KU LEUVEN

STADIUS
Center for Dynamical Systems, Signal Processing and Data Analytics

Citation/Reference	Ante Jukić, Toon van Waterschoot, and Simon Doclo **Adaptive speech dereverberation using constrained sparse multi-channel linear prediction** *IEEE Signal Process. Lett.*, vol. 24, no. 1, pp. 101-105, Jan. 2017.
Archived version	Author manuscript: the content is identical to the content of the submitted paper, but without the final typesetting by the publisher
Published version	https://doi.org/10.1109/LSP.2016.2640939
Journal homepage	http://ieeexplore.ieee.org/xpl/RecentIssue.jsp?punumber=97
Author contact	toon.vanwaterschoot@esat.kuleuven.be *+ 32 (0)16 321927*
IR	ftp://ftp.esat.kuleuven.be/pub/SISTA/vanwaterschoot/abstracts/16-194.html

(article begins on next page)

Adaptive Speech Dereverberation Using Constrained Sparse Multichannel Linear Prediction

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

***Abstract*—In this letter, we present an adaptive speech dereverberation method based on constrained sparse multichannel linear prediction (MCLP), minimizing the mixed $\ell_{2,p}$ norm of the desired component. In order to prevent overestimation of the undesired reverberant component, possibly leading to severe distortions of the output, we propose to use a statistical model for late reverberation to limit the power of the MCLP-based estimate. The resulting constrained optimization problem is solved by using the alternating direction method of multipliers, resulting in two variants of the dereverberation algorithm. Simulation results show that the proposed constraint increases the robustness with respect to parameter selection and improves the usability for dynamic scenarios in comparison to the unconstrained method.**

***Index Terms*—Adaptive filtering, constrained linear prediction, sparsity, speech dereverberation.**

I. Introduction

SPEECH recorded using distant microphones inside an enclosure is often degraded by reverberation, typically resulting in a decreased speech intelligibility and speech recognition performance [1], [2]. In order to reduce these effects, effective dereverberation is required in many applications and several speech dereverberation methods have been proposed [3]–[9].

One of the most popular blind multichannel (MC) speech dereverberation approaches is based on MC linear prediction (MCLP) [5]–[7], [10]–[13]. MCLP-based methods aim to predict the undesired reverberant component in the microphone signals, which is subsequently subtracted from the microphone signals. The prediction filters are typically obtained by maximizing sparsity of the dereverberated signal in the time–frequency domain [7], [12]. Adaptive versions of MCLP-based dereverberation have been proposed in [14] and [15], where the filter updates are based on the recursive least squares (RLS) algorithm [16]. However, since these methods typically do not include additional knowledge about the undesired component, they may lead to a significant overestimation of the undesired component and severe distortions of the output signal [15].

Manuscript received August 24, 2016; revised November 16, 2016; accepted November 30, 2016. Date of publication December 15, 2016; date of current version January 5, 2017. This research was supported in part by the Marie Curie ITN DREAMS Grant ITN-GA-2012-316969, and in part by the Cluster of Excellence 1077 "Hearing4All." The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Balázs Bank.

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

T. van Waterschoot is with the Department of Electrical Engineering, KU Leuven, Leuven 3001, Belgium (e-mail: toon.vanwaterschoot@esat.kuleuven.be).

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

Digital Object Identifier 10.1109/LSP.2016.2640939

In this letter, we propose to integrate additional knowledge about the reverberant component to prevent overestimation of the undesired component. More precisely, we propose to constrain the sparse MCLP-based estimate of the undesired component by using an estimate of the late reverberant power spectral density (PSD) [17], [18]. The resulting constrained optimization problem is solved by using the alternating direction method of multipliers (ADMM) [19], which can be implemented efficiently as a variant of RLS. In contrast to [20], we propose adaptive sparse MCLP based on the more general $\ell_{2,p}$-norm and propose two different variants of the ADMM-based algorithm. Simulations demonstrate the advantages of the proposed constrained methods when the prediction filters need to adapt quickly, e.g., for a moving source, and the optimal forgetting factor is not known.

II. Signal Model

We consider a scenario with a single speech source and M microphones in a reverberant room, with $s(k,n)$ denoting the clean speech signal in the short-time Fourier transform (STFT) domain, where k and n are the frequency and the time frame index. We assume that the mth microphone signal $x_m(k,n)$ can be decomposed as $x_m(k,n) = d_m(k,n) + u_m(k,n)$, where the desired component $d_m(k,n)$ contains the direct speech and early reflections, while the undesired component $u_m(k,n)$ contains the late reflections. In the following, we omit the frequency index k, since the signal will be modeled in each frequency bin independently. The MC model can be written as

$$\mathbf{x}(n) = \mathbf{d}(n) + \mathbf{u}(n) \tag{1}$$

where $\mathbf{x}(n) = [x_1(n), \ldots, x_M(n)]^{\mathsf{T}}$ and $\mathbf{d}(n)$ and $\mathbf{u}(n)$ are defined similarly. As shown in [5], the undesired component $\mathbf{u}(n)$ can be modeled by using MCLP as the sum of filtered (delayed) microphone signals, i.e.

$$\mathbf{u}(n) = \mathbf{G}^{\mathsf{H}}(n)\tilde{\mathbf{x}}_\tau(n) \tag{2}$$

where $\mathbf{G}(n) = [\mathbf{g}_1(n), \ldots, \mathbf{g}_M(n)] \in \mathbb{C}^{ML_g \times M}$ denotes the multiple-input multiple-output prediction filter, with $\mathbf{g}_m(n) \in \mathbb{C}^{ML_g}$ containing L_g taps per microphone, and $\tilde{\mathbf{x}}_\tau(n) \in \mathbb{C}^{ML_g}$ is a signal buffer defined as

$$\begin{aligned}\tilde{\mathbf{x}}_\tau(n) = {} & [x_1(n-\tau), \ldots, x_1(n-\tau-L_g+1), \\ & \ldots, x_M(n-\tau), \ldots, x_M(n-\tau-L_g+1)]^{\mathsf{T}}.\end{aligned} \tag{3}$$

The prediction delay τ ensures preservation of the short-time speech correlation and early reflections in the desired component [5], [11]. The goal of dereverberation is to recover the desired component $\mathbf{d}(n)$, which can be achieved by estimating the undesired component $\mathbf{u}(n)$ in (2) and subtracting it from the

1070-9908 © 2016 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission.
See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.

reverberant microphone signals $\mathbf{x}(n)$, i.e., the estimated desired component is equal to the prediction error [5], [6].

III. Sparse MCLP

By assuming that the prediction filter $\mathbf{G}(n)$ does not change over time, i.e., $\mathbf{G}(n) = \mathbf{G}$, *batch* dereverberation methods based on the signal model in (1) have been derived in [6], [21]. Given a batch of N frames, the prediction filter $\mathbf{G}$ is estimated by maximizing sparsity across time [7], [21], which can be formulated as minimizing the mixed $\ell_{2,p}$-norm of the desired component, i.e., $\min_{\mathbf{G}} \sum_{n=1}^{N} \|\mathbf{d}(n)\|_2^p$, where $\|.\|_2$ is the ℓ_2-norm and $p \leq 1$ is the shape parameter. This optimization problem can be solved using the iteratively reweighted least squares algorithm [21], [22], by approximating the ℓ_p-norm using a weighted ℓ_2-norm, i.e.

$$\hat{\mathbf{G}} = \arg\min_{\mathbf{G}} \sum_{n=1}^{N} w(n) \|\mathbf{d}(n)\|_2^2 \tag{4}$$

where the weights $w(n)$ are set to obtain a first-order approximation of the ℓ_p-norm [22]. For a known $\mathbf{d}(n)$, the weights $w(n)$ can be set to $w(n) = \left(\|\mathbf{d}(n)\|_2^2/M + \varepsilon\right)^{p/2-1}$, where ε is a small positive regularization constant. In this case, the weights $w(n)$ put more emphasis on frames where the desired signal $\mathbf{d}(n)$ should have a relatively small energy, exactly corresponding to the sparsity-promoting behavior of the ℓ_p-norm [7]. Since in practice the true $\mathbf{d}(n)$ is obviously not known, the weights $w(n)$ are usually computed using the estimated $\mathbf{d}(n)$ from the previous iteration [7], [22].

Similarly as in [15], an *adaptive* variant of sparse MCLP, estimating $\mathbf{G}(n)$ for each n, can be derived by incorporating an exponential window in (4), leading to

$$\hat{\mathbf{G}}(n) = \arg\min_{\mathbf{G}(n)} \sum_{t=1}^{n} \gamma^{n-t} w(t) \|\mathbf{d}(t)\|_2^2, \tag{5}$$

with forgetting factor $\gamma \in (0, 1]$. The prediction filter $\hat{\mathbf{G}}(n)$ in (5) can be computed by solving the unconstrained optimization problem

$$\hat{\mathbf{G}}(n) = \arg\min_{\mathbf{G}(n)} F\left(\mathbf{G}(n)\right) \tag{6}$$

with $F : \mathbb{C}^{ML_g \times M} \rightarrow \mathbb{R}$ a quadratic cost function equal to

$$\begin{aligned} F\left(\mathbf{G}(n)\right) \;=\; & \operatorname{tr}\left[\mathbf{G}^{\mathsf{H}}(n)\hat{\mathbf{Q}}(n)\mathbf{G}(n)\right] \\ & -2\Re\left\{\operatorname{tr}\left[\mathbf{G}^{\mathsf{H}}(n)\hat{\mathbf{R}}(n)\right]\right\} \end{aligned} \tag{7}$$

with the matrices $\hat{\mathbf{Q}}(n)$ and $\hat{\mathbf{R}}(n)$ defined as

$$\begin{aligned} \hat{\mathbf{Q}}(n) &= \sum_{t=1}^{n} \gamma^{n-t} w(t) \tilde{\mathbf{x}}_\tau(t) \tilde{\mathbf{x}}_\tau^{\mathsf{H}}(t), \\ \hat{\mathbf{R}}(n) &= \sum_{t=1}^{n} \gamma^{n-t} w(t) \tilde{\mathbf{x}}_\tau(t) \mathbf{x}^{\mathsf{H}}(t) \end{aligned} \tag{8}$$

and $\operatorname{tr}[.]$ denoting the trace. The closed-form solution for the prediction filter in (6) can hence be written as

$$\hat{\mathbf{G}}(n) = \hat{\mathbf{Q}}^{-1}(n)\hat{\mathbf{R}}(n). \tag{9}$$

Algorithm 1: Adaptive Sparse MCLP-Based Speech Dereverberation.

input: $\mathbf{x}(n)$, $\hat{\mathbf{G}}(n-1)$, $\hat{\mathbf{Q}}^{-1}(n-1)$
parameters: forgetting factor γ, shape parameter p
1: $w(n) =$ compute using (10) and Algorithm 2
2: $\hat{\mathbf{k}}(n) = \frac{\hat{\mathbf{Q}}^{-1}(n-1)\tilde{\mathbf{x}}_\tau(n)}{\frac{\gamma}{w(n)} + \tilde{\mathbf{x}}_\tau^{\mathsf{H}}(n)\hat{\mathbf{Q}}^{-1}(n-1)\tilde{\mathbf{x}}_\tau(n)}$
3: $\hat{\mathbf{G}}(n) = \hat{\mathbf{G}}(n-1) + \hat{\mathbf{k}}(n)\left[\mathbf{x}(n) - \hat{\mathbf{G}}^{\mathsf{H}}(n-1)\tilde{\mathbf{x}}_\tau(n)\right]^{\mathsf{H}}$
4: $\hat{\mathbf{Q}}^{-1}(n) = \frac{1}{\gamma}\left[\mathbf{I} - \hat{\mathbf{k}}(n)\tilde{\mathbf{x}}_\tau^{\mathsf{H}}(n)\right]\hat{\mathbf{Q}}^{-1}(n-1)$
5: $\hat{\mathbf{u}}(n) = \hat{\mathbf{G}}^{\mathsf{H}}(n)\tilde{\mathbf{x}}_\tau(n)$
output: $\hat{\mathbf{u}}(n)$, $\hat{\mathbf{G}}(n)$, $\hat{\mathbf{Q}}^{-1}(n)$
6: $\hat{\mathbf{d}}(n) = \mathbf{x}(n) - \hat{\mathbf{u}}(n)$

Algorithm 2: PSD Estimation. All Operations Applied Element-Wise.

input: $\mathbf{x}(n)$
parameters: smoothing constant α, duration of the early part T_d (seconds) and n_d (frames), decay constant $\Delta = \frac{3\ln 10}{T_{60}/T_d}$
1: $\hat{\boldsymbol{\sigma}}_x^2(n) = \alpha\hat{\boldsymbol{\sigma}}_x^2(n-1) + (1-\alpha)|\mathbf{x}(n)|^2$
2: $\hat{\boldsymbol{\sigma}}_r^2(n) = e^{-2\Delta}\hat{\boldsymbol{\sigma}}_x^2(n-n_d)$
3: $\hat{\boldsymbol{\sigma}}_d^2(n) = \alpha\hat{\boldsymbol{\sigma}}_d^2(n-1) + (1-\alpha)\max\left\{|\mathbf{x}(n)|^2 - \hat{\boldsymbol{\sigma}}_r^2(n), 0\right\}$
output: $\hat{\boldsymbol{\sigma}}_r(n)$, $\hat{\boldsymbol{\sigma}}_d(n)$

Since the matrices $\hat{\mathbf{Q}}(n)$ and $\hat{\mathbf{R}}(n)$ in (8) are rank-1 perturbations of $\gamma\hat{\mathbf{Q}}(n-1)$ and $\gamma\hat{\mathbf{R}}(n-1)$, the matrix inversion lemma can be used to obtain an RLS algorithm [16], as given in Algorithm 1. The computational complexity of Algorithm 1 is quadratic in the number of prediction filter coefficients per channel, with $\mathcal{O}\left(M^2L_g^2\right)$ operations.

Since the weights $w(n)$ are related to the power of the desired component, they can be computed using the average PSD of the desired component, i.e.

$$w(n) = \left(\|\hat{\boldsymbol{\sigma}}_d(n)\|_2^2/M + \varepsilon\right)^{p/2-1} \tag{10}$$

with $\hat{\boldsymbol{\sigma}}_d^2(n) = [\hat{\sigma}_{d,1}^2(n), \ldots, \hat{\sigma}_{d,M}^2(n)]^{\mathsf{T}}$ containing the PSDs of the desired component in all microphones. Algorithm 2 describes the PSD estimators based on the exponential decay model for late reverberation [17], [18], [23], which requires an estimate of the reverberation time T_{60}. Please note that due to the recursive averaging these PSD estimators are not very sensitive to the accuracy of the T_{60} estimate.

IV. Constrained Sparse MCLP

For dynamic scenarios, e.g., a moving speaker, where tracking of variations in the acoustic transfer functions between the source and the microphones is required, small values of the forgetting factor are generally preferred [16], [24]. However, small values of the forgetting factor result in a prediction error that approaches zero [16]. Since the output signal $\mathbf{d}(n)$ in (1) is equal to the prediction error, this may result in overestimation of the undesired component $\hat{\mathbf{u}}(n)$ and excessive cancellation of the speech signal [15]. Similarly, small values of the forgetting

factor may lead to ill-conditioning of the matrix $\hat{\mathbf{Q}}(n)$, resulting in an unstable output [15].

To alleviate this overestimation of the undesired component, we propose to incorporate prior knowledge about the undesired reverberation. More specifically, we propose to constrain the power of the MCLP-based estimate of the undesired component by the PSD estimate of the late reverberation based on the exponential decay model [17], [18], leading to the following optimization problem

$$\check{\mathbf{G}}(n) = \arg\min_{\mathbf{G}(n)} F\left(\mathbf{G}(n)\right) \quad \text{subject to} \quad |\mathbf{G}^{\mathsf{H}}(n)\tilde{\mathbf{x}}_\tau(n)|^2 \leq \hat{\boldsymbol{\sigma}}_u^2(n). \quad (11)$$

The vector $\hat{\boldsymbol{\sigma}}_u(n) = [\hat{\sigma}_{u,1}(n), \ldots, \hat{\sigma}_{u,M}(n)]^{\mathsf{T}}$ contains the bounds for the undesired component, and is defined as $\hat{\boldsymbol{\sigma}}_u(n) = \min\{\hat{\boldsymbol{\sigma}}_r(n), |\mathbf{x}(n)|\}$, with $\hat{\boldsymbol{\sigma}}_r(n)$ being the late reverberant PSD estimate, cf. Algorithm 2. By using the constrained optimization problem in (11) instead of the unconstrained optimization problem in (6), the excessive speech cancellation for small values of the forgetting factor γ should be prevented, while the performance for large values of the forgetting factor γ should not be significantly deteriorated.

The optimization problem in (11) can be rewritten by introducing a splitting variable $\mathbf{z}(n)$, i.e.

$$\min_{\mathbf{G}(n),\mathbf{z}(n)} F\left(\mathbf{G}(n)\right) + C\left(\mathbf{z}(n)\right) \quad \text{subject to} \quad \mathbf{G}^{\mathsf{H}}(n)\tilde{\mathbf{x}}_\tau(n) = \mathbf{z}(n) \quad (12)$$

where the inequality constraint in (11) is replaced with a convex barrier function $C : \mathbb{C}^M \rightarrow \mathbb{R}$, which is defined as $C\left(\mathbf{z}(n)\right) = 0$ when the constraint is satisfied, i.e., $|z_m(n)| \leq \hat{\sigma}_{u,m}(n)$ for all m, and $C\left(\mathbf{z}(n)\right) = \infty$ otherwise. Since F and C are convex functions, the optimization problem in (12) can be solved efficiently by applying the ADMM algorithm [19]. The augmented Lagrangian for the problem in (12) can be written as

$$\mathcal{L}\left(\mathbf{G}(n), \mathbf{z}(n), \boldsymbol{\lambda}(n)\right) = F\left(\mathbf{G}(n)\right) + C\left(\mathbf{z}(n)\right) + \frac{\rho}{2}\|\mathbf{G}^{\mathsf{H}}(n)\tilde{\mathbf{x}}_\tau(n) - \mathbf{z}(n) - \boldsymbol{\lambda}(n)\|_2^2 - \frac{\rho}{2}\|\boldsymbol{\lambda}(n)\|_2^2 \quad (13)$$

where ρ is a penalty parameter and $\boldsymbol{\lambda}(n)$ is the so-called dual variable [19]. The ADMM algorithm proceeds by minimizing $\mathcal{L}$ alternately with respect to $\mathbf{G}(n)$ and $\mathbf{z}(n)$, followed by an ascent over $\boldsymbol{\lambda}(n)$, i.e., in the ith iteration we have

$$\check{\mathbf{G}}^i(n) = \mathcal{S}^F_{\rho,\tilde{\mathbf{x}}_\tau(n)}\left(\left(\check{\mathbf{z}}^{i-1}(n) + \boldsymbol{\lambda}^{i-1}(n)\right)^{\mathsf{H}}\right) \quad (14)$$

$$\check{\mathbf{z}}^i(n) = \mathcal{S}^C_{\rho,\mathbf{I}}\left(\left(\check{\mathbf{G}}^i(n)\right)^H \tilde{\mathbf{x}}_\tau(n) - \boldsymbol{\lambda}^{i-1}(n)\right) \quad (15)$$

$$\boldsymbol{\lambda}^i(n) = \boldsymbol{\lambda}^{i-1}(n) + \check{\mathbf{z}}^i(n) - \left(\check{\mathbf{G}}^i(n)\right)^H \tilde{\mathbf{x}}_\tau(n) \quad (16)$$

where $\mathcal{S}^f_{\rho,\mathbf{A}}(\mathbf{v}) = \arg\min_{\mathbf{x}} f(\mathbf{x}) + \frac{\rho}{2}\|\mathbf{A}^{\mathsf{H}}\mathbf{x} - \mathbf{v}\|_2^2$ [25]. Since F is quadratic, $\check{\mathbf{G}}^i(n)$ in (14) is given in a closed form, similar to (9), and can be efficiently computed by applying the matrix inversion lemma [16]. Computing $\check{\mathbf{z}}^i(n)$ in (15) corresponds to a projection on the constraint set, i.e., clipping of the magnitudes. The complete iterative procedure is given in Algorithm 3. The iterative updates in Algorithm 3 can be interpreted as an iterative correction of the unconstrained filter $\hat{\mathbf{G}}(n)$ to obtain the constrained filter $\check{\mathbf{G}}(n)$ satisfying the inequality constraint in (11).

Algorithm 3: ADMM for the Constrained Problem in (11). Operations in Step 6 are Applied Element-Wise.

input: $\mathbf{x}(n)$, $\hat{\mathbf{G}}(n)$, $\hat{\mathbf{u}}(n)$, $\hat{\mathbf{Q}}(n)$ estimated using Algorithm 1, $\hat{\boldsymbol{\sigma}}_u(n) = \min\{\hat{\boldsymbol{\sigma}}_r(n), |\mathbf{x}(n)|\}$ estimated using Algorithm 2

parameters: penalty parameter ρ, number of iterations I

1: initialize: $\check{\mathbf{z}}^0(n) = \mathbf{0}, \boldsymbol{\lambda}^0(n) = \mathbf{0}$

2: $\check{\mathbf{k}}(n) = \frac{\hat{\mathbf{Q}}^{-1}(n)\tilde{\mathbf{x}}_\tau(n)}{\frac{2}{\rho} + \tilde{\mathbf{x}}_\tau^{\mathsf{H}}(n)\hat{\mathbf{Q}}^{-1}(n)\tilde{\mathbf{x}}_\tau(n)}$

3: **for** $i = 1, \ldots, I$ **do**

4: $\quad \check{\mathbf{G}}^i(n) = \hat{\mathbf{G}}(n) + \check{\mathbf{k}}(n)\left[\check{\mathbf{z}}^{i-1}(n) + \boldsymbol{\lambda}^{i-1}(n) - \hat{\mathbf{u}}(n)\right]^{\mathsf{H}}$

5: $\quad \check{\mathbf{u}}^i(n) = \left(\check{\mathbf{G}}^i(n)\right)^{\mathsf{H}} \tilde{\mathbf{x}}_\tau(n)$

6: $\quad \check{\mathbf{z}}^i(n) = \min\left\{\frac{\hat{\boldsymbol{\sigma}}_u(n)}{\left|\check{\mathbf{u}}^i(n) - \boldsymbol{\lambda}^{i-1}(n)\right|}, 1\right\}\left(\check{\mathbf{u}}^i(n) - \boldsymbol{\lambda}^{i-1}(n)\right)$

7: $\quad \boldsymbol{\lambda}^i(n) = \boldsymbol{\lambda}^{i-1}(n) + \check{\mathbf{z}}^i(n) - \check{\mathbf{u}}^i(n)$

8: **end for**

output: $\check{\mathbf{u}}^I(n), \mathbf{z}^I(n)$

9: $\hat{\mathbf{d}}(n) = \mathbf{x}(n) - \check{\mathbf{u}}^I(n)$ ▷ u-variant

10: $\hat{\mathbf{d}}(n) = \mathbf{x}(n) - \check{\mathbf{z}}^I(n)$ ▷ z-variant

The complexity of Algorithm 3 is quadratic and is dominated by the computation of the gain vector $\check{\mathbf{k}}(n)$ with $\mathcal{O}(M^2L_g^2)$ operations, equivalent to the computation of the gain vector in Algorithm 1, with the ADMM complexity being $\mathcal{O}(M^2L_g)$ operations per iteration. Note that the equality constraint in (12) will be satisfied as $i \rightarrow \infty$ [19], while for a relatively small number of iterations $\check{\mathbf{u}}^i(n)$ and $\check{\mathbf{z}}^i(n)$ will not necessarily be equal, and only the latter will definitely satisfy the constraint in (11). Therefore, it is possible to use either $\check{\mathbf{u}}^i(n)$, obtained by using linear filtering, or the splitting variable $\check{\mathbf{z}}^i(n)$ as an estimate of the undesired component for dereverberation, leading to two variants of the dereverberation algorithm. As a special case, the method proposed in [20] can be obtained by using the u-variant of Algorithm 3 with $p = 0$ in Algorithm 1.

V. Simulations

We investigate the performance of the adaptive sparse MCLP method (denoted as ADA) given in Algorithm 1 and the two proposed variants of the constrained adaptive sparse MCLP method (denoted as cADA-u and cADA-z) given in Algorithm 3. The STFT is computed using a 64-ms Hann window with 16-ms shift, and the prediction delay in (3) is set to $\tau = 2$ (corresponding to 32 ms in the time domain). The PSDs of the desired component and the late reverberation are estimated by using Algorithm 2 with $\alpha = 0.5, T_d = 50$ ms. The ADMM penalty parameter is set to $\rho = 10^3$ and the number of iterations is $I = 25$. Please note that using a smaller number of iterations (in the order 10) would not result in a significant performance difference. The forgetting factor γ is varied between 0.75 and 0.999. The prediction filter $\hat{\mathbf{G}}$ is initialized with zeros and $\hat{\mathbf{Q}}$ as the identity matrix. Before processing the test signal an additional 5 s of the speech signal are processed to reduce initialization effects. The performance is evaluated in terms of the frequency-weighted segmental SNR (fwsSNR) [26] and the perceptual evaluation of speech quality (PESQ) measure [27] using the anechoic/close-talking speech signal as the reference. The sampling frequency is $f_s = 16$ kHz. Exemplary audio samples are available online.[1]

[1] www.sigproc.uni-oldenburg.de/audio/ante/spl_2016/audio.html

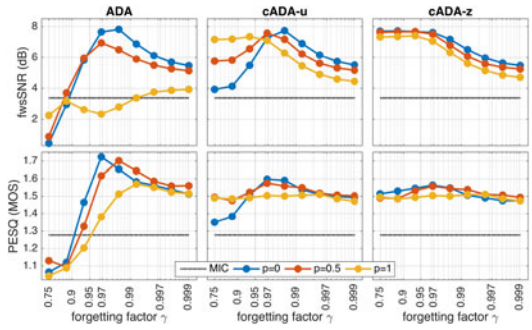


Fig. 1. Instrumental measures versus forgetting factor γ for ADA (left), cADA-u (center), and cADA-z (right).

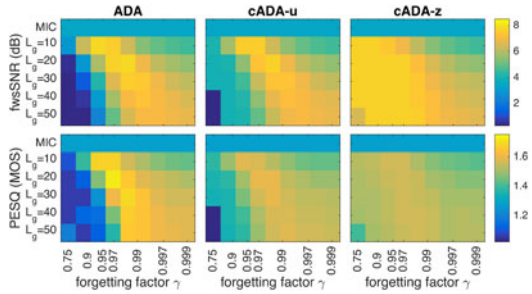


Fig. 2. Instrumental measures for different filter lengths L_g versus forgetting factor γ for ADA (left), cADA-u (center), and cADA-z (right) with $p = 0$.

In the first two experiments we used $M = 2$ room impulse responses measured in a room with $T_{60} \approx 0.7$ s [4], with an intermicrophone distance of approximately 14 cm, and a source–microphone distance of approximately 2 m. The speech source was alternated between two positions, located 45° left and 45° right of the broadside direction. The clean speech signal contained six utterances with a total length of approximately 18 s [28] and the microphone signals were generated by alternating the source position for each utterance. In the first experiment, we evaluate the performance of the considered methods for different shape parameters p with the filter length set to $L_g = 20$. The instrumental measures are shown in Fig. 1. It can be observed that ADA performs similarly for $p = 0$ and $p = 0.5$, with $p = 1$ leading to a decreased performance. In general, the performance of ADA strongly depends on γ, even resulting in a significant performance degradation with respect to the microphone signal for small values of γ due to overestimation of the undesired component. On the other hand, although the proposed cADA-u and cADA-z achieve a somewhat lower best-case performance (in terms of PESQ) than the optimally tuned ADA, both cADA-u and cADA-z are much more robust to the values of the forgetting factor γ and the shape parameter p, since the constraint prevents overestimation of the undesired component. This can be advantageous in practical applications, where the dynamics of the acoustic scenario and the optimal forgetting factor γ are in general unknown.

In the second experiment, we evaluate the performance for different filter lengths L_g with the shape parameter set to $p = 0$. The instrumental measures are shown in Fig. 2. It can be observed that ADA becomes very sensitive to small forgetting factors as the filter length L_g increases. More specifically, combining a large L_g with a small γ results in significant distortions and instability of the output. While cADA-u is less influenced, the performance still drops for small γ, especially for large L_g when the matrix $\hat{\mathbf{Q}}(n)$ becomes ill-conditioned. Finally, it can be observed that cADA-z is quite robust with respect to the the filter length and forgetting factor, since the constraint prevents overestimation and possible instability of the output.

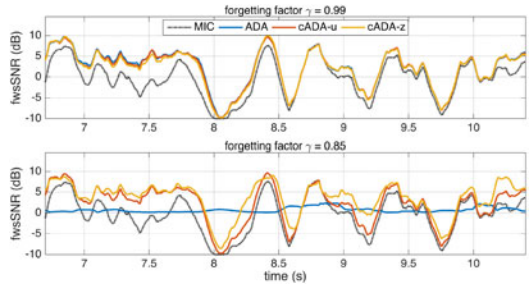


Fig. 3. Smoothed fwsSNR for a moving speaker versus time with $\gamma = 0.99$ (top) and $\gamma = 0.85$ (bottom). The speaker starts walking around 8 s.

TABLE I
PESQ VALUES FOR THE EXPERIMENT WITH A MOVING HUMAN SPEAKER

γ	MIC	ADA	cADA-u	cADA-z
0.99	1.32	1.49	1.47	1.46
0.85	1.32	1.18	1.43	1.43

In the third experiment, we used a recording of a moving human speaker with $M = 2$ microphones with intermicrophone distance of approximately 11 cm in a room with $T_{60} \approx 750$ ms, containing some background noise (cf. [9]). The total length of the recording is approximately 42 s, where the speaker is first static and then starts walking around 8 s. We have set the parameters as $L_g = 20$, $p = 0$, and $\gamma \in \{0.99, 0.85\}$. An excerpt of the frame-wise values of the fwsSNR (averaged across 15 frames) is shown in Fig. 3, while the overall PESQ values are given in Table I. On the one hand, it can be observed in Fig. 3 that for a relatively large forgetting factor $\gamma = 0.99$ all algorithms perform similarly, resulting in improvements compared to the microphone signal for the static part and relatively small improvements for the dynamic part. On the other hand, for a relatively small forgetting factor $\gamma = 0.85$, the unconstrained algorithm ADA results in excessive speech cancellation due to overestimation of the undesired component, which is also noticeable from the PESQ value in Table I. However, the constrained algorithms result in performance improvements for the dynamic part, with cADA-z performing better than cADA-u in terms of fwsSNR.

VI. CONCLUSION

In this letter, we have presented a novel adaptive speech dereverberation method based on constrained sparse MCLP, where a statistical model of late reverberation has been used to constrain the power of the MCLP-based estimate of the undesired reverberant component. The resulting constrained optimization problem is solved by using ADMM, resulting in an efficient implementation. Simulation results show that both proposed increase the robustness with respect to the forgetting factor and the filter length, with the cADA-z variant outperforming the cADA-u variant. Hence, the proposed methods improve the performance of MCLP-based dereverberation in dynamic scenarios, i.e., when the prediction filters need to adapt quickly.

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] T. Yoshioka "Making machines understand us in reverberant rooms: Robustness against reverberation for automatic speech recognition," *IEEE Signal Process. Mag.*, vol. 29, no. 6, pp. 114–126, Nov. 2012.

[3] P. A. Naylor and N. D. Gaubitch, *Speech Dereverberation*. Berlin, Germany: Springer-Verlag, 2010.

[4] K. Kinoshita "A summary of the REVERB challenge: State-of-the-art and remaining challenges in reverberant speech processing research," *EURASIP J. Adv. Signal Process.*, vol. 2016, no. 1, pp. 1–19, 2016.

[5] 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, Language Process.*, vol. 18, no. 7, pp. 1717–1731, Sep. 2010.

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

[7] A. Jukić, T. van Waterschoot, T. Gerkmann, and S. Doclo, "Multi-channel linear prediction-based speech dereverberation with sparse priors," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 23, no. 9, pp. 1509–1520, Sep. 2015.

[8] 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, Language Process.*, vol. 22, no. 8, pp. 1320–1335, Aug. 2014.

[9] B. Schwartz, S. Gannot, and E. Habets, "Online speech dereverberation using Kalman filter and EM algorithm," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 23, no. 2, pp. 394–406, Feb. 2015.

[10] M. Triki and D. T. M. Slock, "Delay and predict equalization for blind speech dereverberation," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, Toulouse, France, May 2006, pp. V-97–V-100.

[11] 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, Language Process.*, vol. 17, no. 4, pp. 534–545, May 2009.

[12] A. Jukić, T. van Waterschoot, T. Gerkmann, and S. Doclo, "A general framework for multi-channel speech dereverberation exploiting sparsity," in *Proc. AES 60th Int. Conf.*, Leuven, Belgium, Feb. 2016, pp. 1–8.

[13] T. Dietzen, A. Spriet, W. Tirry, S. Doclo, M. Moonen, and T. van Waterschoot, "Partitioned block frequency domain Kalman filter for multi-channel linear prediction based blind speech dereverberation," in *Proc. Int. Workshop Acoust. Echo Noise Control*, Xi'an, China, Sep. 2016, pp. 1–5.

[14] T. Yoshioka, H. Tachibana, T. Nakatani, and M. Miyoshi, "Adaptive dereverberation of speech signals with speaker-position change detection," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, Taipei, Taiwan, Apr. 2009, pp. 3733–3736.

[15] T. Yoshioka and T. Nakatani, "Dereverberation for reverberation-robust microphone arrays," in *Proc. Eur. Signal Process. Conf.*, Marrakech, Morocco, Sep. 2013, pp. 1–5.

[16] S. Haykin, *Adaptive Filter Theory*, 3rd ed. Englewood Cliffs, NJ, USA: Prentice-Hall, 2013.

[17] J. D. Polack, "La Transmission De L'energie Sonore Dans Les Salles," Ph.D. dissertation, *Université du Maine, Le Mans, France*, 1988.

[18] K. Lebart, J. M. Boucher, and P. N. Denbigh, "A new method based on spectral subtraction for speech dereverberation," *Acta Acust.*, vol. 87, no. 3, pp. 359–366, May/Jun. 2001.

[19] S. Boyd, N. Parikh, E. Chu, B. Peleato, and J. Eckstein, "Distributed optimization and statistical learning via the alternating direction method of multipliers," *Found. Trends Mach. Learn.*, vol. 3, no. 1, pp. 1–122, 2011.

[20] A. Jukić, Z. Wang, T. van Waterschoot, T. Gerkmann, and S. Doclo, "Constrained multi-channel linear prediction for adaptive speech dereverberation," in *Proc. Int. Workshop Acoust. Echo Noise Control*, Xi'an, China, Sep. 2016, pp. 1–5.

[21] A. Jukić, T. van Waterschoot, T. Gerkmann, and S. Doclo, "Group sparsity for MIMO speech dereverberation," in *Proc. IEEE Workshop Appl. Signal Process. Audio Acoust.*, New Paltz, NY, USA, Oct. 2015, pp. 1–5.

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

[23] 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, Jun. 2009.

[24] T. K. Akino, "Optimum-weighted RLS channel estimation for rapid fading MIMO channels," *IEEE Trans. Wirel. Commun.*, vol. 7, no. 11, pp. 4248–4260, Nov. 2008.

[25] N. Parikh and S. Boyd, "Proximal algorithms," *Found. Trends Mach. Learn.*, vol. 1, no. 3, pp. 127–239, 2014.

[26] Y. Hu and P. C. Loizou, "Evaluation of objective quality measures for speech enhancement," *IEEE/ACM Trans. Audio Speech Language Process.*, vol. 16, no. 1, pp. 229–238, Jan. 2008.

[27] 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," *ITU-T Recommendation P.862, Int. Telecommun. Union, Geneva*, 2001.

[28] T. Robinson, J. Fransen, D. Pye, J. Foote, and S. Renals, "WSJCAM0: A british english speech corpus for large vocabulary continuous speech recognition," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, Detroit, MI, USA, May 1995, pp. 81–84.