KU LEUVEN

STADIUS
Center for Dynamical Systems, Signal Processing and Data Analytics

Citation/Reference	Ante Jukić, Toon van Waterschoot, Timo Gerkmann, and Simon Doclo **A general framework for incorporating time-frequency domain sparsity in multi-channel speech dereverberation** *J. Audio Eng. Soc.*, vol. 65, no. 1/2, pp. 17-30, Jan./Feb. 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/10.17743/jaes.2016.0064
Journal homepage	http://www.aes.org/journal/
Author contact	toon.vanwaterschoot@esat.kuleuven.be *+ 32 (0)16 321927*
IR	ftp://ftp.esat.kuleuven.be/pub/SISTA/vanwaterschoot/abstracts/16-192.html

(article begins on next page)

Journal of the Audio Engineering Society
Vol. 65, No. 1/2, January/February 2017 (© 2017)
DOI: https://doi.org/10.17743/jaes.2016.0064

A General Framework for Incorporating Time-Frequency Domain Sparsity in Multichannel Speech Dereverberation

ANTE JUKIĆ[1], **TOON VAN WATERSCHOOT**,[2] *AES Associate Member*,
(ante.jukic@uni-oldenburg.de) (toon.vanwaterschoot@esat.kuleuven.be)

TIMO GERKMANN,[3] *AES Member*, **AND SIMON DOCLO**,[1] *AES Associate Member*
(gerkmann@informatik.uni-hamburg.de) (simon.doclo@uni-oldenburg.de)

[1] *University of Oldenburg, Department of Medical Physics and Acoustics, and Cluster of Excellence Hearing4All, Oldenburg, Germany*
[2] *KU Leuven, Department of Electrical Engineering (ESAT-STADIUS/ETC), Leuven, Belgium*
[3] *University of Hamburg, Department of Informatics, Hamburg, Germany*

Blind multichannel speech dereverberation methods based on multichannel linear prediction (MCLP) estimate the dereverberated speech component without any knowledge of the room acoustics by estimating and subtracting the undesired reverberant component from the reference microphone signal. In this paper we present a general framework for incorporating sparsity in the time-frequency domain into MCLP-based speech dereverberation. The presented framework enables to use either a wideband or a narrowband signal model with either an analysis or a synthesis sparsity prior for the desired speech component and generalizes state-of-the-art MCLP-based speech dereverberation methods, which is shown both analytically as well as using simulations.

0 INTRODUCTION

Recordings of a speech signal in an enclosed space with microphones placed at a distance from the speech source are typically corrupted with reverberation, caused by reflections against surfaces and objects within the enclosure. While some amount of reverberation can be beneficial, strong reverberation is typically problematic for speech communication applications, resulting in a degraded speech intelligibility and automatic speech recognition performance [1, 2]. Effective speech dereverberation is, hence, a prerequisite for many applications, such as hands-free telephony, voice-based human-machine interfaces, and hearing aids. During the past decades many dereverberation approaches have been developed [3, 4], aiming to remove the unwanted reverberant component from the recorded microphone signals, while at the same time preserving the desired speech component.

In general, multi-microphone techniques are more appealing than single-microphone techniques, since they enable to exploit spatial information in addition to spectro-temporal information. Widely investigated multi-microphone techniques that can achieve perfect dereverberation are based on inverse filtering [5]. Inverse filtering methods can be broadly classified into indirect and direct methods. Indirect methods consist of two steps: in the first step the acoustic transfer functions (ATFs) between the speech source and the microphones are estimated, e.g., using blind system identification [6]; in the second step equalization filters are designed based on the estimated ATFs [5]. Although robust multichannel (MC) equalization techniques have been proposed, in practice their dereverberation performance is often limited due to ATF estimation errors, possibly causing severe distortions in the output signals [7–9]. Direct methods, which will be considered in this paper, estimate dereverberation filters without any knowledge of the ATFs [10–12]. A popular class of direct inverse filtering methods is based on multichannel linear prediction (MCLP), which estimate the desired speech component by predicting the reverberant component through linear filtering of (delayed) reverberant microphone signals and subtracting it from the reference microphone signal.

It is well known that speech signals have a sparse representation in the time-frequency (TF) domain, due to the combined effects of speech pauses and the spectral shape and harmonic structure of speech signals [13, 14]. In the presence of reverberation, the recorded microphone signals exhibit however a lower level of sparsity than the anechoic speech signal, due to spectro-temporal smearing of the speech energy [14]. This property has been exploited for MCLP-based speech dereverberation, by

estimating the desired speech component that is more sparse than the recorded microphone signal [10, 12].

The main goal of this paper is to present a general framework for blind speech dereverberation exploiting sparsity of the speech signal in the TF domain. To model the observed signals, we use a wideband MCLP-based signal model in the time domain, or a narrowband MCLP-based signal model in the TF domain. We derive several optimization problems, combining either a wideband or a narrowband signal model with a sparse analysis or synthesis prior for the speech signal coefficients [15], which can be solved using alternating direction method of multipliers (ADMM) [16]. To transform the time-domain signal into the TF domain we will use the short-time Fourier transform (STFT), although the proposed framework supports general TF transforms by using corresponding analysis/synthesis operators, e.g., adaptive non-stationary Gabor transforms [17]. To promote sparsity, we will use the commonly used weighted ℓ_1-norm, although other sparsity-promoting functions can be used in the presented framework. In addition to the locally computed weights for the weighted ℓ_1-norm [18], we also consider structured weights by using a neighborhood in the TF domain [19] or a low-rank approximation of the speech power spectrogram [20]. The effectiveness of the considered speech dereverberation methods is evaluated using simulations. It is shown that the ADMM-based methods result in a competitive performance and may lead to improvements in certain cases, e.g., for a small number of reweighting iterations. While wideband methods offer more flexibility, it is shown that the narrowband methods achieve a good performance with a relatively low complexity, making them more relevant for practical applications. Moreover, including additional structure in the TF domain, e.g., by using structured weights, can be used to improve the performance of sparsity-based dereverberation methods. Some preliminary results have been presented in [21].

The paper is organized as follows. In Sec. 1 the signal models for the MCLP-based speech dereverberation are introduced. Several optimization problems are formulated in Sec. 2, followed by a discussion on the selection of the sparsity-promoting cost function and the relationship to the existing methods. Using simulations, the performance of all considered methods is evaluated in Sec. 3.

1 SIGNAL MODEL

We consider a fixed source-array geometry with a single speech source in a reverberant environment and M microphones. In the time-domain the m-th microphone signal $x_m(t)$ can be modeled as the convolution of the anechoic speech signal $s(t)$ with a room impulse response (RIR) $r_m(t)$ of length L_r, i.e., $x_m(t) = r_m(t) * s(t)$. The reference microphone signal $x_{\text{ref}}(t)$ can be decomposed into a desired component $d(t)$ and an undesired component $u(t)$ as

$$x_{\text{ref}}(t) = \underbrace{\sum_{l=0}^{L_\tau-1} r_{\text{ref}}(l)s(t-l)}_{d(t)} + \underbrace{\sum_{l=L_\tau}^{L_r-1} r_{\text{ref}}(l)s(t-l)}_{u(t)}, \quad (1)$$

where the desired component is obtained by convolving the anechoic speech signal with the early part of the RIR (consisting of the first L_τ samples) and the undesired component is obtained by convolving the anechoic speech signal with the late part of the RIR (consisting of the remaining samples). The goal of speech dereverberation is then to recover the desired component $d(t)$ consisting of the anechoic speech signal and early reflections, which can be beneficial for speech intelligibility [22]. When multiple microphones are available, it has been shown that in principle perfect dereverberation can be achieved using the multiple-input/output inverse theorem (MINT) [5]. Assuming that the RIRs do not share common zeros and using inverse filters $h_m(t)$ of length $L_h \geq (L_r - 1)/(M-1)$, the anechoic speech signal can be obtained as $s(t) = \sum_{m=1}^{M} h_m(t) * x_m(t)$. By using this, it can be shown that the undesired component $u(t)$ in Eq. (1) can be obtained by convolving the delayed microphone signals with the prediction filters $g_m(t)$, i.e., as

$$u(t) = \sum_{m=1}^{M} \sum_{l=0}^{L_g-1} g_m(l)x_m(t-L_\tau-l), \quad (2)$$

where $g_m(l)$ is the prediction filter related to the m-th microphone [23, 10]. The expression in Eq. (2) ensures that the prediction filters $g_m(l)$ for estimation of the undesired component $u(t)$ exist and can be computed when the RIRs $r_m(t)$ are perfectly known [10]. However, in this paper we aim to estimate the prediction filters blindly, without using any knowledge about the RIRs or the source-array geometry. The prediction delay L_τ should ensure that the direct speech component in the reference microphone cannot be predicted using Eq. (2), i.e., that subtracting the predicted undesired component does not destroy the short-time autocorrelation of the speech signal [24, 10]. If the inter-microphone distances are relatively small (as assumed in this paper), the relative delays between the reference microphone and the other microphones are rather small, i.e., in the order of ms, for all possible source positions. In this case, the required prediction delay only depends on the short-term autocorrelation of the speech signal. A common practice in MCLP-based dereverberation is, hence, to set the prediction delay in the range of 30 to 40 ms [24, 10]. It has been shown in [10] that with a suitable prediction delay and given enough samples, subtracting the undesired component in Eq. (2) from the reference microphone signal does not change the direct component, while possibly altering the early reflections. A block scheme of an MCLP-based speech dereverberation dereverberation system is depicted in Fig. 1.

In the following, we assume that a batch of T time-domain samples is available, where T is much larger than the number of the unknown filter coefficients ML_g. Eq. (2) can then be written in vector form as $\mathbf{u} = \mathbf{X}\mathbf{g}$, where $\mathbf{u} = [u(1), \ldots, u(T)]^\mathsf{T}$ is the undesired component (with $.^\mathsf{T}$ denoting the transpose operator), $\mathbf{g} \in \mathbb{R}^{ML_g}$ is a MC prediction filter composed of the filter coefficients for all channels, i.e.,

$$\mathbf{g} = [\mathbf{g}_1^\mathsf{T}, \ldots, \mathbf{g}_M^\mathsf{T}]^\mathsf{T} \in \mathbb{R}^{ML_g} \quad (3)$$

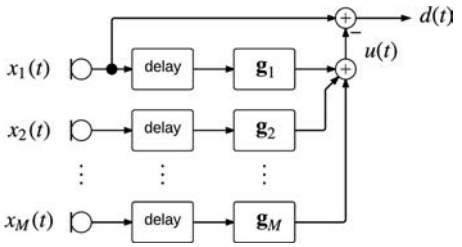


Fig. 1. Block scheme of an MCLP-based dereverberation system with the first microphone selected as the reference.

and

$$\mathbf{X} = [\mathbf{X}_1, \ldots, \mathbf{X}_M] \in \mathbb{R}^{T \times ML_g} \tag{4}$$

is an MC convolution matrix with $\mathbf{X}_m \in \mathbb{R}^{T \times L_g}$ being the convolution matrix of $x_m(t)$ delayed by L_τ samples, i.e.,

$$\mathbf{X}_m = \begin{bmatrix} 0 & 0 & \ldots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ x_m(1) & 0 & \ddots & \vdots \\ x_m(2) & x_m(1) & \ddots & \vdots \\ \vdots & x_m(2) & \ddots & 0 \\ \vdots & \vdots & \ddots & x_m(1) \\ \vdots & \vdots & \ddots & \vdots \\ x_m(T-L_\tau) & \ldots & \ldots & x_m(T-L_\tau-L_g+1) \end{bmatrix}. \tag{5}$$

The wideband signal model in Eq. (1) can hence be written in vector form as

$$\mathbf{x}_{\text{ref}} = \mathbf{d} + \mathbf{X}\mathbf{g}, \tag{6}$$

where $\mathbf{x}_{\text{ref}}$ and $\mathbf{d}$ are defined similarly as $\mathbf{u}$.

While the wideband model in Eq. (6) perfectly holds when the MINT conditions are fulfilled, the prediction filter $\mathbf{g}$ can be very long and dereverberation based on the wideband model can be computationally demanding [10]. In order to reduce the length of the filters, the wideband model in Eq. (6) is commonly approximated in the STFT domain [10–12]. Let $\mathbf{\Psi} \in \mathbb{C}^{T \times KN}$, with $KN > T$, denote the overcomplete frame [25] corresponding to the STFT, relating a time-domain signal with T samples to KN coefficients in the TF domain, corresponding to N time blocks and K frequency bins. The TF coefficients of the time-domain signal $\mathbf{d}$ can be obtained by applying the analysis transform as $\tilde{\mathbf{d}} = \mathbf{\Psi}^{\mathsf{H}}\mathbf{d} \in \mathbb{C}^{KN}$ (with $.^{\mathsf{H}}$ denoting conjugate transpose operator). We will use $\tilde{\mathbf{d}}_k \in \mathbb{C}^N$ to denote a vector containing the N TF coefficients in the k-th frequency bin and $\tilde{d}_{k,n} \in \mathbb{C}$ to denote a single coefficient.[1] For simplicity, we assume that $\mathbf{\Psi}\mathbf{\Psi}^{\mathsf{H}} = \mathbf{I}$ where $\mathbf{I}$ is the identity matrix implying that the inverse STFT can be obtained as $\mathbf{d} = \mathbf{\Psi}\tilde{\mathbf{d}}$ (i.e., $\mathbf{\Psi}$ is a Parseval tight frame [25]). The narrowband signal model is obtained by approximating the time-domain convolution in Eq. (6) in each frequency bin independently, i.e.,

$$\tilde{\mathbf{x}}_{\text{ref},k} = \tilde{\mathbf{d}}_k + \tilde{\mathbf{X}}_k\tilde{\mathbf{g}}_k, \tag{7}$$

where $\tilde{\mathbf{X}}_k \in \mathbb{C}^{N \times M\tilde{L}_g}$ is a MC convolution matrix obtained from the coefficients in the k-th frequency bin delayed by $\tilde{L}_\tau$ time blocks. The prediction filters $\tilde{\mathbf{g}}_k \in \mathbb{C}^{M\tilde{L}_g}$ for the narrow-band model Eq. (7) are typically much shorter than their time-domain counterpart (i.e., $\tilde{L}_g \ll L_g$) and are estimated independently for each frequency [10].

2 TIME-FREQUENCY DOMAIN SPARSITY FOR DEREVERBERATION

Sparsity of speech signals in the TF domain has been extensively exploited in source separation [26, 14, 27], audio inpainting [28], and dereverberation [29, 10, 12]. In general, sparsity of a vector (signal) is related to the magnitude of its elements (samples), e.g., a signal with only a small number of samples with significant magnitude is approximately sparse. Sparsity has typically been used in the following two paradigms: the synthesis sparsity and analysis sparsity [15]. Synthesis sparsity is based on the assumption that a signal can be expressed as a linear combination of a relatively small number of elements from a dictionary. In the considered scenario, this would imply that the time-domain desired speech signal $\mathbf{d}$ can be represented with a relatively small number of estimated coefficients in the TF domain, i.e., $\mathbf{d} \approx \mathbf{\Psi}\tilde{\mathbf{d}}$ with a sparse $\tilde{\mathbf{d}}$. Analysis sparsity is based on the assumption that a signal has a sparse representation when a suitable analysis operator is applied. In the considered scenario, this would imply that the estimated time-domain speech signal $\mathbf{d}$ has a sparse STFT representation, i.e., that $\tilde{\mathbf{d}} = \mathbf{\Psi}^{\mathsf{H}}\mathbf{d}$ is sparse. While both paradigms assume sparsity of the TF coefficients, the synthesis sparsity leads to estimation of the TF coefficients, while the analysis sparsity leads to estimation of the time-domain signal. The paradigms are equivalent only if the analysis operator is equal to the inverse of the synthesis operator [15]. In the considered case this is not fulfilled since the STFT frame $\mathbf{\Psi}$ is overcomplete (i.e., redundant, since $KN > T$) and thus not invertible, and hence the two paradigms differ.

In the remainder of this section we present different formulations of MCLP-based speech dereverberation exploiting sparsity in the TF domain for a fixed sparsity-promoting cost function P. In Secs. 2.1 and 2.2 we first consider the wideband model in Eq. (6) with analysis and synthesis sparsity prior, respectively. In Sec. 2.3 we then consider the narrowband model in Eq. (7) with the synthesis sparsity prior. All obtained optimization algorithms can be efficiently solved using the ADMM algorithm [16], which is briefly reviewed in Appendix A. In Secs. 2.4 and 2.5 we discuss the selection of the sparsity promoting cost function and the relationship of the existing algorithms with the proposed formulations.

[1] In the remainder of the paper all variables related to the TF domain will be denoted with $\tilde{(.)}$.

2.1 Wideband Model and Analysis Sparsity

In this section we consider the problem of speech dereverberation with the analysis sparsity prior and the wideband model in Eq. (6). We estimate the desired speech signal $\mathbf{d}$ in the time domain and enforce its TF coefficients to be sparse in terms of the cost function P, leading to the following optimization problem

$$\begin{array}{ll} \min_{\mathbf{d},\mathbf{g}} & P\left(\boldsymbol{\Psi}^{\mathsf{H}}\mathbf{d}\right) \\ \text{subject to} & \mathbf{d}+\mathbf{X}\mathbf{g}=\mathbf{x}_{\text{ref}}. \end{array} \tag{8}$$

By applying the ADMM algorithm (cf., Appendix A), the obtained problem can be solved using the following iterative update rules

$$\begin{array}{ll} \mathbf{d} & \leftarrow \arg\min_{\mathbf{d}} P\left(\boldsymbol{\Psi}^{\mathsf{H}}\mathbf{d}\right)+\frac{\rho}{2}\|\mathbf{d}+\mathbf{X}\mathbf{g}-\mathbf{x}_{\text{ref}}+\boldsymbol{\mu}\|_2^2, \\ \mathbf{g} & \leftarrow \arg\min_{\mathbf{g}} \|\mathbf{d}+\mathbf{X}\mathbf{g}-\mathbf{x}_{\text{ref}}+\boldsymbol{\mu}\|_2^2, \\ \boldsymbol{\mu} & \leftarrow \boldsymbol{\mu}+\gamma\left(\mathbf{d}+\mathbf{X}\mathbf{g}-\mathbf{x}_{\text{ref}}\right), \end{array} \tag{9}$$

where ρ is the penalty parameter, $\boldsymbol{\mu}$ is the dual variable, and γ is a parameter used for faster convergence (cf., Appendix A). The update for the time domain signal $\mathbf{d}$ corresponds to a generalized Lasso problem [16] and can be efficiently solved using ADMM algorithm as shown in Appendix B. Note that in this case the ADMM algorithm for solving the generalized Lasso is "nested" inside the ADMM algorithm for solving Eq. (8).

The update for the filter $\mathbf{g}$ is a least-squares problem with a closed-form solution given as

$$\mathbf{g} \leftarrow \left(\mathbf{X}^{\mathsf{T}}\mathbf{X}\right)^{-1}\mathbf{X}^{\mathsf{T}}\left(\mathbf{x}_{\text{ref}}-\mathbf{d}-\boldsymbol{\mu}\right)=\mathbf{g}_{\ell_2}-\mathbf{g}_{\text{iter}}, \tag{10}$$

where $\mathbf{g}_{\ell_2}=\left(\mathbf{X}^{\mathsf{T}}\mathbf{X}\right)^{-1}\mathbf{X}^{\mathsf{T}}\mathbf{x}_{\text{ref}}$ is an iteration-independent term, and $\mathbf{g}_{\text{iter}}=\left(\mathbf{X}^{\mathsf{T}}\mathbf{X}\right)^{-1}\mathbf{X}^{\mathsf{T}}\left(\mathbf{d}+\boldsymbol{\mu}\right)$ is an iteration-dependent correction term. The iteration-independent term $\mathbf{g}_{\ell_2}$ is equal to the closed-form solution for the ℓ_2-norm as the cost function in Eq. (8), i.e., $P(\boldsymbol{\Psi}^{\mathsf{H}}\mathbf{d})=\|\boldsymbol{\Psi}^{\mathsf{H}}\mathbf{d}\|_2^2$. From earlier work it is known that such filters typically do not perform well for dereverberation [24, 10, 12]. However, similarly as in [30], the iteration-dependent term $\mathbf{g}_{\text{iter}}$ can be seen as a correction that "sparsifies" the estimate of the desired speech $\mathbf{d}$, which has shown to be crucial for MCLP-based dereverberation [10, 12]. Note that the matrix $\mathbf{X}^{\mathsf{T}}\mathbf{X}$ is the same for all iterations, such that it only needs to be factored once and the factorization can be used for solving the corresponding linear system in subsequent iterations [16]. Moreover, since $\mathbf{X}$ is a block-convolution matrix, both $\mathbf{X}^{\mathsf{T}}\mathbf{X}$ and $\mathbf{X}^{\mathsf{T}}\mathbf{x}_{\text{ref}}$ can be obtained through multichannel correlation. Additionally, the block-Toeplitz structure of $\mathbf{X}^{\mathsf{T}}\mathbf{X}$ can be further exploited to apply a faster solver, similarly as in [30], but generalized to the multichannel case.

2.2 Wideband Model and Synthesis Sparsity

In this section we consider the problem of speech dereverberation with the synthesis sparsity prior and the wideband model in Eq. (6). We estimate the desired speech signal coefficients $\tilde{\mathbf{d}}$ in the TF domain and enforce them to be sparse in terms of the cost function P, leading to the following optimization problem

$$\begin{array}{ll} \min_{\tilde{\mathbf{d}},\mathbf{g}} & P\left(\tilde{\mathbf{d}}\right) \\ \text{subject to} & \boldsymbol{\Psi}\tilde{\mathbf{d}}+\mathbf{X}\mathbf{g}=\mathbf{x}_{\text{ref}}. \end{array} \tag{11}$$

The desired speech signal in the time domain is then obtained by performing the inverse STFT of the estimated coefficients, i.e., $\mathbf{d}=\boldsymbol{\Psi}\tilde{\mathbf{d}}$. By applying the ADMM algorithm (cf., Appendix A), the obtained problem can be solved using the following iterative update rules

$$\begin{array}{ll} \tilde{\mathbf{d}} & \leftarrow \arg\min_{\tilde{\mathbf{d}}} P\left(\tilde{\mathbf{d}}\right)+\frac{\rho}{2}\|\boldsymbol{\Psi}\tilde{\mathbf{d}}+\mathbf{X}\mathbf{g}-\mathbf{x}_{\text{ref}}+\boldsymbol{\mu}\|_2^2 \\ \mathbf{g} & \leftarrow \arg\min_{\mathbf{g}} \|\boldsymbol{\Psi}\tilde{\mathbf{d}}+\mathbf{X}\mathbf{g}-\mathbf{x}_{\text{ref}}+\boldsymbol{\mu}\|_2^2 \\ \boldsymbol{\mu} & \leftarrow \boldsymbol{\mu}+\gamma\left(\boldsymbol{\Psi}\tilde{\mathbf{d}}+\mathbf{X}\mathbf{g}-\mathbf{x}_{\text{ref}}\right), \end{array} \tag{12}$$

where ρ is the penalty parameter, $\boldsymbol{\mu}$ is the dual variable and γ is a parameter used for faster convergence (cf., Appendix A). The update for the STFT coefficients $\tilde{\mathbf{d}}$ corresponds to a Lasso problem [31] and can be efficiently solved using the iterative shrinkage/thresholding algorithm (ISTA), as shown in Appendix C, or using its fast variant (FISTA) [32]. Similarly as in Eq. (10), the update for the prediction filter $\mathbf{g}$ is a least-squares problem with a closed-form solution given as

$$\mathbf{g} \leftarrow \left(\mathbf{X}^{\mathsf{T}}\mathbf{X}\right)^{-1}\mathbf{X}^{\mathsf{T}}\left(\mathbf{x}_{\text{ref}}-\boldsymbol{\Psi}\tilde{\mathbf{d}}-\boldsymbol{\mu}\right)=\mathbf{g}_{\ell_2}-\mathbf{g}_{\text{iter}}, \tag{13}$$

where $\mathbf{g}_{\ell_2}$ is the same iteration-independent term as in Eq. (10), and $\mathbf{g}_{\text{iter}}=\left(\mathbf{X}^{\mathsf{T}}\mathbf{X}\right)^{-1}\mathbf{X}^{\mathsf{T}}\left(\boldsymbol{\Psi}\tilde{\mathbf{d}}+\boldsymbol{\mu}\right)$ is the iteration-dependent term.

2.3 Narrowband Model

In this section we consider the problem of speech dereverberation with the synthesis sparsity prior and the narrowband model in Eq. (7). Similarly as in Sec. 2.2, we estimate the desired speech signal coefficients $\tilde{\mathbf{d}}$ in the TF domain and enforce them to be sparse in terms of the cost function P. In contrast to the wideband model, since the narrowband model is independent across frequencies and assuming that the cost function P is also separable, the speech signal coefficients $\tilde{\mathbf{d}}_k$ can be estimated for each frequency bin k independently. The desired speech signal in the time domain can then be obtained by performing the inverse STFT of the estimated coefficients as $\mathbf{d}=\boldsymbol{\Psi}\tilde{\mathbf{d}}$. The optimization problem in the k-th frequency bin for estimating $\tilde{\mathbf{d}}_k$ can be written as

$$\begin{array}{ll} \min_{\tilde{\mathbf{d}}_k,\tilde{\mathbf{g}}_k} & P\left(\tilde{\mathbf{d}}_k\right) \\ \text{subject to} & \tilde{\mathbf{d}}_k+\tilde{\mathbf{X}}_k\tilde{\mathbf{g}}_k=\tilde{\mathbf{x}}_{\text{ref},k}. \end{array} \tag{14}$$

By applying the ADMM algorithm (cf., Appendix A), the obtained problem can be solved using the following iterative update rules

$$\begin{array}{l} \tilde{\mathbf{d}}_k \leftarrow \arg\min_{\tilde{\mathbf{d}}_k} P\left(\tilde{\mathbf{d}}_k\right)+\frac{\rho}{2}\|\tilde{\mathbf{d}}_k+\tilde{\mathbf{X}}_k\tilde{\mathbf{g}}_k-\tilde{\mathbf{x}}_{\text{ref},k}+\tilde{\boldsymbol{\mu}}_k\|_2^2, \\ \tilde{\mathbf{g}}_k \leftarrow \arg\min_{\tilde{\mathbf{g}}_k} \|\tilde{\mathbf{d}}_k+\tilde{\mathbf{X}}_k\tilde{\mathbf{g}}_k-\tilde{\mathbf{x}}_{\text{ref},k}+\tilde{\boldsymbol{\mu}}_k\|_2^2, \\ \tilde{\boldsymbol{\mu}}_k \leftarrow \tilde{\boldsymbol{\mu}}_k+\gamma\left(\tilde{\mathbf{d}}_k+\tilde{\mathbf{X}}_k\tilde{\mathbf{g}}_k-\tilde{\mathbf{x}}_{\text{ref},k}\right), \end{array} \tag{15}$$

where ρ is the penalty parameter. The update for the STFT coefficients $\tilde{\mathbf{d}}_k$ in the k-th frequency bin is already in the

form of a proximal operator (cf., Eq. (A.4)), and can be immediately written as

$$\tilde{\mathbf{d}}_k \leftarrow S_\rho^P \left(\tilde{\mathbf{x}}_{\mathrm{ref},k} - \tilde{\mathbf{X}}_k \tilde{\mathbf{g}}_k - \tilde{\boldsymbol{\mu}}_k\right), \tag{16}$$

where S_ρ^P is the proximal operator of the cost function P (cf., Eq. (A.4)). Similarly as in Eq. (10) and Eq. (13), the update for the prediction filter $\tilde{\mathbf{g}}_k$ in the k-th frequency bin is a least-squares problem with a closed-form solution given as

$$\tilde{\mathbf{g}}_k \leftarrow \left(\tilde{\mathbf{X}}_k^{\mathsf{H}} \tilde{\mathbf{X}}_k\right)^{-1} \tilde{\mathbf{X}}_k^{\mathsf{H}} \left(\tilde{\mathbf{x}}_{\mathrm{ref},k} - \tilde{\mathbf{d}}_k - \tilde{\boldsymbol{\mu}}_k\right) = \tilde{\mathbf{g}}_{k,\ell_2} - \tilde{\mathbf{g}}_{k,\mathrm{iter}} \tag{17}$$

where $\tilde{\mathbf{g}}_{k,\ell_2}$ is the iteration-independent term, and $\tilde{\mathbf{g}}_{k,\mathrm{iter}} = \left(\tilde{\mathbf{X}}_k^{\mathsf{H}} \tilde{\mathbf{X}}_k\right)^{-1} \tilde{\mathbf{X}}_k^{\mathsf{H}} \left(\tilde{\mathbf{d}}_k + \tilde{\boldsymbol{\mu}}_k\right)$ is the iteration-dependent term. Similarly as before, the matrix $\tilde{\mathbf{X}}_k^{\mathsf{H}} \tilde{\mathbf{X}}_k$ only needs to be factored once and can be used to solve the corresponding linear system in subsequent iterations. Note that this matrix is much smaller than the corresponding matrix in the wide-band model (since $\tilde{L}_g \ll L_g$), and the resulting iterations do not involve STFT analysis/synthesis since all computations are performed in the TF domain.

2.4 Sparsity-Promoting Cost Function

The previously presented methods enforce sparsity of the TF coefficients in terms of the cost function P, i.e., P is applied on the TF-domain coefficients. Hence, an appropriate sparsity-promoting cost function P needs to be selected. Typical cost functions for enforcing sparsity include the ℓ_1-norm, nonconvex ℓ_p-norms with $p \in (0, 1)$, or the ℓ_0-norm (counting the number of non-zero coefficients) and its smoothed variants [33, 34].

The proposed framework can be used with any sparsity-promoting function P, as long as its proximal operator S_ρ^P can be computed (cf., Eq. (A.4)). However, in this work we confine ourselves to the weighted ℓ_1-norm, which is one of the most commonly used sparsity-promoting cost functions [18, 19, 36, 37], and has been shown to be more effective for audio applications than its non-weighted counterpart [36]. The cost function is then defined as

$$P\left(\tilde{\mathbf{d}}\right) = \|\tilde{\mathbf{d}}\|_{\mathbf{w},1} = \sum_{k,n} w_{k,n} |\tilde{d}_{k,n}|, \tag{18}$$

where $\tilde{\mathbf{d}}$ is a vector of coefficients in the TF domain, and $\mathbf{w}$ is a vector of nonnegative weights. The weights $w_{k,n}$ are selected in such a way that the weighted ℓ_1-norm simulates the behavior of the scaling-insensitive ℓ_0-norm [18, 36].

Estimation of a sparse $\tilde{\mathbf{d}}$ using the weighted ℓ_1-norm in Eq. (18) as the cost function is an iterative two-step procedure. First, the weights $\mathbf{w}$ are computed based on the previous estimate of $\tilde{\mathbf{d}}$. Second, an appropriate optimization problem with the cost function in Eq. (18) is solved, and consequently a new estimate of the TF coefficients $\tilde{\mathbf{d}}$ is obtained. All previously presented ADMM-based methods will be employed in such a reweighted procedure in Sec. 3.

The weights $w_{k,n}$ for the weighted ℓ_1-norm in Eq. (18) are typically computed locally, using a single coefficient,

$$w_{k,n} = \frac{\varepsilon}{|\tilde{d}_{k,n}| + \varepsilon}, \tag{19}$$

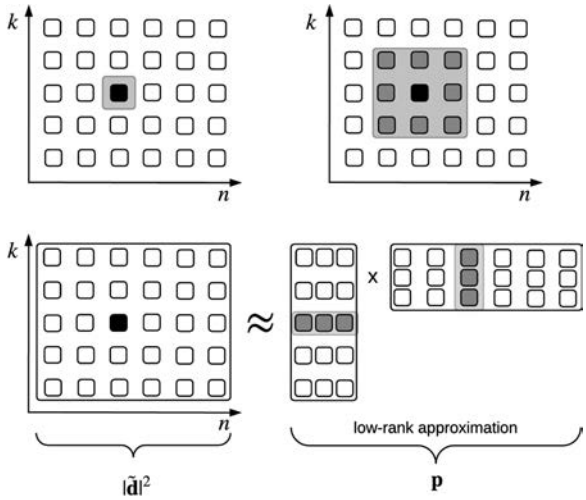


Fig. 2. Computation of the weight $w_{k,n}$ for the coefficient marked with : locally computed weight (top left), weight computed using a neighborhood with dimension 3 across time blocks and frequencies (top right) and weight computed using a low-rank approximation with rank equal to 3 (bottom).

where ε is a small regularization coefficient to prevent division by zero in the denominator and is included in the numerator to ensure that the largest weight is normalized to one [36]. Since in practice the true coefficients $\tilde{d}_{k,n}$ are not available, the weights in Eq. (19) are computed based on an estimate of $\tilde{d}_{k,n}$ from the previous iteration in the reweighting procedure.

To take into account the TF structure of the desired signal, the concept of neighborhoods for shrinkage operators has been introduced in [19], and here we adopt it for computing the weights. Assuming that a neighborhood $\mathcal{N}_{k,n}$ of the coefficient $\tilde{d}_{k,n}$ is defined, the corresponding weight can be computed as a weighted average across the neighborhood, i.e.,

$$w_{k,n} = \frac{\varepsilon}{\sqrt{\sum_{(k',n') \in \mathcal{N}_{k,n}} \eta_{k',n'} |\tilde{d}_{k',n'}|^2} + \varepsilon}, \tag{20}$$

where $\eta_{k',n'}$ are the coefficients of the neighborhood that sum to one. Similarly as in [37, 19], in our simulations we will employ rectangular neighborhoods with equal weights. Alternatively, it is well known that speech spectrograms can be modeled well using a low-rank approximation [38, 39]. Similarly as in [20], the weights can then be obtained by computing a low-rank approximation $\mathbf{p}$ of the power spectrogram $|\tilde{\mathbf{d}}|^2 \in \mathbb{R}_{0+}^{K \times N}$, a nonnegative matrix containing the squared magnitudes of the TF coefficients, i.e., $\mathbf{p} \approx |\tilde{\mathbf{d}}|^2$, and computing the weights as

$$w_{k,n} = \frac{\varepsilon}{\sqrt{p_{k,n}} + \varepsilon}. \tag{21}$$

The three different considered ways of computing weights for Eq. (18) are illustrated in Fig. 2. For illustration we used a 3×3 neighborhood and a rank-3 approximation of $|\tilde{\mathbf{d}}|^2$.

The proximal operator (cf., Eq. (A.4)) for the weighted ℓ_1-norm in Eq. (18) can be computed element-wise using soft thresholding as

$$S_{\rho}^{P}\left(\tilde{d}_{k,n}\right) = \underbrace{\left(1 - \frac{\rho^{-1} w_{k,n}}{|\tilde{d}_{k,n}|}\right)_{+}}_{\text{real-valued gain}} \tilde{d}_{k,n}, \tag{22}$$

where $(G)_+ = \max(G, 0)$ [16]. In the context of speech enhancement [40], the proximal operator in Eq. (22) can be interpreted as applying a real-valued gain on the complex-valued coefficients in $\tilde{\mathbf{d}}$. As noted in [40], in speech enhancement a lower bound $G_{\min}$ on the gain is often introduced, i.e., $(G)_+ = \max(G, G_{\min})$, in order to prevent suppression of small coefficients $\tilde{d}_{k,n}$ to exactly zero. As shown in Appendix D, this corresponds to a cost function P in the form of a Huber function [16], which is quadratic for small magnitudes and equal to a scaled absolute value for large magnitudes, and the transition point depends on the penalty parameter ρ, weight $w_{k,n}$ and the lower bound $G_{\min}$ (cf. (A.12)).

2.5 Relationship to Existing Methods

The wideband signal model has been employed for MCLP-based dereverberation [41, 42, 24, 10], however, without explicitly enforcing sparsity of the desired speech signal. For example, in [41, 24] the time-domain prediction filters were estimated by minimizing the output energy, which is equivalent to using the ℓ_2-norm of $\mathbf{d}$ as the cost function, i.e.,

$$P(\mathbf{d}) = \|\mathbf{d}\|_2^2 = \sum_{t=1}^{T} |d(t)|^2. \tag{23}$$

This is a special case of the formulation in Eq. (8), with the ℓ_2-norm as the cost function and without the analysis operator. In this case, the closed-form solution for the prediction filter is given with $\mathbf{g}_{\ell_2}$, as in Sec. 2.1. In [10] a short-time Gaussian model of the desired time-domain signal was used. The obtained algorithm is equivalent to using the weighted ℓ_2-norm as the cost function, i.e.,

$$P(\mathbf{d}) = \|\mathbf{d}\|_{\mathbf{w},2}^2 = \sum_{t=1}^{T} w(t)|d(t)|^2. \tag{24}$$

This is a special case of the formulation in Eq. (8), with the weighted ℓ_2-norm as the cost function and without the analysis operator. For fixed weights, the obtained weighted least-squares optimization problem has a closed-form solution for the prediction filter. The weights $w(t)$ are computed from the previous estimate of the desired speech signal by averaging the energy in the time-domain across a short window centered at t [10]. When employed in a reweighting procedure, this can be interpreted as promoting sparsity of the desired time-domain signal $\mathbf{d}$. However, originally a single reweighting iteration was used, and it was reported that multiple iterations do not always improve performance [10]. Note that the wideband methods in [24, 10] use a signal-dependent prewhitening step before applying dereverberation.

The narrowband signal model has also been employed for MCLP-based speech dereverberation [10, 11]. The most relevant method is the weighted prediction error (WPE) method [10], and it has shown to be very effective for multichannel speech dereverberation [43, 4]. Based on a locally Gaussian model of the desired speech coefficients, the cost function for the WPE method is equal to the weighted ℓ_2-norm [12], i.e.,

$$P\left(\tilde{\mathbf{d}}_k\right) = \|\tilde{\mathbf{d}}_k\|_{\mathbf{w}_k,2}^2 = \sum_{n} w_{k,n}|\tilde{d}_{k,n}|^2. \tag{25}$$

This is a special case of the formulation in Eq. (8), with the weighted ℓ_2-norm as the cost function. Although it would be possible to use the ADMM algorithm for this cost function, the obtained optimization problem can be solved more straightforwardly using the iteratively reweighted least squares algorithm. For fixed weights, the obtained weighted least-squares optimization problem has a closed-form solution for the prediction filter. The weights $\mathbf{w}_k$ can be computed from the estimate of the desired speech coefficients from the previous iteration as $w_{k,n} \leftarrow \varepsilon / \left(|\tilde{d}_{k,n}|^2 + \varepsilon\right)$ [10, 12], which is similar to Eq. (19), by replacing magnitude with squared magnitude. When employed in a reweighting procedure, the considered weighted ℓ_2-norm cost function simulates the behavior of the ℓ_0-norm [12], in the same way as the weighted ℓ_1-norm in Sec. 2.4. Similarly as described in Sec. 2.4, the weights for the WPE method can be computed using a low-rank approximation of the speech spectrogram [20], or using neighborhood based weights.

3 SIMULATIONS

In this section we evaluate the speech dereverberation performance of the ADMM-based methods proposed in Sec. 2 and the iteratively reweighted least squares-based WPE method. We will consider the wideband model with analysis sparsity (WB-A), the wideband model with synthesis sparsity (WB-S), and the narrowband model (NB) with the weighted ℓ_1-norm as the cost function.

3.1 Setup and Performance Measures

We consider an acoustic scenario with a single speech source and $M = 2$ microphones. We have considered two simulated acoustic systems with RIRs from the REVERB challenge [4]. For the first acoustic system (AC_1) the reverberation time was $T_{60} \approx 500$ ms, while for the second acoustic system (AC_2) the reverberation time was $T_{60} \approx 700$ ms. In both cases the distance between the speech source and the microphones was approximately 2 m. The reverberant microphone signals were obtained by convolving the RIRs with a clean speech signal, and the first microphone was selected as the reference microphone. For the evaluation we used a set of 10 speech samples (5 male and 5 female speakers) with an average length of approximately 5.2 s sampled at $f_s = 16$ kHz.

The performance of the considered dereverberation methods is evaluated in terms of frequency-weighted segmental signal-to-noise ratio (FWSSNR) and PESQ [4]. These instrumental performance measures were selected

because of their correlation with perceptual listening tests when evaluating the quality and the perceived amount of reverberation of processed speech signals [44, 4]. The clean speech signal was used as the reference for evaluating the measures, and the obtained results were averaged over all speech samples [4].

3.2 Implementation Details

The analysis and synthesis STFT was computed using a tight frame $\mathbf{\Psi}$ based on a 64 ms Hamming window with 16 ms window shift. The length of the prediction filters was set to $L_g = 5120$ for the wideband model and $\tilde{L}_g = 20$ for the narrowband model, corresponding to 320 ms in the time domain, which is a typical setting for the considered acoustic systems [43]. The prediction delay was set to $L_\tau = 256$ for the wideband model and $\tilde{L}_\tau = 2$ for the narrowband model, corresponding to 32 ms in the time domain.

The weights $w_{k,n}$ for the weighted ℓ_1-norm in Eq. (18) were computed either locally according to Eq. (19), using a rectangular neighborhood according to Eq. (20), or using a low-rank approximation according to Eq. (21). In all experiments the estimate of the desired speech signal was initialized using the reverberant reference microphone signal, which in turn was also used to compute the initial weights. A small positive constant $\varepsilon = 10^{-8}$ was used to regularize the weights. The low-rank approximation $\mathbf{p}$ in Eq. (21) was computed using nonnegative matrix factorization (NMF) with Itakura-Saito divergence with the rank set to 30 [20].

The maximum number of iterations for the ADMM algorithm was set to 40 with $\gamma = 1.6$, since increasing the number of ADMM iterations did not seem to have a significant influence on the performance, while considerably increasing the computational complexity. The stopping criterion was defined as in [16] with a relative tolerance equal to 10^{-3}. For the generalized Lasso, required for the WB-A method (cf., Sec. 2.1), we used the penalty parameter set to $\delta = 1$ (cf., Appendix B). For the Lasso problem, required for the WB-S method (cf., Sec. 2.2), we used FISTA with the maximum number of iterations set to 40 with early stopping when the relative change of the estimate is smaller than 10^{-3} (cf., Appendix C). In all experiments we used the lower bound $G_{\min} = 0.01$.

3.3 Simulation Results

In the following simulations we will investigate the performance of the considered methods with respect to several parameters. First, we investigate the influence of the penalty parameter ρ for the ADMM-based methods. Second, we investigate the influence of the rectangular neighborhoods for computing the weights. Third, we investigate the performance of the considered methods in the reweighting procedure when using different weights. Finally, we discuss the computational complexity of the methods. Exemplary audio samples for all methods are available online[2], showing

[2] http://www.sigproc.uni-oldenburg.de/audio/ante/tfsp/audio.html

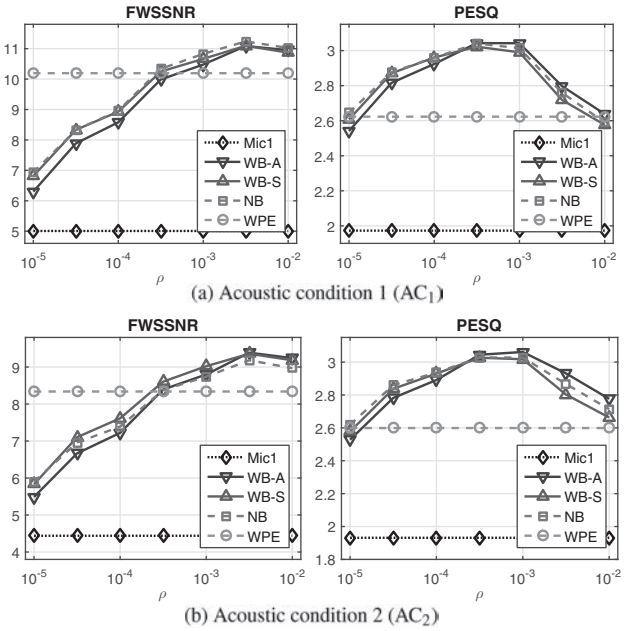


Fig. 3. Instrumental measures vs. penalty parameter ρ (local weights, $i_{\mathrm{RW}} = 1$).

that most processed signals perceptually resemble the clean signal, with some coloration due to the uncontrolled early reflections and hardly audible processing artifacts arising due to the soft thresholding operator.

3.3.1 *Influence of the Penalty Parameter*

While the ADMM algorithm typically converges to modest accuracy after a few tens of iterations, typically the penalty parameter ρ may have a large impact on the convergence, such that an appropriate value needs to be selected. In this section we investigate the influence of the penalty parameter ρ in Eqs. (8), (11), and (14) using locally computed weights as in Eq. (19) and a single reweighting iteration ($i_{\mathrm{RW}} = 1$).[3]

Fig. 3 depicts the obtained instrumental measures for the reverberant reference microphone, the considered ADMM-based methods for different values of the penalty parameter ρ in the set $\{10^{-5}, 5 \cdot 10^{-5}, \ldots, 10^{-2}\}$, and the WPE method. It can be observed that all considered methods result in improvements in terms of the instrumental measures when compared to the reverberant signal at the reference microphone. The WPE method results in significant improvements compared to the reference for all measures and both ACs. It can be observed that the performance obtained using the ADMM-based methods depends on the penalty parameter ρ. Both FWSSNR and PESQ exhibit a similar behavior, with the performance first increasing and then decreasing with ρ. This behavior can be explained by referring to the shape of the proximal operator in Eq. (22).

[3] As will be shown in Sec. 3.3.3, the most significant performance improvement is typically observed after the first reweighting iteration. Hence, it can be assumed that the optimal penalty parameter for $i_{\mathrm{RW}} = 1$ also yields an adequate performance for $i_{\mathrm{RW}} > 1$.

Small values of the penalty parameter ρ result in a relatively high value of the threshold when applying the proximal operator, resulting in a strong suppression of the STFT coefficients and over-suppression of the desired speech signal in each ADMM iteration. Large values of the penalty parameter result in a relatively low value of the threshold, resulting in a weak suppression of the STFT coefficients and a relatively low dereverberation in each ADMM iteration. Overall, it can be observed that it is possible to achieve a better performance using the ADMM-based methods than using the WPE method. Similar behavior with respect to ρ was also observed when using the neighborhood weights in Eq. (20) and NMF weights in Eq. (21). Based on this experimental evidence, for the following experiments we will use $\rho = 10^{-3}$.

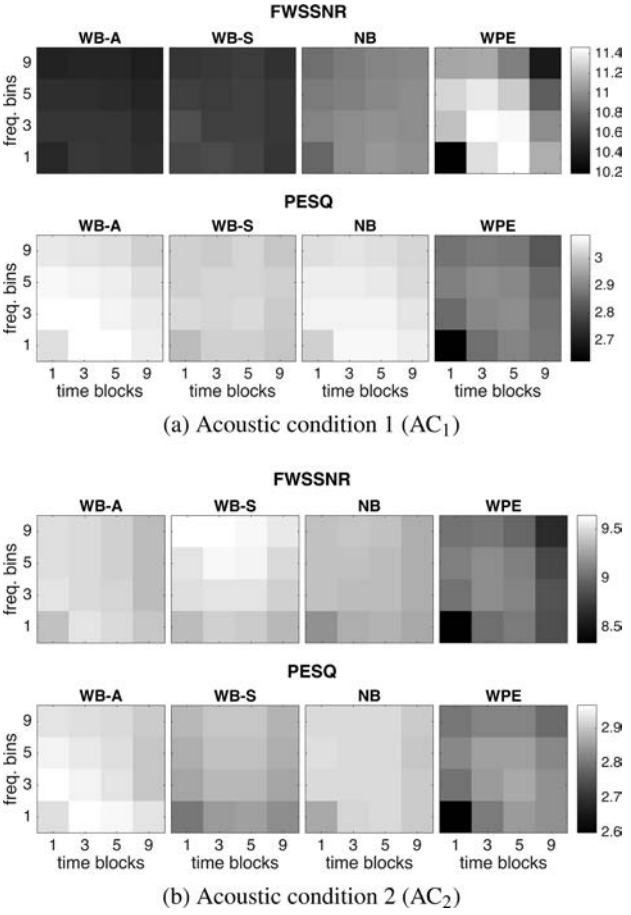
3.3.2 Neighborhood Selection

In this section we investigate the influence of the rectangular neighborhood for computing the weights $w_{k,n}$ as in Eq. (20). For this analysis we consider symmetric rectangular neighborhoods with dimensions across time blocks and frequency bins selected from the set $\{1, 3, 5, 9\}$. The neighborhood coefficients $\eta_{k',n'}$ (cf., Eq. (20)) are set to the same value and sum to 1, i.e., the neighborhood is uniform, and the current coefficient (k, n) is at the center of the neighborhood, i.e., the neighborhood is symmetric. The case of locally computed weights in Eq. (19) obviously corresponds to a neighborhood with both dimensions equal to 1. The obtained performance in terms of instrumental measures is shown in Fig. 4. On the one hand, the depicted results show that only small improvements compared to the local weights are obtained using the considered rectangular neighborhoods for the ADMM-based methods. Typically, relatively small neighborhoods (e.g., with size equal to three) resulted in minor improvements, with the effect typically diminishing for larger sizes. On the other hand, it can be observed that the proposed neighborhoods improve the performance of the WPE method when compared to the locally computed weights. Based on this experimental observation, for the following experiment we used a symmetrical neighborhood with dimensions 3×3 for all methods.

3.3.3 Reweighting Procedure

In this section we evaluate the performance of the considered methods for a varying number of reweighting iterations i_{RW} and different weight computation. For this analysis we set the number of reweighting iterations to $i_{\mathrm{RW}} \in \{1, \ldots, 10\}$. For the ADMM-based methods with a weighted ℓ_1-norm as the cost function in Eq. (18), the weights $w_{k,n}$ are computed either locally according to Eq. (19), using a rectangular 3×3 neighborhood according to Eq. (20), or using NMF-based low-rank approximation according to Eq. (21). For the WPE method, which employs a weighted ℓ_2-norm, the weights are computed analogously, with magnitudes replaced with squared magnitudes.

Figs. 5 and 6 depict the obtained performance in terms of the instrumental measures for AC_1 and AC_2, respectively. It



(a) Acoustic condition 1 (AC_1)

(b) Acoustic condition 2 (AC_2)

Fig. 4. Influence of the size of the rectangular neighborhood across time blocks and frequency bins for the weights $w_{k,n}$ ($i_{\mathrm{RW}} = 1$).

can be observed that for a single reweighting iteration ($i_{\mathrm{RW}} = 1$) and for locally computed weights the ADMM-based methods perform significantly better than the WPE method. The performance of the WPE method is significantly improved when using neighborhood or NMF weights, while only small differences can be observed for the ADMM-based methods, resulting in a similar overall performance for all methods. Moreover, it can be observed that additional reweighting iterations in general improve the performance of all considered methods. The obtained performance typically increases with reweighting iterations up to $i_{\mathrm{RW}} = 5$, with marginal changes for a larger number of iterations. Note that in Fig. 5 a some degradation in terms of PESQ can be observed for WB-S using local weights. However, this is not observed when using neighborhood and NMF weights, indicating that the additional structure in the TF domain, in addition to sparsity, can be beneficial. Overall, the obtained performance for $i_{\mathrm{RW}} = 10$ iterations is relatively similar for all considered methods. This is a consequence of the fact that both the weighted ℓ_1-norm used with the ADMM-methods and the weighted ℓ_2-norm used for the WPE method simulate the behavior of the ℓ_0-norm when using the considered reweighting procedure.

Summarizing the simulation results, we conclude that the ADMM-based methods perform mostly better than the

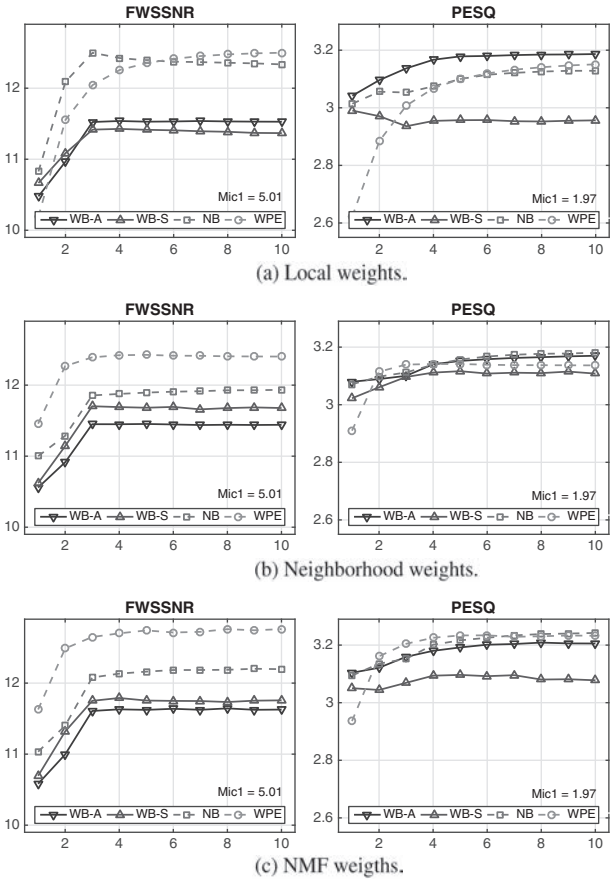


Fig. 5. Influence of the number of reweighting iterations i_{RW} (AC_1): (a) local weights, (b) neighborhood weights, and (c) NMF weights.

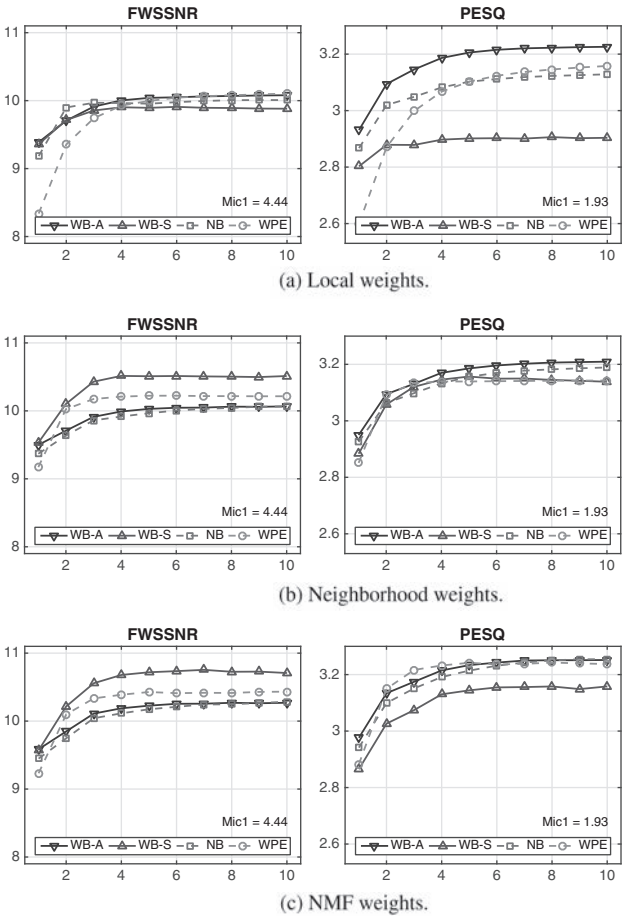


Fig. 6. Influence of the number of reweighting iterations i_{RW} (AC_2): (a) local weights, (b) neighborhood weights, and (c) NMF weights.

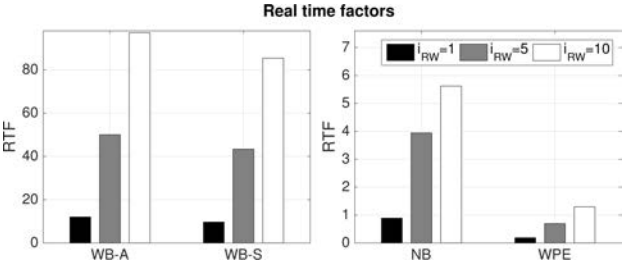


Fig. 7. Average real-time factors for the considered methods.

WPE method for a single reweighting iteration, with the difference being relatively large when using the local weights and relatively small when using structured (neighborhood and NMF) weights. This performance difference can be attributed to the difference in the cost functions, i.e., the weighted ℓ_1-norm employed in the ADMM-based methods, resulting in a sparser solution than the weighted ℓ_2-norm employed in the WPE method (and not to the used optimization algorithm). In general, subsequent reweighting iterations improve the obtained performance, with all methods achieving a similar performance. These similarities can again be attributed to the employed cost function, since both the weighted ℓ_1- and ℓ_2-norm aim to approximate the ℓ_0-norm. Furthermore, the structured weights (i.e., neighborhood- and NMF-based) result in an improved performance for the WPE method, while the effect is much smaller for the ADMM-based methods.

3.3.4 Computational Complexity

In this section we discuss the computational complexity of the considered methods in terms of their real-time factor (RTF), which is defined as the ratio of the computation time and the input duration. All methods have been implemented in Matlab running on a 3,46 GHz Windows 7 machine in single-thread mode. For the ADMM-based methods the linear systems have been solved by factoring the correlation matrix (i.e., $\mathbf{X}^{\mathsf{T}}\mathbf{X}$ in Eq. (10) and Eq. (13) or $\tilde{\mathbf{X}}_k^{\mathsf{H}}\tilde{\mathbf{X}}_k$ in Eq. (17)) once (using Cholesky decomposition), and applying the obtained decomposition to solve the corresponding linear system in the following iterations. Since the WPE method is based on reweighted least squares and hence the matrix of the linear system changes for each reweighting iteration (cf., Eq. (25)), this was not possible for the WPE method. For the narrowband methods (NB and WPE) we sequentially processed all frequency bins, without exploiting parallelization over frequencies.

Fig. 7 depicts the RTFs, averaged across the samples, for the considered methods. On the one hand, the wideband methods result in relatively large RTFs, due to the large dimension (ML_g) of the involved linear systems and the fact that the analysis/synthesis operators $\mathbf{\Psi}^{\mathsf{H}}$ and $\mathbf{\Psi}$ need to be applied in every ADMM iteration (cf., Appendix B

and C). On the other hand, the narrowband methods have much smaller RTFs, due to the smaller dimension ($M\tilde{L}_g$) of the involved linear systems and since the optimization problems are in the TF domain (cf., Eq. (14)) the analysis and synthesis operators need to be applied only once. While the real-time factors for the NB and WPE methods are on the same order of magnitude, the latter was significantly faster (e.g., $\sim$0.9 vs. $\sim$0.2 for $i_{\mathrm{RW}} = 1$). However, it is expected that complexity could be significantly reduced for the ADMM-based methods by exploiting the block-Toeplitz structure [30], which cannot be exploited for the methods based on iteratively reweighted least squares. Note that the complexity of all methods could possibly be further reduced, e.g., by using [45] for fast computation of the correlation matrices.

4 DISCUSSION AND CONCLUSIONS

In this paper we have presented a general framework for multichannel speech dereverberation exploiting sparsity in the time-frequency domain. We have formulated the MCLP-based speech dereverberation as an optimization problem using a cost function that promotes sparsity of the desired speech signal in the time-frequency domain. The presented framework encompasses a wideband or a narrowband signal model as well as an analysis and a synthesis prior for the desired speech signal. While the discussion in this paper has been limited to sparsity in the STFT domain, other time-frequency transforms could be used through a suitable pair of analysis-synthesis operators. We have shown that all resulting optimization problems can be solved using the alternating direction method of multipliers, and that different sparsity-promoting cost functions can used by selecting an appropriate proximal operator.

Simulation results show that the proposed ADMM-based methods using the weighted ℓ_1-norm as the sparsity-promoting cost function perform better than the conventional WPE method for a single reweighting iteration (at a higher computational complexity), and achieve a similar performance for multiple iterations. In addition, we have shown that using neighborhood-based weights for the reweighting iterations can improve the dereverberation performance of the sparsity-based methods.

In conclusion, the narrowband methods appear to be more relevant in practice, since they achieve a good dereverberation performance with a significantly lower computational complexity than the wideband methods. Nevertheless, the wideband methods offer more flexibility in the selection of the TF transform and could be used even when the narrowband model does not hold, e.g., if there is a strong influence between adjacent bands in the TF domain. In addition, the considered reweighting procedure in general improves the dereverberation performance, since the reweighting typically results in a sparser output signal.

The presented work constitutes a flexible and general framework for sparsity-based dereverberation. Further work could therefore include the design of cost functions that exploit additional characteristics of the speech signal and properties of auditory perception, implementation of fast multichannel structure-exploiting linear solvers, and exploration of adaptive time-frequency transforms in the proposed framework.

5 ACKNOWLEDGMENT

This research was supported by the Marie Curie Initial Training Network DREAMS (Grant agreement no. ITN-GA-2012-316969) and by the Cluster of Excellence 1077 "Hearing4All," funded by the German Research Foundation (DFG).

6 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 (2006 Jul.). https://doi.org/10.1121/1.2202888

[2] T. Yoshioka et al., "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 (2012 Oct.). https://doi.org/10.1109/MSP.2012.2205029

[3] P. A. Naylor, and N. D. Gaubitch, *Speech Dereverberation* (Springer, 2010). https://doi.org/10.1007/978-1-84996-056-4

[4] K. Kinoshita et al., "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. 7 (2016 Jan.). https://doi.org/10.1186/s13634-016-0306-6

[5] M. Miyoshi, and Y. Kaneda, "Inverse Filtering of Room Acoustics," *IEEE Trans. Audio Speech Lang. Process.*, vol. 36, no. 2, pp. 145–152 (1988 Feb.). https://doi.org/10.1109/29.1509

[6] A. W. H. Khong, and P. A. Naylor, "Adaptive Blind Multichannel System Identification," in *Speech Dereverberation*, P. A. Naylor, and N. D. Gaubitch (Eds.) (Springer, 2010). https://doi.org/10.1007/978-1-84996-056-4_6

[7] N. D. Gaubitch, and P. A. Naylor, "Equalization of Multichannel Acoustic Systems in Oversampled Subbands," *IEEE Trans. Audio Speech Lang. Process.*, vol. 17, no. 6, pp. 1061–1070 (2009 Aug.). https://doi.org/10.1109/TASL.2009.2015692

[8] W. Zhang, E. A. P. Habets, and P. A. Naylor, "On the Use of Channel Shortening in Multichannel Acoustic System Equalization," *Proc. Int. Workshop Acoustic Echo Noise Control (IWAENC),* Tel Aviv, Israel (2010).

[9] 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 (2013 Sep.). https://doi.org/10.1109/TASL.2013.2260743

[10] T. Nakatani et al., "Speech Dereverberation Based on Variance-Normalized Delayed Linear Prediction," *IEEE Trans. Audio Speech Lang. Process.*,

vol. 18, no. 7, pp. 1717–1731 (2010 Sep.). https://doi.org/10.1109/TASL.2010.2052251

[11] M. Togami et al., "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 (2014 Jul.). https://doi.org/10.1109/TASL.2013.2250960

[12] A. Jukić et al., "Multichannel Linear Prediction-Based Speech Dereverberation with Sparse Priors," *IEEE/ACM Trans. Audio Speech Lang. Process.*, vol. 23, no. 9, pp. 1509–1520 (2015 Sep.). https://doi.org/10.1109/TASLP.2015.2438549

[13] P. Bofil, and M. Zibulevsky, "Underdetermined Blind Source Separation Using Sparse Representations," *Signal Process.,* vol. 81, pp. 2353–2363 (2001). https://doi.org/10.1016/S0165-1684(01)00120-7

[14] S. Makino et al., "Underdetermined Blind Source Separation Using Acoustic Arrays," in *Handbook on Array Processing and Sensor Networks*, S. Haykin, and K. J. R. Liu (Eds.) (John Wiley & Sons, 2010). https://doi.org/10.1002/9780470487068.ch10

[15] M. Elad, P. Milanfar, and R. Rubinstein, "Analysis versus Synthesis in Signal Priors," *Inverse Problems*, vol. 23, no. 3, pp. 947 (2007). https://doi.org/10.1088/0266-5611/23/3/007

[16] S. Boyd et al., "Distributed Optimization and Statistical Learning via the Alternating Direction Method of Multipliers," *Found. Trends Machine Learn.*, vol. 3, no. 1, pp. 1–122 (2011). https://doi.org/10.1561/2200000016

[17] P. Balazs et al., "Adapted and Adaptive Linear Time-Frequency Representations: A Synthesis Point of View," *IEEE Signal Proc. Mag.*, vol. 30, no. 6, pp. 20–31 (2013 Nov.). https://doi.org/10.1109/MSP.2013.2266075

[18] E. J. Candes, M. B. Wakin, and S. P. Boyd, "Enhancing Sparsity by Reweighted ℓ_1 Minimization," *J. Fourier Analysis App.*, vol. 14, no. 5-6, pp. 877–905 (2008). https://doi.org/10.1007/s00041-008-9045-x

[19] M. Kowalski, K. Siedenburg, and M. Dörfler, "Social Sparsity! Neighborhood Systems Enrich Structured Shrinkage Operators," *IEEE Trans. Signal Process.*, vol. 61, no. 10, pp. 2498–2511 (2013 May). https://doi.org/10.1109/TSP.2013.2250967

[20] A. Jukić et al., "Multichannel Linear Prediction-Based Speech Dereverberation with Low-Rank Power Spectrogram Approximation," *Proc. IEEE Int. Conf. Acoust. Speech Signal Process. (ICASSP),* Brisbane, Australia (2015), pp. 96–100. https://doi.org/10.1109/ICASSP.2015.7177939

[21] A. Jukić et al., "A General Framework for Multichannel Speech Dereverberation by Exploiting Sparsity," presented at the *AES 60th International Conference: DREAMS (Dereverberation and Reverberation of Audio, Music, and Speech)* (2016 Jan.), conference paper 9-1.

[22] J. S. Bradley, H. Sasto, and M. Picard, "On the Importance of Early Reflections for Speech in Rooms," *J. Acoust. Soc. Amer.*, vol. 113, no. 6, pp. 3233–3244 (2003 Jun.). https://doi.org/10.1121/1.1570439

[23] D. Gesbert, and P. Duhamel, "Unbiased Blind Adaptive Channel Identification and Equalization," *IEEE Trans. Signal Process.*, vol. 48, no. 1, pp. 148–158 (2000 Jan.). https://doi.org/10.1109/78.815485

[24] K. Kinoshita et al., "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 (2009 May). https://doi.org/10.1109/TASL.2008.2009015

[25] J. Kovacevic, and A. Chebira, "Life Beyond Bases: The Advent of Frames (Part I)," *IEEE Signal Process. Mag.*, vol. 24, no. 4, pp. 86–104 (2007). https://doi.org/10.1109/MSP.2007.4286567

[26] P. Bofil, and M. Zibulevsky, "Blind Separation of More Sources than Mixtures Using Sparsity of Their Short-Time FourierTransform," in *Proc. ICA,* Helsinki, Finland (2000), pp. 87–92.

[27] M. Kowalski, E. Vincent, and R. Gribonval, "Beyond the Narrowband Approximation: Wideband Convex Methods for Under-Determined Reverberant Audio Source Separation," *IEEE Trans. Audio Speech Lang. Process.*, vol. 18, no. 7, pp. 1818–1829 (2010). https://doi.org/10.1109/TASL.2010.2050089

[28] A. Adler et al., "Audio Inpainting," *IEEE Trans. Audio Speech Lang. Process.*, vol. 20, no. 3, pp. 922–932 (2012 Mar.). https://doi.org/10.1109/TASL.2011.2168211

[29] H. Kameoka, T. Nakatani, and T. Yoshioka, "Robust Speech Dereverberation Based on Non-Negativity and Sparse Nature of Speech Spectrograms," *Proc. IEEE Int. Conf. Acoust. Speech Signal Process. (ICASSP),* Taipei, Taiwan (2009), pp. 45–48. https://doi.org/10.1109/ICASSP.2009.4959516

[30] T. L. Jensen et al., "Fast Algorithms for High-Order Sparse Linear Prediction with Applications to Speech Processing," *Speech Commun.*, vol. 76, pp. 143–156 (2016 Feb.). https://doi.org/10.1016/j.specom.2015.09.013

[31] R. Tibshirani, "Regression Shrinkage and Selection via the Lasso," *J. Royal Stat. Soci. Series B*, vol. 58, pp. 267–288 (1996). https://doi.org/10.1111/j.1467-9868.2011.00771.x

[32] A. Beck, and M. Teboulle, "A Fast Iterative Shrinkage-Thresholding Algorithm for Linear Inverse Problems," *SIAM J. Imag. Sci.*, vol. 2, no. 1, pp. 183–202 (2009). https://doi.org/10.1137/080716542

[33] H. Mohimani, M. Babaie-Zadeh, and C. Jutten, "A Fast Approach for Overcomplete Sparse Decomposition Based on Smoothed L0 Norm," *IEEE Trans. Signal Process.*, vol. 57, no. 1, pp. 289–301 (2009 Jan.). https://doi.org/10.1109/tsp.2008.2007606

[34] D. Wipf, and S. Nagarajan, "Iterative Reweighted ℓ_1 and ℓ_2 Methods for Finding Sparse Solutions," *IEEE J. Sel. Topics Signal Process.*, vol. 4, no. 2, pp. 317–329 (2010 Apr.). https://doi.org/10.1109/jstsp.2010.2042413

[35] R. Chartrand, "Shrinkage Mappings and their Induced Penalty Functions," *Proc. IEEE Int. Conf.*

Acoust. Speech Signal Process. (ICASSP), Florence, Italy (2014), pp. 1026–1029. https://doi.org/10.1109/icassp.2014.6853752

[36] S. Arberet et al., "Sparse Reverberant Audio Source Separation via Reweighted Analysis," *IEEE Trans. Audio Speech Lang. Process.*, vol. 21, no. 7, pp. 1391–1402 (2013 Jul.). https://doi.org/10.1109/tasl.2013.2250962

[37] K. Siedenburg, and M. Dörfler, "Audio Denoising by Generalized Time-Frequency Thresholding," presented at the *AES 45th International Conference: Applications of Time-Frequency Processing in Audio* (2012 Mar.), conference paper 5-2.

[38] C. Févotte, N. Bertin, and J.-L. Durrieu, "Nonnegative Matrix Factorization with the Itakura-Saito Divergence: With Application to Music Analysis," *Neural Computation*, vol. 21, no. 3, pp. 793–830 (2009). https://doi.org/10.1162/neco.2008.04-08-771

[39] N. Mohammadiha, P. Smaragdis, and A. Leijon, "Supervised and Unsupervised Speech Enhancement Using Nonnegative Matrix Factorization," *IEEE Trans. Audio Speech Lang. Process.*, vol. 21, no. 10, pp. 2140–2151 (2013 Oct.). https://doi.org/10.1109/tasl.2013.2270369

[40] R. C. Hendriks, T. Gerkmann, and J. Jensen, "DFT-Domain Based Single-Microphone Noise Reduction for Speech Enhancement: A Survey of the State of the Art," *Synthesis Lect. Speech Audio Process.*, vol. 9, no. 1, pp. 1–80 (2013 Jan.). https://doi.org/10.2200/s00473ed1v01y201301sap011

[41] M. Delcroix, T. Hikichi, and M. Miyoshi, "Precise Dereverberation Using Multichannel Linear Prediction," *IEEE Trans. Audio Speech Lang. Process.*, vol. 15, no. 2, pp. 430–440 (2007 Feb.). https://doi.org/10.1109/tasl.2006.881698

[42] T. Nakatani et al., "Speech Dereverberation Based on Maximum-Likelihood Estimation with Time-Varying Gaussian Source Model," *IEEE Trans. Audio Speech Lang. Process.*, vol. 16, no. 8, pp. 1512–1527 (2008 Nov.). https://doi.org/10.1109/tasl.2008.2004306

[43] M. Delcroix et al., "Strategies for Distant Speech Recognition in Reverberant Environments," *EURASIP J. Adv. Signal Process.*, vol. 2015, no. 60 (2015 Jul.). https://doi.org/10.1186/s13634-015-0245-7

[44] S. Goetze et al., "Speech Quality Assessment for Listening-Room Compensation," presented at the *AES 38th International Conference: Sound Quality Evaluation* (2010 Jun.), conference paper 1-1.

[45] T. Yoshioka, T. Nakatani, and M. Miyoshi, "Fast Algorithm for Conditional Separation and Dereverberation," *Proc. European Sig. Process. Conf. (EUSIPCO),* Glasgow, UK (2009), pp. 1432–1436.

[46] M. Fortin, and R. Glowinski, *Augmented Lagrangian Methods* (Elsevier, 1983).

[47] P. Combettes, and V. Wajs, "Signal Recovery by Proximal Forward-Backward Splitting," *Multiscale Modeling and Simulation*, vol. 4, no. 4, pp. 1168–1200 (2005 Nov.). https://doi.org/10.1137/050626090

[48] Y. Lucet, H. H. Bausch, and M. Triens, "The Piecewise Linear-Quadratic Model for Computational Convex Analysis," *Comput. Optim. Appl.*, vol. 43, no. 1, pp. 95–118 (2009 May). https://doi.org/10.1007/s10589-007-9124-y

APPENDIX A: ALTERNATING DIRECTION METHOD OF MULTIPLIERS (ADMM)

In this appendix we present a brief overview of the ADMM algorithm, which can be used to solve the optimization problems in Secs. 2.1, 2.2, and 2.3. A detailed overview of the ADMM algorithm and applications can be found in [16]. The ADMM algorithm is suitable for nonsmooth convex optimization problems of the form

$$\begin{array}{ll} \min_{\mathbf{d},\mathbf{g}} & P(\mathbf{d}) + Q(\mathbf{g}) \\ \text{subject to} & \mathbf{A}\mathbf{d} + \mathbf{B}\mathbf{g} = \mathbf{c}, \end{array} \tag{A.1}$$

where the total cost function conveniently splits for the variables $\mathbf{d}$ and $\mathbf{g}$. The augmented Lagrangian for the constrained optimization problem in Eq. (A.1) can be written as

$$\begin{aligned} L_\rho(\mathbf{d}, \mathbf{g}, \boldsymbol{\mu}) &= P(\mathbf{d}) + Q(\mathbf{g}) \\ &+ \frac{\rho}{2}\|\mathbf{A}\mathbf{d} + \mathbf{B}\mathbf{g} - \mathbf{c} + \boldsymbol{\mu}\|_2^2 - \frac{\rho}{2}\|\boldsymbol{\mu}\|_2^2 \end{aligned} \tag{A.2}$$

where $\boldsymbol{\mu}$ denotes the dual variable and ρ denotes a penalty parameter. The ADMM algorithm is obtained by minimizing the augmented Lagrangian in Eq. (A.2) alternately with respect to $\mathbf{d}$ and $\mathbf{g}$, followed by a dual ascent over $\boldsymbol{\mu}$ [16]. This leads to the following update rules

$$\begin{array}{l} \mathbf{d}^i \leftarrow \arg\min_{\mathbf{d}} P(\mathbf{d}) + \frac{\rho}{2}\|\mathbf{A}\mathbf{d} + \mathbf{B}\mathbf{g}^{i-1} - \mathbf{c} + \boldsymbol{\mu}^{i-1}\|_2^2, \\ \mathbf{g}^i \leftarrow \arg\min_{\mathbf{g}} Q(\mathbf{g}) + \frac{\rho}{2}\|\mathbf{A}\mathbf{d}^i + \mathbf{B}\mathbf{g} - \mathbf{c} + \boldsymbol{\mu}^{i-1}\|_2^2, \\ \boldsymbol{\mu}^i \leftarrow \boldsymbol{\mu}^{i-1} + \gamma\left(\mathbf{A}\mathbf{d}^i + \mathbf{B}\mathbf{g}^i - \mathbf{c}\right), \end{array} \tag{A.3}$$

with i denoting the iteration index. These update rules are iteratively repeated until convergence, with the convergence criterion defined in [16]. For easier readability the iteration index i is not included in the ADMM-based update rules in Sec. 2. The parameter $\gamma \geq 1$ can be used for faster convergence, and should be smaller than $1 + \sqrt{5}/2$ for a convex P [46]. In many applications, the ADMM algorithm converges to modest accuracy after a few tens of iterations, which is often enough in practice [16, 30]. However, the penalty parameter ρ may have a large effect on the convergence of the algorithm and typically depends on the particular choice of P and Q. Hence, when using a finite number of iterations, an appropriate value for the penalty parameter needs to be selected.

An important ingredient of ADMM-based algorithms is the proximal operator. The proximal operator S_ρ^P of the cost function P with the penalty parameter ρ can be defined as

$$S_\rho^P(\mathbf{z}) = \arg\min_{\mathbf{d}} P(\mathbf{d}) + \frac{\rho}{2}\|\mathbf{z} - \mathbf{d}\|_2^2, \tag{A.4}$$

The proximal operator can be seen as a generalization of a projection operator, e.g., if P is an indicator function of a convex set the proximal operator is equal to the Euclidean projection on that set [16]. In many cases the proximal operator can be evaluated efficiently, e.g., when P is the ℓ_1-norm S_ρ^P is the well-known soft thresholding operator.

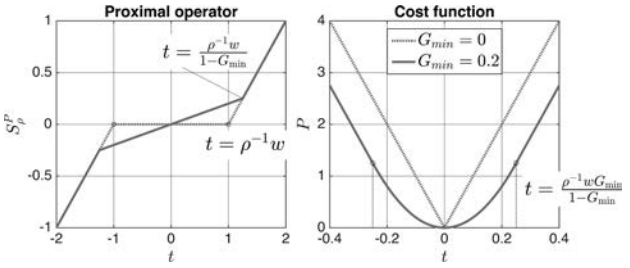


Fig. 8. Proximal operators (left) and the corresponding cost functions (right) on the real axis for $\rho^{-1}w = 1$ and $G_{\min} = \{0, 0.2\}$. For complex-valued inputs the proximal operator and the penalty function are extended using circular symmetry.

APPENDIX B: GENERALIZED LASSO PROBLEM

In this appendix we present a brief overview of the optimization problem for estimation of the desired signal $\mathbf{d}$ in Eq. (8). An optimization problem in the form

$$\min_{\mathbf{d}} \quad P\left(\mathbf{\Psi}^{\mathsf{H}}\mathbf{d}\right) + \frac{\rho}{2}\|\mathbf{d}-\mathbf{z}\|_2^2, \tag{A.5}$$

where P is a weighted ℓ_1-norm is an instance of generalized Lasso [16]. The minimizer of the cost function is equal to the proximal operator of the composition of the analysis operator $\mathbf{\Psi}^{\mathsf{H}}$ and the function P, i.e., the optimal $\mathbf{d}$ is equal to $S_\rho^{P\circ\mathbf{\Psi}^{\mathsf{H}}}(\mathbf{z})$, with $P\circ\mathbf{\Psi}^{\mathsf{H}}$ denoting the composition of the analysis operator $\mathbf{\Psi}^{\mathsf{H}}$ and the cost function P, i.e., $\left(P\circ\mathbf{\Psi}^{\mathsf{H}}\right)(\mathbf{d}) = P\left(\mathbf{\Psi}^{\mathsf{H}}\mathbf{d}\right)$. The optimization problem in Eq. (A.5) can be efficiently solved by applying the ADMM algorithm, resulting in the following iterative updates

$$\begin{aligned}
\mathbf{d} &\leftarrow \tfrac{1}{1+\rho}\left[\mathbf{z}+\rho\mathbf{\Psi}\left(\tilde{\mathbf{u}}-\tilde{\boldsymbol{\mu}}\right)\right],\\
\tilde{\mathbf{u}} &\leftarrow S_{\delta\rho}^{P}\left(\mathbf{\Psi}^{\mathsf{H}}\mathbf{d}+\tilde{\boldsymbol{\mu}}\right),\\
\tilde{\boldsymbol{\mu}} &\leftarrow \tilde{\boldsymbol{\mu}}+\gamma\left(\mathbf{\Psi}^{\mathsf{H}}\mathbf{d}-\tilde{\mathbf{u}}\right),
\end{aligned} \tag{A.6}$$

where $\tilde{\mathbf{u}}$ is the splitting variable satisfying the constraint $\mathbf{\Psi}^{\mathsf{H}}\mathbf{d}-\tilde{\mathbf{u}}=\mathbf{0}$, and δ is the penalty parameter.

APPENDIX C: LASSO PROBLEM

In this appendix we present a brief overview of the optimization problem for estimation of the STFT coefficients $\tilde{\mathbf{d}}$ in Eq. (12). An optimization problem in the form

$$\min_{\tilde{\mathbf{d}}} P\left(\tilde{\mathbf{d}}\right) + \frac{\rho}{2}\|\mathbf{\Psi}\tilde{\mathbf{d}}-\mathbf{z}\|_2^2, \tag{A.7}$$

where P is a weighted ℓ_1-norm is in the form of Lasso and can be efficiently solved by applying the iterative shrinkage/thresholding algorithm (ISTA) [47, 32]. The algorithm consists of iteratively repeating the following update

$$\tilde{\mathbf{d}} \leftarrow S_{\nu\rho}^{P}\left(\tilde{\mathbf{d}}+\frac{1}{\nu}\mathbf{\Psi}^{\mathsf{H}}\left(\mathbf{z}-\mathbf{\Psi}\tilde{\mathbf{d}}\right)\right), \tag{A.8}$$

where ν is the maximum eigenvalue of $\mathbf{\Psi}\mathbf{\Psi}^{\mathsf{H}}$, i.e., in our case $\nu = 1$. An accelerated version of the algorithm, fast ISTA (FISTA) algorithm [32], employs a similar iteration, with almost the same complexity and an improved convergence.

APPENDIX D: SOFT THRESHOLDING WITH A LOWER-BOUND FOR THE GAIN AND THE CORRESPONDING COST FUNCTION

In this appendix we present analytical expressions for the cost function corresponding to the proximal operator in Eq. (22) with a lower bound on the gain. We consider a proximal operator S_ρ^P for a penalty function P of a complex scalar $\tilde{d}$, defined as

$$S_\rho^P\left(\tilde{d}\right) = \max\left(1-\frac{\rho^{-1}w}{|\tilde{d}|}, G_{\min}\right)\cdot\tilde{d}. \tag{A.9}$$

Without the loss of generality (since the function S_ρ^P is circularly symmetric) we can focus only on the positive part of the real axis, denoting the independent variable as t, with $t\in\mathbb{R}, t>0$. As in [35], we define a function $f(t) = \int_0^t S_\rho^P(\xi)\,d\xi, t>0$. For the given mapping in Eq. (A.9), the function f can be written as

$$f(t) = \begin{cases} \frac{1}{2}G_{\min}t^2, & \text{for } t < \frac{\rho^{-1}w}{1-G_{\min}} \\ \frac{1}{2}\left(t-\rho^{-1}w\right)^2 + \frac{1}{2}\frac{\rho^{-2}w^2G_{\min}}{1-G_{\min}}, & \text{for } t \geq \frac{\rho^{-1}w}{1-G_{\min}} \end{cases} \tag{A.10}$$

Following [35], the corresponding cost function can be obtained as $P(t) = \rho\left[f^*(t) - \frac{t^2}{2}\right]$, where f^* is the convex conjugate (Fenchel transform) of f. By observing that f is a convex continuous piecewise quadratic function, its convex conjugate can be obtained using [48] as

$$f^*(t) = \begin{cases} \frac{1}{2}\frac{t^2}{G_{\min}}, & \text{for } t < \frac{\rho^{-1}wG_{\min}}{1-G_{\min}} \\ \rho^{-1}wt + \frac{1}{2}t^2 - \frac{1}{2}\frac{\rho^{-2}w^2G_{\min}}{1-G_{\min}}, & \text{for } t \geq \frac{\rho^{-1}wG_{\min}}{1-G_{\min}} \end{cases} \tag{A.11}$$

Finally, the cost function can be written as

$$P(t) = \begin{cases} \rho\frac{1-G_{\min}}{G_{\min}}\frac{t^2}{2}, & \text{for } t < \frac{\rho^{-1}wG_{\min}}{1-G_{\min}} \\ wt - \frac{1}{2}\frac{\rho^{-1}w^2G_{\min}}{1-G_{\min}}, & \text{for } t \geq \frac{\rho^{-1}wG_{\min}}{1-G_{\min}} \end{cases} \tag{A.12}$$

Without the minimum-gain bound, i.e., when $G_{\min} = 0$, the proximal operator in Eq. (A.9) reduces to soft thresholding, and the cost function is a linear function of t, corresponding to the weighted ℓ_1-norm. When $G_{\min} > 0$ the cost function is quadratic for small values of t and linear for large values of t. The point of transition between the linear and quadratic behavior is modulated with the parameters ρ and w. An illustration of the effect of the lower bound on the proximal operator and the cost function is given in Fig. 8.

THE AUTHORS


Ante Jukić


Toon van Waterschoot


Timo Gerkmann


Simon Doclo

Ante Jukić received the Dipl.-Ing. degree in electrical engineering in 2009 from the University of 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 Ruđer 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 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 (Universit della Svizzera italiana), Switzerland. His research interests are in acoustic signal enhancement, acoustic modeling, audio analysis, and audio reproduction. He has been the Scientific Coordinator of the FP7-PEOPLE Marie Curie Initial Training Network on Dereverberation and Reverberation of Audio, Music, and Speech (DREAMS). Dr. van Waterschoot has been serving as an Associate Editor for the *Journal of the Audio Engineering Society (AES)* 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 member of the IEEE Audio and Acoustic Signal Processing Technical Committee (AASP-TC). He has been serving as an Area Chair for Speech Processing at the European Signal Processing Conference (EUSIPCO 2010, 2013–2015), and as General Chair of the 60th AES Conference in Leuven, Belgium, 2016. He is a member of the AES, the Acoustical Society of America, EURASIP, and IEEE.

•

Timo Gerkmann studied electrical engineering and information sciences at the universities of Bremen and Bochum, Germany. He received his Dipl.-Ing. degree in 2004 and his Dr.-Ing. degree in 2010 both in electrical engineering and information sciences from 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. From 2011 to 2015 he was a professor for speech signal processing at the Universität Oldenburg, Oldenburg, Germany. During 2015 to 2016 he was a Principal Scientist for Audio & Acoustics at Technicolor Research & Innovation in Hanover, Germany. Since 2016 he is a professor for signal processing at the University of Hamburg, Germany. His research interests are on digital signal processing algorithms for speech and audio applied to communication devices, hearing instruments, audio-visual media, and human-machine interfaces. Timo Gerkmann is a Senior Member of the IEEE.

•

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