Multi-Channel Linear Prediction-Based Speech Dereverberation With Sparse Priors

Ante Jukić, *Student Member, IEEE*, Toon van Waterschoot, *Member, IEEE*, Timo Gerkmann, *Senior Member, IEEE*, and Simon Doclo, *Senior Member, IEEE*

***Abstract*—The quality of speech signals recorded in an enclosure can be severely degraded by room reverberation. In this paper, we focus on a class of blind batch methods for speech dereverberation in a noiseless scenario with a single source, which are based on multi-channel linear prediction in the short-time Fourier transform domain. Dereverberation is performed by maximum-likelihood estimation of the model parameters that are subsequently used to recover the desired speech signal. Contrary to the conventional method, we propose to model the desired speech signal using a general sparse prior that can be represented in a convex form as a maximization over scaled complex Gaussian distributions. The proposed model can be interpreted as a generalization of the commonly used time-varying Gaussian model. Furthermore, we reformulate both the conventional and the proposed method as an optimization problem with an ℓ_p-norm cost function, emphasizing the role of sparsity in the considered speech dereverberation methods. Experimental evaluation in different acoustic scenarios show that the proposed approach results in an improved performance compared to the conventional approach in terms of instrumental measures for speech quality.**

***Index Terms*—Multi-channel linear prediction, sparse priors, speech dereverberation, speech enhancement.**

I. Introduction

CAPTURING a speech signal within an enclosed space with microphones placed at a distance from the speech source typically results in recordings corrupted by reverberation, caused by acoustic reflections against the walls and other surfaces within the enclosure. While moderate levels of reverberation can be beneficial, in most cases it results in a decreased speech intelligibility and automatic speech recognition performance [1]–[4]. Hence, effective solutions for dereverberation are required to improve speech intelligibility, perceptual speech quality, and the performance of automatic speech recognition systems in several speech communication applications, such as teleconferencing, hands-free telephony, voice-controlled systems and hearing aids [3]–[5].

In the last decades, several single- and multi-microphone dereverberation approaches have been proposed, which can be broadly classified into acoustic channel equalization, spectral enhancement and probabilistic model-based approaches [6]. Acoustic channel equalization techniques aim to reshape the estimated room impulse responses (RIRs) between the speaker and the microphone array [7]. Although in theory perfect dereverberation can be achieved using multi-channel equalization, in practice the performance may be severely limited by the poor estimation accuracy of the RIRs, requiring robust equalization techniques [8]–[11]. Other speech dereverberation approaches are based on spectral enhancement [12]–[14], where the clean speech spectral coefficients are estimated by applying a (real-valued) gain to the reverberant spectral coefficients. The gain function requires an estimate of the late reverberant spectral variance [15], which is typically based on a statistical room acoustics model. In addition, several probabilistic model-based speech dereverberation approaches have been recently proposed [16]–[21]. Dereverberation is performed by estimating all unknown model parameters, e.g., in a maximum likelihood sense, where either an autoregressive or a convolutive (moving average) transfer function model for the acoustic transfer functions is assumed and the clean speech spectral coefficients are typically modeled using a Gaussian distribution with a time-varying variance.

For a noiseless scenario with a single speech source a blind batch, i.e., utterance-based, speech dereverberation method based on variance-normalized delayed multi-channel linear prediction (MCLP) has been proposed in [16], [17]. Its efficient time-frequency-domain implementation is often referred to as the weighted prediction error (WPE) method [16], [17], [22]. This method assumes an autoregressive model of the reverberation process, i.e., it is assumed that the reverberant component at a certain time can be predicted from the previous samples of the reverberant microphone signals. The desired speech signal can then be estimated as the prediction error, i.e., speech dereverberation boils down to estimation of the parameters of the MCLP model. An additional delay is introduced in the MCLP model in order to prevent distortion of the short-time correlation of the speech signal, thereby only suppressing late

Manuscript received November 27, 2014; revised April 02, 2015; accepted May 09, 2015. Date of publication June 01, 2015; date of current version June 04, 2015. This work was supported in part by the Marie Curie Initial Training Network DREAMS under Grant ITN-GA-2012-316969, and in part by the Research Foundation Flanders (FWO-Vlaanderen) and the Cluster of Excellence 1077 “Hearing4All,” funded by the German Research Foundation (DFG). The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Yunxin Zhao.

A. Jukić, T. Gerkmann, and S. Doclo are with the Department of Medical Physics and Acoustics, University of Oldenburg, 26111 Oldenburg, Germany (e-mail: ante.jukic@uni-oldenburg.de; timo.gerkmann@unioldenburg.de; simon.doclo@uni-oldenburg.de).

T. van Waterschoot is with the Department of Electrical Engineering (ESAT), Stadius Center for Dynamical Systems, Signal Processing, and Data Analytics (STADIUS), KU Leuven, 3000 Leuven, Belgium (e-mail: toon.vanwaterschoot@esat.kuleuven.be).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TASLP.2015.2438549

reverberation [17], [23]. Conventionally, the complex-valued short-time Fourier transform (STFT) domain coefficients of the desired speech signal are modeled using a time-varying Gaussian (TVG) model, under the assumption that the STFT coefficients can be modeled locally (i.e., in each time-frequency bin) using a complex Gaussian distribution with an unknown variance. Speech dereverberation using WPE is then performed by estimating the unknown parameters of the MCLP and TVG models in a maximum-likelihood (ML) sense.

In this paper, we aim to provide a different view on MCLP-based speech dereverberation in the STFT domain. Firstly, we present a general sparse prior for the desired speech signal and use ML estimation to estimate the parameters of the MCLP model [24]. The sparse prior is formulated using a convex representation that is based on a locally Gaussian model [25]–[27]. The obtained model for the desired speech signal can be interpreted as a TVG model with an additional hyperprior on the unknown variance. To derive a practical algorithm, we focus on sparse priors in the family of complex generalized Gaussian (CGG) distributions [28], resulting in the WPE-CGG method for speech dereverberation. In the presented framework, we show that the conventional WPE method can be considered as a special case which is based on a prior that strongly promotes sparsity of the estimated speech signal. Secondly, we reformulate the WPE-CGG method as an optimization problem with a cost function given as the ℓ_p-norm of the desired speech signal. Furthermore, we show that the WPE-CGG method is equivalent to an iteratively reweighted least-squares procedure applied to ℓ_p-norm minimization [29]. From this perspective, the conventional WPE method corresponds to the case $p = 0$. In the experimental section we evaluate the performance of the conventional and the proposed methods for different acoustic scenarios using several instrumental speech quality measures. The obtained results show that the speech enhancement performance can be consistently improved. While the improvements are mild, these come with no additional computational cost, and are consistent with the derived theoretical insights.

The paper is organized as follows. In Section II the problem of speech dereverberation using MCLP in the STFT domain is formulated. The conventional method for MCLP-based speech dereverberation, based on a TVG model for the speech signal, is presented in Section III. Our proposed method using a general sparse prior for the desired speech signal is presented in Section IV. In Section V both the conventional and the proposed methods are reformulated as a minimization of the ℓ_p-norm of the desired speech signal. Simulation results are presented in Section VI.

II. Problem Formulation

We consider an acoustic scenario where a single static speech source in an enclosure is captured by M microphones. Let $s[t]$ denote the clean speech signal in the time domain, with t denoting the discrete-time index. The noiseless reverberant speech signal observed at the m-th microphone, $m \in \{1, \ldots, M\}$, can be modeled in the time domain as

$$x_m[t] = \sum_{l=0}^{L_r-1} r_m[l] s[t-l], \qquad (1)$$

where $r_m[l]$ denotes the RIR between the source and the m-th microphone with length L_r. The RIRs in the time-domain model in (1) are typically very long, and dereverberation is often performed in the STFT domain [16], [19], [23].

The time-doman model in (1) can be approximated in the STFT domain using the convolutive transfer function approximation [30]–[32]. Let $s(k,n)$ denote the clean speech signal in the STFT domain with time frame index $n \in \{1, \ldots, N\}$ and frequency bin index $k \in \{1, \ldots, K\}$, with N and K denoting the number of time frames and frequency bins. The reverberant speech signal observed at the m-th microphone can be represented in the STFT domain using a convolutive (moving average) transfer function model as

$$x_m(k,n) = \sum_{l=0}^{L_h-1} h_m(k,l) s(k,n-l) + e_m(k,n), \qquad (2)$$

where $h_m(k,l)$ models the acoustic transfer function (ATF) between the speech source and the m-th microphone in frequency bin k with length L_h time frames, and the additive term $e_m(k,n)$ represents the modeling error at the m-th microphone. The model in (2) is practically interesting because the time-domain convolution is divided into a set of convolutions in the time-frequency domain, and has been used in various applications [15], [16], [20], [31], [32]. This model can significantly reduce the computational complexity due to shorter ATFs and the possibility of independent processing in each frequency bin. Additionally, certain statistical properties of the speech signal can be more naturally exploited in the time-frequency domain. For example, while speech signals are not necessarily sparse in the time domain, they are typically sparse in the time-frequency domain, a fact that has been exploited for dereverberation [33], [34]. Blind dereverberation using the model in (2) can be formulated as a joint blind estimation of the ATFs and the STFT coefficients of the speech signal [20].

To avoid joint estimation of the ATFs and the STFT coefficients of the speech signal, further simplifications have been used in the literature. As in [16], [17], by disregarding the noise and assuming $e_m(k,n) = 0$, the convolutive model in (2) can be simplified, and the signal at the arbitrarily chosen reference microphone (e.g., $m = 1$) can be written in the MCLP form as

$$x_1(k,n) = d(k,n) + \sum_{m=1}^{M} \sum_{l=0}^{L_g-1} g_m(k,l) x_m(k,n-\tau-l), \qquad (3)$$

where L_g is the number of the prediction coefficients $g_m(k,l)$ for each channel, and τ is the prediction delay. The first term in (3) represents the desired speech signal at the reference microphone

$$d(k,n) = d_1(k,n) = \sum_{l=0}^{\tau-1} h_1(k,l) s(k,n-l), \qquad (4)$$

which consists of the direct speech signal and early reflections determined by the prediction delay τ[17]. The second term in (3) models the late reverberation, which is predicted using the prediction coefficients and the delayed past observations on all M microphones. The MCLP model in (3) can be written as

$$\mathbf{x}_1(k) = \mathbf{d}(k) + \sum_{m=1}^{M} \mathbf{X}_{m,\tau}(k) \mathbf{g}_m(k), \qquad (5)$$

with

$$\mathbf{x}_m(k) = [x_m(k,1),\ldots,x_m(k,N)]^T \in \mathbb{C}^N,$$
$$\mathbf{d}(k) = [d(k,1),\ldots,d(k,N)]^T \in \mathbb{C}^N,$$
$$\mathbf{g}_m(k) = [g_m(k,0),\ldots,g_m(k,L_g-1)]^T \in \mathbb{C}^{L_g},$$

and $\mathbf{X}_{m,\tau}(k) \in \mathbb{C}^{N\times L_g}$ denoting a convolution matrix constructed using $\mathbf{x}_m(k)$ delayed for τ frames. Furthermore, the matrices $\mathbf{X}_{m,\tau}(k)$ and vectors $\mathbf{g}_m(k)$ can be stacked as

$$\mathbf{X}_\tau(k) = [\mathbf{X}_{1,\tau}(k),\ldots,\mathbf{X}_{M,\tau}(k)], \tag{6}$$
$$\mathbf{g}(k) = \left[\mathbf{g}_1^T(k),\ldots,\mathbf{g}_M^T(k)\right]^T, \tag{7}$$

to form a multi-channel convolution matrix $\mathbf{X}_\tau(k)$ and a multi-channel prediction vector $\mathbf{g}(k)$. The MCLP model can now be written more compactly as

$$\mathbf{x}_1(k) = \mathbf{d}(k) + \mathbf{X}_\tau(k)\mathbf{g}(k). \tag{8}$$

From the MCLP model in (8), it follows that the problem of speech dereverberation can be formulated as a blind estimation of the desired speech signal $\mathbf{d}(k)$ from the reverberant observations $\mathbf{x}_m(k), \forall m,k$. Using (8), the desired speech signal can be estimated as

$$\boxed{\hat{\mathbf{d}}(k) = \mathbf{x}_1(k) - \mathbf{X}_\tau(k)\hat{\mathbf{g}}(k),} \tag{9}$$

with $\hat{(.)}$ denoting an estimated value. The desired speech signal can be interpreted as the prediction error in the delayed linear prediction model [17]. Therefore, dereverberation can be performed by calculating the multi-channel prediction vector estimate $\hat{\mathbf{g}}(k)$ for each frequency bin k and applying (9).

Note that in the following we will work in each frequency bin independently, so the index k will be omitted where possible for notational convenience.

III. Conventional MCLP-Based Dereverberation Using TVG Model

Several MCLP-based speech dereverberation methods have been proposed using a TVG model for the desired signal [16], [17], [19], [20], [22]. More specifically, the desired signal $d(k,n)$ in each time-frequency bin is modeled as a zero-mean random variable by means of a circular complex Gaussian distribution with an unknown and time-varying variance. The probability density function for the desired signal can then be written as

$$\mathcal{N}_\mathbb{C}(d(k,n);0,\lambda(k,n)) = \frac{1}{\pi\lambda(k,n)}e^{-\frac{|d(k,n)|^2}{\lambda(k,n)}}, \tag{10}$$

where the variance $\lambda(k,n)$ is considered to be an unknown parameter that needs to be estimated. The TVG model was introduced by arguing that it can model any signal with a time-varying power spectrum [17], [22]. Since the TVG model does not include any dependency across frequencies and it is assumed that the STFT coefficients are independent across time, the likelihood function for the complete time range at a single frequency bin, with the index k omitted, can be written as

$$\mathcal{L}(\mathbf{g},\boldsymbol{\lambda}) = \prod_{n=1}^{N}\mathcal{N}_\mathbb{C}(d(n);0,\lambda(n)), \tag{11}$$

with unknown variances $\boldsymbol{\lambda} = [\lambda(1),\ldots,\lambda(N)]^T$ and the prediction vector $\mathbf{g}$ [17]. Note that the desired signal $d(n)$ in (11) depends on the prediction vector $\mathbf{g}$ as in (9). The assumption that the coefficients of the desired speech signal are independent across time is a simplification that has been successfully employed in dereverberation [16], [17], [20], but also in other speech enhancements methods [35]. The prediction vector $\mathbf{g}$ and the variances $\boldsymbol{\lambda}$ are estimated by maximizing the likelihood in (11) with respect to the unknown parameters, i.e., minimizing the negative log-likelihood by solving the following optimization problem

$$\boxed{\min_{\boldsymbol{\lambda}>0,\mathbf{g}} \sum_{n=1}^{N}\left(\frac{|d(n)|^2}{\lambda(n)} + \log\pi\lambda(n)\right).} \tag{12}$$

Since the joint minimization of (12) with respect to the prediction vector $\mathbf{g}$ and the variances $\boldsymbol{\lambda}$ can not be performed analytically, it was proposed in [17] to use an alternating optimization procedure. The original problem in (12) is split into two subproblems that can be solved more easily. The two subproblems are solved in an alternating fashion, and the whole procedure is repeated iteratively. While this results in simple update rules, there is no guarantee that the alternating procedure will lead to the globally optimal solution (cf. Section V).

Estimation of $\mathbf{g}$: In the first step, the cost function in (12) is minimized with respect to the prediction vector $\mathbf{g}$. Assuming that the variances $\boldsymbol{\lambda}$ are fixed (to the values from the i-th iteration[1]) a least-squares (LS) problem is obtained for estimating the prediction vector

$$\hat{\mathbf{g}}^{(i+1)} = \arg\min_{\mathbf{g}} \sum_{n=1}^{N}\frac{|d(n)|^2}{\hat{\lambda}^{(i)}(n)} = \arg\min_{\mathbf{g}} \mathbf{d}^H\mathcal{D}^{-1}_{\hat{\boldsymbol{\lambda}}^{(i)}}\mathbf{d}, \tag{13}$$

where $\mathcal{D}_{\hat{\boldsymbol{\lambda}}^{(i)}} = diag(\hat{\boldsymbol{\lambda}}^{(i)})$. By combining (8) and (13), the optimal prediction vector $\hat{\mathbf{g}}^{(i+1)}$ can be computed as

$$\boxed{\hat{\mathbf{g}}^{(i+1)} = \left(\mathbf{X}_\tau^H\mathcal{D}^{-1}_{\hat{\boldsymbol{\lambda}}^{(i)}}\mathbf{X}_\tau\right)^{-1}\mathbf{X}_\tau^H\mathcal{D}^{-1}_{\hat{\boldsymbol{\lambda}}^{(i)}}\mathbf{x}_1.} \tag{14}$$

Estimation of λ: In the second step, the cost function in (12) is minimized with respect to the variances in $\boldsymbol{\lambda}$, assuming now that the prediction vector is fixed to $\hat{\mathbf{g}}^{(i+1)}$. The estimate $\hat{\mathbf{d}}^{(i+1)}$ can be calculated using (9) and the optimal variance is obtained as

$$\hat{\lambda}^{(i+1)}(n) = \arg\min_{\lambda(n)>0}\frac{|\hat{d}^{(i+1)}(n)|^2}{\lambda(n)} + \log\pi\lambda(n). \tag{15}$$

The solution to this optimization problem is given as $\hat{\lambda}^{(i+1)}(n) = |\hat{d}^{(i+1)}(n)|^2$, or in short as

$$\boxed{\hat{\boldsymbol{\lambda}}^{(i+1)} = |\hat{\mathbf{d}}^{(i+1)}|^2,} \tag{16}$$

where the absolute value and the power are applied elementwise. In practice, to prevent division by zero a small positive constant $\varepsilon_{\min}$ is included as a lower bound for the estimated variance as

$$\hat{\boldsymbol{\lambda}}^{(i+1)} = \max\{|\hat{\mathbf{d}}^{(i+1)}|^2, \varepsilon_{\min}\}. \tag{17}$$

[1] In the following $(.)^{(i)}$ denotes the value of a variable at the i-th iteration.

This alternating procedure is repeated until a convergence criterion is satisfied or a maximum number of iterations is exceeded. The method is typically initialized by setting the variances as

$$\hat{\boldsymbol{\lambda}}^{(0)} = |\mathbf{x}_1|^2, \tag{18}$$

that is equivalent to setting the initial estimate of the desired speech signal as $\hat{\mathbf{d}}^{(0)} = \mathbf{x}_1$. The presented method is often referred to as the weighted prediction error (WPE) [16], [17]. The WPE method has been modified to include pre-trained log-spectral priors in [22], and a time-varying Laplacian model for the desired speech signal has been used in [36]. Recently, several methods based on auto-regressive modeling have been proposed, aiming to address noisy [37], [38] and time-varying acoustic scenarios [19] with multiple sources [18], [19], [39].

IV. MCLP-Based Dereverberation Using a General Sparse Prior

It is widely accepted that the STFT coefficients of speech signals can be well modeled using sparse priors. This holds both locally, by observing the STFT coefficients in a single time-frequency bin [40]–[42], as well as globally, when observing the distribution of the STFT coefficients in a single frequency bin [43]. Although the real and imaginary parts of the complex-valued STFT coefficients are often assumed to be independent to simplify computations, it has been observed that the distribution of the complex-valued speech coefficients is actually approximately circular [44], [45]. In this section we model the desired speech coefficients in a single frequency bin using a sparse circular prior, and combine it with the MCLP model in (5). The proposed prior can be interpreted as a generalization of the TVG model (cf. Section III), obtained by adding a hyperprior for the variance. A similar approach can be used with other local models (e.g., the locally Laplacian model in [36]). In Section IV-A we present a convex representation of a sparse prior, and use it for MCLP-based dereverberation in Section IV-B. In Section IV-C we formulate dereverberation using a complex generalized Gaussian distribution, and relate the proposed method to the conventional method based on TVG model in Section IV-D.

A. Convex Representation of a Sparse Prior

Intuitively, a prior is considered to be sparse when it is super-Gaussian, i.e., it exhibits a higher peak at the origin and heavier tails than the corresponding Gaussian prior. Here we consider a general circular sparse prior for a complex-valued random variable z that can be represented as

$$\rho(z) = e^{-f(|z|)}. \tag{19}$$

In general, $\rho(.)$ can represent a proper sparse prior (e.g., a probability density), or an *improper* (non-integrable) sparse prior. Formally, it can be shown that when $f'(t)/t$ is decreasing on $t \in (0, \infty)$, with $f'(.)$ denoting the derivative of $f(.)$, the prior will be super-Gaussian, i.e., sparse [25]. In this case, $\rho(z)$ can be conveniently represented as a maximization over scaled Gaussians with different variances, i.e.,

$$\rho(z) = \max_{\lambda > 0} \mathcal{N}_{\mathbb{C}}(z; 0, \lambda)\psi(\lambda), \tag{20}$$

where $\psi(.)$ is a scaling function that can be interpreted as a hyperprior on the variance λ[25], [27]. This representation of a sparse prior is often referred to as the convex type due to its roots in convex analysis [25]. Obviously, the scaling function $\psi(.)$ in (20) is related to $f(.)$ in (19), but the scaling function is typically not required explicitly in practical algorithms [25]. For completeness, the form of the hyperprior $\psi(.)$ for a given sparse prior $\rho(.)$ is given in Appendix A.

B. Speech Dereverberation Using a General Sparse Prior

We now propose to model the STFT coefficients of the desired speech signal using the circular sparse prior $\rho(d(n)) = e^{-f(|d(n)|)}$ with its convex representation given as

$$\rho(d(n)) = \max_{\lambda(n) > 0} \mathcal{N}_{\mathbb{C}}(d(n); 0, \lambda(n))\psi(\lambda(n)). \tag{21}$$

This can be interpreted as a generalization of the TVG model, with an additional hyperprior on the variance $\lambda(n)$ determined by the scaling function $\psi(.)$. Similarly as in the conventional method, the prediction vector $\mathbf{g}$ can be estimated by maximizing the likelihood formed using (21) as

$$\max_{\mathbf{g}} \prod_{n=1}^{N} \max_{\lambda(n) > 0} \mathcal{N}_{\mathbb{C}}(d(n); 0, \lambda(n))\psi(\lambda(n)). \tag{22}$$

This is equivalent to minimizing negative log-likelihood with respect to the prediction vector $\mathbf{g}$ and the variances $\boldsymbol{\lambda}$, i.e.,

$$\boxed{\min_{\boldsymbol{\lambda} > 0, \mathbf{g}} \sum_{n=1}^{N} \left(\frac{|d(n)|^2}{\lambda(n)} + \log \pi\lambda(n) - \log \psi(\lambda(n)) \right),} \tag{23}$$

with $d(n)$ depending on $\mathbf{g}$ through (9). By comparing (23) with the optimization problem in (12), the obtained problem contains an additional term that depends on the scaling function $\psi(.)$. The likelihood can again be maximized by applying an alternating optimization procedure.

Estimation of $\mathbf{g}$: Assuming that the variances $\boldsymbol{\lambda}$ are fixed, the same LS problem is obtained as in the conventional method, with the solution given by (14).

Estimation of $\boldsymbol{\lambda}$: Assuming that the prediction vector is fixed to $\hat{\mathbf{g}}^{(i+1)}$, the variances can be obtained by solving the following problem

$$\hat{\lambda}^{(i+1)}(n) = \operatorname*{arg\,min}_{\lambda(n) > 0} \frac{|\hat{d}^{(i+1)}(n)|^2}{\lambda(n)} + \log \pi\lambda(n) - \log \psi(\lambda(n)). \tag{24}$$

For a general sparse prior in (19), the solution is equal to (for details we refer to Appendix B)

$$\hat{\lambda}^{(i+1)}(n) = \frac{2|\hat{d}^{(i+1)}(n)|}{f'\left(|\hat{d}^{(i+1)}(n)|\right)}, \tag{25}$$

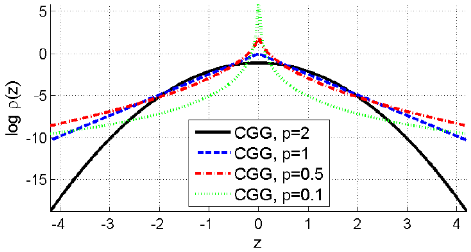

Fig. 1. Logarithm of the CGG prior $\rho(.)$ in (26) for different values of the shape parameter p and variance fixed to 1. Note that the plot shows only values on the real axis (i.e., imaginary part of z is 0), and the prior is circular.

Note that although the optimization problem in (24) includes the scaling function $\psi(.)$, the optimal $\lambda(n)$ for this subproblem depends only on $f(.)$, so the scaling function $\psi(.)$ does not need to be given explicitly (cf. Appendix B).

C. Complex Generalized Gaussian Prior

As an example of a parametric circular zero-mean super-Gaussian prior, in the remainder of the paper we will consider the complex generalized Gaussian (CGG) prior given as [28]

$$\rho(z) = \frac{p}{2\pi\gamma\Gamma(2/p)} e^{-\frac{|z|^p}{\gamma^{p/2}}}, \tag{26}$$

with the scale parameter $\gamma > 0$, the shape parameter $0 < p \leq 2$, and $\Gamma(.)$ denoting the Gamma function. The circular Gaussian distribution is obtained by setting $p = 2$, while smaller values of the shape parameter result in more sparse priors, i.e., a higher peak at zero and heavier tails. This can also be seen from the plot of $\log \rho(z)$ in Fig. 1. Since the CGG prior can be written in the form (19) with $f(.)$ given as

$$f(t) = \frac{t^p}{\gamma^{p/2}} - \log \frac{p}{2\pi\gamma\Gamma(2/p)}, \tag{27}$$

it can be represented using a convex representation in the form (20).

In the case of a CGG prior for the desired signal, the optimal value of $\lambda(n)$ in iteration $(i+1)$ can be written using (25) and (27) as

$$\hat{\lambda}^{(i+1)}(n) = \frac{2\gamma^{p/2}}{p} |d^{(i+1)}(n)|^{2-p}. \tag{28}$$

This expression depends on the shape and scaling parameters of the CGG prior in (26). However, since the estimation of $\mathbf{g}$ using (14), and hence also the estimate of the desired speech signal $\mathbf{d}$ using (9), is invariant to a scaling of the variances $\boldsymbol{\lambda}$, the update in (28) can be simplified to

$$\boxed{\hat{\lambda}^{(i+1)}(n) = |\hat{d}^{(i+1)}(n)|^{2-p},} \tag{29}$$

which depends only on the shape parameter $p \in (0, 2)$ of the CGG prior. In practice, a small positive constant ε is included as a lower bound for the estimated variance to prevent division by zero, i.e.,

$$\hat{\boldsymbol{\lambda}}^{(i+1)} = \max\{|\hat{\mathbf{d}}^{(i+1)}|^{2-p}, \varepsilon_{\min}\}, \tag{30}$$

This method will be referred to as WPE-CGG, which is summarized in Algorithm 1.

Algorithm 1 WPE with a CGG prior.

```
parameters: Filter length $L_g$ and prediction delay $\tau$ in (3), shape parameter $p$ in (26), regularization parameter $\varepsilon_{\min}$, maximum number of iterations $i_{\max}$, tolerance $\eta$
input: $\mathbf{x}_m(k), \forall m, k$
for all$k$ do
    $i \leftarrow 0$
    $\hat{\boldsymbol{\lambda}}^{(0)}(n) \leftarrow |\mathbf{x}_1(n)|^{2-p}$
    repeat
    $\hat{\mathbf{g}}^{(i+1)} \leftarrow \text{calculate as in} (14)$
    $\hat{\mathbf{d}}^{(i+1)} \leftarrow \mathbf{x}_1 - \mathbf{X}_\tau \hat{\mathbf{g}}^{(i+1)}$
    $\hat{\boldsymbol{\lambda}}^{(i+1)} \leftarrow \max\{|\hat{\mathbf{d}}^{(i+1)}|^{2-p}, \varepsilon_{\min}\}$
    $i \leftarrow i + 1$
    until $\|\hat{\mathbf{d}}^{(i+1)} - \hat{\mathbf{d}}^{(i)}\|_2 / \|\hat{\mathbf{d}}^{(i)}\|_2 < \eta$ or $i > i_{\max}$
end for
```

D. Relation to the Conventional Method

It should be noted that the variance update (16) in the conventional method corresponds to setting $p = 0$ in the proposed update (29). When comparing the optimization problem in (12) with the proposed optimization problem in (23), it can be seen that the conventional method is obtained by setting the scaling function $\psi(.)$ equal to a constant value in the proposed method. Hence, for the conventional method the prior for the desired signal, as interpreted in the proposed framework with the scaling function $\psi(.)$ in (20) set to 1, is equal to

$$\rho(d(n)) = \max_{\lambda(n)>0} \frac{e^{-\frac{|d(n)|^2}{\lambda(n)}}}{\pi\lambda(n)} = \frac{e^{-1}}{\pi|d(n)|^2} \propto \frac{1}{|d(n)|^2} \tag{31}$$

since the maximum is attained when $\lambda(n) = |d(n)|^2$. The obtained prior can also be represented in the form (19) as

$$f(t) = \log t^2 + \text{const}. \tag{32}$$

Note that (31) is an *improper* prior since it is not integrable. In addition, it strongly favors values of the desired signal that are close to the origin, i.e., it is a strong sparse prior for the desired signal. This type of sparsity-promoting prior was used previously in various signal processing applications [26], [27], [29], [46]. Although the conventional WPE method was originally derived with the TVG model as the starting point, under the assumption of a locally Gaussian model, this interpretation highlights the underlying role of the sparse prior (31) on the desired speech signal. Similarly, other dereverberation methods based on the TVG model can be formulated using sparsity-promoting cost functions, e.g., [18], [19], [39].

V. Reformulation as ℓ_p-Norm Minimization

In this section we reformulate the conventional WPE and the proposed WPE-CGG methods for estimating the prediction vector $\mathbf{g}$ in terms of an ℓ_p-norm minimization problem, aiming to provide a better understanding of the cost functions underlying the proposed methods, and relating them to the problem of sparse recovery. For a general prior $\rho(.)$, and independent coefficients $d(n)$, the likelihood function is equal to

$$\mathcal{L}(\mathbf{g}) = \prod_{n=1}^{N} \rho\left(d(n)\right). \tag{33}$$

For a sparse prior $\rho(.)$ in the form (19), the ML estimate of the prediction vector $\mathbf{g}$ can hence be obtained by minimizing the negative log-likelihood, i.e.,

$$\hat{\mathbf{g}} = \arg\min_{\mathbf{g}} \sum_{n=1}^{N} f\left(|d(n)|\right). \tag{34}$$

For $\rho(.)$ being a CGG prior as in (26), this ML estimate can be obtained, using (27), as a solution of the following problem

$$\boxed{\min_{\mathbf{g}} \|\mathbf{d}\|_p^p,} \tag{35}$$

where $\|.\|_p$ is the ℓ_p-norm[2] defined as $\|\mathbf{d}\|_p = (\sum_{n=1}^{N} |d(n)|^p)^{1/p}$.

For the conventional method with the prior $\rho(.)$ given in (31), the ML estimate of the prediction vector is obtained, using (32), as

$$\min_{\mathbf{g}} \sum_{n=1}^{N} \log |d(n)|. \tag{36}$$

This logarithmic cost function is often used in signal processing problems as an approximation of the ℓ_0-norm, counting the number of non-zero entries in a vector [29], [46], [47]. The ℓ_0-norm is related to the previously defined ℓ_p-norm through $\|\mathbf{d}\|_0 = \lim_{p\to 0} \sum_{n=1}^{N} |d(n)|^p$. The logarithmic penalty is related to the ℓ_0-norm through $\lim_{p\to 0} \frac{1}{p} \sum_{n=1}^{N} (|d(n)|^p - 1) = \sum_{n=1}^{N} \log |d(n)|$[46]. Moreover, the set of local minima of the optimization problem in (36) corresponds to the set of local minima of the optimization problem [46]

$$\min_{\mathbf{g}} \|\mathbf{d}\|_0. \tag{37}$$

Using (8) the desired speech signal can be further expressed as

$$\mathbf{d} = \mathbf{\Omega}\mathbf{u}, \tag{38}$$

with

$$\mathbf{\Omega} = [\mathbf{x}_1, -\mathbf{X}_\tau], \quad \mathbf{u} = \left[1, \mathbf{g}^T\right]^T, \tag{39}$$

where $\mathbf{u}$ is equivalent to the prediction vector $\mathbf{g}$. Now the optimization problem (35) can be rewritten directly in terms of the prediction vector $\mathbf{u}$ as

$$\boxed{\min_{\mathbf{u}} \|\mathbf{\Omega}\mathbf{u}\|_p^p \quad \text{s.t.} \quad \mathbf{e}_1^T\mathbf{u} = 1,} \tag{40}$$

where $\mathbf{e}_1 = [1, 0, \ldots, 0]^T$. Optimization problems in this form are addressed in the context of the cosparse analysis problem [48]–[50]. In that setting, the matrix $\mathbf{\Omega}$ is the analysis matrix that transforms the unknown variable (i.e., the prediction vector $\mathbf{u}$) to the domain where the sparsity is enforced (i.e., the prediction error $\mathbf{d}$). By solving the problem in (40) an estimate of the prediction vector $\mathbf{u}$ is computed that results in a sparse prediction error, i.e., the desired speech signal, $\mathbf{d}$, with sparsity quantified by means of the ℓ_p-norm. Also, a similar optimization problem was considered in the context of sparse linear prediction in the time domain [51], applied for modeling and coding of speech signals.

The analytically derived sparsity-promoting cost function can be easily justified in the context of dereverberation. Intuitively, reverberation makes the recorded speech signal less sparse than the clean speech signal in the STFT domain. Therefore, on the one hand it is reasonable to enforce an estimate of the desired speech signal whose STFT coefficients are sparser than the STFT coefficients of the reverberant recording. On the other hand, the direct path and early reflections should be preserved in the estimated desired speech signal, which is enforced by using the MCLP model with the prediction delay in (3), resulting in the optimization problem in (40) with a structured analysis matrix $\mathbf{\Omega}$.

In summary, both the conventional method and the proposed method based on CGG priors can be interpreted as iterative optimization methods that aim to compute a minimum of the optimization problem in (35)/(40) corresponding to WPE-CGG for $0 < p \leq 2$ and to the conventional method when $p \to 0$.

A. Iteratively Reweighted LS for ℓ_p-norm Minimization

Note that the optimization problem in (35) is non-convex for $p < 1$, and iterative optimization methods can in general converge only to a local minimum. However, non-convex cost functions often result in a sparser estimated signal than using a convex cost function (e.g., for $p \geq 1$) [29]. Several optimization methods for ℓ_p-norm minimization have been proposed that transform the non-convex problem into a series of convex problems [29], [46], [47]. Here we employ the iteratively reweighted LS (IRLS) method for ℓ_p-norm minimization [29], [46], and show that the obtained method is equivalent to the conventional method and the method based on a CGG prior.

The basic idea in IRLS is to replace the ℓ_p-norm minimization problem with a series of ℓ_2-norm minimization subproblems [29], [49], [52]. Each ℓ_2-norm minimization subproblem can be solved easily, and the solution in one iteration is used to modify the subproblem in the next iteration. More specifically, the ℓ_p-norm cost function in (40) is replaced by a weighted ℓ_2-norm cost function in the i-th iteration as [29]

$$\hat{\mathbf{u}}^{(i+1)} = \arg\min_{\mathbf{u}} \mathbf{u}^H\mathbf{\Omega}^H\mathbf{W}^{(i)}\mathbf{\Omega}\mathbf{u} \quad \text{s.t.} \quad \mathbf{e}_1^T\mathbf{u} = 1, \tag{41}$$

with a real-valued diagonal weighting matrix $\mathbf{W}^{(i)} = \text{diag}(\mathbf{w}^{(i)})$, where $\mathbf{w}^{(i)} = [w^{(i)}(1), \ldots, w^{(i)}(N)]^T$ are the weights. The LS optimization problem in (41) has a closed-form solution

$$\hat{\mathbf{u}}^{(i+1)} = \left(\mathbf{e}_1^T\left(\mathbf{\Omega}^H\mathbf{W}^{(i)}\mathbf{\Omega}\right)^{-1}\mathbf{e}_1\right)^{-1}\left(\mathbf{\Omega}^H\mathbf{W}^{(i)}\mathbf{\Omega}\right)^{-1}\mathbf{e}_1 \tag{42}$$

[2] Note that for $p < 1$ the ℓ_p-norm is actually not a norm, e.g., it does not satisfy the triangle inequality.

that is equivalent to estimating the prediction vector $\hat{\mathbf{g}}^{(i+1)}$ in (14). The estimate of the desired signal in the $(i+1)$-th iteration is given using (38) as $\hat{\mathbf{d}}^{(i+1)} = \mathbf{\Omega}\hat{\mathbf{u}}^{(i+1)}$.

As in [29], [49], [52], the weights are updated in each iteration as

$$w^{(i+1)}(n) = \frac{1}{|\hat{d}^{(i+1)}(n)|^{2-p}}, \tag{43}$$

so that the cost function in (41) is a first-order approximation of the cost function in (40). The updates (42) and (43) result in an iterative method for minimizing (40). To avoid division by zero in (43), the optimization problem is typically regularized by adding a small positive value [29], [49], i.e.,

$$w^{(i+1)}(n) = \left(|\hat{d}^{(i+1)}(n)|^2 + \varepsilon^{(i+1)}\right)^{\frac{p}{2}-1}. \tag{44}$$

When the role of ε is just to avoid division by zero the method is called unregularized IRLS [29]. Setting ε to a larger value can be used to make the linear system in (42) better conditioned. In practice, a regularization strategy where ε is initialized with a large value and then gradually decreased has been shown to be effective in avoiding local minima for $p < 1$[29]. In this case the method is called regularized IRLS. Various strategies for updating the regularization parameter in iteratively reweighted algorithms have been investigated in [46].

By comparing the obtained update for the weights in (43) with the variance update in (29), it can be seen that the weights are equal to the inverse of the variances. With this in mind, the obtained LS problem in (41) is equivalent to the LS problem in (13), i.e., they result in the same prediction vector if the weights are calculated in the same way. The difference between these methods is the weight regularization strategy that is performed by adding a small ε in IRLS, or using ε as a lower bound in WPE-CGG.

The outline of the complete dereverberation algorithm using regularized IRLS (r-IRLS) method in each frequency bin k is given in Algorithm 2. For each frequency bin k the matrix $\mathbf{\Omega}$ is normalized with the maximum magnitude of the STFT coefficients of the reference microphone signal $\mathbf{x}_1$. In this way the values of the regularization parameter ε for r-IRLS can be set independently of the magnitudes of the coefficients in the given frequency bin. The r-IRLS for minimization of (40) is implemented similarly as in [29]. The updates (42) and (44) are iterated until the relative change of the ℓ_2-norm of the output is smaller than the tolerance η. In that case the regularization parameter ε is reduced 10 times, and the tolerance parameter is updated to $\sqrt{\varepsilon}/100$. The unregularized IRLS (u-IRLS) is implemented by omitting the reduction of the regularization parameter ε and tolerance η. Additionally, since $p < 1$ results in a non-convex problem in (40), initialization of the algorithm can influence the final estimate. More details on the initialization are given in Section VI-B.

VI. Experiments

In this section, the results of several experiments for different acoustic scenarios and different numbers of microphones are presented. The results obtained using the conventional WPE method (cf. Section III) and the proposed WPE-CGG method (cf. Section IV) and the IRLS algorithm applied on the ℓ_p-norm minimization problem (cf. Section V) are compared. The considered acoustic systems and the used performance measures are introduced in Section IV-A. The implementation details of the different methods are described in Section IV-B. The performance of the MCLP-based speech dereverberation is evaluated for different values of the shape parameter p, corresponding to different sparse CGG priors for the desired speech signal, is evaluated in Section VI-C. The dereverberation performance for different acoustic scenarios with $M = 2$ microphones is evaluated in Section VI-D. The dereverberation performance using different numbers of microphones is evaluated in Section VI-E, and for different number of iterations in Section VI-F.

Algorithm 2 WPE using the IRLS algorithm. For r-IRLS, the parameter ε is initialized with a relatively large value and gradually reduced. For u-IRLS, the parameter ε is initialized as $\varepsilon_{\min}$. $\|\mathbf{x}\|_\infty$ denotes the maximum absolute value of the elements in $\mathbf{x}$.

```
parameters: Filter length $L_g$ and prediction delay $\tau$ in (3), shape parameter $p$ in (44), regularization parameters $\varepsilon_{\text{init}}, \varepsilon_{\min}$, maximum number of iterations $i_{\max}$, tolerance $\eta_{\text{init}}$
input: $\mathbf{x}_m(k), \forall m, k$
for all $k$ do
    $i \leftarrow 0$
    $\varepsilon^{(0)} \leftarrow \varepsilon_{\text{init}}, \eta^{(0)} \leftarrow \eta_{\text{init}}$
    $\mathbf{\Omega} \leftarrow$ construct as in (39)
    $\mathbf{\Omega} \leftarrow \mathbf{\Omega}/\kappa$, with $\kappa = \|\mathbf{x}_1\|_\infty$
    $\hat{\mathbf{d}}^{(0)} \leftarrow \mathbf{x}_1/\kappa$
    repeat
    $\mathbf{W}^{(i)} \leftarrow$ calculate as in (44)
    $\hat{\mathbf{u}}^{(i+1)} \leftarrow$ calculate as in (42)
    $\hat{\mathbf{d}}^{(i+1)} \leftarrow \mathbf{\Omega}\hat{\mathbf{u}}^{(i+1)}$
if      $\|\hat{\mathbf{d}}^{(i+1)} - \hat{\mathbf{d}}^{(i)}\|_2/\|\hat{\mathbf{d}}^{(i)}\|_2 < \eta^{(i)}$ then
        $\varepsilon^{(i+1)} \leftarrow \varepsilon^{(i)}/10, \eta^{(i+1)} \leftarrow \sqrt{\varepsilon^{(i)}}/100$
    else
        $\varepsilon^{(i+1)} \leftarrow \varepsilon^{(i)}, \eta^{(i+1)} \leftarrow \eta^{(i)}$
    end if
    $i \leftarrow i + 1$
    untill $\varepsilon^{(i)} < \varepsilon_{\min}$ or $i > i_{\max}$
    $\hat{\mathbf{d}} \leftarrow \kappa\hat{\mathbf{d}}^{(i)}$
end for
```

A. Acoustic Systems and Performance Measures

We consider an acoustic scenario with a single speech source and M omni-directional microphones placed at a distance of about 2.3 m from the source. In Section VI-C, Section VI-D, and Section VI-F a scenario with $M = 2$ microphones is considered, while in Section VI-E the number of microphones is set to $M \in \{1, 2, 4\}$. Three different rooms with reverberation time RT_{60} of approximately $\{450, 550, 750\}$ ms were used in the experiments. The distance between the source and the microphones is approximately 2.3 m, and the direct-to-reverberant ratio (DRR) for the reference microphone is DRR

$\in \{9.7, -4.1, -3.8\}$ dB for each of the rooms. The RIRs between the source and the microphones have been measured using the swept-sine technique, and the sampling frequency is set to 16 kHz. The reverberant observations are generated by convolving the measured RIRs with clean (anechoic) speech utterances. Influence of noise has not been considered in the experiments, since the main goal is to evaluate the dereverberation performance, and joint dereverberation and denoising remains a topic for future work. We have used a set of utterances from 40 different speakers (20 male and 20 female), where the average length of the speech samples is approximately 4.2 s. The dereverberation performance is evaluated in terms of different instrumental measures: cepstral distance (CD), perceptual evaluation of speech quality score (PESQ), frequency-weighted segmental signal-to-noise ratio (FWSSNR), and speech-to-reverberation modulation energy ratio (SRMR) [53]. For the intrusive measures (CD, PESQ, FWSSNR), the clean speech signal is used as a ground-truth signal. In the following we present the improvements of the considered instrumental measures when compared to the input signal on the reference microphone. The reported values are obtained by averaging the improvements over all utterances.

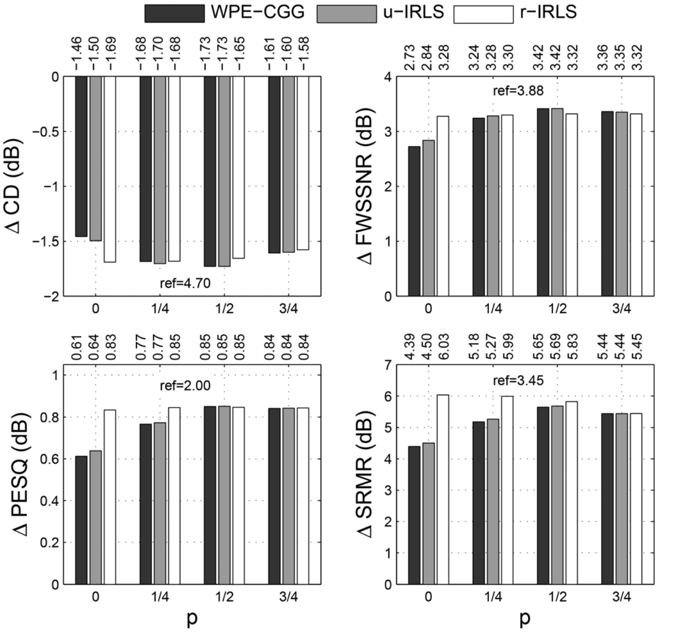

Fig. 2. Results for an acoustic system with $RT_{60} \approx 750$ ms and $M = 2$ microphones for values of the shape parameter $p \in \{0, 1/4, 1/2, 3/4\}$. The reported values are obtained as the averaged improvements over all utterances. The average values calculated for the reference microphone signal are denoted as "ref".

B. Implementation Details

In all experiments the STFT has been calculated using a 64 ms Hamming window with 16 ms shift. The prediction delay in (3) is set to $\tau = 2$ frames in all experiments. The length of the prediction vector L_g in (3) is set to $L_g \in \{35, 15, 8\}$ for $M \in \{1, 2, 4\}$ microphones. While the length L_g could be set depending on the reverberation time, here we used the fixed length for each number of microphones. These settings are similar to the ones used in [54].

The WPE-CGG method is implemented as in Algorithm 1, with the conventional WPE corresponding to the case with $p = 0$. The variance estimate is regularized with the lower bound set to $\varepsilon_{\min} = 10^{-8}$ for all frequency bins k, and the tolerance on the change of the relative ℓ_2-norm of the estimated desired signal is set to $\eta = 10^{-6}$. The u-IRLS minimizing (40) is implemented by fixing the regularization parameter in (44) to $\varepsilon_{\text{init}} = 10^{-8}$. Since the matrix $\mathbf{\Omega}$ is normalized with the maximum magnitude of the STFT coefficients of the reference microphone signal, the regularization parameter $\varepsilon_{\min}$ is always much smaller than the magnitudes of the coefficients in the given frequency bin, and therefore it only serves to avoid division by zero. The tolerance on the change of the relative ℓ_2-norm of the estimated desired signal is set to $\eta = 10^{-6}$. The r-IRLS minimizing (40) is implemented with the initial value for the regularization parameter ε set to $\varepsilon_{\text{init}} = 0.1$ and the minimum value $\varepsilon_{\min} = 10^{-8}$. The same final tolerance $\eta = 10^{-6}$ applies for r-IRLS because $\varepsilon_{\min} = 10^{-8}$ (cf. Algorithm 2).

Since the problem in (40) is non-convex for $p < 1$, the presented algorithms only converge to a local minimum, and the final estimate may heavily depend on the initialization. In compressive sensing the IRLS method is typically initialized with the solution of (40) for $p = 2$ (i.e., the least-squares solution). However, as shown in [17], the least squares solution is not effective for dereverberation, and results in a signal that is even more reverberant than the microphone signal. This occurs because the least squares solution results in a minimum-energy estimate of the desired speech signal with typically many non-zero coefficients. Therefore, the least squares solution is often a poor initialization for the iterative algorithm in the context of dereverberation. In our experiments initializing with the least-squares solution also resulted in a decreased dereverberation performance for the WPE-CGG and u-IRLS methods, whereas the r-IRLS method was in general less affected by initialization (due to the regularization). Therefore, in all experiments we initialized the desired signal $\hat{\mathbf{d}}^{(0)}$ with the reference microphone signal $\mathbf{x}_1$ (or its normalized version).

C. Evaluation for Different Values of the Shape Parameter p

In this section we investigate speech dereverberation performance for different values of the shape parameter p. We consider a scenario with $M = 2$ microphones in a room with $RT_{60} \approx 750$ ms, and compare the WPE-CGG, u-IRLS, and r-IRLS methods for $p \in \{0, 1/4, 1/2, 3/4\}$. The conventional WPE method corresponds to WPE-CGG with $p = 0$. Typical number of iterations for convergence of the WPE-CGG and r-IRLS methods was between 50 and 100, while the r-IRLS method required more iterations, typically between 300 and 400. The improvements of the considered instrumental measures for each value of the shape parameter p are presented in Fig. 2. It can be observed that the performance of the employed optimization methods depends on p. As expected, the performance of the WPE-CGG and the u-IRLS is very similar. It can be observed that both methods perform best for $p = 1/2$, achieving almost identical results. For smaller values of the shape parameter (e.g., $p = 0$ corresponding to the conventional WPE) and also for higher values of the shape parameter (e.g., $p = 3/4$) both methods achieve lower performance. Note

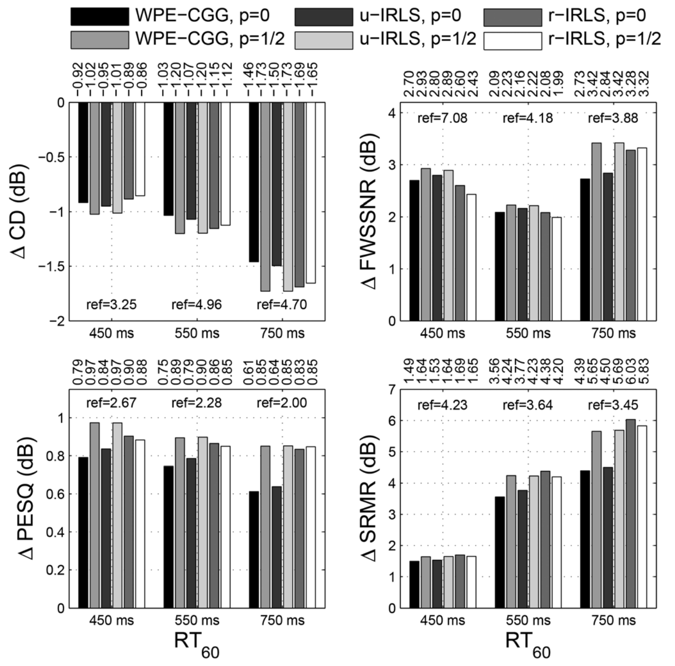

Fig. 3. Results for different acoustic systems with $M = 2$ microphones and $RT_{60} \approx \{450, 550, 750\}$ ms. The reported values are obtained as the averaged improvements over all utterances. The average values calculated for the reference microphone signal are denoted as "ref".

that the used values of p are not optimal in any sense, and are selected to illustrate the effect of the selected cost function on the performance. In the experiments both small values (close to 0), and large values (close to 1) of p resulted in a decreased performance. The r-IRLS is less sensitive to selection of the parameter p due to the regularization strategy, although by increasing the value of the parameter p the performance starts to decrease. However, the regularization strategy also results in a significantly higher number of iterations. These observations are similar with the observed performance of the unregularized and regularized methods in the context of sparse recovery [29].

D. Evaluation in Different Acoustic Scenarios

In this section we investigate the performance in different acoustic scenarios after convergence of the iterative algorithms. We consider a setup with $M = 2$ microphones in rooms with $RT_{60} \approx \{450, 550, 750\}$ ms. In the following, we compare WPE-CGG, u-IRLS and r-IRLS for $p \in \{0, 1/2\}$. The improvements of the considered instrumental measures are presented in Fig. 3. It can be observed that WPE-CGG and u-IRLS with $p = 1/2$ outperforms the case with $p = 0$ in all evaluated measures for all scenarios. The results in Fig. 3 suggest that the performance improvement for the evaluated measures with $p = 1/2$, when compared to $p = 0$, is higher for longer reverberation times. Similar as in the previous experiment, the r-IRLS method is slightly better with $p = 0$ than with $p = 1/2$ for all scenarios, performing similarly to the unregularized methods with $p = 1/2$.

E. Evaluation for Different Number of Microphones

In this section we investigate the performance for different numbers of microphones. We consider a setup in a room with $RT_{60} \approx 750$ ms, with $M \in \{1, 2, 4\}$ microphones. The performance of the WPE-CGG is evaluated, with $p \in \{0, 1/2\}$. The improvements of the evaluated measures are presented in Fig. 4, and it is again visible that $p = 1/2$ outperforms $p = 0$ in all of the evaluated measures. While both algorithms perform better with larger number of microphones, in all cases $p = 1/2$ performs better than $p = 0$.

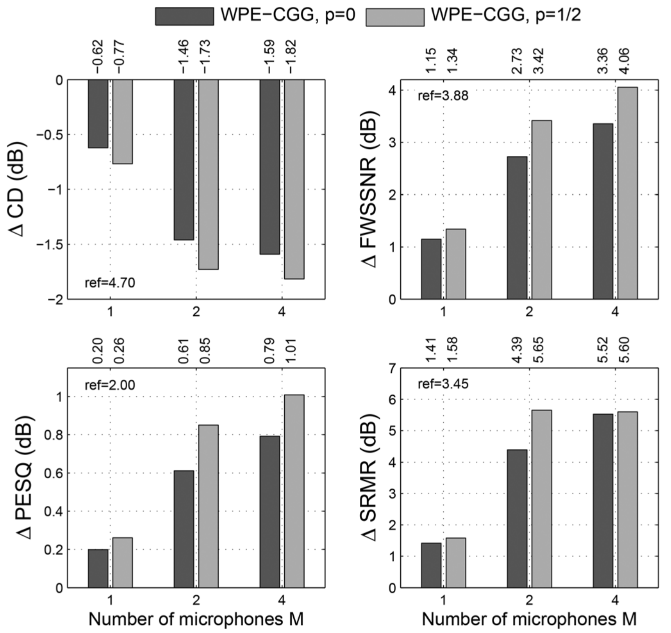

Fig. 4. Results for different acoustic systems with $RT_{60} \approx 750$ ms and $M = \{1, 2, 4\}$ microphones. The reported values are obtained as the averaged improvements over all utterances. The average values calculated for the reference microphone signal are denoted as "ref".

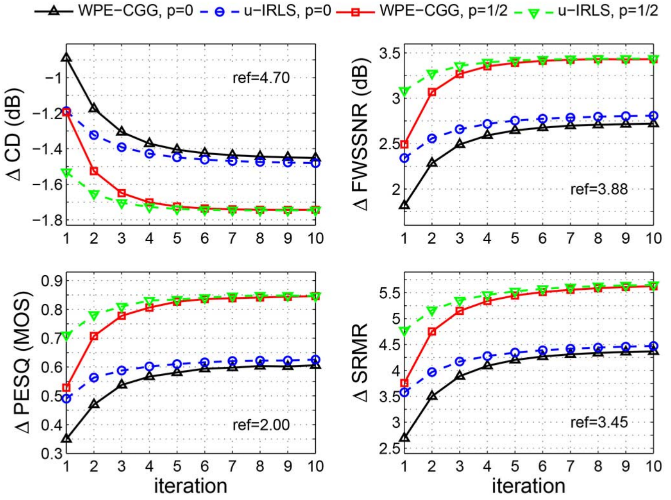

Fig. 5. Results for different number of iterations for an acoustic system with $RT_{60} \approx 750$ ms and $M = 2$ microphones. The reported values are obtained as the averaged improvements over all utterances. The average values calculated for the reference microphone signal are denoted as "ref".

F. Evaluation for Different Number of Iteration

In this section we investigate the iteration-wise performance of the WPE-CGG and u-IRLS methods for $p \in \{0, 1/2\}$. The r-IRLS method is not included in the comparison since it typically requires many more iterations due to the reduction update for the regularization parameter ε. The values of the considered instrumental measures after each iteration are presented in Fig. 5. It can be observed that the results become stable after relatively small number of iterations (up to 10). Also, it can be observed that $p = 1/2$ results in a better performance than $p = 0$ for any number of iterations, with the u-IRLS method converging slightly faster than the WPE-CGG method.

VII. Conclusion

In this paper we have presented a novel MCLP-based speech dereverberation method, based on a sparse prior for modeling the desired speech signal, with a special emphasis on circular priors from the complex generalized Gaussian family. The proposed model can be interpreted as a generalization of the TVG model, with an additional hyperprior on the unknown variances. It has also been shown that the underlying prior in the conventional WPE method strongly promotes sparsity of the desired speech signal, and can be obtained as a special case of the proposed WPE-CGG method with $p = 0$. Furthermore, the proposed method has been reformulated as an optimization problem with the cost function equal to ℓ_p-norm on the desired speech signal. In addition, we have shown that solving this optimization problem by an iteratively reweighted LS scheme results in an equivalent set of updates.

The experimental results for various acoustic scenarios show that the instrumentally predicted speech enhancement performance can be consistently improved in the proposed framework, by setting p to an appropriate value. While the improvements are mild, it is important to keep in mind that these come at virtually no cost with just a small modification of the weight/variance update. As we have analytically shown using the ℓ_p-norm-based formulation, speech dereverberation is achieved by exploiting the fact that the desired speech signal is more sparse than the reverberant recordings in the STFT domain. Furthermore, the highlighted role of sparsity-promoting cost functions suggests also that different cost functions and sparse recovery methods could be applied to achieve speech dereverberation. These insights could be useful not only for the considered MCLP-based dereverberation method but also for other speech enhancement methods.

Appendix A
Convex Representation of a Sparse Prior

We are interested in a circular sparse prior $\rho(z) = e^{-f(|z|)}$ that can be represented in the form (20) for a certain function $\psi(.)$. Due to the circular symmetry of $\rho(.)$, and analogously as in [25], we can write

$$-\log \rho(t) = \inf_{\lambda > 0} \frac{t^2}{\lambda} - \log \frac{\psi(\lambda)}{\pi\lambda}, \tag{45}$$

for $t \in (0, \infty)$. By introducing a function $g(.)$ such that $\rho(t) = e^{-g(t^2)}$, i.e., $f(t) = g(t^2)$, we can write

$$g(t) = \inf_{\lambda > 0} \frac{t}{\lambda} - \log \frac{\psi(\lambda)}{\pi\lambda}. \tag{46}$$

Using results in [25], [55] it follows that $\rho(.)$ has a convex type representation (20) if $g(.)$ is concave on $(0, \infty)$. Then it holds that

$$\psi(\lambda) = \pi\lambda e^{g_c\left(\lambda^{-1}\right)}, \tag{47}$$

where $g_c(.)$ is the concave conjugate of $g(.)$[55]. The condition on $g(.)$ is equivalent to $f'(t)/t$ being non-increasing on $(0, \infty)$ [26], [27], [55].

Appendix B
Variance Estimation

In the variance estimation step we need to solve the optimization problem in (24), which can be written using (47) in the following form

$$\hat{\lambda} = \arg\min_{\lambda > 0} \frac{t^2}{\lambda} - g_c\left(\lambda^{-1}\right), \tag{48}$$

for some $t \geq 0$, with $g_c(\lambda^{-1}) = \log \psi(\lambda) - \log \pi\lambda$ following from (47). Hence, the optimal variance $\hat{\lambda}$ is equal to

$$\hat{\lambda}^{-1} = (g_c')^{-1}(t^2), \tag{49}$$

where $(g_c')^{-1}(.)$ is the inverse function of $g_c'(.)$. Using $g'(t) = \arg\min_u tu - g_c(u)$[25], [55] it follows that $(g_c')^{-1}(t) = g'(t)$, and using $f(t) = g(t^2)$ the optimal λ can be written as

$$\hat{\lambda} = \frac{1}{g'(t^2)} = \frac{2t}{f'(t)}. \tag{50}$$

References

[1] R. Beutelmann and T. Brand, "Prediction of speech intelligibility in spatial noise and reverberation for normal-hearing and hearing-impaired listeners," *J. Acoust. Soc. Amer.*, vol. 120, no. 1, pp. 331–342, Jul. 2006.

[2] M. Omologo, P. Svaizer, and M. Matassoni, "Environmental conditions and acoustic transduction in hands-free speech recognition," *Speech Commun.*, vol. 25, no. 1–3, pp. 75–95, Aug. 1998.

[3] A. Sehr, "Reverberation modeling for robust distant-talking speech recognition," Ph.D. dissertation, Friedrich-Alexander-Univ. Erlangen-Nürenberg, Erlangen, Germany, Oct. 2009.

[4] R. Maas, E. A. P. Habets, A. Sehr, and W. Kellermann, "On the application of reverberation suppression to robust speech recognition," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, Kyoto, Japan, Mar. 2012, pp. 297–300.

[5] M. Jeub, M. Schafer, T. Esch, and P. Vary, "Model-based dereverberation preserving binaural cues," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 18, no. 7, pp. 1732–1745, Sep. 2010.

[6] P. A. Naylor and N. D. Gaubitch*, Speech Dereverberation*. New York, NY, USA: Springer, 2010.

[7] M. Miyoshi and Y. Kaneda, "Inverse filtering of room acoustics," *IEEE Trans. Acoust. Speech Signal Process.*, vol. 36, no. 2, pp. 145–152, Feb. 1988.

[8] A. Mertins, T. Mei, and M. Kallinger, "Room impulse response shortening/reshaping with infinity- and p-norm optimization," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 18, no. 2, pp. 249–259, Feb. 2010.

[9] W. Zhang, E. A. P. Habets, and P. A. Naylor, "On the use of channel shortening in multichannel acoustic system equalization," in *Proc. Int. Workshop Acoust. Echo Noise Control (IWAENC)*, Tel Aviv, Israel, Sep. 2010.

[10] I. Kodrasi, S. Goetze, and S. Doclo, "Regularization for partial multichannel equalization for speech dereverberation," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 21, no. 9, pp. 1879–1890, Sep. 2013.

[11] I. Kodrasi, T. Gerkmann, and S. Doclo, "Frequency-domain single-channel inverse filtering for speech dereverberation: Theory and practice," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, Florence, Italy, May 2014, pp. 5177–5181.

[12] K. Lebart, J. M. Boucher, and P. N. Denbigh, "A new method based on spectral subtraction for speech dereverberation," *Acta Acoust.*, vol. 87, pp. 359–366, 2001.

[13] T. Gerkmann, "Cepstral weighting for speech dereverberation without musical noise," in *Proc. Eur. Signal Process. Conf. (EUSIPCO)*, Barcelona, Spain, Sep. 2011.

[14] B. Cauchi, I. Kodrasi, R. Rehr, S. Gerlach, A. Jukić, T. Germann, S. Doclo, and S. Goetze, "Joint dereverberation and noise reduction using beamforming and a single-channel speech enhancement scheme," in *Proc. REVERB Workshop*, Florence, Italy, May 2014.

[15] E. A. P. Habets, S. Gannot, and I. Cohen, "Late reverberant spectral variance estimation based on a statistical model," *IEEE Signal Process. Lett.*, vol. 16, no. 9, pp. 770–773, Sep.. 2009.

[16] T. Nakatani, T. Yoshioka, K. Kinoshita, M. Miyoshi, and B.-H. Juang, "Blind speech dereverberation with multi-channel linear prediction based on short time fourier transform representation," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, Las Vegas, NV, USA, May 2008, pp. 85–88.

[17] T. Nakatani, T. Yoshioka, K. Kinoshita, M. Miyoshi, and B. H. Juang, "Speech dereverberation based on variance-normalized delayed linear prediction," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 18, no. 7, pp. 1717–1731, Sep. 2010.

[18] T. Yoshioka and T. Nakatani, "Generalization of multi-channel linear prediction methods for blind MIMO impulse response shortening," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 20, no. 10, pp. 2707–2720, Dec. 2012.

[19] M. Togami, Y. Kawaguchi, R. Takeda, Y. Obuchi, and N. Nukaga, "Optimized speech dereverberation from probabilistic perspective for time varying acoustic transfer function," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 21, no. 7, pp. 1369–1380, Jul. 2013.

[20] B. Schwartz, S. Gannot, and E. A. P. Habets, "Multi-microphone speech dereverberation using expectation-maximization and kalman smoother," in *Proc. Eur. Signal Process. Conf. (EUSIPCO)*, Marrakech, Morocco, Sep. 2013.

[21] D. Schmid, G. Enzner, S. Malik, D. Kolossa, and R. Martin, "Variational Bayesian inference for multichannel dereverberation and noise reduction," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 22, no. 8, pp. 1320–1335, Aug. 2014.

[22] Y. Iwata and T. Nakatani, "Introduction of speech log-spectral priors into dereverberation based on Itakura-Saito distance minimization," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, Kyoto, Japan, May 2012, pp. 245–248.

[23] K. Kinoshita, M. Delcroix, T. Nakatani, and M. Miyoshi, "Suppression of late reverberation effect on speech signal using long-term multiple-step linear prediction," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 17, no. 4, pp. 534–545, May 2009.

[24] A. Jukić, T. van Waterschoot, T. Gerkmann, and S. Doclo, "Speech dereverberation with multi-channel linear prediction and sparse priors for the desired signal," in *Proc. Joint Workshop Hands-Free Speech Commun. Microphone Arrays (HSCMA)*, Nancy, France, May 2014, pp. 23–26.

[25] J. A. Palmer, K. Kreutz-Delgado, D. P. Wipf, and B. D. Rao, "Variational EM algorithms for non-gaussian latent variable models," in *Advances in Neural Information Processing Systems 18*. Cambridge, MA, USA: MIT Press, 2006, pp. 1059–1066.

[26] S. D. Babacan, R. Molina, M. N. Do, and A. K. Katsaggelos, "Bayesian blind deconvolution with general sparse image priors," in *Proc. Eur. Conf. Comput. Vis. (ECCCV)*, Florence, Italy, Oct. 2012, pp. 341–355.

[27] D. Wipf and H. Zhang, "Analysis of bayesian blind deconvolution," in *Proc. Int. Conf. Energy Minimizat. Meth. Comput. Vis. Pattern Recogn. (EMMCVPR)*, Lund, Sweden, Aug. 2013, pp. 40–53.

[28] M. Novey, T. Adali, and A. Roy, "A complex generalized Gaussian distribution - characterization, generation, and estimation," *IEEE Trans. Signal Process.*, vol. 58, no. 3, pp. 1427–1433, 2010.

[29] R. Chartrand and W. Yin, "Iteratively reweighted algorithms for compressive sensing," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, Las Vegas, NV, USA, May 2008, pp. 3869–3872.

[30] Y. Avargel and I. Cohen, "System identification in the short-time Fourier transform domain with crossband filtering," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 15, no. 4, pp. 1305–1319, May 2007.

[31] R. Talmon, I. Cohen, and S. Gannot, "Relative transfer function identification using convolutive transfer function approximation," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 17, no. 4, pp. 546–555, May 2009.

[32] R. Talmon, I. Cohen, and S. Gannot, "Convolutive transfer function generalized sidelobe canceler," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 17, no. 7, pp. 1420–1434, Sep. 2009.

[33] H. Kameoka, T. Nakatani, and T. Yoshioka, "Robust speech dereverberation based on non-negativity and sparse nature of speech spectrograms," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, Taipei, Taiwan, Apr. 2009, pp. 45–48.

[34] T. van Waterschoot, B. Defraene, M. Diehl, and M. Moonen, "Embedded optimization algorithms for multi-microphone dereverberation," in *Proc. Eur. Signal Process. Conf. (EUSIPCO)*, Marrakech, Morocco, Sep. 2013.

[35] R. Hendriks, T. Gerkmann, and J. Jensen, "Dft-domain based single-microphone noise reduction for speech enhancement: A survey of the state of the art," *Synth. Lectures Speech Audio Process.*, vol. 9, no. 1, pp. 1–80, Jan. 2013.

[36] A. Jukić and S. Doclo, "Speech dereverberation using weighted prediction error with Laplacian model of the desired signal," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, Florence, Italy, May 2014, pp. 5172–5176.

[37] M. Togami and Y. Kawaguchi, "Noise robust speech dereverberation with Kalman smoother," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, Vancouver, BC, Canada, May 2013, pp. 7447–7451.

[38] N. Ito, S. Araki, and T. Nakatani, "Probabilistic integration of diffuse noise suppression and dereverberation," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, Florence, Italy, May 2014, pp. 5167–5171.

[39] T. Yoshioka, T. Nakatani, M. Miyoshi, and H. G. Okuno, "Blind separation and dereverberation of speech mixtures by joint optimization," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 19, no. 1, pp. 69–84, Jan. 2011.

[40] J. Porter and S. Boll, "Optimal estimators for spectral restoration of noisy speech," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, San Diego, CA, USA, Mar. 1984, vol. 9, pp. 53–56.

[41] R. Martin, "Speech enhancement using MMSE short time spectral estimation with gamma distributed speech priors," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, Orlando, FL, USA, May 2002, pp. I–253.

[42] T. Gerkmann and R. Martin, "Empirical distributions of DFT-domain speech coefficients based on estimated speech variances," in *Proc. Int. Workshop Acoust. Echo Noise Control (IWAENC)*, Tel Aviv, Israel, Sep. 2010.

[43] I. Tashev and A. Acero, "Statistical modeling of the speech signal," in *Proc. Int. Workshop Acoust. Echo Noise Control (IWAENC)*, Tel Aviv, Israel, Sep. 2010.

[44] R. Martin, "Speech enhancement based on minimum mean-square error estimation and supergaussian priors," *IEEE Trans. Speech Audio Process.*, vol. 13, no. 5, pp. 845–856, Aug. 2005.

[45] T. Lotter and P. Vary, "Speech enhancement by MAP spectral amplitude estimation using a super-gaussian speech model," *EURASIP J. Appl. Signal Process.*, vol. 2005, pp. 1110–1126, 2005.

[46] D. Wipf and S. Nagarajan, "Iterative reweighted l1 and l2 methods for finding sparse solutions," *IEEE J. Sel. Topic Signal Process.*, vol. 4, no. 2, pp. 317–329, Apr. 2010.

[47] E. J. Candes, M. B. Wakin, and S. P. Boyd, "Enhancing sparsity by reweighted ℓ_1 minimization," *J. Fourier Anal. Applicat.*, vol. 14, no. 5–6, pp. 877–905, 2008.

[48] S. Nam, M. E. Davies, M. Elad, and R. Gribonval, "The cosparse analysis model and algorithms," *Appl. Comput. Harmon. Anal.*, vol. 34, no. 1, pp. 30–56, 2013.

[49] R. Chartrand, E. Y. Sidky, and X. Pan, "Nonconvex compressive sensing for X-ray CT: An algorithm comparison," in *Proc. Asilomar Conf. Signals, Syst. Comput. (ASILOMAR)*, Pacific Grove, CA, USA, Nov. 2013.

[50] R. Giryes, S. Nam, M. Elad, R. Gribonval, and M. Davies, "Greedy-like algorithms for the cosparse analysis model," in *Linear Algebra and its Applicat.*, Jan. 2014, pp. 22–60.

[51] D. Giacobello, M. G. Christensen, M. N. Murthi, S. H. Jensen, and M. Moonen, "Sparse linear prediction and its applications to speech processing," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 20, no. 5, pp. 1644–1657, Jul. 2012.

[52] B. D. Rao and K. Kreutz-Delgado, "An affine scaling methodology for best basis selection," *IEEE Trans. Signal Process.*, vol. 47, no. 1, pp. 187–200, Jan. 1999.

[53] K. Kinoshita, M. Delcroix, T. Yoshioka, E. Habets, R. Haeb-Umbach, V. Leutnat, A. Sehr, W. Kellermann, R. Maas, S. Gannot, and B. Raj, "The REVERB challenge: A common evaluation framework for dereverberation and recognition of reverberant speech," in *Proc. IEEE Workshop Appls. Signal Process. Audio Acoust. (WASPAA)*, New Paltz, NY, USA, Oct. 2013, pp. 1–4.

[54] M. Delcroix, T. Yoshioka, A. Ogawa, Y. Kubo, M. Fujimoto, I. Nobutaka, K. Kinoshita, M. Espi, T. Hori, T. Nakatani, and A. Nakamura, "Linear prediction-based dereverberation with advanced speech enhancement and recognition technologies for the REVERB challenge," in *Proc. REVERB Workshop*, Florence, Italy, May 2014.

[55] R. T. Rockafellar, *Convex Analysis*. Princeton, NJ, USA: Princeton Univ. Press, 1970.

Ante Jukić (S'10) received the Dipl.-Ing. degree in electrical engineering in 2009 from the University of Zagreb, Zagreb, Croatia. Since 2013 he is with the Signal Processing Group at the University of Oldenburg, Germany, working on speech dereverberation. Previously, he was with the Rudjer Bošković Institute and Xylon, both in Zagreb, Croatia. His research interests include acoustic signal processing, sparse signal processing, and machine learning for data enhancement and analysis.

Toon van Waterschoot (S'04–M'12) received the M.Sc. degree (2001) and the Ph.D. degree (2009) in electrical engineering, both from KU Leuven, Belgium. He is currently a tenure-track Assistant Professor at KU Leuven, Belgium. He has previously held teaching and research positions with the Antwerp Maritime Academy, Belgium (2002), the Institute for the Promotion of Innovation through Science and Technology in Flanders (IWT), Belgium (2003-2007), KU Leuven, Belgium (2008-2009), Delft University of Technology, The Netherlands (2010–2011), and the Research Foundation - Flanders (FWO), Belgium (2011–2014). Since 2005, he has been a Visiting Lecturer at the Advanced Learning and Research Institute of the University of Lugano (Universita della Svizzera Italiana), Switzerland. His research interests are in acoustic signal enhancement, acoustic modeling, audio analysis, and audio reproduction. Dr. van Waterschoot has been serving as an Associate Editor for the *Journal of the Audio Engineering Society* and for the *EURASIP Journal on Audio, Music, and Speech Processing*, and as a Guest Editor for *Signal Processing*. He has been a Nominated Officer for the European Association for Signal Processing (EURASIP), and a Scientific Coordinator of the FP7-PEOPLE Marie Curie Initial Training Network on Dereverberation and Reverberation of Audio, Music, and Speech (DREAMS). He has been serving as an Area Chair for Speech Processing at the European Signal Processing Conference (EUSIPCO 2010, 2013–2015), and will be the General Chair of the 60th AES Conference to be held in Leuven, Belgium, 2016. He is a member of the Audio Engineering Society, the Acoustical Society of America, EURASIP, and IEEE.

Timo Gerkmann (S'08–M'10–SM'15) studied electrical engineering at the universities of Bremen and Bochum, Germany. He received his Dipl.-Ing. degree in 2004 and his Dr.-Ing. degree in 2010 both at the Institute of Communication Acoustics (IKA) at the Ruhr-Universität Bochum, Bochum, Germany. In 2005, he spent six months with Siemens Corporate Research in Princeton, NJ, USA. During 2010 to 2011 Dr. Gerkmann was a Postdoctoral Researcher at the Sound and Image Processing Lab at the Royal Institute of Technology (KTH), Stockholm, Sweden. Since 2011, he has been a Professor for Speech Signal Processing at the Universität Oldenburg, Oldenburg, Germany. His main research interests are digital speech and audio processing, including speech enhancement, dereverberation, modeling of speech signals, speech recognition, and hearing devices. Timo Gerkmann is a Senior Member of the IEEE.

Simon Doclo (S'95–M'03–SM'13) 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 has been 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 the *EURASIP Journal on Advances in Signal Processing*.