sparsity-promoting penalty function in the filter design such that the output signal exhibits spectro-temporal characteristics of a clean signal. While simulation results in [14, 24–26] have shown using instrumental performance measures that all proposed methods effectively increase the robustness of the RMCLS and PMINT techniques, an extensive instrumental and perceptual comparison of the performance of all these methods is lacking.

The objective of this paper is threefold. First, using channel-based and signal-based instrumental performance measures, the reverberant energy suppression and the perceptual speech quality of the different robust extensions of the RMCLS and PMINT techniques are compared for various RIR perturbation levels. While instrumental performance measures indicate that the regularized RMCLS technique yields the largest reverberant energy suppression, different conclusions can be drawn about the perceptual speech quality based on different instrumental performance measures. Second, in order to determine the most perceptually advantageous technique, the overall speech quality is evaluated using subjective listening tests, showing that the regularized and sparsity-promoting PMINT techniques yield the best perceptual speech quality. Third, the correlation between the instrumental and the perceptual results is analyzed, showing the advantage of signal-based performance measures over channel-based performance measures as well as the necessity to develop more reliable instrumental performance measures to evaluate the perceptual speech quality of signals dereverberated using equalization techniques.

The paper is organized as follows. In Sec. 2 the RMCLS and PMINT techniques for acoustic multichannel equalization as well as the different proposed methods to increase their robustness against RIR perturbations are briefly reviewed. In Sec. 3 the considered acoustic scenarios and the algorithmic settings are described. In Secs. 4 and 5 the performance of the different techniques is compared using instrumental performance measures and subjective listening tests. Finally, Sec. 6 discusses the correlation between the instrumental and the perceptual results.

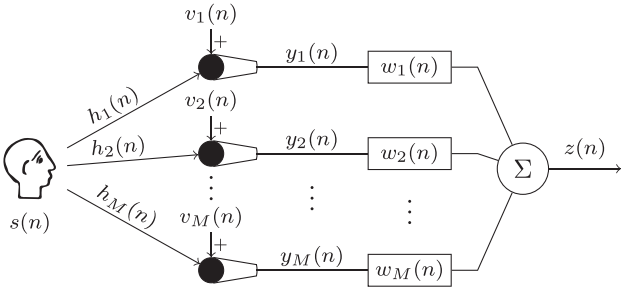

Fig. 1. Acoustic system configuration.

2 ROBUST ACOUSTIC MULTICHANNEL EQUALIZATION

2.1 Configuration and Notation

Consider the acoustic system depicted in Fig. 1, consisting of a single speech source in a reverberant room and M microphones. The mth microphone signal $y_m(n)$, $m = 1, \ldots, M$, at time index n is given by

$$y_m(n) = \underbrace{\sum_{l=0}^{L_h-1} h_m(l)s(n-l)}_{x_m(n)} + v_m(n) = x_m(n) + v_m(n), \quad (1)$$

where $h_m(l)$, $l = 0, \ldots, L_h - 1$, are the coefficients of the time-invariant RIR between the speech source and the mth microphone, $s(n)$ is the clean speech signal, $x_m(n)$ is the reverberant speech component, and $v_m(n)$ is the additive noise component. Since this paper aims to investigate the dereverberation performance of acoustic multichannel equalization techniques, in the following it is assumed that $v_m(n) = 0$, hence $y_m(n) = x_m(n)$.

Using the filter-and-sum structure in Fig. 1, the output signal $z(n)$ is equal to the sum of the filtered microphone signals, i.e.,

$$z(n) = \sum_{m=1}^{M} \sum_{l=0}^{L_w-1} w_m(l)x_m(n-l), \quad (2)$$

where $w_m(l)$, $l = 0, \ldots, L_w - 1$, are the coefficients of the L_w-taps long filter applied to the mth microphone signal. In vector notation, the RIR $\mathbf{h}_m$ and the filter $\mathbf{w}_m$ can be described as

$$\mathbf{h}_m = [h_m(0) \ldots h_m(L_h-1)]^T, \mathbf{w}_m = [w_m(0) \ldots w_m(L_w-1)]^T.$$

Using the ML_w-dimensional stacked filter vector $\mathbf{w} = [\mathbf{w}_1^T \ldots \mathbf{w}_M^T]^T$, the equalized impulse response (EIR) vector $\mathbf{c}$ of length $L_c = L_h + L_w - 1$, i.e., $\mathbf{c} = [c(0) \ldots c(L_c - 1)]^T$, can be expressed as

$$\mathbf{c} = \mathbf{H}\mathbf{w}, \quad (3)$$

where $\mathbf{H}$ denotes the $L_c \times ML_w$-dimensional multichannel convolution matrix of the RIRs, i.e., $\mathbf{H} = [\mathbf{H}_1 \dots \mathbf{H}_M]$, and

$$\mathbf{H}_m = \begin{bmatrix} h_m(0) & 0 & \dots & 0 \\ h_m(1) & h_m(0) & \ddots & \vdots \\ \vdots & h_m(1) & \ddots & 0 \\ h_m(L_h-1) & \vdots & \ddots & h_m(0) \\ 0 & h_m(L_h-1) & \ddots & h_m(1) \\ \vdots & \ddots & \ddots & \vdots \\ 0 & \dots & 0 & h_m(L_h-1) \end{bmatrix}. \quad (4)$$

Defining the L_c–dimensional clean speech vector $\mathbf{s}(n) = [s(n) \dots s(n-L_c+1)]^T$ and the L_w–dimensional mth reverberant signal vector $\mathbf{x}_m(n) = [x_m(n) \dots x_m(n-L_w+1)]^T$, with $\mathbf{x}_m(n) = \mathbf{H}_m^T \mathbf{s}(n)$, the output signal $z(n)$ can be expressed as

$$z(n) = \sum_{m=1}^{M} \mathbf{w}_m^T \mathbf{x}_m(n) = \sum_{m=1}^{M} \mathbf{w}_m^T \mathbf{H}_m^T \mathbf{s}(n) = \underbrace{\mathbf{w}^T \mathbf{H}^T}_{\mathbf{c}^T} \mathbf{s}(n). \quad (5)$$

As indicated by Eq. (5), the dereverberation performance of the speech enhancement system fully depends on the EIR vector $\mathbf{c}$. For conciseness, the time index n will be omitted when possible in the remainder of this paper.

2.2 Acoustic Multichannel Equalization

Acoustic multichannel equalization techniques assume that measurements or estimates of the RIRs are available. Such techniques aim at speech dereverberation by designing a reshaping filter $\mathbf{w}$ such that the (weighted) EIR in Eq. (3) is equal to a (weighted) dereverberated target EIR. Since the true RIRs are typically not available in practice, the reshaping filter is designed using the perturbed multichannel convolution matrix $\hat{\mathbf{H}}$ constructed from the available RIRs $\hat{\mathbf{h}}_m$. This matrix is equal to $\hat{\mathbf{H}} = \mathbf{H} + \mathbf{E}$, where $\mathbf{E}$ represents the convolution matrix of the RIR perturbations arising due to, e.g., temperature fluctuations [15], source-microphone geometry mismatches [16, 17], RIR estimation errors from blind and supervised system identification methods [18, 19], or microphone transfer function mismatches. It should be noted that microphone transfer function mismatches result in convolutive RIR perturbations instead of additive perturbations. However, the techniques discussed in the remainder of this paper are independent of the type of RIR perturbations present in the system, as long as a model is available to characterize these perturbations.

In this paper we will focus on the RMCLS [13] and PMINT [14] techniques, which compute the filter $\mathbf{w}$ as the solution to

$$\mathbf{W}\hat{\mathbf{H}}\mathbf{w} = \mathbf{W}\mathbf{c}_t, \quad (6)$$

with $\mathbf{W}$ an $L_c \times L_c$–dimensional diagonal weighting matrix and $\mathbf{c}_t$ the L_c–dimensional target EIR. The definition of the weighting matrix $\mathbf{W}$ and the target EIR $\mathbf{c}_t$ for the RMCLS and PMINT techniques is presented in Tables 1 and 2 respectively, where τ denotes a delay, L_d denotes the length of the direct path and early reflections, $\mathbf{I}$ denotes the $L_c \times L_c$–dimensional identity matrix, and $p \in \{1, \dots, M\}$, i.e., for the PMINT technique, the direct path and the early reflections of the target EIR are controlled by the first part of one of the available RIRs. Without loss of generality, other desired EIRs could also be used instead, as long as they are perceptually close to the true RIRs. From these definitions of $\mathbf{W}$ and $\mathbf{c}_t$, it can be observed that on the one hand, the RMCLS technique does not constrain all taps of the EIR, aiming only at suppressing the reverberant tail, while on the other hand, the PMINT technique constrains all taps of the EIR, aiming at suppressing the reverberant tail and preserving the perceptual speech quality. For more details on these techniques, we refer to [13, 14].

Table 1. Definition of the weighting matrix $\mathbf{W}$ in Eq. (6) for the RMCLS and PMINT techniques.

Technique	Weighting matrix $\mathbf{W}$
RMCLS	$\mathrm{diag}[\underbrace{1 \dots 1}_{\tau}\ \underbrace{1\,0 \dots 0}_{L_d}\ 1 \dots 1]^T$
PMINT	$\mathbf{I}$

Table 2. Definition of the target EIR $\mathbf{c}_t$ in Eq. (6) for the RMCLS and PMINT techniques.

Technique	Target EIR $\mathbf{c}_t$
RMCLS	$[\underbrace{0 \dots 0}_{\tau}\ 1\,0 \dots 0]^T$
PMINT	$[\underbrace{0 \dots 0}_{\tau}\ \underbrace{\hat{h}_p(0) \dots \hat{h}_p(L_d-1)}_{L_d}\ 0 \dots 0]^T$

The filter solving Eq. (6) is computed by minimizing the least-squares cost function[1]

$$J_{\mathrm{LS}} = \|\mathbf{W}(\hat{\mathbf{H}}\mathbf{w} - \mathbf{c}_t)\|_2^2. \quad (7)$$

As shown in [11, 14], assuming that the RIRs $\hat{\mathbf{h}}_m$ do not share any common zeros and using $L_w \geq \lceil \frac{L_h-1}{M-1} \rceil$, with $\lceil \cdot \rceil$ the ceiling operator, the filter minimizing Eq. (7) is equal to

$$\mathbf{w}_{\mathrm{LS}} = (\mathbf{W}\hat{\mathbf{H}})^+ \mathbf{W}\mathbf{c}_t, \quad (8)$$

where $\{\cdot\}^+$ denotes the matrix pseudo-inverse. When the true RIRs are available, i.e., $\hat{\mathbf{H}} = \mathbf{H}$, this filter yields perfect dereverberation performance, i.e., $\mathbf{W}\mathbf{H}\mathbf{w}_{\mathrm{LS}} = \mathbf{W}\mathbf{c}_t$ [14]. However, in the presence of RIR perturbations, i.e., $\hat{\mathbf{H}} \neq \mathbf{H}$, this filter typically fails to achieve dereverberation, i.e., $\mathbf{W}\mathbf{H}\mathbf{w}_{\mathrm{LS}} \neq \mathbf{W}\mathbf{c}_t$, possibly even causing large distortions in the output signal [14].

2.3 Increasing Robustness Against RIR Perturbations

In this section several methods that have been proposed to increase the robustness of the RMCLS and PMINT techniques are briefly reviewed. Furthermore, insights on the computational complexity of these different methods are provided.

[1] Strictly speaking, the cost function in Eq. (7) is a weighted least-squares cost function for $\mathbf{W} \neq \mathbf{I}$.

2.3.1 Decreasing the Reshaping Filter Length

In [24] it was analytically shown that using a shorter reshaping filter length than conventionally used, i.e., $L_w < \lceil \frac{L_h-1}{M-1} \rceil$, decreases the condition number of the matrix $\mathbf{W}\hat{\mathbf{H}}$. As analytically shown in [27], a smaller condition number yields a better-conditioned least-squares optimization criterion, with the resulting least-squares solution $\mathbf{w}_{\text{LS}}$ in Eq. (8) being less sensitive to perturbations in $\mathbf{W}\hat{\mathbf{H}}$.

2.3.2 Incorporating Regularization

In [14] it was proposed to increase the robustness of the RMCLS and PMINT techniques by incorporating regularization in the filter design, such that the distortion energy due to RIR perturbations is reduced. The regularized least-squares cost function is given by

$$J_{\text{RLS}} = \|\mathbf{W}(\hat{\mathbf{H}}\mathbf{w} - \mathbf{c}_t)\|_2^2 + \delta \mathbf{w}^T \mathbf{R}_\mathbf{e} \mathbf{w}, \tag{9}$$

with $\mathbf{R}_\mathbf{e}$ denoting the matrix modeling the perturbations, i.e., $\mathbf{R}_\mathbf{e} = \mathcal{E}\{\mathbf{E}^T\mathbf{E}\}$, where $\mathcal{E}$ denotes the expected value operator, and δ is a regularization parameter providing a trade-off between the minimization of the least-squares error J_{LS} and the distortion energy due to RIR perturbations $\mathbf{w}^T\mathbf{R}_\mathbf{e}\mathbf{w}$. The regularized least-squares filter minimizing Eq. (9) is given by

$$\mathbf{w}_{\text{RLS}} = [(\mathbf{W}\hat{\mathbf{H}})^T(\mathbf{W}\hat{\mathbf{H}}) + \delta \mathbf{R}_\mathbf{e}]^{-1}(\mathbf{W}\hat{\mathbf{H}})^T \mathbf{W}\mathbf{c}_t. \tag{10}$$

2.3.3 Incorporating Sparsity-Promoting Penalty Functions

In [25, 26] it was proposed to increase the robustness of the RMCLS and PMINT techniques by incorporating penalty functions that promote sparsity of the output signal in the short-time Fourier transform (STFT) domain, such that the output signal exhibits characteristics of a clean speech signal. The L_z-dimensional output signal vector $\mathbf{z} = [z(n) \ \ldots \ z(n - L_z + 1)]^T$ can be expressed as

$$\mathbf{z} = \mathbf{X}\mathbf{w}, \tag{11}$$

where $\mathbf{X}$ denotes the $L_z \times ML_w$-dimensional multichannel convolution matrix of the microphone signals, i.e., $\mathbf{X} = [\mathbf{X}_1 \ \mathbf{X}_2 \ \ldots \ \mathbf{X}_M]$, and

$$\mathbf{X}_m = \begin{bmatrix} x_m(n) & \cdots & x_m(n - L_w + 1) \\ x_m(n-1) & \cdots & x_m(n - L_w) \\ \vdots & \ddots & \vdots \\ x_m(n - L_z + 1) & \cdots & x_m(n - L_w - L_z + 2) \end{bmatrix}. \tag{12}$$

The sparsity-promoting least-squares cost function is then given by

$$J_{\text{SLS}} = \|\mathbf{W}(\hat{\mathbf{H}}\mathbf{w} - \mathbf{c}_t)\|_2^2 + \eta f_{\text{SP}}(\mathbf{\Psi}\mathbf{z}), \tag{13}$$

where f_{SP} denotes a sparsity-promoting penalty function and η is a weighting parameter providing a trade-off between the minimization of the least-squares error J_{LS} and the penalty function value $f_{\text{SP}}(\mathbf{\Psi}\mathbf{z})$. The operator $\mathbf{\Psi}$ denotes the STFT operator transforming the L_z-dimensional time-domain vector $\mathbf{z}$ into the $L_{\tilde{z}}$-dimensional time-frequency domain vector $\tilde{\mathbf{z}}$ consisting of the STFT coefficients of the output signal, with $\tilde{\mathbf{z}} = \mathbf{\Psi}\mathbf{z}$. Since no closed-form expression is available for the filter minimizing the cost function in Eq. (13), the sparsity-promoting least-squares filter can be computed using, e.g., the iterative alternating direction method of multipliers (ADMM) algorithm [28]. Introducing the auxiliary variable $\mathbf{a}$ such that the optimization problem in Eq. (13) is split into simpler sub-problems, the ADMM algorithm computes the sparsity-promoting least-squares filter using the following update rules [25, 26] until a termination criterion is satisfied (cf., Sec. 3):

$$\begin{aligned} \mathbf{w}^{(i+1)} = {} & [2(\mathbf{W}\hat{\mathbf{H}})^T(\mathbf{W}\hat{\mathbf{H}}) + \rho \mathbf{X}^T\mathbf{X}]^{-1} \\ & \times [2(\mathbf{W}\hat{\mathbf{H}})^T(\mathbf{W}\mathbf{c}_t) + \rho \mathbf{X}^T \mathbf{\Psi}^H(\mathbf{a}^{(i)} - \mathbf{\lambda}^{(i)})], \end{aligned} \tag{14}$$

$$\mathbf{a}^{(i+1)} = S_{\eta/\rho}(\mathbf{\Psi}\mathbf{X}\mathbf{w}^{(i+1)} + \mathbf{\lambda}^{(i)}), \tag{15}$$

$$\mathbf{\lambda}^{(i+1)} = \mathbf{\lambda}^{(i)} + \mathbf{\Psi}\mathbf{X}\mathbf{w}^{(i+1)} - \mathbf{a}^{(i+1)}. \tag{16}$$

In Eqs. (14)–(16), $\{\cdot\}^{(i)}$ denotes the variable in the ith iteration, $\mathbf{\lambda}$ is the so-called dual (splitting) variable, $\rho > 0$ is the ADMM penalty parameter, and $S_{\eta/\rho}$ denotes the proximal mapping of the used sparsity-promoting penalty function [29]. Simulation results in [26] have shown that incorporating an l_0-norm, l_1-norm, or weighted l_1-norm sparsity-promoting penalty function significantly increases the robustness of the RMCLS and PMINT techniques, with the weighted l_1-norm penalty function yielding the best performance. Hence, in this paper we only consider the weighted l_1-norm penalty function, defined as

$$f_{\text{SP}}(\tilde{\mathbf{z}}) = \|\text{diag}\{\mathbf{u}\}\tilde{\mathbf{z}}\|_1 = \sum_{q=0}^{L_{\tilde{z}}-1} |u(q)\tilde{z}(q)|, \tag{17}$$

with $u(q) > 0$, $q = 0, \ldots, L_{\tilde{z}} - 1$, denoting user-defined scalar weights. In order to preserve the spectro-temporal structure of a typical speech signal, the weights $u(q)$ are defined as [26]

$$u(q) = \frac{1}{|\tilde{x}_p(q)| + \zeta}, \quad q = 0, \ \ldots, \ L_{\tilde{z}} - 1, \tag{18}$$

with $\tilde{x}_p(q)$ the STFT coefficients of the p-th microphone signal, where $p \in \{1, \ldots, M\}$, and $\zeta > 0$ a small positive scalar used to avoid division by 0.

While simulation results in [14, 24–26] have shown that all proposed methods are effective in increasing the robustness of the RMCLS and PMINT techniques against RIR perturbations, an extensive instrumental and perceptual comparison of the performance of all these methods to determine the most robust and perceptually advantageous technique is lacking.

2.3.4 Computational Complexity Considerations

The computational complexity of all considered methods is at most cubic[2], since matrix multiplications and matrix inversions account for the dominant operations in all reshaping filter computations, cf., Eqs. (8), (10), and (14)–(16). The complexity of using a shorter reshaping filter length is $O(n_r^3)$, where n_r denotes the number of rows of the matrix

[2] This upper bound may be tightened when exploiting the fact that the matrices involved are symmetric or Toeplitz.

Table 3. Characteristics of the considered acoustic systems.

System	T_{60} [ms]	d_{sm} [m]	d_{im} [m]	L_h
S_1	450	3	0.05	3600
S_2	610	2	0.04	4880

$\mathbf{W}\hat{\mathbf{H}}$ when $L_w < \left\lceil \frac{L_h-1}{M-1} \right\rceil$. The complexity of using regularization is $O(n_c^3)$, where n_c denotes the number of columns of the matrix $\mathbf{W}\hat{\mathbf{H}}$ when $L_w = \left\lceil \frac{L_h-1}{M-1} \right\rceil$. Finally, the complexity of using a sparsity-promoting penalty function is $O(L_z^3)$, where L_z denotes the length of the output signal vector. Since typically $n_r < n_c \ll L_z$, decreasing the reshaping filter length results in the lowest computational complexity, whereas incorporating a sparsity-promoting penalty function results in the highest computational complexity. In addition, the execution of the sparsity-promoting method takes a significantly longer time than the execution of the other methods due to the multiple number of iterations.

3 ACOUSTIC SCENARIOS AND ALGORITHMIC SETTINGS

This section describes the considered acoustic scenarios and the algorithmic settings for which the performance of the robust extensions of the RMCLS and PMINT techniques is evaluated.

3.1 Acoustic Scenarios

We considered two different reverberant acoustic systems with a single speech source and $M = 4$ omnidirectional microphones. For each acoustic system, Table 3 presents the reverberation time T_{60}, the source-microphone distance d_{sm}, the inter-microphone distance d_{im}, and the RIR length L_h at a sampling frequency $f_s = 8$ kHz. The RIRs between the speech source and the microphones were measured using the swept-sine technique [30] and the reverberant signals were generated by convolving two sentences of clean speech (approximately 4 s long) from the HINT database [31] with the measured RIRs. In order to simulate RIR perturbations, the measured RIRs were perturbed by proportional Gaussian distributed errors as proposed in [32], such that a desired normalized projection misalignment (NPM), defined as

$$\text{NPM} = 10\log_{10} \frac{\left\| \mathbf{h}_m - \frac{\mathbf{h}_m^T \hat{\mathbf{h}}_m}{\hat{\mathbf{h}}_m^T \hat{\mathbf{h}}_m} \hat{\mathbf{h}}_m \right\|_2^2}{\|\mathbf{h}_m\|_2^2}, \tag{19}$$

is obtained. The considered NPMs are $\text{NPM}_1 = -33$ dB and $\text{NPM}_2 = -15$ dB, with NPM_1 representing a moderate perturbation level and NPM_2 representing a high perturbation level. Hence, the performance of all considered techniques is evaluated for four different acoustic scenarios, i.e., S_1-NPM_1, S_2-NPM_1, S_1-NPM_2, and S_2-NPM_2.

3.2 Algorithmic Settings

As previously mentioned, we investigate the performance of the following techniques (cf., Sec. 2.2):

- L-RMCLS, i.e., the RMCLS technique using a shorter reshaping filter length;
- R-RMCLS, i.e., the regularized RMCLS technique;
- S-RMCLS, i.e., the weighted l_1-norm sparsity-promoting RMCLS technique;
- L-PMINT, i.e., the PMINT technique using a shorter reshaping filter length;
- R-PMINT, i.e., the regularized PMINT technique; and
- S-PMINT, i.e., the weighted l_1-norm sparsity-promoting PMINT technique.

For the R-RMCLS, S-RMCLS, R-PMINT, and S-PMINT techniques the reshaping filter length is set to $L_w = \left\lceil \frac{L_h-1}{M-1} \right\rceil$, i.e., $L_w = 1200$ for the system S_1 and $L_w = 1627$ for the system S_2. As shown in [11], this filter length is the minimum length required for perfect dereverberation performance. For all techniques, the delay is set to $\tau = 90$ and the length of the direct path and early reflections is set to $L_d = 0.01 \times f_s$ (i.e., 10 ms), cf., Tables 1 and 2. The target EIR $\mathbf{c}_t$ for the robust extensions of the PMINT technique is set to the direct path and early reflections of the first RIR $\hat{\mathbf{h}}_1$, i.e., $p = 1$. For the regularized techniques, the matrix $\mathbf{R}_{\mathbf{e}}$ modeling the RIR perturbations is set to $\mathbf{R}_{\mathbf{e}} = \mathbf{I}$. For the sparsity-promoting techniques, the STFT is computed using a 32 ms Hamming window with 50% overlap between successive frames. As in [25], the variables $\mathbf{w}$, $\mathbf{a}$, and $\boldsymbol{\lambda}$ are initialized with $[1\,0\ldots0]^T$ and the termination criterion is set to either the number of iterations exceeding 150 or the change in the filter norm dropping below 10^{-3}. The weights in Eq. (18) are computed using the STFT coefficients of the first microphone signal $\tilde{x}_1(q)$, i.e., $p = 1$.

The considered reshaping filter lengths L_w for the L-RMCLS and L-PMINT techniques, regularization parameters δ for the R-RMCLS and R-PMINT techniques, and weighting and penalty parameters η and ρ for the S-RMCLS and S-PMINT techniques are

$$L_w \in \left\{ 500, 600, \ldots, \left\lceil \frac{L_h - 1}{M - 1} \right\rceil \right\}, \tag{20}$$

$$\delta \in \{10^{-7}, 10^{-6}, \ldots, 10^{-1}, 1, 3, 5, 7, 10\}, \tag{21}$$

$$\eta \in \{10^{-7}, 10^{-6}, 10^{-5}, 10^{-4}\}, \tag{22}$$

$$\rho \in \{10^{-7}, 10^{-6}, \ldots, 10^{-1}\}. \tag{23}$$

As in [14, 24], the optimal reshaping filter length, the optimal regularization parameter, and the optimal weighting and penalty parameters used in the following simulations are intrusively selected from Eqs. (20)–(23) as the parameters maximizing the perceptual evaluation of speech quality (PESQ) score [33] for each technique and each acoustic scenario (cf., Sec. 4 for details on the PESQ score computation). It should be noted that the computation of the PESQ score for selecting the optimal parameters is an intrusive procedure that is not applicable in practice, since knowledge of the true RIRs is required in order to compute the reference signal and the resulting EIR.

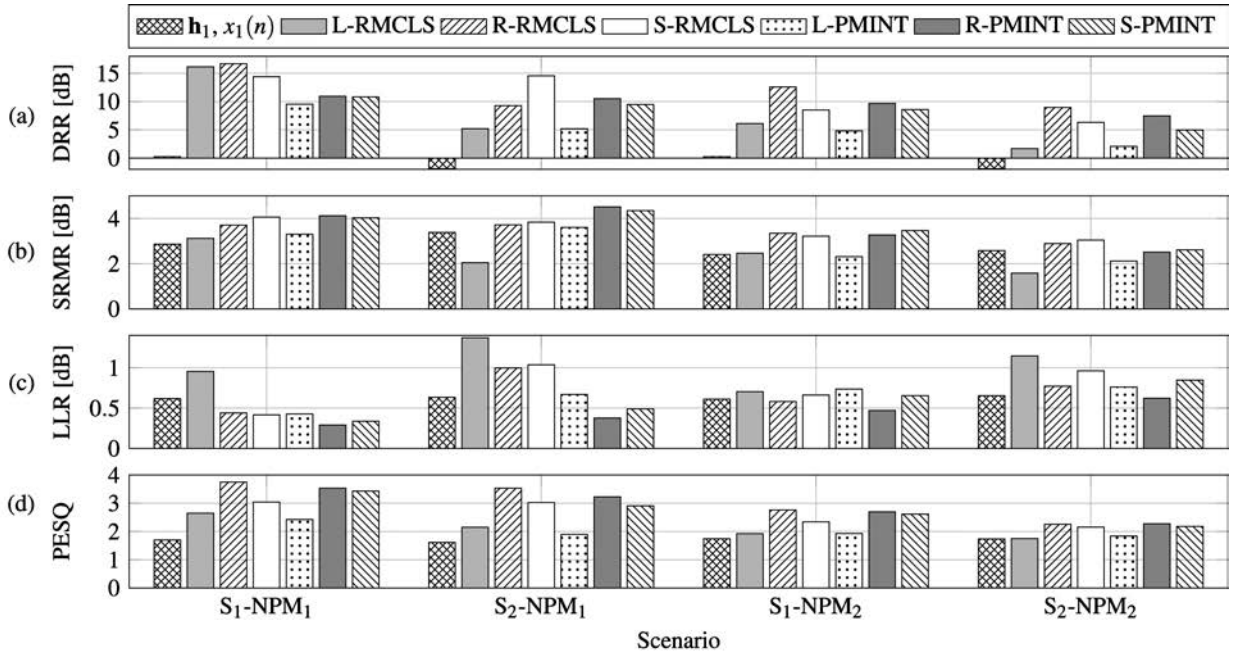

Fig. 2. Instrumental measures for the robust extensions of the RMCLS and PMINT techniques for all considered acoustic scenarios: (a) DRR, (b) SRMR, (c) LLR, and (d) PESQ.

4 INSTRUMENTAL EVALUATION

In this section the performance of the different considered techniques is evaluated by means of commonly used instrumental performance measures, i.e., direct-to-reverberant ratio (DRR) [6], speech-to-reverberation modulation energy ratio (SRMR) [34], log likelihood ratio (LLR) [35], and PESQ [33]. The channel-based DRR measure has been shown to correlate well with the perceived amount of reverberation for unprocessed signals [36], whereas the signal-based SRMR, LLR, and PESQ measures have been shown to correlate well with the perceived overall quality of signals processed by speech enhancement algorithms for dereverberation and noise reduction [37]. While the SRMR measure is a non-intrusive measure, the LLR and PESQ measures are intrusive measures comparing the output signal to a (dereverberated) reference signal. The reference signal employed in this evaluation is the clean speech signal convolved with the direct path and the early reflections (up to 10 ms) of the true RIR $\mathbf{h}_1$. Note that a higher DRR, a higher SRMR, a lower LLR, and a higher PESQ score indicate a better performance.

Fig. 2a depicts the obtained DRR for the input RIR $\mathbf{h}_1$ and for the EIRs $\mathbf{c}$ obtained using the robust extensions of the RMCLS and PMINT techniques. The following conclusions can be drawn by comparing the presented DRR values:

- All techniques improve the DRR in comparison to the input RIR $\mathbf{h}_1$.
- The robust extensions of the RMCLS technique generally yield a similar or higher DRR than the robust extensions of the PMINT technique. This is to be expected since the robust extensions of the RMCLS technique relax the constraints on the filter design and aim only at suppressing the late reverberation, whereas the robust extensions of the PMINT technique also aim at preserving the perceptual speech quality (which is not reflected by the DRR measure).
- The R-RMCLS technique typically yields the highest DRR for the considered scenarios (except for the scenario S_2-NPM_1, where the S-RMCLS technique yields the highest DRR).
- The R-PMINT technique typically yields a higher DRR than the S-PMINT technique (except for the scenario S_1-NPM_1, where the R-PMINT and S-PMINT techniques yield a similar DRR).
- The L-RMCLS and L-PMINT techniques yield the lowest DRR out of all considered robust extensions. This is not surprising since these techniques simply use a shorter reshaping filter length, without explicitly taking into account the structure of the RIR perturbations or the characteristics of the output speech signal.
- The performance of all considered techniques is generally higher for the system S_1 than for the system S_2. This can be explained by the higher reverberation time of the system S_2, leading to a larger number of perturbed RIR taps to be reshaped, and hence, an increased sensitivity of all considered techniques to RIR perturbations.

Figs. 2b–2d depict the obtained SRMR, LLR, and PESQ scores for the reverberant microphone signal $x_1(n)$ and for the output signals $z(n)$ obtained using the robust

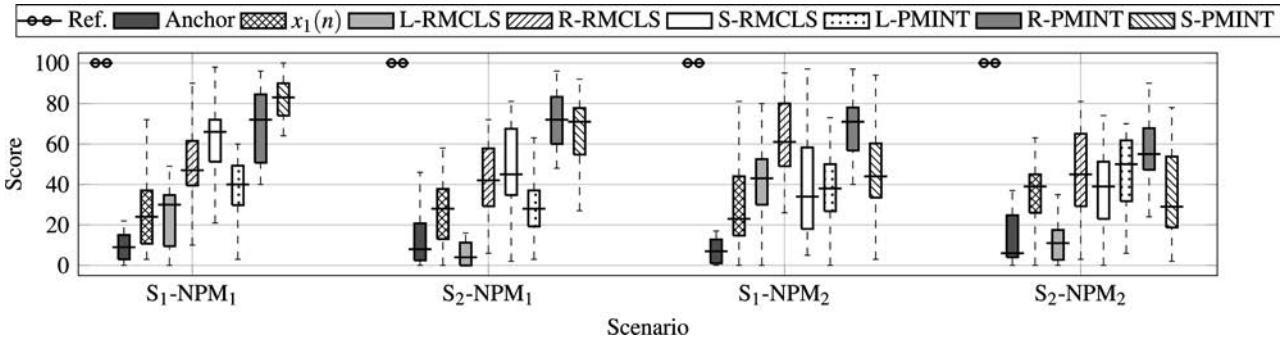

Fig. 3. MUSHRA scores for the reverberant microphone signal $x_1(n)$ and for the output signals $z(n)$ obtained using the robust extensions of the RMCLS and PMINT techniques for all considered acoustic scenarios. In addition, the scores of the hidden reference and the anchor are displayed. For each box, the central mark is the median, the edges of the box are the 25th and the 75th percentiles, and the whiskers extend to 1.5 times the interquartile range from the median.

extensions of the RMCLS and PMINT techniques. The following conclusions can be drawn by comparing the presented instrumental measures:

- Not all techniques improve the overall quality in comparison to the reverberant signal $x_1(n)$, e.g., the L-RMCLS technique yields a lower SRMR for scenarios S_2-NPM$_1$ and S_2-NPM$_2$, the L-PMINT technique yields a lower SRMR for the scenario S_2-NPM$_2$, and the R-PMINT technique is the only technique consistently improving the LLR for all scenarios.
- The robust extensions of the PMINT technique generally yield a similar or better SRMR and LLR than the robust extensions of the RMCLS technique. Surprisingly, the robust extensions of the RMCLS technique generally yield a similar or better PESQ score than the robust extensions of the PMINT technique, implying that PESQ does not appear to reflect the better preservation of the early reflections achieved by the robust extensions of the PMINT technique but puts more emphasis instead on the better reverberant tail suppression achieved by the robust extensions of the RMCLS technique.
- The R-PMINT technique typically yields the best SRMR and LLR (except for the scenario S_2-NPM$_2$, where the S-RMCLS technique yields the best SRMR), whereas the R-RMCLS technique typically yields the best PESQ score.
- As expected, the L-RMCLS and L-PMINT techniques typically yield the lowest performance in terms of all instrumental performance measures.
- The performance of all considered techniques is generally higher for the system S_1 than for the system S_2.

In summary, based on instrumental performance measures it can be said that incorporating regularization and sparsity-promoting penalty functions is more advantageous to increase the robustness of equalization techniques than using a shorter reshaping filter length. Furthermore, as expected, the robust extensions of the RMCLS technique achieve a larger reverberant energy suppression (as evaluated using the DRR measure) than the robust extensions of the PMINT technique, with the R-RMCLS technique typically yielding the best performance. However, when comparing the perceptual speech quality achieved by the different techniques, different conclusions can be derived depending on the used instrumental performance measure, highlighting the necessity of conducting subjective listening tests.

5 PERCEPTUAL RESULTS

The perceptual evaluation is based on a multi stimulus test with hidden reference and anchor (MUSHRA) using the specifications given in [38]. The evaluation is conducted for the reverberant microphone signal $x_1(n)$ and for the output signals $z(n)$ obtained using all considered techniques. In addition to these signals, a hidden reference and an anchor are presented to the subjects. The hidden reference has been generated as the clean speech signal convolved with the direct path and the early reflections (up to 10 ms) of the true RIR $\mathbf{h}_1$. The anchor has been generated as the low-pass filtered microphone signal $x_1(n)$ with a cut-off frequency of 3 kHz. Sound samples for each considered acoustic scenario can be found at bit.ly/mushrasamples. The signals are diotically presented to the subjects through headphones (Sennheiser HDA 200) using an RME Fireface UFX sound card, with all signals normalized in amplitude. A total of 21 self-reported normal hearing subjects who are familiar with speech processing participated in the listening tests. The subjects evaluated the signals in terms of the attribute "overall speech quality" on a scale from 0 to 100. Prior to the actual evaluation, the subjects were trained to familiarize themselves with the task and the signals under test. Furthermore, they could adjust the sound volume to a comfortable level. The order of presentation of signals and scenarios were randomized between all subjects.

Fig. 3 depicts the obtained MUSHRA scores for the reverberant microphone signal and for the output signals obtained using the robust extensions of the RMCLS and PMINT techniques. For completeness, the obtained MUSHRA scores for the reference and the anchor are also depicted, illustrating that the reference is correctly identified for all scenarios and that the anchor is typically rated as having the worst perceptual speech quality (except for the

scenario S_2-NPM_1, where the L-RMCLS technique yields a worse quality). In general it can be observed that the rating variability between subjects (as shown by the whiskers in each boxplot) is rather large. This is commonly the case for listening tests evaluating the overall speech quality achieved by dereverberation algorithms, e.g., [2, 3, 39]. Since the artifacts and distortions produced by the considered techniques are quite different, the perception of these artifacts and distortions by different subjects is also rather different.

For the moderate RIR perturbation level ($NPM_1 = -33$ dB), it can be observed that all proposed techniques typically improve the perceptual speech quality in comparison to the reverberant microphone signal (except for the L-RMCLS technique yielding a worse perceptual speech quality and the L-PMINT technique yielding a similar perceptual speech quality for the system S_2). Furthermore, out of the different methods proposed to increase the robustness of equalization techniques, incorporating a sparsity-promoting penalty function yields the best perceptual speech quality, whereas using a shorter reshaping filter length yields the worst perceptual speech quality. Finally, it can be observed that due to the better preservation of the early reflections, the robust extensions of the PMINT technique yield a better perceptual speech quality than the robust extensions of the RMCLS technique, with the S-PMINT technique yielding the best perceptual speech quality.

For the higher RIR perturbation level ($NPM_2 = -15$ dB), it can be observed that more techniques fail to improve the perceptual speech quality in comparison to the reverberant microphone signal, i.e., the L-RMCLS, S-RMCLS, and S-PMINT techniques yield a similar or worse quality for the system S_2. Furthermore, it can be observed that, similarly as before, the robust extensions of the PMINT technique yield a similar or better perceptual speech quality than the robust extensions of the RMCLS technique (except for the S-PMINT technique yielding a worse quality than the S-RMCLS technique for the system S_2). However, unlike for $NPM_1 = -33$ dB, it can now be observed that the R-PMINT technique results in the best perceptual speech quality, outperforming the S-PMINT technique. Incorporating regularization in the RMCLS and PMINT techniques yields the best perceptual speech quality, whereas incorporating sparsity-promoting penalty functions typically yields the worst perceptual speech quality. Hence, while incorporating sparsity-promoting penalty functions seems to be very advantageous to increase the robustness and the perceptual speech quality in the presence of moderate RIR perturbation levels, the performance of sparsity-promoting techniques seems to deteriorate more rapidly with increasing perturbation levels than the performance of regularized techniques. This is not unexpected, since sparsity-promoting penalty functions only rely on general spectro-temporal characteristics of clean speech signals, whereas regularization aims at explicitly modeling and suppressing the level of RIR perturbations.

To determine whether the previously discussed results are statistically significant, a statistical analysis has been conducted. Since the data are not normally distributed, a Friedman's test [40] with the factor "technique" has been performed for the different considered scenarios. As summarized in Table 4, the statistical analysis shows a significant influence of the factor "technique" for all scenarios. To determine the sources of significance, a Wilcoxon signed-rank test [41] has been separately conducted for each scenario. The obtained results for each acoustic scenario are presented in Tables 5–8, with the ticks representing a statistically significant difference, i.e., $p < 0.05$, and the crosses representing no statistically significant difference, i.e., $p < 0.05$. The presented results are obviously symmetric across the diagonal since such entries correspond to the same pair comparison. Table 5 shows that for the system S_1 and the moderate RIR perturbation level NPM_1 only the regularized and the sparsity-promoting techniques yield a statistically significant improvement in comparison to the reverberant microphone signal. Furthermore, it can be observed that the S-PMINT technique is the only technique yielding a statistically significant improvement in comparison to all other techniques. Table 6 shows that for the system S_2 and the moderate RIR perturbation level NPM_1 only the R-PMINT, S-PMINT, and R-RMCLS techniques yield a statistically significant improvement in comparison to the reverberant microphone signal. Furthermore, it can be observed that these techniques yield the most statistically significant improvements in comparison to other techniques. Table 7 shows that for the system S_1 and the high RIR perturbation level NPM_2 the R-PMINT technique and the robust extensions of the RMCLS technique yield a statistically significant improvement in comparison to the reverberant microphone signal. Furthermore, it can be observed that the R-PMINT and the R-RMCLS techniques yield the most statistically significant improvements in comparison to other techniques. Finally, Table 8 shows that for the system S_2 and the high RIR perturbation level NPM_2 only the R-PMINT technique yields a statistically significant improvement in comparison to the reverberant microphone signal. Furthermore, the R-PMINT technique also yields the most statistically significant improvements in comparison to other techniques.

Table 4. Results of the Friedman's test for all considered scenarios. The variable χ^2 denotes the Friedman's chi square statistic and the value $\rho < 0.001$ indicates the significance of the results.

Scenario	χ^2	ρ
S_1-NPM_1	140	$\rho < 0.001$
S_2-NPM_1	88	$\rho < 0.001$
S_1-NPM_2	119	$\rho < 0.001$
S_2-NPM_2	83	$\rho < 0.001$

In summary, even though the statistical significance criterion is not always satisfied, the trend of the results confirm that the robust extensions of the PMINT technique yield a better perceptual speech quality than the robust extensions of the RMCLS technique. Furthermore, the S-PMINT technique results in the best perceptual speech quality for moderate RIR perturbation levels, whereas the R-PMINT

Table 5. Wilcoxon signed-rank test for scenario S_1-NPM_1. The ticks represent a statistically significant difference and the crosses represent no statistically significant difference.

	$x_1(n)$	L-RMCLS	R-RMCLS	S-RMCLS	L-PMINT	R-PMINT	S-PMINT
$x_1(n)$		✗	√	√	✗	√	√
L-RMCLS	✗		√	√	✗	√	√
R-RMCLS	√	√		✗	✗	√	√
S-RMCLS	√	√	✗		√	✗	√
L-PMINT	✗	✗	✗	√		√	√
R-PMINT	√	√	√	✗	√		√
S-PMINT	√	√	√	√	√	√	

Table 6. Wilcoxon signed-rank test for scenario S_2-NPM_1. The ticks represent a statistically significant difference and the crosses represent no statistically significant difference.

	$x_1(n)$	L-RMCLS	R-RMCLS	S-RMCLS	L-PMINT	R-PMINT	S-PMINT
$x_1(n)$		✗	√	✗	✗	√	√
L-RMCLS	✗		√	✗	✗	√	✗
R-RMCLS	√	√		√	√	✗	√
S-RMCLS	✗	✗	√		✗	√	✗
L-PMINT	✗	✗	√	✗		√	✗
R-PMINT	√	√	✗	√	√		√
S-PMINT	✗	✗	√	✗	✗	√	

Table 7. Wilcoxon signed-rank test for scenario S_1-NPM_2. The ticks represent a statistically significant difference and the crosses represent no statistically significant difference.

	$x_1(n)$	L-RMCLS	R-RMCLS	S-RMCLS	L-PMINT	R-PMINT	S-PMINT
$x_1(n)$		√	√	√	✗	√	✗
L-RMCLS	√		√	√	√	√	√
R-RMCLS	√	√		✗	√	√	√
S-RMCLS	√	√	✗		√	√	√
L-PMINT	✗	√	√	√		√	√
R-PMINT	√	√	√	√	√		✗
S-PMINT	✗	√	√	√	√	✗	

Table 8. Wilcoxon signed-rank test for scenario S_2-NPM_2. The ticks represent a statistically significant difference and the crosses represent no statistically significant difference.

	$x_1(n)$	L-RMCLS	R-RMCLS	S-RMCLS	L-PMINT	R-PMINT	S-PMINT
$x_1(n)$		√	✗	✗	✗	√	✗
L-RMCLS	√		√	√	√	√	√
R-RMCLS	✗	√		✗	✗	✗	✗
S-RMCLS	✗	√	✗		✗	√	✗
L-PMINT	✗	√	✗	✗		✗	✗
R-PMINT	√	√	✗	√	✗		√
S-PMINT	✗	√	✗	✗	✗	√	

Table 9. Absolute value of the Pearson product-moment correlation coefficient between the perceptual ratings and the instrumental performance measures.

Measure	S_1-NPM_1	S_2-NPM_1	S_1-NPM_2	S_2-NPM_2
DRR	0.25	0.61	0.81	0.44
SRMR	**0.93**	**0.97**	0.62	0.52
LLR	0.77	0.71	0.75	**0.87**
PESQ	0.76	0.69	**0.82**	0.56

technique results in the best perceptual speech quality for high RIR perturbation levels.

6 CORRELATION ANALYSIS BETWEEN INSTRUMENTAL AND PERCEPTUAL RESULTS

When comparing the instrumental evaluation results in Sec. 4 with the perceptual evaluation results in Sec. 5, it can be observed that not all perceptual results can be well predicted by the instrumental performance measures. On the one hand, the instrumental performance measures accurately predicted that (1) the regularized and sparsity-promoting techniques generally outperform the techniques using a shorter reshaping filter length and that (2) the performance of all considered techniques for the system S_1 is typically better than for the system S_2. On the other hand, the instrumental performance measures failed to predict (1) the consistent perceptual advantage of the robust extensions of the PMINT technique over the robust extensions of the RMCLS technique as well as (2) the perceptual advantage of the S-PMINT technique over the R-PMINT technique for moderate RIR perturbation levels.

Table 9 shows the correlation between the perceptual ratings and the instrumental performance measures as determined by the Pearson product-moment correlation coefficient (PPMCC), computed as

$$\text{PPMCC} = \frac{\sum_j (a_j - \bar{a})(b_j - \bar{b})}{\sqrt{\sum_j (a_j - \bar{a})^2 (b_j - \bar{b})^2}}, \tag{24}$$

where a_j and b_j denote the perceptual and instrumental ratings for the j-th sound sample and $\bar{a}$ and $\bar{b}$ denote the respective mean values. It can be observed that for each scenario at least one instrumental performance measure yields a high correlation to the perceptual ratings, with the signal-based performance measures typically yielding a higher correlation than the channel-based DRR measure. Furthermore, it can be observed that except for the LLR measure, the correlation for all other measures strongly depends on the considered scenario. The correlation for the DRR measure varies between 0.25 and 0.81, which is to be expected since a purely energy-based measure cannot always reflect how the remaining distortions in the late reverberant tail are perceived. For the moderate RIR perturbation level NPM_1, the SRMR measures shows a very high correlation to the perceptual ratings, whereas for the higher perturbation level NPM_2, the correlation values significantly decrease. Hence, it appears that the SRMR measure, which has been primarily developed and optimized on unprocessed reverberant signals, can very well predict the quality of signals with little or no distortions but does not reflect the distortions introduced by equalization techniques. Furthermore, the auditory-based PESQ measure does not always appear to reflect the distortions introduced by equalization techniques, e.g., yielding a low correlation of 0.56 for the scenario with most distortions in the output signal, i.e., S_2-NPM_2. Finally, it appears that a relatively simple linear prediction coefficient based distance measure such as LLR reflects the distortions introduced by equalization techniques more reliably than all other measures over all considered scenarios, with correlation values varying between 0.71 and 0.87.

In summary, while instrumental performance measures are certainly a valuable tool when designing speech dereverberation techniques, the impact of distortions and artifacts caused by acoustic multichannel equalization techniques can only be truly assessed using subjective listening tests. Since the considered instrumental measures are incapable of accurately predicting the perceptual ratings, further development of instrumental performance measures is required.

7 CONCLUSION

In this paper we have evaluated the performance of several robust extensions of acoustic multichannel equalization based on RMCLS and PMINT by means of instrumental performance measures and subjective listening tests. Instrumental performance measures show that the regularized RMCLS technique yields the largest reverberant energy suppression. Subjective listening tests show that the robust extensions of the PMINT technique yield the best perceptual speech quality, with the sparsity-promoting PMINT technique yielding the best quality for moderate RIR perturbation levels and the regularized PMINT technique yielding the best quality for high RIR perturbation levels. A correlation analysis between the instrumental and perceptual results shows that signal-based performance measures typically yield a higher correlation than channel-based performance measures when evaluating the perceptual quality of signals processed by acoustic multichannel equalization techniques. Furthermore, the provided correlation analysis highlights the need to develop more accurate instrumental performance measures, reliably reflecting the distortions introduced by acoustic multichannel equalization techniques.

8 REFERENCES

[1] R. Beutelmann and T. Brand, "Prediction of Speech Intelligibility in Spatial Noise and Reverberation for Normal-Hearing and Hearing-Impaired Listeners," *J. Acous. Soc. Am.*, vol. 120, no. 1, pp. 331–342 (2006 Jul.). http://dx.doi.org/10.1121/1.2202888

[2] A. Warzybok, I. Kodrasi, J. O. Jungmann, E. A. P. Habets, T. Gerkmann, A. Mertins, S. Doclo, B. Kollmeier, and S. Goetze, "Subjective Speech Quality and Speech Intelligibility Evaluation of Single-Channel Dereverberation Algorithms," *Proc. International Workshop on Acoustic Echo*

and Noise Control, Antibes, France (Sep. 2014), pp. 333–337. http://dx.doi.org/10.1109/IWAENC.2014.6954313

[3] S. Goetze, E. Albertin, J. Rennies, E. Habets, and K.-D. Kammeyer, "Speech Quality Assessment for Listening-Room Compensation," *J. Audio Eng. Soc.*, vol. 62, pp. 386–399 (2014 Jun.). http://dx.doi.org/10.17743/jaes.2014.0025

[4] T. Yoshioka, A. Sehr, M. Delcroix, K. Kinoshita, R. Maas, T. Nakatani, and W. Kellermann, "Making Machines Understand Us in Reverberant Rooms: Robustness against Reverberation for Automatic Speech Recognition," *IEEE Signal Processing Magazine*, vol. 29, no. 6, pp. 114–126 (2012 Nov.). http://dx.doi.org/10.1109/MSP.2012.2205029

[5] F. Xiong, B. Meyer, N. Moritz, R. Rehr, J. Anemüller, T. Gerkmann, S. Doclo, and S. Goetze, "Front-End Technologies for Robust ASR in Reverberant Environments— Spectral Enhancement-Based Dereverberation and Auditory Modulation Filterbank Features," *EURASIA J. Advances in Sig. Proc.*, vol. 2015, no. 1 (2015). http://dx.doi.org/10.1186/s13634-015-0256-4

[6] P. A. Naylor and N. D. Gaubitch, Eds., *Speech Dereverberation* (London, UK: Springer, 2010). http://dx.doi.org/10.1007/978-1-84996-056-4

[7] E. A. P. Habets, S. Gannot, and I. Cohen, "Late Reverberant Spectral Variance Estimation Based on a Statistical Model," *IEEE Signal Processing Letters*, vol. 16, no. 9, pp. 770–774 (2009 Sep.). http://dx.doi.org/10.1109/LSP.2009.2024791

[8] A. Kuklasiński, S. Doclo, T. Gerkmann, S. H. Jensen, and J. Jensen, "Multichannel PSD Estimators for Speech Dereverberation—A Theoretical and Experimental Comparison," *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing*, Brisbane, Australia (Apr. 2015), pp. 91–95. http://dx.doi.org/10.1109/ICASSP.2015.7177938

[9] T. Nakatani, T. Yoshioka, K. Kinoshita, M. Miyoshi, and B.-H. Juang, "Speech Dereverberation Based on Variance-Normalized Delayed Linear Prediction," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, no. 7, pp. 1717–1731 (2010 Sep.). http://dx.doi.org/10.1109/TASL.2010.2052251

[10] A. Jukić, T. Van Waterschoot, T. Gerkmann, and S. Doclo, "Multichannel Linear Prediction-Based Speech Dereverberation with Sparse Priors," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 23, no. 9, pp. 1509–1520 (2015 Sep.). http://dx.doi.org/10.1109/TASLP.2015.2438549

[11] M. Miyoshi and Y. Kaneda, "Inverse Filtering of Room Acoustics," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 36, no. 2, pp. 145–152 (1988 Feb.). http://dx.doi.org/10.1109/29.1509

[12] M. Kallinger and A. Mertins, "Multichannel Room Impulse Response Shaping—A Study," *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing*, Toulouse, France (May 2006), pp. 101–104. http://dx.doi.org/10.1109/ICASSP.2006.1661222

[13] F. Lim, W. Zhang, E. A. P. Habets, and P. A. Naylor, "Robust Multichannel Dereverberation Using Relaxed Multichannel Least Squares," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 22, no. 9, pp. 1379–1390 (2014 Jun.). http://dx.doi.org/10.1109/TASLP.2014.2329632

[14] I. Kodrasi, S. Goetze, and S. Doclo, "Regularization for Partial Multichannel Equalization for Speech Dereverberation," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 21, no. 9, pp. 1879–1890 (2013 Sep.). http://dx.doi.org/10.1109/TASL.2013.2260743

[15] G. W. Elko, E. Diethorn, and T. Gänsler, "Room Impulse Response Variation Due to Thermal Fluctuation and its Impact on Acoustic Echo Cancellation," *Proc. International Workshop on Acoustic Echo and Noise Control*, Kyoto, Japan (Sep. 2003), pp. 67–70.

[16] D. Cole, M. Moody, and S. Sridharan, "Position-Independent Enhancement of Reverberant Speech," *J. Audio Eng. Soc.*, vol. 45, pp. 142–147 (1997 Mar.).

[17] B. D. Radlovic, R. C. Williamson, and R. A. Kennedy, "Equalization in an Acoustic Reverberant Environment: Robustness Results," *IEEE Transactions on Speech and Audio Processing*, vol. 8, no. 3, pp. 311–319 (2000 May). http://dx.doi.org/10.1109/89.841213

[18] M. A. Haque and T. Hasan, "Noise Robust Multichannel Frequency-Domain LMS Algorithms for Blind Channel Identification," *IEEE Signal Processing Letters*, vol. 15, pp. 305–308 (2008 Feb.). http://dx.doi.org/10.1109/LSP.2008.917803

[19] X. Lin, A. W. H. Khong, and P. A. Naylor, "A Forced Spectral Diversity Algorithm for Speech Dereverberation in the Presence of Near-Common Zeros," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 3, pp. 888–899 (2012 Mar.). http://dx.doi.org/10.1109/TASL.2011.2168208

[20] F. Lim and P. Naylor, "Statistical Modelling of Multichannel Blind System Identification Errors," *Proc. International Workshop on Acoustic Echo and Noise Control*, Antibes, France (Sep. 2014), pp. 119–123. http://dx.doi.org/10.1109/IWAENC.2014.6953350

[21] S. Goetze, "*On the Combination of Systems for Listening-Room Compensation and Acoustic Echo Cancellation in Hands-Free Telecommunication Systems*," Ph.D. dissertation, Dept. of Telecommunications, University of Bremen, Bremen, Germany (2013).

[22] H. Hacihabibouglu and Z. Cvetkovic, "Multichannel Dereverberation Theorems and Robustness Issues," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 2, pp. 67–689 (2012 Feb.). http://dx.doi.org/10.1109/TASL.2011.2164528

[23] I. Arweiler and J. M. Buchholz, "The Influence of Spectral Characteristics of Early Reflections on Speech Intelligibility," *J. Acous. Soc. Am.*, vol. 130, no. 2, pp. 996–1005 (2011 Aug.). http://dx.doi.org/10.1121/1.3609258

[24] I. Kodrasi and S. Doclo, "The Effect of Inverse Filter Length on the Robustness of Acoustic Multichannel Equalization," *Proc. European Signal Processing Conference*, Bucharest, Romania (Aug. 2012).

[25] I. Kodrasi, A. Jukić, and S. Doclo, "Robust Sparsity-Promoting Acoustic Multichannel Equalization for Speech Dereverberation," *Proc. IEEE International*

Conference on Acoustics, Speech, and Signal Processing, Shanghai, China (Mar. 2016), pp. 166–170. http://dx.doi.org/10.1109/ICASSP.2016.7471658

[26] I. Kodrasi and S. Doclo, "Signal-Dependent Penalty Functions for Robust Acoustic Multichannel Equalization," *IEEE Transactions on Audio, Speech, and Language Processing*, manuscript under review (2016 Mar.).

[27] P. Wedin, "Perturbation Theory for Pseudo-Inverses," *BIT Numerical Mathematics*, vol. 13, no. 2, pp. 217–232 (1973 Jun.). http://dx.doi.org/10.1007/BF01933494

[28] S. P. Boyd, N. Parikh, E. Chu, B. Peleato, and J. Eckstein, "Distributed Optimization and Statistical Learning via the Alternating Direction Method of Multipliers," *Foundations and Trends in Machine Learning*, vol. 3, no. 1, pp. 1–122 (2011 Jan.). http://dx.doi.org/10.1561/2200000016

[29] N. Parikh and S. P. Boyd, "Proximal Algorithms," *Foundations and Trends in Optimization*, vol. 1, no. 3, pp. 127–239 (2014 Jan.). http://dx.doi.org/10.1561/2400000003

[30] A. Farina, "Simultaneous Measurement of Impulse Response and Distortion with a Swept-Sine Technique," presented at the *108th Convention of the Audio Engineering Society* (2000 Feb.), convention paper 5093.

[31] M. Nilsson, S. D. Soli, and A. Sullivan, "Development of the Hearing in Noise Test for the Measurement of Speech Reception Thresholds in Quiet and in Noise," *J. Acous. Soc. Am.*, vol. 95, no. 2, pp. 1085–1099 (1994 Feb.). http://dx.doi.org/10.1121/1.408469

[32] W. Zhang and P. A. Naylor, "An Algorithm to Generate Representations of System Identification Errors," *Research Letters in Signal Processing*, vol. 2008 (Jan. 2008). http://dx.doi.org/10.1155/2008/529291

[33] ITU-T, "*Perceptual Evaluation of Speech Quality (PESQ), an Objective Method for End-to-End Speech Quality Assessment of Narrowband Telephone Networks and Speech Codecs P.862*," International Telecommunications Union (ITU-T) Recommendation (Feb. 2001).

[34] J. F. Santos, M. Senoussaoui, and T. H. Falk, "An Improved Non-Intrusive Intelligibility Metric for Noisy and Reverberant Speech," *Proc. International Workshop on Acoustic Echo and Noise Control*, Antibes, France (Sep. 2014), pp. 55–59. http://dx.doi.org/10.1109/IWAENC.2014.6953337

[35] Y. Hu and P. C. Loizou, "Evaluation of Objective Quality Measures for Speech Enhancement," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 16, no. 1, pp. 229–238 (2008 Jan.). http://dx.doi.org/10.1109/TASL.2007.911054

[36] B. Cauchi, H. Javed, T. Gerkmann, S. Doclo, S. Goetze, and P. Naylor, "Perceptual and Instrumental Evaluation of the Perceived Level of Reverberation," *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing*, Shanghai, China (Mar. 2016), pp. 629–633. http://dx.doi.org/10.1109/ICASSP.2016.7471751

[37] K. Kinoshita, M. Delcroix, S. Gannot, E. A. P. Habets, R. Haeb-Umbach, W. Kellermann, V. Leutnant, R. Maas, T. Nakatani, B. Raj, A. Sehr, and T. Yoshioka, "A Summary of the REVERB Challenge: State-of-the-Art and Remaining Challenges in Reverberant Speech Processing Research," *EURASIA J. Advan. Sig. Proc.*, vol. 2016, no. 7 (2016). http://dx.doi.org/10.1186/s13634-016-0306-6

[38] ITU-T, "*Method for the Subjective Assessment of Intermediate Quality Levels of Coding Systems*," International Telecommunications Union (ITU-T) Recommendation (Jan. 2003).

[39] B. Cauchi, I. Kodrasi, R. Rehr, S. Gerlach, A. Jukić, T. Gerkmann, S. Doclo, and S. Goetze, "Combination of MVDR Beamforming and Single-Channel Spectral Processing for Enhancing Noisy and Reverberant Speech," *EURASIA J. Advan. Sig. Proc.*, vol. 2015, no. 1 (2015). http://dx.doi.org/10.1186/s13634-015-0242-x

[40] M. Friedman, "The Use of Ranks to Avoid the Assumption of Normality Implicit in the Analysis of Variance," *J. Amer. Statistical Assn.*, vol. 32, no. 200, pp. 675–701 (1937 Dec.). http://dx.doi.org/10.1080/01621459.1937.10503522

[41] J. D. Gibbons and S. Chakrabort, *Nonparametric Statistical Inference* (New York, USA: Marcel Dekker, Inc., 2003).

THE AUTHORS


Ina Kodrasi


Benjamin Cauchi


Stefan Goetze


Simon Doclo

Ina Kodrasi received the Master of Science degree in communications, systems, and electronics in 2010 from Jacobs University Bremen, Germany, and her Ph.D. degree in 2015 from the University of Oldenburg, Germany. From 2010 to 2015 she worked as an early stage researcher at the Signal Processing Group of the University of Oldenburg. From 2010 to 2011 she was also with the Fraunhofer Institute for Digital Media Technology, project group Hearing, Speech and Audio Technology in Oldenburg where she worked on microphone array beamforming. Since December 2015 she is a postdoctoral researcher at the Signal Processing Group of the University of Oldenburg in the field of speech dereverberation and noise reduction.

•

Benjamin Cauchi received the Bachelor in Musicology from the University of Rouen, France. He studied in the Music School of Cardiff University, UK, and received the M.Sc. in acoustic and signal processing applied to music from the University Pierre and Marie Curie, Paris VI, France. His Master thesis, conducted in the IRCAM, in Paris, focused on auditory scene classification. He worked as a research assistant in Fraunhofer IDMT, Oldenburg, Germany, and Imperial College London, UK, before starting his Ph.D. as an ESR Marie Curie fellow in the project group Hearing, Speech, and Audio Technology of the Fraunhofer Institute for Digital Media Technology in Oldenburg, Germany, and the Signal Processing Group of the University of Oldenburg. His research work focuses on the design of speech enhancement algorithms for noisy and reverberant environments and on the prediction of the quality of processed speech.

•

Stefan Goetze is head of Audio System Technology for Assistive Systems at the Fraunhofer Institute for Digital Media Technology IDMT, project group Hearing, Speech and Audio (HSA) in Oldenburg, Germany. He received his Dipl.-Ing. in 2004 and Dr.-Ing. in 2013 at the University of Bremen, Germany, where he worked as a research engineer from 2004 to 2008. His research interests are sound pick/up, processing and enhancement, such as noise reduction, acoustic echo cancellation and dereverberation, as well as assistive technologies, human-machine-interaction, detection and classification of acoustic events and automatic speech recognition. Stefan Goetze is lecturer at the University of Bremen since 2007 and is leading and partially coordinating national as well as international research projects in the field of acoustics for ambient assisted living (AAL) technologies. He is a member of IEEE and an associate member of the AES.

•

Simon Doclo received the M.Sc. degree in electrical engineering and the Ph.D. degree in applied sciences from the Katholieke Universiteit Leuven, Belgium, in 1997 and 2003. From 2003 to 2007 he was a Postdoctoral Fellow with the Research Foundation - Flanders at the Electrical Engineering Department (Katholieke Universiteit Leuven) and the Adaptive Systems Laboratory (McMaster University, Canada). From 2007 to 2009 he was a Principal Scientist with NXP Semiconductors at the Sound and Acoustics Group in Leuven, Belgium. Since 2009 he is a full professor at the University of Oldenburg, Germany, and scientific advisor for the project group Hearing, Speech and Audio Technology of the Fraunhofer Institute for Digital Media Technology. His research activities center around signal processing for acoustical and biomedical applications, more specifically microphone array processing, active noise control, acoustic sensor networks, and hearing aid processing. Prof. Doclo received the Master Thesis Award of the Royal Flemish Society of Engineers in 1997 (with Erik De Clippel), the Best Student Paper Award at the International Workshop on Acoustic Echo and Noise Control in 2001, the EURASIP Signal Processing Best Paper Award in 2003 (with Marc Moonen), and the IEEE Signal Processing Society 2008 Best Paper Award (with Jingdong Chen, Jacob Benesty, Arden Huang). He was member of the IEEE Signal Processing Society Technical Committee on Audio and Acoustic Signal Processing (2008–2013) and Technical Program Chair for the IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA) in 2013. Prof. Doclo has served as guest editor for several special issues (*IEEE Signal Processing Magazine*, Elsevier Signal Processing) and is associate editor for *IEEE/ACM Transactions on Audio, Speech and Language Processing* and *EURASIP Journal on Advances in Signal Processing*.