Speech Dereverberation Based on Convex Optimization Algorithms for Group Sparse Linear Prediction

Giacobello, Daniele; Jensen, Tobias Lindstrøm

Published in:
IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2018

DOI (link to publication from Publisher):
10.1109/ICASSP.2018.8462560

Publication date:
2018

Document Version
Early version, also known as pre-print

SPEECH DEREVERBERATION BASED ON CONVEX OPTIMIZATION ALGORITHMS FOR GROUP SPARSE LINEAR PREDICTION

Daniele Giacobello[1] *and Tobias Lindstrøm Jensen*[2]

[1]Sonos Inc., Santa Barbara, CA, USA
[2]Signal and Information Processing, Department of Electronic Systems, Aalborg, Denmark

ABSTRACT

In this paper, we consider methods for improving far-field speech recognition using dereverberation based on sparse multi-channel linear prediction. In particular, we extend successful methods based on nonconvex iteratively reweighted least squares, that look for a sparse desired speech signal in the short-term Fourier transform domain, by proposing sparsity promoting convex functions. Furthermore, we show how to improve performance by applying regularization into both the reweighted least squares and convex methods. We evaluate the methods using large scale simulations by mimicking the application scenarios of interest. The experiments show that the proposed convex formulations and regularization offer improvements over existing methods with added robustness and flexibility in fairly different acoustic scenarios.

1. INTRODUCTION

Speech dereverberation has become an integral component of front-end processing techniques for automatic speech recognition (ASR). In particular, the recent advent of *smart* loudspeakers like the Amazon Echo, Google Home, and Sonos One, has pushed the robustness required in far-field ASR, as the user expects the same level of performance in multiple condition, including being at different distances in different acoustic environments [1]. This makes dereverberation one of the most prominent algorithm for enabling far-field human-computer interaction [2].

Several approaches have been proposed for speech dereverberation (see, e.g., [3] and references therein). The ones based on acoustic equalization, notably [4], can theoretically, achieve perfect dereverberation, being based on the estimation of the inverse of the room impulse responses (RIRs) between the source and microphones. These problems are, however, ill-conditioned and highly sensitive to RIR estimation errors [5]. More robust blind dereverberation methods based on multi-channel linear prediction (MCLP), applied in the short-term Fourier transform (STFT) [6, 7], have received attention lately as they do not require a priori knowledge of the room acoustic and are relatively easy and cheap to implement [1].

MCLP approaches in the STFT domain assume that, for each frequency bin, the reverberant component can be predicted from previous samples. Conventional MCLP is based on the minimization of the ℓ_2 norm of the error between observed and predicted signal, the desired speech. This is consistent with the maximum-likelihood assumption of the error being i.i.d. complex Gaussian with unknown variance [8]. In [6], the weighted prediction error (WPE) method proposes to estimate the variance allowing for a better model that, in turn, gives better performance in terms of dereverberation [2]. Based on different statistical assumptions, the work in [7] sets out to find a sparse desired speech signal belonging to the complex generalized Gaussian (CGG) distribution [9], resulting in the WPE-CGG method. By employing the iteratively reweighted least-squares (IRLS) algorithm (see, e.g., [10]) to solve the sparse approximation problem in [7], it also shows that the WPE method is equivalent to the WPE-CGG when the reweighting is done targeting the ℓ_0 *quasi* norm [11]. It is then interesting to see the MCLP problem from an empirical perspective; the STFT of clean speech is widely accepted to be sparse or belonging to a heavy-tailed distribution [12], while the reverberant speech appears like a *blurred* version of it. Modeling the problem in a MCLP framework allows to estimate a sparse component, the desired speech, while modeling the reverberation as a convolutive process which is approximated by the predicted speech. Both works in [6] and [7] were then extended to the multiple-input multiple-out (MIMO) case, in [13] and [14], respectively, which makes them desirable to work with other type of speech enhancement algorithm in commercial devices.

In this paper, we propose to revisit the MIMO MCLP scheme presented in [14] and propose new solutions based on convex formulations. In [14], approximations of mixed quasi norms are considered to enforce group sparsity across time where the groups are the magnitude of the desired signal across all the channels. In this work, we focus on the convex formulations of the $\ell_{1,2}$ and $\ell_{1,1}$ mixed norms on the desired signal. To make sure that a proper model is chosen for the predictor, we also include a model order selection criteria by imposing sparsity on the predictor. The paper is organized as follows. In Section 2 and 3, we give a overview of the MCLP framework and the current solutions available. In Section 4, we present our convex methods to solve the sparse MCLP problem. In Section 5, we present our simulation framework and the experimental results providing our conclusions.

2. FUNDAMENTALS

We consider an acoustic system composed of one speech point source and M microphones. The signal at the $m-$th microphone at time n is

$$x_m(n) = \sum_{m=1}^{M} r_m(n) * s(n) + e_m(n), \tag{1}$$

where $s(n)$ is the clean speech signal, $r_m(n)$ is the RIR between the speech source and the $m-$th microphone, and $*$ is the convolution operator. We focus our attention on so-called *utterance-based*

This work was partly supported by the Danish Council for Independent Research, Technology and Production Sciences. Grant no. 4005-00122. The work of D. Giacobello was performed during a research stay at Aalborg Universitet, Denmark funded under the same grant. The work of T. L. Jensen was performed while employed at Aalborg University.

Code for the proposed convex algorithms is available at `https://github.com/giacobello/`.

batch processing techniques where a full reverberant speech file is processed all at once [15]. Denoting $s(k,n)$ as the STFT of the clean speech, with frame index $n \in \{1,\dots,N\}$ and frequency bin index $k \in \{1,\dots,K\}$, the reverberant speech signal at the $m-$th microphone becomes

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

where $h_m(k,l)$ models the acoustic transfer function between the speech source and the m-th microphone in the k-th frequency bin with length L_h. The model in (2) divides the time-domain convolution in (1) into a set of convolution in the time-frequency domain and has been widely adopted in the dereverberation literature [13]. Given the general assumption of ignoring the noise term, as done in [13, 14], we can rewrite (2) as

$$x_m(k,n) = \underbrace{\sum_{l=0}^{\tau-1} h_m(k,l)s(k,n-l)}_{d_m(k,n)} + \underbrace{\sum_{l=\tau}^{L_g-1} h_m(k,l)s(k,n-l)}_{r_m(k,n)},$$

where the first term $d_m(k,n)$ is the desired speech and the second term $r_m(k,n)$ is the reverberation term. Notice that the term τ is a *delay* [6] allows for modeling the direct speech and the early reflections which generally do not give issues in terms of recognition accuracy or speech quality and intelligibility [2]. Borrowing the notation in [14], using M prediction filters of length L_g, the desired speech signal can be rewritten as

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

where $g_{m,i}(k,l)$ is the $l-$th prediction coefficient between the $i-$th and the $m-$th channel. The equivalent model in matrix notation is:

$$\mathbf{D}(k) = \mathbf{X}(k) - \mathbf{X}_\tau(k)\mathbf{G}(k), \quad (4)$$

where

$$\begin{aligned}
\mathbf{D}(k) &= [\mathbf{d}_1(k),\cdots,\mathbf{d}_M(k)] \in \mathbb{C}^{N\times M},\\
\mathbf{d}_m(k) &= [d_m(k,1),\cdots,d_m(k,N)]^T \in \mathbb{C}^{N\times 1},\\
\mathbf{X}(k) &= [\mathbf{x}_1(k),\cdots,\mathbf{x}_M(k)] \in \mathbb{C}^{N\times M},\\
\mathbf{x}_m(k) &= [x_m(k,1),\cdots,x_m(k,N)]^T \in \mathbb{C}^{N\times 1},\\
\mathbf{X}_\tau(k) &= [\mathbf{X}_{\tau,1}(k),\cdots,\mathbf{X}_{\tau,M}(k)] \in \mathbb{C}^{N\times ML_g},
\end{aligned}$$

and $\mathbf{X}_{\tau,m}(k) \in \mathbb{C}^{N\times L_g}$ is the convolution matrix of $x_m(k,n-\tau)$. The prediction matrix is

$$\mathbf{G}(k) = [\mathbf{g}_1(k),\cdots,\mathbf{g}_M(k)] \in \mathbb{C}^{ML_g\times M}, \quad (5)$$

$$\begin{aligned}
\text{with } \mathbf{g}_m(k) = [&g_{m,1}(k,0),\cdots,g_{m,1}(k,L_g-1),\dots\\
&g_{m,M}(k,0),\cdots,g_{m,M}(k,L_g-1)]^T \in \mathbb{C}^{ML_g\times 1}.
\end{aligned} \quad (6)$$

3. MIMO MULTICHANNEL LINEAR PREDICTION

The prediction coefficients matrix $\mathbf{G}$ in (4) is then found by solving the following optimization problem:

$$\hat{\mathbf{G}} = \underset{\mathbf{G}}{\operatorname{argmin}} \parallel \mathbf{X} - \mathbf{X}_\tau\mathbf{G} \parallel_{p,q}^q + \alpha \parallel \mathbf{G}\parallel_{r,s}^s, \quad (7)$$

where $\|\cdot\|_{p,q}^q$ is defined as the $\ell_{p,q}$ norm of a matrix $\mathbf{V} \in \mathbb{C}^{n\times m}$:

$$\|\mathbf{V}\|_{p,q} = \left(\sum_{i=1}^{n} \|\mathbf{V}_{i,:}\|_p^q\right)^{1/q} \quad (8)$$

and $\|\mathbf{V}_{i,:}\|_p$ is the ℓ_p norm of the $i-$th row-vector $\mathbf{V}_{i,:}$. We have omitted the frequency index k and we will continue to do so for the remainder of the paper for clarity and conciseness. The choice of the norms p, q, r, s, and the regularization term α will engender different type of solutions with different meanings. Solving the problem with $\alpha = 0$ in an element-wise least-squares sense (Frobenius norm)

$$\hat{\mathbf{G}} = \underset{\mathbf{G}}{\operatorname{argmin}} \|\mathbf{X} - \mathbf{X}_\tau\mathbf{G}\|_{2,2}^2 = \underset{\mathbf{G}}{\operatorname{argmin}} \|\mathbf{X} - \mathbf{X}_\tau\mathbf{G}\|_{\mathrm{F}}^2\,, \quad (9)$$

is equivalent to solve for each of the M microphones separately:

$$\hat{\mathbf{g}}_m = \underset{\mathbf{g}_m}{\operatorname{argmin}} \|\mathbf{x}_m - \mathbf{X}_\tau\mathbf{g}_m\|_2^2 = \left(\mathbf{X}_\tau^H\mathbf{X}_\tau\right)^{-1}\mathbf{X}_\tau^H\mathbf{x}_m, \quad (10)$$

thus $\hat{\mathbf{G}} = [\hat{\mathbf{g}}_1,\dots,\hat{\mathbf{g}}_M]$ and $\hat{\mathbf{d}}_m = \mathbf{x}_m - \mathbf{X}_\tau\hat{\mathbf{g}}_m$. In order to obtain a sparse residual (i.e., desired signal $\mathbf{D}$) in the MIMO linear prediction case, it was proposed to replace the $\ell_{2,2}$ norm by solving the $\ell_{2,1}$ norm through the approximation of the ℓ_1 norm provided by the IRLS algorithm [14]. A summary of the algorithm is shown in Algorithm 1. It is interesting to notice that when $q = 0$, the WPE method in [13] with *Scaled Identity Matrix Method* is equivalent to the IRLS algorithm [14]. The closed form solution of the weighted element-wise $\ell_{2,2}$ norm problem is

$$\hat{\mathbf{G}} = \left(\mathbf{X}_\tau^H\mathbf{W}\mathbf{X}_\tau\right)^{-1}\mathbf{X}_\tau^H\mathbf{W}\mathbf{X}. \quad (11)$$

It is then clear that both approaches presented in [14] and [13], seek sparsity in the STFT domain for the estimated desired speech, one explicitly and one implicitly. Both methods rely on nonconvex optimization methods based on IRLS ([10] to solve a sparse approximation problem). By using the ℓ_1 norm as a better approximation for sparsity (the so-called ℓ_0 norm) than the reweighted ℓ_2 norm [16,17], we set out to improve the estimation of $\hat{\mathbf{G}}$ and thus achieving better dereverberation.

3.1. Regularization and Model Order Selection

The objective of the term $\alpha\|\mathbf{G}\|_{r,s}^s$ in (7) is twofold. Firstly to act as a regularization term, given that often very closed spaced microphones compose the array, the matrix $\mathbf{X}_\tau^H\mathbf{X}_\tau$ might be close to be singular. Secondly to act as a model order selection penalization term, considering that, if the order L_g is not chosen appropriately, the estimation of $\hat{\mathbf{G}}$ might suffer from ill-conditioning. This second scenario is particularly interesting as L_g might be fixed but the acoustic properties of the space in which the algorithm is deployed might vary dramatically.

In particular, if we see the sparse approximation problem ℓ_0 norm as a minimum description length (MDL) constraint [18], then solving the ℓ_0 norm with the IRLS algorithm, we simply obtain iterations solving

$$\underset{\mathbf{G}}{\text{minimize}} \parallel \operatorname{diag}(\mathbf{w_D})^{1/2}\mathbf{D}\|_2^2 + \alpha\| \operatorname{diag}(\mathbf{w_G})^{1/2}\mathbf{G}\|_2^2, \quad (12)$$

a Tikhonov regularized version of the problem in Algorithm 1 with closed form solution

$$\hat{\mathbf{G}} = \left(\mathbf{X}_\tau^H \operatorname{diag}(\mathbf{w_D})\mathbf{X}_\tau + \alpha \operatorname{diag}(\mathbf{w_G})\right)^{-1}\mathbf{X}_\tau^H \operatorname{diag}(\mathbf{w_D})\mathbf{X}, \quad (13)$$

where $\mathbf{w}_{\mathbf{D},n} = \left(\|\mathbf{d}_n\|_2^2 + \epsilon\right)^{q/2-1}$ and $\mathbf{w}_{\mathbf{G},m} = \left(\|\mathbf{g}_m\|_2^2 + \epsilon\right)^{q/2-1}$.

Algorithm 1 WPE with Scaled Identity Matrix

Inputs: speech segment $\mathbf{X}$, approximation norm $0 \leq q \leq 1$
Outputs: predictor $\hat{\mathbf{G}}^i$, dereverberated multichannel signal $\hat{\mathbf{D}}^i$
$i = 0$, initial weights $w_m^0 = \left(\|\mathbf{x}_n\|_2^2 + \epsilon\right)^{q/2-1}, \forall n$
while halting criterion false **do**
$\mathbf{W}^i = \text{diag}\left(\mathbf{w}^i\right)$
$\hat{\mathbf{G}}^i = \text{argmin}_{\mathbf{G}} \|\mathbf{W}_i^{1/2}\left(\mathbf{X} - \mathbf{X}_\tau \mathbf{G}\right)\|_2^2$
$\hat{\mathbf{D}}^{i+1} = \mathbf{X} - \mathbf{X}_\tau \hat{\mathbf{G}}^i$
$\mathbf{w}^{i+1} = \left(\|\mathbf{d}_n\|_2^2 + \epsilon\right)^{q/2-1}, \forall n$
$i \leftarrow i + 1$
end while

4. CONVEX FORMULATIONS

We may also make convex formulations of (11) and (13), in-particular we consider problems on the form ($q = r = s = 1$)

$$\hat{\mathbf{G}} = \underset{G}{\text{argmin}} \| \mathbf{X} - \mathbf{X}_\tau \mathbf{G} \|_{p,1}^1 + \alpha \|\mathbf{G}\|_{1,1}^1. \tag{14}$$

We consider the cases $p = 1$ and $p = 2$, where $p = 1$ is an element-wise ℓ_1 formulation, while $p = 2$ is a group LASSO formulation [19]. To solve these problem we use alternating-direction methods of multipliers (ADMM) [20, 21], first presented in [22, 23].

4.1. Least Absolute Devation (LAD)

For $p = 1$, (14) becomes the element-wise regularized least-sum-of-absolute problem,

$$\hat{\mathbf{G}} = \underset{G}{\text{argmin}} \|\mathbf{X} - \mathbf{X}_\tau \mathbf{G} \|_{1,1}^1 + \alpha \|\mathbf{G}\|_{1,1}^1 . \tag{15}$$

Similar to the least-squares formulation in (9), this problem is separable, and can be solved as

$$\hat{\mathbf{g}}_m = \underset{\mathbf{g}_m}{\text{argmin}} \|\mathbf{x}_m - \mathbf{X}_\tau \mathbf{g}_m\|_1 + \alpha \|\mathbf{g}_m\|_1, \quad m = 1, \ldots, M. \tag{16}$$

Since $\|\mathbf{x}_1\|_1 + \|\mathbf{x}_2\|_1 = \left\| \begin{bmatrix} \mathbf{x}_1 \\ \mathbf{x}_2 \end{bmatrix} \right\|_1$, the problem in (16) can be seen as a non-regularized least-sum-of-absolute problem

$$\hat{\mathbf{g}}_m = \underset{\mathbf{g}_m}{\text{argmin}} \left\| \begin{bmatrix} \mathbf{x}_m \\ 0 \end{bmatrix} - \begin{bmatrix} \mathbf{X}_\tau \\ \alpha \mathbf{I} \end{bmatrix} \mathbf{g}_m \right\|_1, \; m = 1, \ldots, M \tag{17}$$

with new data and coefficient matrices. The ADMM algorithm for this particular type of formulation is known, see, e.g., [24] or the overview work [20, 21].

4.2. Group LASSO (GL)

For $p = 2$, the problem (14) is not separable and the ADMM algorithm is more complicated. The basic algorithm is shown in Algorithm 2. The function $\mathcal{S}_t$ is the proximity operator

$$\mathcal{S}_t\left(\begin{bmatrix} \mathbf{V}_1 \\ \mathbf{V}_2 \end{bmatrix}\right) = \underset{\mathbf{U_1}, \mathbf{U_2}}{\text{argmin}} t\|\mathbf{U}_1\|_{1,1} + t\|\mathbf{U}_2\|_{2,1} + \frac{1}{2} \left\| \begin{bmatrix} \mathbf{V}_1 - \mathbf{U}_1 \\ \mathbf{V}_2 - \mathbf{U}_2 \end{bmatrix} \right\|_{2,2}^2 .$$

This subproblem is however separable with the solution

$$\mathcal{S}_t \left(\begin{bmatrix} \mathbf{V}_1 \\ \mathbf{V}_2 \end{bmatrix} \right) = \begin{bmatrix} \mathbf{E}_t(\mathbf{V}_1) \odot \mathbf{V}_1 \\ \mathbf{D}_t(\mathbf{V}_2)\mathbf{V}_2 \end{bmatrix}, \tag{18}$$

where $\mathbf{D}_t(\mathbf{V}) = \text{diag}([(1 - t/\|\mathbf{V}_{1,:}\|_2)_+ \cdots (1 - t/\|\mathbf{V}_{n,:}\|_2)_+]$, $\{\mathbf{E}_t(\mathbf{V})\}_{i,j} = (1 - t/|\mathbf{V}_{i,j}|)_+$, with the operator $(a)_+ = \max(a, 0)$, $a \in \mathbb{R}$. Note that soft-thresholding is applied on $\mathbf{V}_1$ and block soft-thresholding on $\mathbf{V}_2$.

Algorithm 2 ADMM for the (regularized) GL formulation

Inputs: speech segment $\mathbf{X}$, regularization parameter α
Outputs: predictor $\hat{\mathbf{G}}^i$, dereverberated multichannel signal $\hat{\mathbf{D}}^i$
$i = 0$, $\mathbf{Z}^0 = 0$, $\mathbf{\Lambda}^0 = 0$
while halting criterion false **do**
$\hat{\mathbf{G}}^i = (\mathbf{X}_\tau^H \mathbf{X}_\tau + \alpha \mathbf{I})^{-1} \begin{bmatrix} \alpha \mathbf{I} & \mathbf{X}_\tau^H \end{bmatrix} \left(\mathbf{Z}^i + \begin{bmatrix} \mathbf{0} \\ \mathbf{X} \end{bmatrix} - \mathbf{\Lambda}^i \right)$
$\mathbf{R}^i = \begin{bmatrix} \alpha \hat{\mathbf{G}}^i \\ \mathbf{X}_\tau \hat{\mathbf{G}}^i - \mathbf{X} \end{bmatrix}$
$\mathbf{Z}^{i+1} = \mathcal{S}_t(\mathbf{R}^i + \mathbf{\Lambda}^i)$
$\mathbf{\Lambda}^{i+1} = \mathbf{\Lambda}^i + \mathbf{R}^i - \mathbf{Z}^{i+1}$
$i \leftarrow i + 1$
end while

4.3. Resource considerations

The main difference in terms of computation and memory requirements, between IRLS-type methods and the convex methods presented, arises from solving the two systems of equation with coefficient matrices

$$\mathbf{X}_\tau^H \, \text{diag}(\mathbf{w}_\mathbf{D}^i) \mathbf{X}_\tau + \text{diag}(\mathbf{w}_\mathbf{G}^i) \in \mathbb{C}^{ML_g \times ML_g}, \tag{19}$$

and

$$\mathbf{X}_\tau^H \mathbf{X}_\tau + \alpha \mathbf{I} \in \mathbb{C}^{ML_g \times ML_g}, \tag{20}$$

respectively. In particular, the coefficient matrix in (19) changes for each $i-$th IRLS iteration while (20) remains constant in LAD and GL. Furthermore, the diagonal reweighting will often destroy matrix structures.

If the signals are windowed using the autocorrelation method [18], then (20) can be permuted to a block Toeplitz matrix with $M \times M$ blocks and solved using, e.g., the classical block-Levinson method [25] in $\mathcal{O}(M^3 L_g^2)$ (and can be formed only once in a direct manner in $\mathcal{O}(NM^2 L_g)$. Differently, to form and solve (19) requires $\mathcal{O}(N(ML_g)^2 + (ML_g)^3)$ operations using Cholesky factorization. Thus, while IRLS methods are selected for their simple implementation, they often have asymptotically higher resource requirements when compared to the convex formulations with *splitting* type methods. Similar considerations are presented in [26, 27] for other signal processing problems.

5. EXPERIMENTAL ANALYSIS

We evaluated the performance of the methods presented simulating artificial utterances that mimic real use cases specific for voice enabled smart speakers. Similarly to the experiments presented in [28, 29], we use a room configuration generator. Our target hardware is known, in particular, we considered a 6-microphone circular array of 72 mm diameter. This will be assumed suspended in space for simplicity in the simulation.

5.1. Scenario Generator

We used COMSOL®to generate the six room impulse responses (RIRs) from a omnidirectional point source to each of the microphone populating the array. COMSOL®models the room acoustic by solving directly the Helmholtz equation, or the scalar wave equation, using the finite element method [30]. The size of the room was set to have a uniform distribution of width, length, and height between 3 to 8 m, 3 to 10 m, 2 to 4 m, respectively. The distance

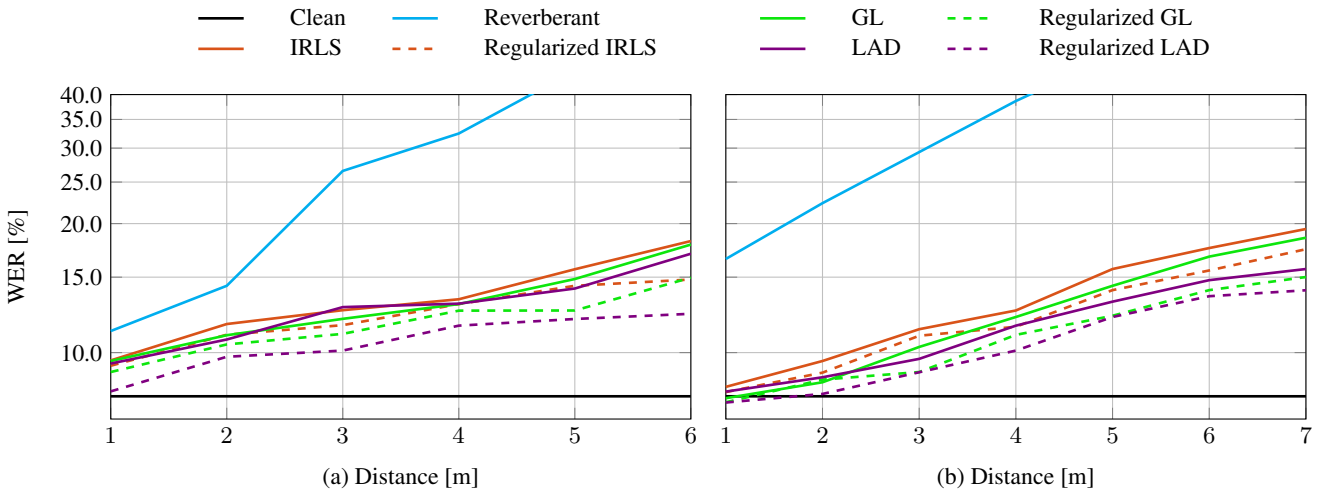

Fig. 1. Results for small room (a) and large room (b) as a function of distance for all the considered methods.

between the center of the array and the point source was chosen between 1 and 7 m with azimuth, θ, and elevation, ϕ, randomly selected in the interval $[-180, 180]$ and $[45, 135]$ degrees. Both source and microphones are assumed to be at least 0.25 m away from the wall. Focusing only on the dereverberation algorithm, we assume only diffuse HVAC noise at SNR uniformly distributed between 10 to 30 dB at the center of the array (roughly similar at each microphone). COMSOL® does not allow to choose a overall T_{60} for the room, so we tune the reflection coefficient for each of the surfaces of the cuboid to obtain a reverberation time for each room included between 300 and 700 ms with a skewed distribution towards the higher values. Pieces of furniture were also used and distributed in the room to make the RIRs more realistic. We generated 1000 rooms with this method.

5.2. Results

We evaluated the dereverberation performance of the proposed methods in terms of word error rate (WER) by processing the speech through an ASR engine. The engine was trained using the Librispeech 100hrs corpus [31]. The set is composed of 100 hours of clean speech, 125 male, 125 female speakers) derived from audiobooks data. We trained a ASR Kaldi baseline following the s5 recipe for Librispeech 100hrs with the same language model. We chose to train only on clean speech as we focused our analysis on the dereverberation performance of the algorithm. This Kaldi model used composite mel frequency cepstral coefficient (MFCC) features over which linear discriminant analysis (LDA), maximum likelihood linear transform (MLLT), and speaker adaption transform (SAT) transformations were applied to generate 40-dimensional features used during training; the DNN architecture itself consists of 4 p-norm layers with 3486 outputs corresponding to the context dependent clusters [32].

We compared IRLS, LAD, GL, and their regularized versions. The dereverberation algorithm was based on an analysis-modification-synthesis scheme with 50% overlapping Hamming windows of 32ms length. The order of the predictor was $L_g = 10$, the prediction delay $\tau = 2$, and the convolution matrices $\mathbf{X}_{\tau,m}$, $m = 1, \ldots, M$ were generated using the autocorrelation-type windowing [18]. The halting criterion for the IRLS was the Frobenius norm of the difference between the solution prediction error at step i and $i + 1$ to be lower than 10^{-3}. The IRLS was run with $q = 0$ in Algorithm 1, making it equivalent to WPE. The regularization parameter was chosen as $\alpha = 0.1$, which showed empirically to provide a good tradeoff between level of regularization and accuracy of the solution. The ADMM was run for 100 iterations.

We test using the Librispeech test partition [31]. Each file is processed with the RIRs and deconvolved in the STFT domain. Each output channel is run through the ASR engine and the best output is chosen. We split the results into two main cases: small room and large room. These were defines as as rooms with volume below and above 90 m^3, respectively. Furthermore, we analyzed only small rooms with $T_{60} < 400$ms and large rooms with $T_{60} \geq 400$ms. This was done as with randomization of the reflection coefficients to obtain reasonable T_{60} values, some of the rooms had unrealistic acoustic. About 840 out of 1000 were then used for the final analysis, split equally in the two groups.

The results are shown in Figure 1. Firstly, it can be seen that both LAD and GL generally perform better than IRLS demonstrating overall that solving directly the ℓ_1 norm with convex tools as solution to the sparse approximation problem is a good idea. Secondly, the regularized methods perform much better than their non-regularized counterparts, meaning that with a slight add in computational cost, we see improvements both in small and large rooms for all distances showing less dependence on the choice of L_g. While the proposed algorithm does not explicitly model the noise, the system is reliable also in moderate noise conditions as the one considered. This means that, while joint denoising and dereverberation approaches might achieve slight better results, a two step procedure is also a valuable alternative. Also, using multicondition training of the ASR engine is generally required for good performance.

6. CONCLUSIONS

We proposed two algorithms, LAD and GL, that solve the sparse MCLP-based MIMO speech dereverberation problem by applying the convex ℓ_1 norm rather than using nonconvex IRLS-type approaches. We also introduced the concept of regularization in the MCLP optimization problem for added robustness. The dereverberation performance showed a clear improvement over state-of-the-art nonconvex methods for sparse approximation. We also emphasized that, while IRLS methods are simpler to implement, they often require higher computation and memory per iteration than the proposed splitting type convex method.

7. REFERENCES

[1] B. Li *et. al.*, "Acoustic modeling for Google home," in *Interspeech*, 2017.

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

[3] P. Naylor and N. D. Gaubitch, *Speech dereverberation*, Springer Science & Business Media, 2010.

[4] M. Miyoshi and Y. Kaneda, "Inverse filtering of room acoustics," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 36, no. 2, pp. 145–152, 1988.

[5] I. Kodrasi, S. Goetze, and S. Doclo, "Regularization for partial multichannel equalization for speech dereverberation," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 21, no. 9, pp. 1879–1890, 2013.

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

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

[8] D. Giacobello, M.G. Christensen, M.N. Murthi, S.H. Jensen, and M. Moonen, "Sparse linear prediction and its applications to speech processing," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 5, pp. 1644–1657, 2012.

[9] S. Nadarajah, "A generalized normal distribution," *Journal of Applied Statistics*, vol. 32, no. 7, pp. 685–694, 2005.

[10] I. Daubechies, R. DeVore, M. Fornasier, and C.S. Güntürk, "Iteratively reweighted least squares minimization for sparse recovery," *Communications on Pure and Applied Mathematics*, vol. 63, no. 1, pp. 1–38, 2010.

[11] D. Wipf and S. Nagarajan, "Iterative reweighted ℓ_1 and ℓ_2 methods for finding sparse solutions," *IEEE Journal of Selected Topics in Signal Processing*, vol. 4, no. 2, pp. 317–329, 2010.

[12] S. Gazor and W. Zhang, "Speech probability distribution," *IEEE Signal Processing Letters*, vol. 10, no. 7, pp. 204–207, 2003.

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

[14] A. Jukić, T. van Waterschoot, T. Gerkmann, and S. Doclo, "Group sparsity for MIMO speech dereverberation," in *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, 2015.

[15] K. Kinoshita, M. Delcroix, T. Yoshioka, T. Nakatani, A. Sehr, W. Kellermann, and R. Maas, "The REVERB challenge: A common evaluation framework for dereverberation and recognition of reverberant speech," in *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, 2013.

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

[17] D. Giacobello, M. G. Christensen, M. N. Murthi, S. H. Jensen, and M. Moonen, "Enhancing sparsity in linear prediction of speech by iteratively reweighted 1-norm minimization," in *IEEE International Conference on Acoustics Speech and Signal Processing (ICASSP)*, 2010.

[18] P. Stoica and R.L. Moses, *Spectral analysis of signals*, Pearson Prentice Hall, Upper Saddle River, NJ, 2005.

[19] J. Friedman, T. Hastie, and R. Tibshirani, *The elements of statistical learning*, Springer Series in Statistics, New York, NY, 2001.

[20] 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.

[21] D. P. Bertsekas, *Convex Optimization Algorithms*, Athena Scientific, 2015.

[22] R. Glowinski and A. Marroco, "Sur l'approximation, par éléments finis d'ordre un, et la résolution, par pénalisation-dualité d'une classe de problémes de Dirichlet non linéaires," *ESAIM: Mathematical Modelling and Numerical Analysis - Modélisation Mathématique et Analyse Numérique*, vol. 9, no. 2, pp. 41–76, 1975.

[23] D. Gabay and B. Mercier, "A dual algorithm for the solution of nonlinear variational problems via finite element approximation," *Computers & Mathematics with Applications*, vol. 2, no. 1, pp. 17–40, 1976.

[24] Y. Zhang J. Yang, "Alternating direction algorithms for ℓ_1-problems in compressive sensing," *SIAM Journal of Scientific Computing*, vol. 33, no. 1, pp. 250–278, 2011.

[25] R. A. Wiggins and E. A. Robinson, "Recursive solution to the multichannel filtering problem," *Journal of Geophysical. Research*, vol. 70, no. 8, pp. 1885–1891, 1965.

[26] D. O'Connor and L. Vandenberghe, "Primal-dual decomposition by operator splitting and applications to image deblurring," *SIAM Journal on Imaging Sciences*, vol. 7, no. 3, pp. 1724–1754, 2014.

[27] T. L. Jensen, D. Giacobello, T. van Waterschoot, and M. G. Christensen, "Fast algorithms for high-order sparse linear prediction with applications to speech processing," *Speech Communication*, vol. 76, pp. 143 – 156, 2016.

[28] C. Kim, A. Misra, K. Chin, T. Hughes, A. Narayanan, T. Sainath, and M. Bacchiani, "Generation of large-scale simulated utterances in virtual rooms to train deep-neural networks for far-field speech recognition in Google home," in *Interspeech*, 2017.

[29] D. Giacobello, J. Wung, R. Pichevar, and J. Atkins, "Tuning methodology for speech enhancement algorithms using a simulated conversational database and perceptual objective measures," in *Joint Workshop on Hands-free Speech Communication and Microphone Arrays (HSCMA)*, 2014.

[30] H. Kuttruff, *Room acoustics*, CRC Press, 2016.

[31] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: an ASR corpus based on public domain audio books," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2015.

[32] D. Povey and *et. al.*, "The Kaldi speech recognition toolkit," in *IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*, 2011.