A general framework for multi-channel speech dereverberation exploiting sparsity

Ante Jukić[1], Toon van Waterschoot[2], Timo Gerkmann[1], Simon Doclo[1] *

[1] *University of Oldenburg, Department of Medical Physics and Acoustics, and the Cluster of Excellence Hearing4All, Oldenburg, Germany*

[2] *KU Leuven, Department of Electrical Engineering (ESAT-STADIUS/ETC), Leuven, Belgium*

Correspondence should be addressed to Ante Jukić (`ante.jukic@uni-oldenburg.de`)

ABSTRACT

We consider the problem of blind multi-channel speech dereverberation without the knowledge of room acoustics. The dereverberated speech component is estimated by subtracting the undesired component, estimated using multi-channel linear prediction (MCLP), from the reference microphone signal. In this paper we present a framework for MCLP-based speech dereverberation by exploiting sparsity in the time-frequency domain. The presented framework uses a wideband or a narrowband signal model and a sparse analysis or synthesis model for the desired speech component. The proposed problems involving a reweighted ℓ_1-norm, are solved in a flexible optimization framework. The obtained results are comparable to the state of the art, motivating further extensions exploiting sparsity and speech structure.

1. INTRODUCTION

Effective speech dereverberation is very important for many speech communication applications, such as hands-free telephony, voice-controlled systems or hearing aids [1]. While moderate levels of reverberation can be beneficial, severe reverberation significantly decreases speech intelligibility and automatic speech recognition performance. The field of speech dereverberation has been very active over the last decade, and many speech dereverberation methods have been proposed, e.g., based on acoustic multichannel equalization [2, 3], spectral enhancement [4, 5], and probabilistic models [6–9].

It is widely accepted that speech signals have a sparse representation in the time-frequency (TF) domain, a fact extensively exploited in source separation [10–14], audio inpainting [15, 16], and dereverberation [9, 17, 18]. While many speech enhancement methods employ a narrowband signal model [19], operating in each frequency bin independently, a wideband signal model has been recently employed to improve source separation in reverberant environments [14, 20]. In this paper we present a general framework for speech dereverberation using multi-channel linear prediction (MCLP) that exploits sparsity in the TF domain. We present different formulations with a wideband or a narrowband signal model and sparse analysis or synthesis model for the speech signal [21]. In this work, we will use a reweighted ℓ_1-norm for promoting sparsity, but the presented formulations allow other sparsity-promoting cost functions. As opposed to the locally computed weights [9], we also employ neighborhoods in the TF domain [22], and a low-rank approximation of the power spectrogram [23]. Using simulations we compare the proposed methods with similar state-of-the-art methods. The presented framework provides a transparent view on sparsity-based speech dereverberation and generalizes existing methods.

2. SIGNAL MODEL

We consider a fixed source-array geometry with a single speech source and M microphones. The time-domain reverberant signal at the m-th microphone can be modeled as the convolution of the clean speech signal $s(t)$ with a room impulse response (RIR) $r_m(t)$ of length L_r. The reference microphone signal can be decomposed as

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

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

where $d(t)$ represents the desired speech component at the reference microphone, consisting of the direct path and early reflections, and $u(t)$ represents the undesired speech component at the reference microphone, consisting of the late reflections. The goal is then to recover the desired speech component $d(t)$ that includes early reflections, preserved due to the delay L_τ, which can be beneficial for speech intelligibility [25]. When $M > 1$ microphones are available, it can be shown that using MINT [2] the undesired part can be obtained by filtering the delayed microphone signals, i.e., $u(t)$ can be written as

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

where $g_m(l)$ denotes the l-th tap of the prediction filter, of length L_g, related to the m-th microphone [6]. Assuming that T time-domain samples are observed, using vector notation we can write

$$\mathbf{x}_{\text{ref}} = \mathbf{d} + \underbrace{\mathbf{Xg}}_{\mathbf{u}}, \tag{3}$$

where $\mathbf{x}_{\text{ref}} = [x_{\text{ref}}(1), \ldots, x_{\text{ref}}(T)]^T$, and $\mathbf{d}$ and $\mathbf{u}$ are defined similarly. The matrix $\mathbf{X} \in \mathbb{R}^{T \times ML_g}$ is a multi-channel (block-) convolution matrix obtained from the microphone signals delayed by L_τ, and $\mathbf{g} \in \mathbb{R}^{ML_g}$ is a multi-channel prediction filter composed of the filter coefficients for all M channels.

While the wideband model in (3) holds perfectly if the MINT conditions are fulfilled, the filters $\mathbf{g}$ can be very long and dereverberation based on the wideband model (3) can be computationally demanding, therefore the wideband model (3) is often approximated in the STFT domain [6, 9, 26, 27]. Let $\mathbf{\Psi} \in \mathbb{C}^{T \times KN}$ denote the STFT frame for which $\mathbf{\Psi}\mathbf{\Psi}^H = \mathbf{I}$ holds, i.e., $\mathbf{\Psi}$ is a tight (overcomplete) frame with $T < KN$, where N denotes the number of time indices and K denotes the number of frequency bins in the TF domain. The TF coefficients of the time-domain signal $\mathbf{d}$ can be obtained as $\tilde{\mathbf{d}} = \mathbf{\Psi}^H \mathbf{d} \in \mathbb{C}^{KN}$, and we use $\tilde{\mathbf{d}}_k \in \mathbb{C}^N$ to denote a vector containing only the TF coefficients corresponding to the k-th frequency bin and $\tilde{d}_{k,n}$ to denote a single coefficient.[1] The narrowband signal model is obtained by approximating the time-domain convolution in (3) in each frequency bin independently [28], and can be written analogously as before as

$$\tilde{\mathbf{x}}_{\text{ref},k} = \tilde{\mathbf{d}}_k + \tilde{\mathbf{X}}_k \tilde{\mathbf{g}}_k, \tag{4}$$

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 $\tilde{L}_\tau$ frames. The prediction filters $\tilde{\mathbf{g}}_k \in \mathbb{C}^{M\tilde{L}_g}$ for the narrow-band model (4) are typically much shorter, i.e., $\tilde{L}_g << L_g$, and are estimated independently for each frequency.

3. PROPOSED METHODS

Sparsity of speech signals in the STFT domain is well known and often exploited for various speech enhancement tasks [9, 13, 14, 22]. This property naturally leads to two modeling paradigms, i.e., the sparse synthesis model and the sparse analysis model [21]. The idea behind the synthesis sparsity is that the desired 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 $\mathbf{d} = \mathbf{\Psi}\tilde{\mathbf{d}}$ with sparse $\tilde{\mathbf{d}}$. The idea behind the analysis sparsity is that the desired signal has a sparse representation when an analysis operator is applied. In the considered scenario this would imply that $\tilde{\mathbf{d}} = \mathbf{\Psi}^H \mathbf{d}$ is sparse. While both models assume sparsity of the STFT coefficients, the models are equivalent only if the analysis operator is equal to the inverse of the synthesis operator [21]. In the considered case this is not fulfilled since the STFT frame $\mathbf{\Psi}$ is overcomplete and the two models differ.

In the remainder of this section we present a brief overview of the alternating directions method of multipliers (ADMM) and the proposed formulations for a fixed sparsity-promoting cost function P, followed by a discussion on the selection of the sparsity promoting cost function and relation to other algorithms.

3.1. Alternating direction method of multipliers

The ADMM algorithm [29] is suitable for nonsmooth convex optimization problems of the form

$$\min_{\mathbf{d},\mathbf{g}} \quad P(\mathbf{d}) + Q(\mathbf{g}) \tag{5}$$

$$\text{subject to} \quad \mathbf{Ad} + \mathbf{Bg} = \mathbf{c}, \tag{6}$$

where the cost function can be conveniently split for the variables $\mathbf{d}$ and $\mathbf{g}$. The augmented Lagrangian for the problem (5) can be written as

$$L_\rho(\mathbf{d}, \mathbf{g}, \boldsymbol{\mu}) = P(\mathbf{d}) + Q(\mathbf{g}) + \frac{\rho}{2} \|\mathbf{Ad} + \mathbf{Bg} - \mathbf{c} + \boldsymbol{\mu}\|_2^2, \tag{7}$$

where $\boldsymbol{\mu}$ is a dual (splitting) variable and ρ is a penalty parameter. The ADMM algorithm minimizes L_ρ in (7) alternately with respect to $\mathbf{d}$ and $\mathbf{g}$, followed by a dual

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

ascent over the dual variable $\boldsymbol{\mu}$ [29], leading to the following update rules

$$\mathbf{d} \leftarrow \arg\min_{\mathbf{d}} P(\mathbf{d}) + \frac{\rho}{2}\|\mathbf{Ad}+\mathbf{Bg}-\mathbf{c}+\boldsymbol{\mu}\|_2^2, \quad (8)$$

$$\mathbf{g} \leftarrow \arg\min_{\mathbf{g}} Q(\mathbf{g}) + \frac{\rho}{2}\|\mathbf{Ad}+\mathbf{Bg}-\mathbf{c}+\boldsymbol{\mu}\|_2^2, \quad (9)$$

$$\boldsymbol{\mu} \leftarrow \boldsymbol{\mu} + \gamma(\mathbf{Ad}+\mathbf{Bg}-\mathbf{c}), \quad (10)$$

where $\gamma \in [1, (1+\sqrt{5})/2)$ can be used for faster convergence [30, 31]. An important ingredient of ADMM-based algorithms is the proximal mapping operator [32]. For a cost function P we define the proximal mapping S_λ^P of P as

$$S_\lambda^P(\mathbf{z}) = \arg\min_{\mathbf{d}} \lambda P(\mathbf{d}) + \frac{1}{2}\|\mathbf{z}-\mathbf{d}\|_2^2, \quad (11)$$

which can often be evaluated very efficiently [32]. More details about the ADMM algorithm can be found in [29].

3.2. Wideband model and analysis sparsity

First, we consider the estimation of the desired speech signal $\mathbf{d}$ in the time domain by enforcing the STFT coefficients of $\mathbf{d}$ to be sparse, subject to the wideband model in (3). The optimization problem for estimating $\mathbf{d}$ can be written as

$$\begin{aligned} \min_{\mathbf{d},\mathbf{g}} \quad & P\left(\boldsymbol{\Psi}^H\mathbf{d}\right) \\ \text{subject to} \quad & \mathbf{d}+\mathbf{Xg} = \mathbf{x}_{\text{ref}}. \end{aligned} \quad (12)$$

By applying the ADMM algorithm the obtained problem can be solved using the following update rules

$$\mathbf{d} \leftarrow \arg\min_{\mathbf{d}} P\left(\boldsymbol{\Psi}^H\mathbf{d}\right) + \frac{\rho}{2}\|\mathbf{d}+\mathbf{Xg}-\mathbf{x}_{\text{ref}}+\boldsymbol{\mu}\|_2^2, \quad (13)$$

$$\mathbf{g} \leftarrow \arg\min_{\mathbf{g}} \|\mathbf{d}+\mathbf{Xg}-\mathbf{x}_{\text{ref}}+\boldsymbol{\mu}\|_2^2, \quad (14)$$

$$\boldsymbol{\mu} \leftarrow \boldsymbol{\mu} + \gamma(\mathbf{d}+\mathbf{Xg}-\mathbf{x}_{\text{ref}}). \quad (15)$$

The update for $\mathbf{d}$ corresponds to a generalized Lasso problem [29] with the update rule

$$\mathbf{d} \leftarrow S_{1/\rho}^{P\circ\boldsymbol{\Psi}^H}\left(\mathbf{x}_{\text{ref}}-\mathbf{Xg}-\boldsymbol{\mu}\right), \quad (16)$$

which can be computed using ADMM as shown in Appendix B. The update for $\mathbf{g}$ can be written as

$$\mathbf{g} \leftarrow \left(\mathbf{X}^H\mathbf{X}\right)^{-1}\mathbf{X}^H\left(\mathbf{x}_{\text{ref}}-\mathbf{d}-\boldsymbol{\mu}\right) = \mathbf{g}_{\ell_2} - \mathbf{g}_{\text{iter}}, \quad (17)$$

where $\mathbf{g}_{\ell_2} = \left(\mathbf{X}^H\mathbf{X}\right)^{-1}\mathbf{X}^H\mathbf{x}_{\text{ref}}$ is iteration-independent and equal to the solution for the squared ℓ_2-norm as the cost function in (12), i.e., $P(\boldsymbol{\Psi}^H\mathbf{d}) = \|\boldsymbol{\Psi}^H\mathbf{d}\|_2^2$, and $\mathbf{g}_{\text{iter}} = \left(\mathbf{X}^H\mathbf{X}\right)^{-1}\mathbf{X}^H\left(\mathbf{d}+\boldsymbol{\mu}\right)$ is the iteration-dependent correction term. Note that the matrix $\mathbf{X}^H\mathbf{X}$ needs to be factored only once, and the cached factorization can be used to solve the corresponding linear system in subsequent iterations [29]. Since $\mathbf{X}$ is a block-convolution matrix, both $\mathbf{X}^H\mathbf{X}$ and $\mathbf{X}^H\mathbf{x}_{\text{ref}}$ can be efficiently computed using correlation. The computationally most demanding step is the evaluation of the STFT and its inverse at each iteration when solving (16).

3.3. Wideband model and synthesis sparsity

Next we consider the estimation of the STFT coefficients $\tilde{\mathbf{d}}$ of the desired speech signal by enforcing these coefficients to be sparse, subject to the wideband model in (3). The desired speech signal in the time domain is finally obtained by performing the inverse STFT of the estimated coefficients as $\mathbf{d} = \boldsymbol{\Psi}\tilde{\mathbf{d}}$. The optimization problem for estimating $\tilde{\mathbf{d}}$ can be written as

$$\begin{aligned} \min_{\tilde{\mathbf{d}},\mathbf{g}} \quad & P\left(\tilde{\mathbf{d}}\right) \\ \text{subject to} \quad & \boldsymbol{\Psi}\tilde{\mathbf{d}}+\mathbf{Xg} = \mathbf{x}_{\text{ref}}. \end{aligned} \quad (18)$$

The obtained problem can again be solved by applying the ADMM algorithm, resulting in the following iterations

$$\tilde{\mathbf{d}} \leftarrow \arg\min_{\tilde{\mathbf{d}}} P\left(\tilde{\mathbf{d}}\right) + \frac{\rho}{2}\|\boldsymbol{\Psi}\tilde{\mathbf{d}}+\mathbf{Xg}-\mathbf{x}_{\text{ref}}+\boldsymbol{\mu}\|_2^2 \quad (19)$$

$$\mathbf{g} \leftarrow \arg\min_{\mathbf{g}} \|\boldsymbol{\Psi}\tilde{\mathbf{d}}+\mathbf{Xg}-\mathbf{x}_{\text{ref}}+\boldsymbol{\mu}\|_2^2 \quad (20)$$

$$\boldsymbol{\mu} \leftarrow \boldsymbol{\mu} + \gamma\left(\boldsymbol{\Psi}\tilde{\mathbf{d}}+\mathbf{Xg}-\mathbf{x}_{\text{ref}}\right). \quad (21)$$

The update for $\tilde{\mathbf{d}}$ corresponds to a Lasso problem [33] (cf. Appendix A), and can be solved by applying the fast iterative shrinkage/thresholding algorithm (FISTA) [34, 35]. Similarly as in (17), the update for $\mathbf{g}$ can be written as

$$\mathbf{g} \leftarrow \left(\mathbf{X}^H\mathbf{X}\right)^{-1}\mathbf{X}^H\left(\mathbf{x}_{\text{ref}}-\boldsymbol{\Psi}\tilde{\mathbf{d}}-\boldsymbol{\mu}\right) = \mathbf{g}_{\ell_2} - \mathbf{g}_{\text{iter}}, \quad (22)$$

where $\mathbf{g}_{\ell_2}$ is the iteration-independent term, and $\mathbf{g}_{\text{iter}} = \left(\mathbf{X}^H\mathbf{X}\right)^{-1}\mathbf{X}^H\left(\boldsymbol{\Psi}\tilde{\mathbf{d}}+\boldsymbol{\mu}\right)$ is the iteration-dependent term. The computationally most demanding step is the evaluation of the STFT and its inverse at each iteration of FISTA.

3.4. Narrowband model

Finally, we consider the estimation of the STFT coefficients $\tilde{\mathbf{d}}$ of the desired speech signal by enforcing these coefficients to be sparse, subject to the narrowband model in (4). Since the signal model is independent across frequencies, we obtain a set of K optimization problems which can be solved independently. The optimization problem for estimating the STFT coefficients in the k-th frequency bin can be written as

$$\begin{aligned} \min_{\tilde{\mathbf{d}}_k,\tilde{\mathbf{g}}_k} \quad & P\left(\tilde{\mathbf{d}}_k\right) \\ \text{subject to} \quad & \tilde{\mathbf{d}}_k + \tilde{\mathbf{X}}_k\tilde{\mathbf{g}}_k = \tilde{\mathbf{x}}_{\text{ref},k}. \end{aligned} \tag{23}$$

The obtained problem can be solved by applying the ADMM algorithm, resulting in the following update rules

$$\tilde{\mathbf{d}}_k \leftarrow \arg\min_{\tilde{\mathbf{d}}_k} P\left(\tilde{\mathbf{d}}_k\right) + \frac{\rho}{2}\|\tilde{\mathbf{d}}_k + \tilde{\mathbf{X}}_k\tilde{\mathbf{g}}_k - \tilde{\mathbf{x}}_{\text{ref},k} + \tilde{\boldsymbol{\mu}}_k\|_2^2, \tag{24}$$

$$\tilde{\mathbf{g}}_k \leftarrow \arg\min_{\tilde{\mathbf{g}}_k} \|\tilde{\mathbf{d}}_k + \tilde{\mathbf{X}}_k\tilde{\mathbf{g}}_k - \tilde{\mathbf{x}}_{\text{ref},k} + \tilde{\boldsymbol{\mu}}_k\|_2^2, \tag{25}$$

$$\tilde{\boldsymbol{\mu}}_k \leftarrow \tilde{\boldsymbol{\mu}}_k + \gamma\left(\tilde{\mathbf{d}}_k + \tilde{\mathbf{X}}_k\tilde{\mathbf{g}}_k - \tilde{\mathbf{x}}_{\text{ref},k}\right). \tag{26}$$

The update for $\tilde{\mathbf{d}}_k$ can be immediately written as

$$\tilde{\mathbf{d}}_k = S^P_{1/\rho}\left(\tilde{\mathbf{x}}_{\text{ref},k} - \tilde{\mathbf{X}}_k\tilde{\mathbf{g}}_k - \tilde{\boldsymbol{\mu}}_k\right), \tag{27}$$

and the update for $\tilde{\mathbf{g}}_k$ can be written as

$$\tilde{\mathbf{g}}_k \leftarrow \left(\tilde{\mathbf{X}}_k^H\tilde{\mathbf{X}}_k\right)^{-1}\tilde{\mathbf{X}}_k^H\left(\tilde{\mathbf{x}}_{\text{ref}_k} - \tilde{\mathbf{d}}_k - \tilde{\boldsymbol{\mu}}_k\right) = \tilde{\mathbf{g}}_{k,\ell_2} - \tilde{\mathbf{g}}_{k,\text{iter}}, \tag{28}$$

where $\tilde{\mathbf{g}}_{k,\ell_2}$ is the iteration-independent term, and $\tilde{\mathbf{g}}_{k,\text{iter}} = \left(\tilde{\mathbf{X}}_k^H\tilde{\mathbf{X}}_k\right)^{-1}\tilde{\mathbf{X}}_k^H\left(\tilde{\mathbf{d}}_k + \tilde{\boldsymbol{\mu}}_k\right)$ is the iteration-dependent term. Similarly as before, the matrix $\tilde{\mathbf{X}}_k^H\tilde{\mathbf{X}}_k$ needs to be factored once and later used to solve the corresponding linear system in subsequent iterations. Note that this matrix is much smaller than the corresponding matrix in the wideband model, and the resulting iterations do not involve the STFT.

3.5. Sparsity promoting cost function

In order to promote sparsity of the STFT coefficients we need to select an appropriate cost function P. Typical cost functions include the ℓ_1-norm, nonconvex ℓ_p-norms with $p \in (0,1)$, or ℓ_0-norm, counting the number of nonzero elements of a vector. Alternatively, it is possible to define a proximal mapping such that the corresponding function is suitable for promoting sparsity [36]. In this paper we use a weighted ℓ_1-norm as the cost function, which can be written as

$$P\left(\tilde{\mathbf{d}}\right) = \|\tilde{\mathbf{d}}\|_{\mathbf{w},1} = \sum_{k,n} w_{k,n}|\tilde{d}_{k,n}|, \tag{29}$$

with nonnegative weights $\mathbf{w}$, which should represent information about the significant coefficients in the desired signal. Recovery of a sparse $\tilde{\mathbf{d}}$ using the cost function (29) is an iterative two-step procedure: (*i*) weights $\mathbf{w}$ are computed based on the previous estimate of $\tilde{\mathbf{d}}$, (*ii*) a new estimate of $\tilde{\mathbf{d}}$ is obtained by solving an appropriate optimization problem. All of the previously proposed algorithms are employed in such a reweighted procedure.

The proximal mapping for the weighted ℓ_1-norm in (29) can be computed element-wise using soft thresholding [35] as

$$S^P_\lambda\left(\tilde{d}_{k,n}\right) = \left(1 - \lambda\frac{w_{k,n}}{|\tilde{d}_{k,n}|}\right)_+ \tilde{d}_{k,n}, \tag{30}$$

where $(G)_+ = \max(G,0)$ [29]. In the context of speech enhancement [19], the proximal mapping in (30) can be interpreted as a gain function. As noted in [19, 37], in speech enhancement a lower bound on the gain is often introduced, i.e., $(G)_+ = \max(G, G_{\text{min}})$, to prevent suppression of the small coefficients $\tilde{d}_{k,n}$ to a value of exactly zero. It can be shown that this corresponds to a cost function P in the form of a Huber function [29], which is quadratic for small arguments and equal to a scaled absolute value for large arguments.

The weights $\mathbf{w}$ in (29) are often computed in such a way that the obtained weighted norm simulates behavior of the ℓ_0-norm [20, 38]. To achieve scaling sensitivity, as with the ℓ_0-norm, the weights can be computed as

$$w_{k,n} = \left(|\tilde{d}_{k,n}|^2 + \varepsilon\right)^{-1/2}, \tag{31}$$

where ε is a small regularization coefficient to prevent division by zero. In (31) the weight $w_{k,n}$ is computed locally using the true coefficient $\tilde{d}_{k,n}$. Since in practice the true coefficients are not available, the weights are computed based on the estimate of $\tilde{d}_{k,n}$ from the previous iteration. To take into account the TF structure of the desired signal, the concept of neighborhood was introduced in [22]. Assuming that a neighborhood $\mathcal{N}_{k,n}$ of the coefficient $\tilde{d}_{k,n}$ is defined, the corresponding weight can be computed as

$$w_{k,n} = \left(\sum_{(k',n')\in\mathcal{N}_{k,n}} \eta_{k',n'}|\tilde{d}_{k',n'}|^2 + \varepsilon\right)^{-1/2}, \tag{32}$$

where $\eta_{k',n'}$ are coefficients of the neighborhood that sum to one. Similarly as in [22, 24], we employ rectangular neighborhoods with equal weights. Note that here the neighborhood is used to compute the weights for the reweighted ℓ_1-norm, and not to derive a structured shrinkage operator as in [22].

Alternatively, it is well-known that speech spectrograms can be modeled well using a low-rank approximation [39,40]. Similarly as in [23], the weights can 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, and computing the weights as

$$w_{k,n} = \left(p_{k,n}+\varepsilon\right)^{-1/2}. \tag{33}$$

3.6. Relation to existing methods

Several MCLP-based speech dereverberation methods have been proposed employing the narrowband signal model in (4) and a locally Gaussian model of the desired speech signal, e.g., [6,26,27,41]. This typically results in an iteratively reweighted least-squares minimization, with the cost function being equal to a weighted ℓ_2-norm [23], i.e.,

$$P\left(\tilde{\mathbf{d}}_k\right) = \|\tilde{\mathbf{d}}_k\|_{\mathbf{w}_k,2}^2 = \sum_n w_{k,n}|\tilde{d}_{k,n}|^2 \tag{34}$$

The weights for the current iteration are computed based on the estimate of the desired speech signal from the previous iteration as $w_{k,n} \leftarrow \left(|\tilde{d}_{k,n}|^2+\varepsilon\right)^{-1}$ [6,9]. The prediction filters are obtained by solving the following problem

$$\min_{\tilde{\mathbf{g}}_k} \quad \|\tilde{\mathbf{x}}_{\mathrm{ref},k} - \tilde{\mathbf{X}}_k\tilde{\mathbf{g}}_k\|_{\mathbf{w}_k,2}^2, \tag{35}$$

having a closed-form solution. The estimated desired signal coefficients in the k-th frequency bin are computed as $\tilde{\mathbf{d}}_k \leftarrow \tilde{\mathbf{x}}_{\mathrm{ref},k} - \tilde{\mathbf{X}}_k\tilde{\mathbf{g}}_k$. This algorithm is known as the weighted prediction error (WPE). In [23] it was proposed to compute the weights using a low-rank approximation of the power spectrogram, and here we additionally propose to compute the weights for WPE using neighborhoods, similarly as in (32). A locally Laplacian model for the desired speech signal coefficients was used in [42] resulting in an iteratively reweighted ℓ_1 minimization. On the other hand, the wide-band signal model has been employed in MCLP-based dereverberation [6, 43, 44], however without explicitly enforcing sparsity of the desired signal.

4. SIMULATIONS

The performance of the presented methods for multi-microphone dereverberation has been evaluated through several simulations. We have considered an acoustic system with RIRs from the REVERB challenge [45]. The acoustic system (AC) consists of $M = 2$ microphones in a room with a reverberation time (T_{60}) of approximately 700 ms, and the distance between the source and the microphones was approximately 2 m. The microphone signals were generated by convolving the RIRs with a clean speech sample (length $\sim$ 6 s) from [45], and the sampling frequency in all experiments was $f_s = 16$ kHz. The dereverberation performance was evaluated using cepstral distance (CD), perceptual evaluation of speech quality (PESQ), and frequency-weighted segmental signal-to-noise ratio (FWS), with the clean speech as the reference signal [45]. In the following we compare the WPE method and the three proposed methods based on ADMM: wideband model with analysis sparsity (WB-A), wideband model with synthesis sparsity (WB-S), and the narrowband model (NB).

4.1. Implementation details

The STFT was computed using a tight frame 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. The prediction delay was set to $L_\tau = 256$ for the wideband model and $\tilde{L}_\tau = 2$ for the narrowband model. The weights were computed in three different ways: locally, using a rectangular neighborhood, and using a low-rank approximation. In all experiments we initialize the estimation of $\tilde{\mathbf{d}}$ using the reference microphone signal. A small positive constant $\varepsilon = 10^{-8}$ was added to the weights and they were additionally normalized so that the largest weight is equal to 1. The neighborhood-based weights were computed using a neighborhood which includes the current coefficient and 2 adjacent coefficients on each side across the frequencies. In general, neighborhoods across the temporal direction typically resulted in a decreased dereverberation performance, due to the fact that such neighborhoods preserve persistence of the signal [24], which can be counterproductive for dereverberation. A low-rank approximation was computed as in [23] using NMF with Itakura-Saito divergence with the rank set to 30. The maximal number of iterations of the ADMM for the proposed methods was set to 50, with the stopping criteria as in [29] and a relative tolerance of 10^{-3}. The parameter ρ was selected from the set $\left\{10^{-4}, 5\cdot10^{-4},\ldots,1\right\}$ intrusively, as the one yielding the highest improvement in terms of PESQ score. In the simulations we used a minimum gain equal to $G_{\min} = 0.01$. Setting the minimum gain to 0 resulted in a higher dereverberation performance but also a higher speech distortion, which was reflected in the instrumental measures. Setting the minimum gain to a higher value (e.g. 0.1) typically resulted in decreased dereverberation performance. For the Lasso problem 20 iterations of FISTA were performed, while for the generalized Lasso 20 iterations of ADMM with the penalty $\rho = 1$ were performed.

4.2. Evaluation

In the Table 1 we state the results obtained using a different number of reweighting iterations. For a single reweighting iteration ($i_{\mathrm{RW}} = 1$), the proposed methods outperform the baseline WPE when locally computed weights are used. The use of neighborhood-based weights increases the performance for most of the methods (except WB-S), with WPE achieving the highest gains when compared to the local weights. The use of the low-rank approximation-based weights results in further improvements for most of the methods. The methods based on the narrowband approximation, i.e., WPE and NB, result in a similar performance while the wideband WB-A seems to perform somewhat better.

Using five reweighting iterations ($i_{\mathrm{RW}} = 5$) WB-S performs somewhat worse while other achieve a similar performance when locally computed weights are used. Neighborhood-based and low-rank approximation-based weights improve the performance for WPE, with minor differences for the other methods. Overall, all of the evaluated methods achieve a similar dereverberation performance, with WB-S not improving in the experiment when multiple reweighting iterations are performed. In terms of computational complexity the narrowband methods are clearly preferred: average times for one reweighting iteration for 6 s of speech with $M = 2$ microphones are about 1.5 s, 3 s, 2 min and 3 min for WPE, NB, WB-S and WB-A (with all algorithms implemented in Matlab). As noted earlier, the wideband methods are much more complex due to the evaluation of the STFT and its inverse in the iterative procedure [14, 20].

5. CONCLUSION

In this paper, we have presented a general framework for MCLP-based speech dereverberation exploiting sparsity of the speech signal. Three different formulations have been proposed, i.e., using a wideband signal model and an analysis or a synthesis model for the desired speech signal, and using a narrowband signal model. The obtained optimization problems are solved using ADMM, offering flexibility for additional constraints or alternative sparsity promoting functions. Additionally, as a step in the direction of exploiting the speech signal structure we include a concept of structured weights, using neighborhood or low-rank approximation, which further improve speech dereverberation performance.

Table 1: Obtained results for the considered AC. Values of CD, FWS and PESQ for the reference microphone were 4.4, 5.3 and 1.9, respectively.

Alg.	Weights	ΔCD		ΔFWS		ΔPESQ	
i_{RW}		1	5	1	5	1	5
WPE	local	1.3	2.3	3.3	4.4	0.7	1.2
	neigh.	1.8	2.5	3.8	4.8	1.1	1.4
	NMF	1.9	2.6	4.1	5.0	1.1	1.5
NB	local	1.5	2.1	2.2	4.2	1.1	1.2
	neigh.	1.8	2.1	3.0	3.4	1.2	1.4
	NMF	1.9	2.1	3.1	3.4	1.2	1.4
WB-S	local	1.6	1.6	3.0	3.0	1.1	1.1
	neigh.	1.6	1.6	2.8	2.8	1.1	1.2
	NMF	1.6	1.6	2.9	2.9	1.1	1.1
WB-A	local	1.8	2.3	3.4	4.5	1.2	1.4
	neigh.	1.9	2.3	3.7	4.5	1.3	1.4
	NMF	2.1	2.3	3.9	4.6	1.3	1.4

6. APPENDIX

6.1. Lasso

An optimization problem in the form

$$\min_{\tilde{\mathbf{d}}} \lambda P\left(\tilde{\mathbf{d}}\right) + \frac{1}{2}\|\mathbf{\Psi}\tilde{\mathbf{d}} - \mathbf{z}\|_2^2, \tag{36}$$

where P is a weighted ℓ_1-norm is in the form of Lasso and can be solved by applying a forward-backward algorithm [32, 34, 35], also known as the iterative shrinkage/thresholding algorithm (ISTA), consisting of the following iteration

$$\tilde{\mathbf{d}} \leftarrow \mathcal{S}^{P}_{\lambda/\nu}\left(\tilde{\mathbf{d}} + \frac{1}{\nu}\mathbf{\Psi}^H\left(\mathbf{z} - \mathbf{\Psi}\tilde{\mathbf{d}}\right)\right), \tag{37}$$

with ν being the maximum eigenvalue of $\mathbf{\Psi}\mathbf{\Psi}^H$, i.e., $\nu = 1$. An accelerated fast ISTA (FISTA) algorithm [35] employs a similar iteration, with the same complexity and improved convergence.

6.2. Generalized Lasso

An optimization problem in the form

$$\min_{\mathbf{d}} \quad \lambda P\left(\mathbf{\Psi}^H\mathbf{d}\right) + \frac{1}{2}\|\mathbf{d} - \mathbf{z}\|_2^2, \tag{38}$$

where P is a weighted ℓ_1-norm is an instance of generalized Lasso [29]. The minimizer of the cost function is equal to the proximal mapping of the composition of the analysis operator $\mathbf{\Psi}^H$ and the function P, i.e., the optimal $\mathbf{d}$ is equal to $\mathcal{S}^{P\circ\mathbf{\Psi}^H}_{\lambda}(\mathbf{z})$, where $\circ$ denotes the function composition. Applying ADMM results in the following iterations

$$\mathbf{d} \leftarrow \frac{1}{1+\rho}\left[\mathbf{z}+\rho\mathbf{\Psi}\left(\tilde{\mathbf{u}}-\tilde{\boldsymbol{\mu}}\right)\right], \quad (39)$$

$$\tilde{\mathbf{u}} \leftarrow S^{P}_{\lambda/\rho}\left(\mathbf{\Psi}^{H}\mathbf{d}+\tilde{\boldsymbol{\mu}}\right), \quad (40)$$

$$\tilde{\boldsymbol{\mu}} \leftarrow \tilde{\boldsymbol{\mu}}+\gamma\left(\mathbf{\Psi}^{H}\mathbf{d}-\tilde{\mathbf{u}}\right), \quad (41)$$

where $\tilde{\mathbf{u}}$ is a splitting variable satisfying the constraint $\mathbf{\Psi}^{H}\mathbf{d}-\tilde{\mathbf{u}}=\mathbf{0}$.

7. REFERENCES

[1] P. A. Naylor and N. D. Gaubitch, *Speech Dereverberation*, Springer, 2010.

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

[3] 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, Sept. 2013.

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

[5] 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, June 2009.

[6] 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, Sept. 2010.

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

[8] B. Schwartz, S. Gannot, and E.A.P. Habets, “Online Speech Dereverberation Using Kalman Filter and EM Algorithm,” *IEEE/ACM Trans. Audio Speech Lang. Process.*, vol. 23, no. 2, pp. 394–406, Feb 2015.

[9] 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 Lang. Process.*, vol. 23, no. 9, pp. 1509–1520, 2015.

[10] P. Bofil and M. Zibulevsky, “Blind separation of more sources than mixtures using sparsity of their short-time Fourier transform,” in *Proc. ICA*, Helsinki, Finland, June 2000, pp. 87–92.

[11] M. Zibulevsky and B. Pearlmutter, “Blind source separation by sparse decomposition in a signal dictionary,” *Neural Computation*, vol. 13, no. 4, pp. 863–882, 2001.

[12] P.G. Georgiev, F. Theis, and A. Cichocki, “Sparse component analysis and blind source separation for underdetermined mixtures,” *IEEE Trans. Neural Networks*, vol. 16, no. 4, pp. 992–996, July 2005.

[13] S. Makino, S. Araki, S. Winter, and H. Sawada, “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.

[14] 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.

[15] A. Adler *et al.*, “Audio inpainting,” *IEEE Trans. Audio Speech Lang. Process.*, vol. 20, no. 3, pp. 922–932, Mar. 2012.

[16] B. Defraene, N. Mansour, S. De Hertogh, T. van Waterschoot, M. Diehl, and M. Moonen, “Declipping of audio signals using perceptual compressed sensing,” *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 21, no. 12, pp. 2627–2637, Dec 2013.

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

[18] T. van Waterschoot, B. Defraene, M. Diehl, and M. Moonen, “Embedded optimization algorithms for multi-microphone dereverberation,” in *Proc. European Signal Process. Conf. (EUSIPCO)*, Marrakech, Morocco, Sept. 2013.

[19] 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 Lectures on Speech and Audio Processing*, vol. 9, no. 1, pp. 1–80, Jan. 2013.

[20] S. Arberet, P. Vandergheynst, R.E. Carrillo, J.-P. Thiran, and Y. Wiaux, “Sparse reverberant audio source separation via reweighted analysis,” *IEEE Trans. Audio Speech Lang. Process.*, vol. 21, no. 7, pp. 1391–1402, July 2013.

[21] M. Elad, P. Milanfar, and R. Rubinstein, “Analysis versus synthesis in signal priors,” *Inverse problems*, vol. 23, no. 3, pp. 947, 2007.

[22] 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, May 2013.

[23] A. Jukić, N. Mohammadiha, T. van Waterschoot, T. Gerkmann, and S. Doclo, "Multi-channel linear prediction-based speech dereverberation with low-rank power spectrogram approximation," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, Brisbane, Australia, Apr. 2015, pp. 96–100.

[24] K. Siedenburg and M. Dörfler, "Audio denoising by generalized time-frequency thresholding," in *Audio Eng. Soc. Conf.: Appl. Time-Frequency Process. Audio*, Mar. 2012.

[25] 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, June 2003.

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

[27] M. Togami, Y. Kawaguchi, R. Takeda, Y. Obuchi, and N. Nukaga, "Optimized speech dereverberation from probabilistic perspective for time varying acoustic transfer function," *IEEE Trans. Audio Speech Lang. Process.*, vol. 21, no. 7, pp. 1369–1380, July 2014.

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

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

[30] M. Fortin and R. Glowinski, *Augmented Lagrangian Methods*, Elsevier, 1983.

[31] R. Chartrand, "Nonconvex compressive sensing for x-ray ct: an algorithm comparison," in *Proc. Asilomar Conf. Signals, Syst. Computers (ASILOMAR)*, Pacific Grove, CA, USA, Nov. 2013, pp. 665–669.

[32] N. Parikh and S. Boyd, "Proximal algorithms," *Foundations and Trends in Machine Learning*, vol. 1, no. 3, pp. 127–239, 2014.

[33] R. Tibshirani, "Regression shrinkage and selection via the lasso," *Journal of the Royal Statistical Society, Series B*, vol. 58, pp. 267–288, 1996.

[34] P. Combettes and V. Wajs, "Signal recovery by proximal forwardbackward splitting," *Multiscale Modeling and Simulation*, vol. 4, no. 4, pp. 1168–1200, Nov. 2005.

[35] A. Beck and M. Teboulle, "A fast iterative shrinkage-thresholding algorithm for linear inverse problems," *SIAM J. Imaging Sciences*, vol. 2, no. 1, pp. 183–202, 2009.

[36] R. Chartrand, "Shrinkage mappings and their induced penalty functions," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, Florence, Italy, May 2014, pp. 1026–1029.

[37] D. Malah, R.V. Cox, and A.J. Accardi, "Tracking speech-presence uncertainty to improve speech enhancement in non-stationary noise environments," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, Mar 1999, vol. 2, pp. 789–792 vol.2.

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

[39] C. Fevotte, 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.

[40] 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, Oct 2013.

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

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

[43] T. Nakatani, B.-H. Juang, T. Yoshioka, K. Kinoshita, M. Delcroix, and M. Miyoshi, "Speech dereverberation based on maximum-likelihood estimation with time-varying gaussian source model," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 16, no. 8, pp. 1512–1527, Nov 2008.

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

[45] K. Kinoshita *et al.*, "The REVERB challenge: A common evaluation framework for dereverberation and recognition of reverberant speech," in *Proc. IEEE Workshop Appl. Signal Process. Audio Acoust. (WASPAA)*, New Paltz, USA, Oct. 2013.