GROUP SPARSITY FOR MIMO SPEECH DEREVERBERATION

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
ante.jukic@uni-oldenburg.de

ABSTRACT

Reverberation can severely affect the speech signals recorded in a room, possibly leading to a significantly reduced speech quality and intelligibility. In this paper we present a batch algorithm employing a signal model based on multi-channel linear prediction in the short-time Fourier transform domain. Aiming to achieve multiple-input multiple-output (MIMO) speech dereverberation in a blind manner, we propose a cost function based on the concept of group sparsity. To minimize the obtained nonconvex function, an iteratively reweighted least-squares procedure is used. Moreover, it can be shown that the derived algorithm generalizes several existing speech dereverberation algorithms. Experimental results for several acoustic systems demonstrate the effectiveness of nonconvex sparsity-promoting cost functions in the context of dereverberation.

Index Terms— speech dereverberation, multi-channel linear prediction, group sparsity

1. INTRODUCTION

Recordings of a speech signal in an enclosed space with microphones placed at a distance from the speaker are typically affected by reverberation, which is caused by reflections of the sound against the walls and objects in the enclosure. While moderate levels of reverberation may be beneficial, in severe cases it typically results in a decreased speech intelligibility and automatic speech recognition performance [1, 2]. Therefore, effective dereverberation is required for various speech communication applications, such as hands-free telephony, hearing aids, or voice-controlled systems [2, 3]. Many dereverberation methods have been proposed during the last decade [3], such as methods based on acoustic multi-channel equalization [4, 5], spectral enhancement [6, 7], or probabilistic modeling [8–13].

Several dereverberation methods employ the multi-channel linear prediction (MCLP) model to estimate the clean speech signal [8–10, 14]. The main idea of MCLP-based methods is to decompose the reverberant microphone signals into a desired and an undesired component, which can be predicted from the previous samples of all microphone signals. Estimation of the prediction coefficients for a multiple-input single-output dereverberation system, with multiple microphones and a single output signal, has been formulated using a time-varying Gaussian model in [8], while generalized sparse priors have been used in [14]. A generalization of [8] to a multiple-input multiple-output (MIMO) dereverberation system, based on a time-varying multivariate Gaussian model, has been proposed in [9] and is referred to as the generalized weighted prediction error (GWPE) method. The GWPE method has been extended for a time-varying acoustic scenario in [10], as well as for joint dereverberation and suppression of diffuse noise [15].

In this paper, we consider a MIMO system and formulate the estimation of the prediction filters using a cost function based on the concept of group sparsity [16–18]. It is well known that speech signals are sparse in the short-time Fourier transform (STFT) domain and that reverberation decreases sparsity [19–21]. The main idea of the proposed cost function is to estimate the prediction coefficients that make the estimated desired speech signal in the STFT domain more sparse than the observed reverberant microphone signals. Using the concept of mixed norms [22], the proposed cost function takes into account the group structure of the coefficients across the microphones. More specifically, the cost function aims to estimate prediction coefficients that make the STFT coefficients of the desired speech signal sparse over time, whilst taking into account the spatial correlation between the channels. The obtained nonconvex problem is solved using the iteratively reweighted least squares method [23]. Furthermore, the derived batch algorithm generalizes several previously proposed speech dereverberation algorithms [8, 9, 14]. The performance of the proposed method is evaluated for several acoustic systems, and the obtained results show the nonconvex cost functions outperform the convex case.

2. SIGNAL MODEL

We consider a single speech source recorded using M microphones in a noiseless scenario. Let $s(k,n)$ denote the clean speech signal in the STFT domain, with $k \in \{1,\dots,K\}$ the frequency bin index and $n \in \{1,\dots,N\}$ the time frame index. The STFT coefficients of the observed noiseless reverberant signal at the m-th microphone $x_m(k,n)$ can be modeled as

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

where the L_h coefficients $h_m(k,l)$ represents the convolutive transfer function between the source and m-th microphone [12, 13], and $e_m(k,n)$ models the error of the approximation in a single band [24]. Several dereverberation algorithms are based on an autoregressive model of reverberation, subsequently using MCLP to estimate the undesired reverberation [8–10, 25]. Assuming the model in (1) holds perfectly and the error term can be disregarded, e.g., as in [8, 9], the reverberant signal at the m-th microphone can be written as

$$x_m(k,n) = d_m(k,n) + r_m(k,n). \qquad (2)$$

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).

The first term $d_m(k,n) = \sum_{l=0}^{\tau-1} h_m(k,l)s(k,n-l)$, with τ being a parameter, models the desired speech signal at the m-th microphone consisting of the direct speech and early reflections, which can be useful in speech communication [26]. The second term $r_m(k,n) = \sum_{l=\tau}^{L_h-1} h_m(k,l)s(k,n-l)$ models the remaining undesired reverberation. When $M > 1$ the undesired term at time frame n can be predicted from the previous microphone samples on all M microphones delayed by τ, as used in, e.g., [8–10]. Using M prediction filters of length L_g the undesired term $r_m(k,n)$ can be written as

$$r_m(k,n) = \sum_{i=1}^{M} \sum_{l=0}^{L_g-1} x_i(k,n-\tau-l)g_{m,i}(k,l), \tag{3}$$

where $g_{m,i}(k,l)$ is the l-th prediction coefficient between the i-th and the m-th channel. The signal model in (2) can be rewritten in vector notation as

$$\mathbf{x}_m(k) = \mathbf{d}_m(k) + \tilde{\mathbf{X}}_\tau(k)\mathbf{g}_m(k), \tag{4}$$

with vectors

$$\mathbf{x}_m(k) = [x_m(k,1),\ldots,x_m(k,N)]^T,$$
$$\mathbf{d}_m(k) = [d_m(k,1),\ldots,d_m(k,N)]^T,$$

and the multi-channel convolution matrix

$$\tilde{\mathbf{X}}_\tau(k) = \left[\tilde{\mathbf{X}}_{\tau,1}(k),\ldots,\tilde{\mathbf{X}}_{\tau,M}(k)\right]$$

where $\tilde{\mathbf{X}}_{\tau,m}(k) \in \mathbb{C}^{N\times L_g}$ is the convolution matrix of $\mathbf{x}_m(k)$ delayed for τ samples. The vector $\mathbf{g}_m(k) \in \mathbb{C}^{ML_g}$ contains the prediction coefficients $g_{m,i}(k,l)$ between the m-th channel and all other M channels. In the following we omit the frequency bin index k, since the model in (4) is applied in each frequency bin independently. Defining the M-channel input matrix $\mathbf{X} = [\mathbf{x}_1,\ldots,\mathbf{x}_M]$, the M-channel output matrix $\mathbf{D} = [\mathbf{d}_1,\ldots,\mathbf{d}_M]$, the prediction coefficients in $\mathbf{G} = [\mathbf{g}_1,\ldots,\mathbf{g}_M]$, and using (4), a MIMO signal model in each frequency bin can be written as

$$\mathbf{X} = \mathbf{D} + \tilde{\mathbf{X}}_\tau\mathbf{G}, \tag{5}$$

The problem of speech dereverberation, i.e., estimation of the desired speech signal $\mathbf{D}$, is now reduced to the estimation of the prediction coefficients $\mathbf{G}$ for predicting the undesired reverberation.

3. GROUP SPARSITY FOR SPEECH DEREVERBERATION

In this section we formulate speech dereverberation as an optimization problem with a cost function promoting group-sparsity, and propose to solve it using iteratively reweighted least squares (IRLS). We start with defining mixed norms and briefly review their relationship to group sparsity.

3.1. Mixed norms and group sparsity

Mixed norms are often used in the context of sparse signal processing [18,22]. Let $\mathbf{D} \in \mathbb{C}^{N\times M}$ be a matrix with elements $d_{n,m}$, with the elements of its n-th row contained in a (column) vector $\mathbf{d}_{n,:}$, i.e., $\mathbf{d}_{n,:} = [d_{n,1},\ldots,d_{n,M}]^T$. Let $p \geq 1$, and $\mathbf{\Phi} \in \mathbb{C}^{M\times M}$ be a positive definite matrix. We define the mixed norm $\ell_{\mathbf{\Phi};2,p}$ of the matrix $\mathbf{D}$ as

$$\|\mathbf{D}\|_{\mathbf{\Phi};2,p} = \left(\sum_{n=1}^{N} \|\mathbf{d}_{n,:}\|_{\mathbf{\Phi};2}^p\right)^{1/p}, \tag{6}$$

where $\|\mathbf{d}_{n,:}\|_{\mathbf{\Phi};2} = \sqrt{\mathbf{d}_{n,:}^H\mathbf{\Phi}^{-1}\mathbf{d}_{n,:}}$ is the $\ell_{\mathbf{\Phi};2}$ norm of the vector $\mathbf{d}_{n,:}$. The role of the matrix $\mathbf{\Phi}$ is to model the correlation structure within each group, i.e., row of $\mathbf{D}$. When $\mathbf{\Phi} = \mathbf{I}$ we denote the corresponding mixed norm as $\ell_{2,p}$. In words, the mixed $\ell_{\mathbf{\Phi};2,p}$ norm of $\mathbf{D}$ is composed of the inner $\ell_{\mathbf{\Phi};2}$ norm applied on the rows of $\mathbf{D}$ in the first step, and the outer ℓ_p norm applied on the vector composed of the values obtained in the first step. Intuitively, the inner $\ell_{\mathbf{\Phi};2}$ norm measures the energy of the coefficients in each row, while the outer ℓ_p norm is applied on the obtained energies and measures the number of rows with significant energies, i.e., the mixed norm $\ell_{\mathbf{\Phi};2,p}$ provides a measure of group sparsity of $\mathbf{D}$, with groups being the rows of $\mathbf{D}$. Therefore, minimization of (6) aims at estimating a matrix $\mathbf{D}$ that has some rows with a significant energy (in terms of the $\ell_{\mathbf{\Phi};2}$ norm) and the remaining rows have a small energy.

Mixed norms generalize the usual matrix and vector norms [22, 27], e.g., $\ell_{2,2}$ is the Frobenius norm of a matrix. A commonly used mixed norm is $\ell_{2,1}$, which is well known as Group-Lasso [16] or joint sparsity [17], and it is often used in sparse regression with the goal of keeping or discarding entire groups (here rows) of elements in a matrix [27]. Similarly as in the case of a vector norm, for $p \in [0,1)$ in (6) the obtained functional is not a norm since it is not convex. Still, we will refer to $\ell_{\mathbf{\Phi};2,p}$ for $p < 1$ as a norm.

3.2. Proposed formulation

In this paper we propose to estimate the prediction coefficients $\mathbf{G}$ by solving the following optimization problem

$$\begin{aligned} \min_{\mathbf{G}} \quad & \|\mathbf{D}\|_{\mathbf{\Phi};2,p}^p = \sum_{n=1}^{N} \|\mathbf{d}_{n,:}\|_{\mathbf{\Phi};2}^p \\ \text{subject to} \quad & \mathbf{D} = \mathbf{X} - \tilde{\mathbf{X}}_\tau\mathbf{G} \end{aligned} \tag{7}$$

for $p \leq 1$. The motivation behind the proposed cost function is to estimate such prediction filters $\mathbf{G}$ that result in some rows with significant energy in $\mathbf{D}$, and suppress the coefficients in the remaining rows. For $p = 1$ and $\mathbf{\Phi} = \mathbf{I}$ the cost function in (7) is the $\ell_{2,1}$ norm as in Group-Lasso, with the groups being defined across the M channels. While for $p = 1$ the cost function in (7) is convex, it is known that nonconvex penalty functions can be more useful in enforcing sparsity [28].

The proposed cost function for speech dereverberation with multiple microphones is motivated with the following common assumptions in the context of multi-channel speech processing. Firstly, due to reverberation, the STFT-domain coefficients of the microphone signals are less sparse than the STFT-domain coefficients of the corresponding clean speech signal [19–21]. Therefore, it is reasonable to estimate prediction filters that result in an estimate of the desired speech signal that is more sparse than the microphone signals. Secondly, for relatively small arrays it is plausible to assume that at a given time frame the speech signal is present or absent simultaneously on all channels [9]. Therefore, it is reasonable to formulate estimation of the prediction filters using a cost function promoting group sparsity as in (7), with the groups defined across the channels and the matrix $\mathbf{\Phi}$ capturing the spatial correlation between the channels. The prediction filters obtained

by solving (7) aim to estimate the desired speech signal coefficients $\mathbf{D}$ that are more sparse than the reverberant speech coefficients in $\mathbf{X}$, by simultaneously keeping or discarding the coefficients across the channels. Therefore, the undesired reverberation will be suppressed, with the spatial correlation (group structure) being taken into account.

3.3. Nonconvex minimization using IRLS

A class of algorithms for solving ℓ_p norm minimization problems is based on iteratively reweighted least squares [23]. The idea is to replace the original cost function with a series of convex quadratic problems. Namely, in every iteration the ℓ_p norm is approximated by a weighted ℓ_2 norm [23]. The same idea is applied here, i.e., the $\ell_{\mathbf{\Phi};2,p}$ norm in (7) is approximated with a weighted $\ell_{\mathbf{\Phi};2,2}$ norm. Therefore, in the i-th iteration the ℓ_p norm of the energies of the rows of $\mathbf{D}$ is replaced by a weighted ℓ_2 norm, resulting in the following approximation

$$\sum_{n=1}^{N} \|\mathbf{d}_{n,:}\|_{\mathbf{\Phi};2}^{p} \approx \sum_{n=1}^{N} w_n^{(i)} \|\mathbf{d}_{n,:}\|_{\mathbf{\Phi};2}^{2} = \mathrm{tr}\left\{\mathbf{W}^{(i)}\mathbf{D}\mathbf{\Phi}^{-T}\mathbf{D}^{H}\right\}, \tag{8}$$

where $\mathbf{W}^{(i)}$ is a diagonal matrix with the weights $w_n^{(i)}$ on its diagonal, and $\mathrm{tr}\{.\}$ denoting the trace operator. Similarly as in [23], the weights $w_n^{(i)}$ are selected such that the approximation in (8) is a first-order approximation of the corresponding $\ell_{\mathbf{\Phi};2,p}$ cost function, and therefore the n-th weight can be expressed as $w_n^{(i)} = \|\mathbf{d}_{n,:}\|_{\mathbf{\Phi};2}^{p-2}$. In the i-th iteration, the weights $w_n^{(i)}$ are computed from the previous estimate of the desired speech signal $\mathbf{D}^{(i-1)}$, i.e., as $w_n^{(i)} = \|\mathbf{d}_{n,:}^{(i-1)}\|_{\mathbf{\Phi};2}^{p-2}$. To prevent a division by zero, a small positive constant ε can be included in the weight update [23]. Given the weights $w_n^{(i)}$, the optimization problem using approximation in (8) can be written as

$$\min_{\mathbf{G}} \mathrm{tr}\left\{\left(\mathbf{X} - \tilde{\mathbf{X}}_\tau\mathbf{G}\right)^{H}\mathbf{W}^{(i)}\left(\mathbf{X} - \tilde{\mathbf{X}}_\tau\mathbf{G}\right)\mathbf{\Phi}^{-T}\right\}, \tag{9}$$

with the solution for the prediction filters given as

$$\mathbf{G}^{(i)} = \left(\tilde{\mathbf{X}}_\tau^{H}\mathbf{W}^{(i)}\tilde{\mathbf{X}}_\tau\right)^{-1}\tilde{\mathbf{X}}_\tau^{H}\mathbf{W}^{(i)}\mathbf{X}. \tag{10}$$

Note that the obtained solution does not depend on the matrix $\mathbf{\Phi}$. However, the choice of $\mathbf{\Phi}$ affects the calculation of the weights $w_n^{(i)}$, and can therefore influence the final estimate. Additionally, the matrix $\mathbf{\Phi}$, capturing the spatial (within-group) correlation, can be updated using the current estimate $\mathbf{D}^{(i)}$ of the desired speech signal as

$$\mathbf{\Phi}^{(i)} = \frac{1}{N}\sum_{n=1}^{N} w_n^{(i)}\mathbf{d}_{n,:}^{(i)}\mathbf{d}_{n,:}^{(i)H} = \frac{1}{N}\mathbf{D}^{(i)T}\mathbf{W}^{(i)}\mathbf{D}^{(i)*}, \tag{11}$$

with $(.)^*$ denoting complex conjugate. This update can be obtained by minimizing the cost function in (9) with an additional term $(N \log\det\mathbf{\Phi})$. The obtained expression can be interpreted as a maximum-likelihood estimator of $\mathbf{\Phi}$ when $\mathbf{d}_{n,:}$ is modeled using a zero-mean complex Gaussian distribution with covariance $w_n^{-1}\mathbf{\Phi}$, as commonly used in speech enhancement and group sparse learning [29]. The complete algorithm for solving (7) using IRLS is outlined in Algorithm 1.

Algorithm 1 MIMO speech dereverberation with group sparsity using IRLS.

parameters: Filter length L_g and prediction delay τ in (3), p in (7), regularization parameter ε, maximum number of iterations $i_{\max}$, tolerance η
input: STFT coefficients of the observed signals $\mathbf{X}(k), \forall k$
for all k **do**
  $i \leftarrow 0$, $\mathbf{D}^{(0)} \leftarrow \mathbf{X}$, $\mathbf{\Phi}^{(0)} \leftarrow \mathbf{I}$
  repeat
    $i \leftarrow i + 1$
    $w_n^{(i)} \leftarrow \left(\|\mathbf{d}_{n,:}^{(i-1)}\|_{\mathbf{\Phi}^{(i-1)};2}^{2} + \varepsilon\right)^{p/2-1}, \forall n$
    $\mathbf{G}^{(i)} \leftarrow \left(\tilde{\mathbf{X}}_\tau^{H}\mathbf{W}^{(i)}\tilde{\mathbf{X}}_\tau\right)^{-1}\tilde{\mathbf{X}}_\tau^{H}\mathbf{W}^{(i)}\mathbf{X}$
    $\mathbf{D}^{(i)} = \mathbf{X} - \tilde{\mathbf{X}}_\tau\mathbf{G}^{(i)}$
    if estimate $\mathbf{\Phi}$ **then** $\mathbf{\Phi}^{(i)} \leftarrow \frac{1}{N}\mathbf{D}^{(i)T}\mathbf{W}^{(i)}\mathbf{D}^{(i)*}$
  until $\|\mathbf{D}^{(i)} - \mathbf{D}^{(i-1)}\|_F / \|\mathbf{D}^{(i)}\|_F < \eta$ or $i \geq i_{\max}$
end for

3.4. Relation to existing methods

The GWPE method in [9] was derived based on a locally Gaussian model for the multi-channel desired signal, with the variances being unknown and time and frequency varying. The obtained optimization problem was formulated using a cost function based the on Hadamard-Fischer mutual correlation, which favors temporally uncorrelated random vectors. An appropriate auxiliary (majorizing) function was used to derive a practical algorithm based on alternating optimization. By comparing Algorithm 1 with the updates in [9], it can be seen that the GWPE method corresponds to the proposed method when $p = 0$, i.e., to the minimization of the $\ell_{\mathbf{\Phi};2,0}$ norm in (7). Furthermore, if an $\ell_{p,p}$ norm is used as the cost function in (7) the proposed method is reduced to a multiple-input single-output method [14] applied M times to generate M outputs, with each microphone being selected as the reference exactly once. In this case, the group structure is disregarded and the resulting cost function is equal to the ℓ_p norm applied element-wise on $\mathbf{D}$, meaning that the prediction coefficients for each output are calculated independently. The special case of $p = 0$ corresponds to the variance-normalized MCLP proposed originally in [8]. The considered MCLP-based algorithms have in common that the used cost functions promote sparsity of the desired speech signal coefficients to achieve dereverberation.

4. EXPERIMENTAL EVALUATION

We performed several simulations to investigate the dereverberation performance of the proposed method. We have considered two acoustic systems with a single speech source and measured RIRs taken from the REVERB challenge [30]. The first acoustic system (AC_1) consists of $M = 2$ microphones in a room with a reverberation time of $T_{60} \approx 500\,\mathrm{ms}$, and the second acoustic system (AC_2) consists of $M = 4$ microphones in a room with a reverberation time of $T_{60} \approx 700\,\mathrm{ms}$, with the distance between the source and the microphones being approximately $2\,\mathrm{m}$ in both cases. We have considered both noiseless and noisy scenario, with the latter obtained using the background noise provided in the REVERB challenge. The proposed method was tested on 20 different speech sentences (uttered by different speakers) taken from the WSJCAM0 corpus [31], with an average length of approximately $7\,\mathrm{s}$. The

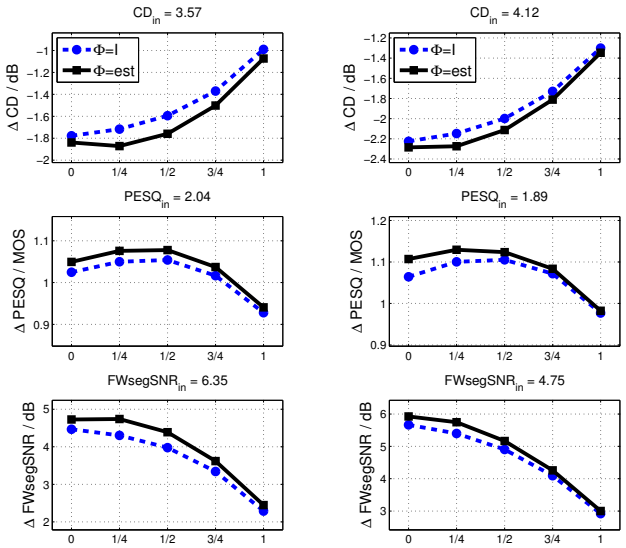
performance was evaluated in terms of the following instrumental speech quality measures: cepstral distance (CD), perceptual evaluation of speech quality (PESQ), and frequency-weighted segmental signal-to-noise ratio (FWsegSNR) [30]. The measures were evaluated with the clean speech signal as the reference. Note that lower values of CD indicate better performance.

The STFT was computed using a tight frame based on a 64 ms Hamming window with 16 ms shift. The length of the prediction filters in (3) was set to $L_g = 20$ for $M = 2$ microphones, and $L_g = 10$ for $M = 4$ microphones, similarly as in [25]. The prediction delay τ in (3) was set to 2, the maximum number of iterations was $i_{\max} = 20$ with the stopping criterion set to $\eta = 10^{-4}$, and the regularization parameter was fixed to $\varepsilon = 10^{-8}$.

In the first experiment we evaluate the dereverberation performance in the noiseless case in AC_1 and AC_2 for different values of the parameter p in the proposed cost function in (7). Additionally, we evaluate the performance of the method with a fixed correlation matrix $\mathbf{\Phi} = \mathbf{I}$, and with an estimated correlation matrix $\mathbf{\Phi}$ as in (11). To quantify the dereverberation performance, we average improvements of the evaluated measures over the M microphones and over all speech sentences. The obtained improvements are shown in Fig. 1. Firstly, it can be seen that the dereverberation performance exhibits a similar trend when using the fixed correlation matrix $\mathbf{\Phi} = \mathbf{I}$ or the estimated correlation matrix, with the latter performing better. Secondly, it can be seen that the dereverberation performance highly depends on the cost function in the proposed approach, i.e., on the parameter p. It can be observed that the performance deteriorates as the cost function comes closer to the convex case, i.e., as the parameter p approaches $p = 1$. In general, nonconvex cost functions, which promote sparsity more aggressively, achieve better performance, i.e., for p closer to 0. Additionally, mild improvements can be observed for values of p slightly higher than zero, as also observed in the case of a multiple-input single-output algorithm in [14]. In the second experiment we evaluate the dereverberation performance in the presence of noise. The microphone signals are obtained by adding noise to the reverberant signals to achieve a desired value of reverberant signal-to-noise ratio (RSNR). In this experiment we use the background noise provided in the REVERB challenge, which was recorded in the same room and with the same array as the corresponding RIRs, and was caused mainly by the air conditioning system [30]. In this case we show only the performance of the method with the estimated correlation matrix, since it performed better in the previous experiment. Again, the improvements of the evaluated measures are averaged over the M microphones and over all speech sentences, with the results for $p \in \{0, 1/4, 1\}$ shown in Fig. 2. The proposed algorithm does not explicitly model the noise, and the improvements are achieved by dereverberation while the noise component is typically not affected, similar as in [8]. This is due to the fact that noise is typically less predictable than reverberation, and therefore the estimated prediction filters capture almost exclusively the latter. Similarly as in the previous experiment, the achieved performance highly depends on the convexity of the cost function, with the nonconvex cost functions performing significantly better than the convex case.

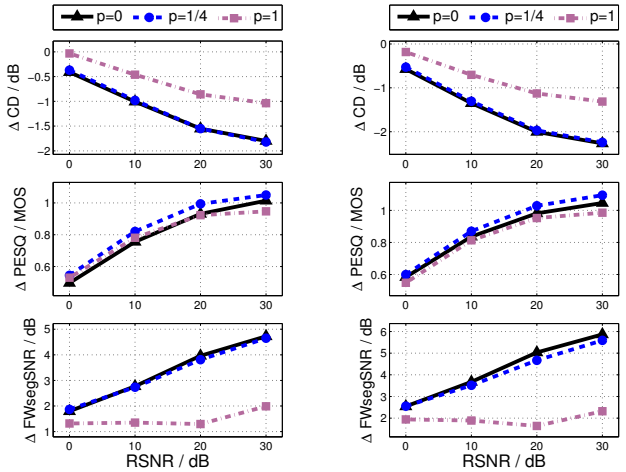
5. CONCLUSION

In this paper we have presented a formulation of the MCLP-based MIMO speech dereverberation problem based on the concept of group sparsity. The obtained nonconvex optimization problem is solved using iteratively reweighted least squares, with the derived algorithm generalizing several previously proposed MCLP-based methods. The dereverberation performance of the proposed method is evaluated in several acoustic scenarios, with and without noise and for different reverberation times, and the experimental results show the effectiveness of the nonconvex cost functions. Moreover, the presented formulation clearly highlights the role of sparsity in the STFT domain, and can be used to combine dereverberation with other sparsity-based enhancement algorithms, e.g., [27].



(a) AC_1: $T_{60} \approx 500$ ms, $M = 2$ (b) AC_2: $T_{60} \approx 700$ ms, $M = 4$

Figure 1: Improvements of the speech quality measures for the noiseless scenario in AC_1 (left) and AC_2 (right) vs. parameter p of the cost function. The correlation matrix $\mathbf{\Phi}$ was fixed to $\mathbf{I}$ or estimated using (11).



(a) AC_1: $T_{60} \approx 500$ ms, $M = 2$ (b) AC_2: $T_{60} \approx 700$ ms, $M = 4$

Figure 2: Improvements of the speech quality measures for the noisy scenario in the AC_1 (left) and the AC_2 (right) vs. RSNR. The correlation matrix $\mathbf{\Phi}$ was estimated using (11).

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, July 2006.

[2] A. Sehr, *Reverberation Modeling for Robust Distant-Talking Speech Recognition*, Ph.D. thesis, Friedrich-Alexander-Universität Erlangen-Nürenberg, Erlangen, Oct. 2009.

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

[4] 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.

[5] 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.

[6] 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.

[7] 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.

[8] 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.

[9] 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.

[10] 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, July 2014.

[11] D. Schmid *et al.*, "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.

[12] A. Jukić, T. van Waterschoot, T. Gerkmann, and S. Doclo, "Speech dereverberation with convolutive transfer function approximation using MAP and variational deconvolution approaches," in *Proc. Int. Workshop Acoustic Echo Noise Control (IWAENC)*, Antibes-Juan Les Pins, France, Sept. 2014, pp. 51–55.

[13] 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.

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

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

[16] M. Yuan and Y. Lin, "Model selection and estimation in regression with grouped variables," *J. Royal Stat. Soc. Serie B*, vol. 68, no. 1, pp. 49–67, 2006.

[17] M. Fornasier and H. Rahut, "Recovery algorithm for vector-valued data with joint sparsity constraints," *SIAM J. Num. Anal.*, vol. 46, no. 2, pp. 577–613, 2008.

[18] M. Kowalski and B. Torrésani, "Structured sparsity: from mixed norms to structured shrinkage," in *SPARS'09*, Saint-Malo, France, Apr. 2009.

[19] 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.

[20] K. Kumatani *et al.*, "Beamforming With a Maximum Negentropy Criterion," vol. 17, no. 5, pp. 994–1008.

[21] 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.

[22] A. Benedek and R. Panzone, "The space L^p with mixed norm," *Duke Math. J.*, vol. 28, no. 3, pp. 301–324, 1961.

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

[24] 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.

[25] M. Delcroix *et al.*, "Linear prediction-based dereverberation with advanced speech enhancement and recognition technologies for the REVERB challenge," in *Proc. REVERB Challenge Workshop*, Florence, Italy, May 2014.

[26] J. S. Bradley, H. Sato, 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.

[27] 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.

[28] R. Chartrand, "Exact Reconstruction of Sparse Signals via Nonconvex Minimization," *IEEE Signal Process. Lett.*, vol. 14, no. 10, pp. 707–710, Oct. 2007.

[29] Z. Zhao and B. D. Rao, "Sparse Signal Recovery With Temporally Correlated Source Vectors Using Sparse Bayesian Learning," *IEEE J. Sel. Topic Signal Process.*, vol. 5, no. 5, pp. 912–926, Sept. 2011.

[30] 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.

[31] T. Robinson *et al.*, "WSJCAM0: A British English Speech Corpus For Large Vocabulary Continuous Speech Recognition," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, Detroit, USA, May 1995, pp. 81–84.