ROBUST SPARSITY-PROMOTING ACOUSTIC MULTI-CHANNEL EQUALIZATION FOR SPEECH DEREVERBERATION

Ina Kodrasi, Ante Jukić, Simon Doclo

University of Oldenburg, Department of Medical Physics and Acoustics,
and Cluster of Excellence Hearing4All, Oldenburg, Germany
ina.kodrasi@uni-oldenburg.de

ABSTRACT

This paper presents a novel signal-dependent method to increase the robustness of acoustic multi-channel equalization techniques against room impulse response (RIR) estimation errors. Aiming at obtaining an output signal which better resembles a clean speech signal, we propose to extend the acoustic multi-channel equalization cost function with a penalty function which promotes sparsity of the output signal in the short-time Fourier transform domain. Two conventionally used sparsity-promoting penalty functions are investigated, i.e., the l_0-norm and the l_1-norm, and the sparsity-promoting filters are iteratively computed using the alternating direction method of multipliers. Simulation results for several RIR estimation errors show that incorporating a sparsity-promoting penalty function significantly increases the robustness, with the l_1-norm penalty function outperforming the l_0-norm penalty function.

Index Terms— speech dereverberation, sparsity, robustness, RIR estimation errors

1. INTRODUCTION

Acoustic multi-channel equalization techniques aim at speech dereverberation by reshaping the estimated room impulse responses (RIRs) between the speech source and the microphone array [1–4]. Although in theory perfect dereverberation can be achieved when multiple microphones are available [1], such an approach poses the practical challenge of achieving robustness against errors in the estimated RIRs [3–5]. Since the estimated RIRs typically differ from the true RIRs [6, 7], acoustic multi-channel equalization techniques may fail to achieve dereverberation and may cause distortions in the output signal [3, 6].

Several methods have been proposed to increase the robustness of equalization techniques against RIR estimation errors, such as, e.g., relaxing the constraints on the filter design by using a weighted least-squares or an energy-based optimization criterion [2, 4], incorporating regularization to reduce the filter energy [3, 5], or using a shorter filter length to improve the conditioning of the underlying optimization criteria [8]. To the best of our knowledge, all proposed methods are *signal-independent* methods, i.e., they rely only on the estimated RIRs for the filter design, without incorporating knowledge of the resulting output signal.

In this paper we propose a *signal-dependent* method to increase the robustness of equalization techniques by incorporating the output signal in the filter design, with the aim of enforcing it to exhibit spectro-temporal characteristics of a clean speech signal. Given the successful exploitation of the sparse nature of clean speech in several speech enhancement techniques [9–12], we propose to promote sparsity of the output signal in the short-time Fourier transform (STFT) domain by extending the acoustic multi-channel equalization cost function with an l_0-norm or l_1-norm penalty function. The sparsity-promoting filters are iteratively computed using the alternating direction method of multipliers (ADMM), since it is a well-suited algorithm for solving large scale problems with sparsity-promoting penalty functions [13]. Simulation results show that promoting sparsity of the output signal significantly increases the robustness of equalization techniques against RIR estimation errors, with the l_1-norm penalty function outperforming the l_0-norm penalty function.

This work was supported by a Grant from the GIF, the German-Israeli Foundation for Scientific Research and Development, the Cluster of Excellence 1077 "Hearing4All", funded by the German Research Foundation (DFG), and the Marie Curie Initial Training Network DREAMS (Grant no. 316969).

2. CONFIGURATION AND NOTATION

Consider the acoustic system depicted in Fig. 1, consisting of a single speech source and M microphones. The m-th microphone signal $x_m(n)$ at time index n is given by

$$x_m(n) = s(n) * h_m(n), \qquad m = 1, \ \ldots, \ M, \tag{1}$$

with $s(n)$ the clean speech signal, $h_m(n)$ the RIR between the speech source and the m-th microphone, and $*$ denoting convolution. Using the filter-and-sum structure in Fig. 1, the output signal $z(n)$ is equal to the sum of the filtered microphone signals, i.e.,

$$z(n) = \sum_{m=1}^{M} x_m(n) * w_m(n) = s(n) * \underbrace{\sum_{m=1}^{M} h_m(n) * w_m(n)}_{c(n)}, \tag{2}$$

with $w_m(n)$ the filter applied to the m-th microphone signal and $c(n)$ the equalized impulse response (EIR) between the speech source and the output of the system. In vector notation, the RIR $\mathbf{h}_m$ and the filter $\mathbf{w}_m$ are given by

$$\mathbf{h}_m = [h_m(0) \ldots h_m(L_h{-}1)]^T, \ \mathbf{w}_m = [w_m(0) \ldots w_m(L_w{-}1)]^T, \tag{3}$$

where L_h and L_w denote the RIR length and the filter length, respectively. Furthermore, the L_c-dimensional EIR vector $\mathbf{c}$, with

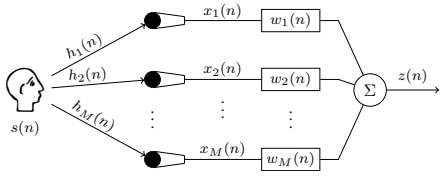

Fig. 1. Acoustic system configuration.

$L_c = L_h + L_w - 1$, is given by $\mathbf{c} = [c(0) \dots c(L_c-1)]^T$. Using the ML_w-dimensional stacked filter vector $\mathbf{w} = [\mathbf{w}_1^T \dots \mathbf{w}_M^T]^T$ and the $L_c \times ML_w$-dimensional multi-channel convolution matrix of the RIRs $\mathbf{H} = [\mathbf{H}_1 \dots \mathbf{H}_M]$, with $\mathbf{H}_m$ the $L_c \times L_w$-dimensional convolution matrix of $\mathbf{h}_m$, the output signal in (2) can be expressed as

$$z(n) = \sum_{m=1}^{M} \mathbf{w}_m^T \underbrace{\mathbf{H}_m^T \mathbf{s}(n)}_{\mathbf{x}_m(n)} = \underbrace{\mathbf{w}^T \mathbf{H}^T}_{\mathbf{c}^T} \mathbf{s}(n), \tag{4}$$

with $\mathbf{s}(n) = [s(n) \dots s(n-L_c+1)]^T$ the L_c-dimensional clean speech vector, $\mathbf{x}_m(n) = [x_m(n) \dots x_m(n-L_w+1)]^T$ the L_w-dimensional signal vector at the m-th microphone, and

$$\mathbf{c} = \mathbf{H}\mathbf{w}. \tag{5}$$

Alternatively, the output signal in (4) can also be expressed as

$$z(n) = \sum_{m=1}^{M} \mathbf{x}_m^T(n)\mathbf{w}_m = \mathbf{x}^T(n)\mathbf{w}, \tag{6}$$

with $\mathbf{x}(n) = [\mathbf{x}_1^T(n) \dots \mathbf{x}_M^T(n)]^T$ the ML_w-dimensional stacked signal vector. Based on (6), the L_z-dimensional output signal vector $\mathbf{z}(n) = [z(n) \dots z(n-L_z+1)]^T$ can be written as

$$\mathbf{z}(n) = \mathbf{X}(n)\mathbf{w}, \tag{7}$$

with $\mathbf{X}(n)$ the $L_z \times ML_w$-dimensional multi-channel convolution matrix of the microphone signals, i.e., $\mathbf{X}(n) = [\mathbf{X}_1(n) \dots \mathbf{X}_M(n)]$, where $\mathbf{X}_m(n)$ denotes the $L_z \times L_w$-dimensional convolution matrix of $\mathbf{x}_m(n)$. For conciseness, the time index n will be omitted when possible in the remainder of this paper.

3. ACOUSTIC MULTI-CHANNEL EQUALIZATION

Acoustic multi-channel equalization techniques aim at speech dereverberation by designing the filter $\mathbf{w}$ such that the resulting EIR $\mathbf{c}$ in (5) resembles a dereverberated target EIR $\mathbf{c}_t$. Since the true RIRs are generally not available in practice [6, 7], such techniques design $\mathbf{w}$ using the estimated multi-channel convolution matrix $\hat{\mathbf{H}}$ (constructed from the estimated RIRs $\hat{\mathbf{h}}_m$) instead of the true multi-channel convolution matrix $\mathbf{H}$.

In this paper we will focus on the partial multi-channel equalization technique based on the multiple-input/output inverse theorem (PMINT) proposed in [3], which has been shown to be a perceptually advantageous technique. The PMINT technique aims at simultaneously suppressing late reverberation and preserving perceptual speech quality by using a target EIR $\mathbf{c}_t$ whose late reflection taps are equal to $\mathbf{0}$, while the remaining taps are equal to the direct path and early reflections of one of the estimated RIRs, i.e.,

$$\mathbf{c}_t = [\hat{h}_q(0) \dots \hat{h}_q(L_d-1)\ 0 \dots 0]^T, \tag{8}$$

with L_d the length of the direct path and early reflections and $q \in \{1, \dots, M\}$. The PMINT filter is computed by minimizing the least-squares cost function

$$J_{\text{P}}(\mathbf{w}) = \|\hat{\mathbf{H}}\mathbf{w} - \mathbf{c}_t\|_2^2. \tag{9}$$

As shown in [1,3], assuming that the RIRs do not share any common zeros and using $L_w \geq \left\lceil \frac{L_h-1}{M-1} \right\rceil$, with $\lceil \cdot \rceil$ the ceiling operator, the PMINT filter minimizing the cost function in (9) is equal to

$$\mathbf{w}_{\text{P}} = \hat{\mathbf{H}}^{+}\mathbf{c}_t, \tag{10}$$

where $\{\cdot\}^+$ denotes the matrix pseudo-inverse. When the true RIRs are available, i.e., $\hat{\mathbf{H}} = \mathbf{H}$, the PMINT filter yields perfect dereverberation performance, i.e., $\mathbf{H}\mathbf{w}_{\text{P}} = \mathbf{c}_t$ [3]. However, in the presence of RIR estimation errors, i.e., $\hat{\mathbf{H}} \neq \mathbf{H}$, the PMINT filter fails to achieve dereverberation, i.e., $\mathbf{H}\mathbf{w}_{\text{P}} \neq \mathbf{c}_t$, causing distortions in the output signal [3].

Several methods have been proposed to increase the robustness of equalization techniques against RIR estimation errors, e.g., relaxing the constraints on the filter design using a weighted least-squares or an energy-based optimization criterion [2, 4], incorporating regularization [3, 5], or decreasing the filter length L_w [8]. To the best of our knowledge, all proposed methods are *signal-independent* methods, i.e., they only use the estimated convolution matrix $\hat{\mathbf{H}}$ for the filter design without incorporating knowledge of the resulting output signal $\mathbf{z}$. The objective of this paper is to explore the potential of increasing the robustness of acoustic multi-channel equalization techniques by incorporating a *signal-dependent* penalty function, which enforces the output signal $\mathbf{z}$ to exhibit spectro-temporal characteristics of a clean speech signal. Although the proposed method is discussed as an extension of the PMINT technique, it can also be applied to increase the robustness of other (more robust and signal-independent) acoustic multi-channel equalization techniques.

4. SPARSITY-PROMOTING ACOUSTIC MULTI-CHANNEL EQUALIZATION

While in principle any well-defined characteristic of clean speech could be exploited, we propose to use a penalty function which promotes sparsity of the output signal in the STFT domain. The advantage of promoting sparsity of the output signal is expected to be twofold. First, it is widely accepted that clean speech is sparse in the STFT domain [9, 14, 15]. Empirical observations, e.g., in [9], have shown that when clean speech is corrupted by reverberation (and noise), the STFT coefficients are less sparse than the STFT coefficients of clean speech. Hence, promoting sparsity of the output signal yields a signal which better resembles a clean speech signal. Second, in the presence of RIR estimation errors, non-robust equalization techniques introduce distortions (i.e., non-zero STFT coefficients) in the output signal [3]. By sparsifying the STFT representation of the output signal it is expected that these distortions are reduced.

4.1. Cost function formulation

The STFT coefficients of the output signal are computed as

$$Z(k,l) = \sum_{n=0}^{K-1} w_{\text{STFT}}(n) z(lR+n) e^{-\frac{j2\pi kn}{K}}, \tag{11}$$

where $k = 0, \dots, K-1$, denotes the frequency bin index with K the frame size, $l = 0, \dots, L-1$, denotes the frame index with L the total number of frames, $w_{\text{STFT}}(n)$ denotes the STFT analysis window, and R denotes the frame shift. Similarly as in [16], we define the STFT operator $\mathbf{\Psi} \in \mathcal{C}^{L_{\tilde{z}} \times L_z}$ which transforms the L_z-dimensional time-domain vector $\mathbf{z}$ into the $L_{\tilde{z}}$-dimensional frequency-domain vector $\tilde{\mathbf{z}}$, i.e., $\tilde{\mathbf{z}} = \mathbf{\Psi}\mathbf{z}$, consisting of all STFT coefficients $Z(k,l)$, with $L_{\tilde{z}} = K \times L$ (i.e., $\tilde{\mathbf{z}}$ denotes the stacked vector of the columns of the spectrogram of $\mathbf{z}$). When using a tight STFT analysis window, the inverse STFT operator $\mathbf{\Psi}^H \in \mathcal{C}^{L_z \times L_{\tilde{z}}}$ is such that $\mathbf{\Psi}^H\mathbf{\Psi} = \mathbf{I}$, with $\mathbf{I}$ the $L_z \times L_z$-dimensional identity matrix.

The proposed sparsity-promoting PMINT cost function is then defined by adding a penalty function to (9), i.e.,

$$J_{\text{S-P}}(\mathbf{w}) = J_{\text{P}}(\mathbf{w}) + \eta f_{\text{S}}(\tilde{\mathbf{z}}) = \|\hat{\mathbf{H}}\mathbf{w} - \mathbf{c}_t\|_2^2 + \eta f_{\text{S}}(\mathbf{\Psi}\mathbf{X}\mathbf{w}), \tag{12}$$

with $f_{\text{s}}(\tilde{\mathbf{z}})$ a sparsity-promoting penalty function and η a weighting parameter providing a trade-off between the minimization of the least-squares error and the sparsity of the STFT coefficients of the output signal. For the penalty function $f_{\text{s}}(\tilde{\mathbf{z}})$, we propose two commonly used sparsity-promoting norms, i.e., the l_0-norm[1] $f_{\text{s}}^0(\tilde{\mathbf{z}})$ and the l_1-norm $f_{\text{s}}^1(\tilde{\mathbf{z}})$, defined as

$$f_{\text{s}}^0(\tilde{\mathbf{z}}) = |i : \tilde{z}(i) \neq 0|, \qquad f_{\text{s}}^1(\tilde{\mathbf{z}}) = \sum_{i=0}^{L_{\tilde{z}}-1} |\tilde{z}(i)|, \tag{13}$$

with the l_1-norm differing from the l_0-norm by penalizing the large coefficients of $\tilde{\mathbf{z}}$ more than the small coefficients. It should be noted that the l_0-norm is non-convex and it is well known that optimization problems with non-convex penalty functions are typically hard (if not impossible) to solve exactly, particularly for large scale problems [17]. The l_1-norm on the other hand can be viewed as a convex relaxation of the l_0-norm, and efficient methods have been proposed to solve optimization problems with l_1-norm penalty functions [18, 19].

4.2. Cost function optimization based on the alternating direction method of multipliers (ADMM)

Since no closed-form expression is available for the filter minimizing the cost function in (12), iterative optimization algorithms are required. We have chosen to use the ADMM algorithm, since it is a well-suited algorithm for solving large scale optimization problems of the form in (12) [13]. Within the ADMM framework, the minimization of the sparsity-promoting cost function in (12) is reformulated as

$$\min_{\mathbf{w}} \left[\|\hat{\mathbf{H}}\mathbf{w} - \mathbf{c}_t\|_2^2 + \eta f_{\text{s}}(\tilde{\mathbf{a}}) \right] \quad \text{subject to} \quad \mathbf{\Psi X w} = \tilde{\mathbf{a}}, \tag{14}$$

introducing the auxiliary variable $\tilde{\mathbf{a}}$ such that the optimization problem in (12) can be split into simpler sub-problems. The augmented Lagrangian of (14) is equal to

$$\mathcal{L} = \|\hat{\mathbf{H}}\mathbf{w} - \mathbf{c}_t\|_2^2 + \eta f_{\text{s}}(\tilde{\mathbf{a}}) + \frac{\rho}{2}\|\mathbf{\Psi X w} + \boldsymbol{\lambda} - \tilde{\mathbf{a}}\|_2^2, \tag{15}$$

with $\boldsymbol{\lambda}$ the $L_{\tilde{z}}$-dimensional dual variable and $\rho > 0$ the penalty parameter. The ADMM algorithm minimizes (15) alternately with respect to the variables $\mathbf{w}$ and $\tilde{\mathbf{a}}$, followed by a dual ascent over the variable $\boldsymbol{\lambda}$. The ADMM update rules are given by:

$$\mathbf{w}^{(i+1)} = \arg\min_{\mathbf{w}} \left[\|\hat{\mathbf{H}}\mathbf{w} - \mathbf{c}_t\|_2^2 + \frac{\rho}{2}\|\mathbf{\Psi X w} + \boldsymbol{\lambda}^{(i)} - \tilde{\mathbf{a}}^{(i)}\|_2^2 \right], \tag{16}$$

$$\tilde{\mathbf{a}}^{(i+1)} = \arg\min_{\tilde{\mathbf{a}}} \left[\eta f_{\text{S}}(\tilde{\mathbf{a}}) + \frac{\rho}{2}\|\mathbf{\Psi X w}^{(i+1)} + \boldsymbol{\lambda}^{(i)} - \tilde{\mathbf{a}}\|_2^2 \right], \tag{17}$$

$$\boldsymbol{\lambda}^{(i+1)} = \boldsymbol{\lambda}^{(i)} + \mathbf{\Psi X w}^{(i+1)} - \tilde{\mathbf{a}}^{(i+1)}, \tag{18}$$

where $\{\cdot\}^{(i)}$ denotes the variable in the i-th iteration. The original minimization problem in (12) is hence decomposed into simpler sub-problems, which are solved in an alternating fashion using the update rules in (16), (17), and (18) until a convergence criterion is satisfied or a maximum number of iterations is exceeded (cf. Section 5.1). In the following, the update rules for the filter and the auxiliary variable in (16) and (17) are presented.

1) Filter update rule: Minimizing (16) yields

$$\mathbf{w}^{(i+1)} = \underbrace{(2\hat{\mathbf{H}}^T\hat{\mathbf{H}} + \rho\mathbf{X}^T\mathbf{X})}_{\mathbf{C}}{}^{-1} [\underbrace{2\hat{\mathbf{H}}^T\mathbf{c}_t}_{\mathbf{b}_1} + \rho \underbrace{\mathbf{X}^T\mathbf{\Psi}^H(\tilde{\mathbf{a}}^{(i)} - \boldsymbol{\lambda}^{(i)})}_{\mathbf{b}_2^{(i)}}], \tag{19}$$

where the variables $\mathbf{C}$, $\mathbf{b}_1$, and $\mathbf{b}_2$ are introduced to highlight that only the variable $\mathbf{b}_2$ is iteration-dependent, whereas the variables $\mathbf{C}$ and $\mathbf{b}_1$ can be pre-computed. Although (19) appears to require a matrix inversion in each iteration, the filter update can be efficiently computed by, e.g., storing the LU-factorization of $\mathbf{C}$ and using forward and backward substitution.

2) Auxiliary variable update rule: The solution to (17) is the proximal mapping of the sparsity-promoting penalty function, which exists in closed-form for the l_0- and l_1-norm penalty functions [19,20]. Defining the variable $\mathbf{b}^{(i)} = \mathbf{\Psi X w}^{(i+1)} + \boldsymbol{\lambda}^{(i)}$ to simplify the notation, the proximal mapping for the l_0-norm is the element-wise *hard thresholding* map, i.e.,

$$\tilde{a}_j^{(i+1)} = \left(|b_j^{(i)}| - \frac{\eta}{\rho} \right)_+ \frac{b_j^{(i)}}{|b_j^{(i)}| - \frac{\eta}{\rho}}, \tag{20}$$

with $\{\cdot\}_j$ denoting the j-th element of a vector and $(T)_+ = \max(T, 0)$. The proximal mapping for the l_1-norm penalty function is the element-wise *soft thresholding* map, i.e.,

$$\tilde{a}_j^{(i+1)} = \left(|b_j^{(i)}| - \frac{\eta}{\rho} \right)_+ \frac{b_j^{(i)}}{|b_j^{(i)}|}. \tag{21}$$

In summary, using the filter update rule in (19), the auxiliary variable update rule in (20) or (21), and the dual variable update rule in (18) until a termination criterion is satisfied, the sparsity-promoting PMINT filter can be computed.

5. SIMULATIONS

In this section the dereverberation performance of the sparsity-promoting PMINT technique using the l_0- and l_1-norm penalty functions is compared to the PMINT technique for several RIR estimation errors.

5.1. Algorithmic settings and performance measures

We consider an acoustic system with a single speech source and $M = 4$ microphones in a room with reverberation time $T_{60} \approx 610$ ms [21]. The source-microphone distance is 2 m, the inter-microphone distance is 4 cm, and the RIR length is $L_h = 4880$ at a sampling frequency of 8 kHz. To generate the microphone signals, 10 sentences (approximately 17 s long) from the HINT database [22] have been convolved with the measured RIRs.

To simulate RIR estimation errors, the measured RIRs have been perturbed by adding scaled white noise as proposed in [23], such that a desired level of normalized projection misalignment (NPM) between the true and the estimated RIRs is generated. The considered NPM values are NPM $\in \{-33\text{ dB}, -27\text{ dB}, -21\text{ dB}, -15\text{ dB}\}$.

For all considered techniques the filter length is $L_w = 1600$ and the target response $\mathbf{c}_t$ is set to the direct path and early reflections of $\hat{\mathbf{h}}_1$, with the length of the direct path and early reflections corresponding to 10 ms (cf. (8)). Furthermore, for the sparsity-promoting PMINT technique, the weighting and penalty parameters are empirically set to $\eta = 10^{-4}$ and $\rho = 10^{-1}$.

In order to reduce the computational complexity, the sparsity-promoting filters are computed using only the first 2 sentences of the microphone signals (approximately 3 s long). However, the complete output signal has been used for the evaluation. The STFT is computed using a 32 ms Hamming window with 50% overlap between successive frames. The frame size is $K = 256$ and the total number of frames is $L = 208$. For the ADMM algorithm, the variables $\mathbf{w}$, $\tilde{\mathbf{a}}$, and $\boldsymbol{\lambda}$ are initialized with $[1\ 0 \ldots 0]^T$. Furthermore, the termination criterion is set to either the number of iterations

[1] Note that the l_0-norm is not a norm in the strict mathematical sense, since it does not satisfy all properties of a norm.

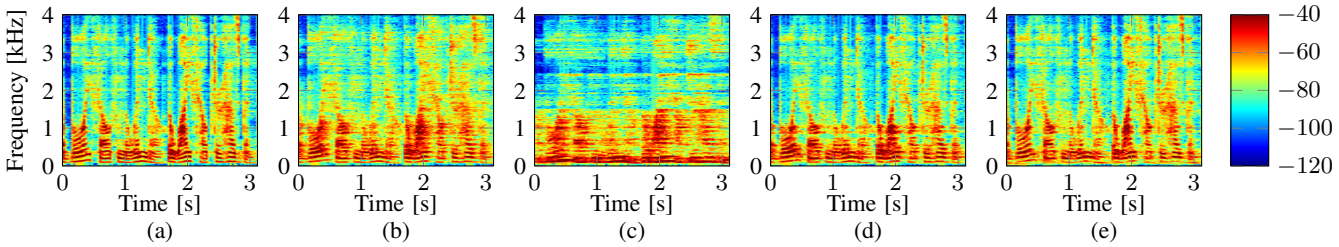

Fig. 2. Spectrogram of (a) reference signal $s_r(n)$, (b) input signal $x_1(n)$, and output signal $z(n)$ obtained using (c) PMINT, (d) l_0-PMINT, (e) l_1-PMINT (NPM $= -33$ dB).

Table 1. Performance in terms of ΔDRR, ΔPESQ, and ΔCD of the PMINT, l_0-PMINT, and l_1-PMINT techniques for several NPMs. The input DRR, PESQ, and CD are -1.89 dB, 2.15, and 4.11 dB.

NPM [dB]	-33	-27	-21	-15
		ΔDRR [dB]		
PMINT	-18.47	-18.16	-17.43	-15.41
l_0-PMINT	3.74	2.76	2.34	0.22
l_1-PMINT	**8.15**	**8.04**	**7.68**	**3.78**
		ΔPESQ		
PMINT	-0.41	-0.50	-0.29	-0.43
l_0-PMINT	0.07	0.05	-0.02	-0.08
l_1-PMINT	**0.36**	**0.42**	**0.32**	**0.03**
		ΔCD [dB]		
PMINT	1.84	1.96	1.88	1.90
l_0-PMINT	-0.68	-0.29	0.05	0.39
l_1-PMINT	**-0.90**	**-0.84**	**-0.50**	**0.14**

exceeding a maximum number of iterations or the relative change in the solution norm dropping below a tolerance, i.e.,

$$i+1 > i_{\max} \quad \text{or} \quad \frac{\|\mathbf{w}^{(i+1)} - \mathbf{w}^{(i)}\|_2}{\|\mathbf{w}^{(i)}\|_2} < \epsilon, \tag{22}$$

with $i_{\max} = 150$ and $\epsilon = 10^{-3}$.

The dereverberation performance is evaluated in terms of the reverberant energy suppression and the perceptual speech quality improvement. The reverberant energy suppression is evaluated using the direct-to-reverberant-ratio improvement (ΔDRR) [24] between the resulting EIR $\mathbf{c}$ and the true RIR $\mathbf{h}_1$. The improvement in perceptual speech quality is evaluated using the improvement in PESQ (ΔPESQ) [25] and in cepstral distance (ΔCD) [26] between the output signal $z(n)$ and the microphone signal $x_1(n)$. The reference signal $s_r(n)$ employed for PESQ and cepstral distance is the clean speech signal convolved with the direct path and early reflections of the true RIR $\mathbf{h}_1$. It should be noted that an improvement in perceptual speech quality is indicated by a positive ΔPESQ and a negative ΔCD.

5.2. Results

Table 1 presents the ΔDRR, ΔPESQ, and ΔCD values obtained using the PMINT and the sparsity-promoting PMINT techniques with l_0-norm (l_0-PMINT) and l_1-norm (l_1-PMINT) penalty functions for several NPMs. It can be observed that, as expected, the PMINT technique fails to achieve dereverberation in the presence of RIR estimation errors and introduces distortions in the output signal, worsening the DRR, PESQ, and CD values. Furthermore, it can be observed that incorporating the l_0-norm penalty function significantly improves the robustness of the PMINT technique, typically yielding a slight improvement (for NPM $= -33$ dB and NPM $= -27$ dB) or a similar performance as the input signal (for NPM $= -21$ dB and NPM $= -15$ dB) in terms of all performance measures. Finally, it can be observed that incorporating the l_1-norm penalty function yields the best performance in terms of the ΔDRR, ΔPESQ, and ΔCD measures, outperforming the l_0-norm penalty function and resulting in a considerable improvement in comparison to the input signal.

To better illustrate the advantages of promoting sparsity of the output signal in the STFT domain, Fig. 2 presents the spectrograms of the reference signal $s_r(n)$, the microphone signal $x_1(n)$, and the output signal $z(n)$ obtained using the PMINT, l_0-PMINT, and l_1-PMINT techniques for an exemplary scenario of NPM $= -33$ dB. Comparing Figs. 2(a) and 2(b), it can be observed that due to the spectro-temporal smearing effect of reverberation, the spectrogram of $x_1(n)$ is significantly less sparse than the spectrogram of $s_r(n)$. Furthermore, Fig. 2(c) shows that due to the distortions introduced in the output signal by the non-robust PMINT technique, the spectrogram of the output signal $z(n)$ obtained using PMINT is significantly less sparse than the spectrogram of $x_1(n)$. On the contrary, Fig. 2(d) shows that incorporating the l_0-norm penalty function sparsifies the spectrogram of the output signal, largely suppressing the distortions introduced by the PMINT technique as well as slightly suppressing the reverberant energy (e.g., around 1 kHz). Finally, Fig. 2(e) shows that incorporating the l_1-norm penalty function results in an even sparser spectrogram than the l_0-norm, significantly suppressing both the distortions introduced by the PMINT technique as well as the reverberant energy.

In summary, these simulation results confirm that incorporating a sparsity-promoting penalty function in equalization techniques yields a significant increase in robustness against RIR estimation errors. Incorporating such penalty functions in other (more robust signal-independent) equalization techniques as well as investigating other sparsity-promoting penalty functions, e.g., the weighted l_1-norm [27], remain topics for future investigation.

6. CONCLUSION

In this paper we have presented a novel signal-dependent method to increase the robustness of acoustic equalization techniques against RIR estimation errors. We have proposed to incorporate an l_0- and l_1-norm penalty function, aiming at promoting sparsity in the output signal and reducing distortions generated by non-robust equalization techniques. The sparsity-promoting filters have been iteratively computed using the alternating direction method of multipliers. Simulation results for several RIR estimation errors have validated the effectiveness of incorporating sparsity-promoting penalty functions to increase the robustness of equalization techniques, with the l_1-norm penalty function outperforming the l_0-norm penalty function.

7. REFERENCES

[1] 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, Feb. 1988.

[2] M. Kallinger and A. Mertins, "Multi-channel room impulse response shaping - a study," in *Proc. International Conference on Acoustics, Speech, and Signal Processing*, Toulouse, France, May 2006, pp. 101–104.

[3] 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, Sept. 2013.

[4] F. Lim, W. Zhang, E. A. P. Habets, and P. A. Naylor, "Robust multichannel dereverberation using relaxed multichannel least squares," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 22, no. 9, pp. 1379–1390, June 2014.

[5] T. Hikichi, M. Delcroix, and M. Miyoshi, "Inverse filtering for speech dereverberation less sensitive to noise and room transfer function fluctuations," *EURASIP Journal on Advances in Signal Processing*, vol. 2007, 2007.

[6] B. D. Radlovic, R. C. Williamson, and R. A. Kennedy, "Equalization in an acoustic reverberant environment: robustness results," *IEEE Transactions on Speech and Audio Processing*, vol. 8, no. 3, pp. 311–319, May 2000.

[7] K. Hasan and P. A. Naylor, "Analyzing effect of noise on LMS-type approaches to blind estimation of SIMO channels: robustness issue," in *Proc. European Signal Processing Conference*, Florence, Italy, Sept. 2006.

[8] I. Kodrasi and S. Doclo, "The effect of inverse filter length on the robustness of acoustic multichannel equalization," in *Proc. European Signal Processing Conference*, Bucharest, Romania, Aug. 2012.

[9] S. Makino, S. Araki, 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, Hoboken, USA, 2010.

[10] H. Kameoka, T. Nakatani, and T. T. Yoshioka, "Robust speech dereverberation based on non-negativity and sparse nature of speech spectrograms," in *Proc. International Conference on Acoustics, Speech, and Signal Processing*, Apr. 2009, pp. 45–48.

[11] T. Van Waterschoot, B. Defraene, M. Diehl, and M. Moonen, "Embedded optimization algorithms for multi-microphone dereverberation," in *Proc. European Signal Processing Conference*, Marrakech, Morocco, Sept. 2013.

[12] 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, Sept. 2015.

[13] S. P. 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, Jan. 2011.

[14] I. Tashev and A. Acero, "Statistical modeling of the speech signal," in *Proc. International Workshop on Acoustic Echo and Noise Control*, Tel Aviv, Israel, Sept. 2010.

[15] T. Gerkmann and R. Martin, "Empirical distributions of DFT-domain speech coefficients based on estimated speech variances," in *Proc. International Workshop on Acoustic Echo and Noise Control*, Tel Aviv, Israel, Sept. 2010.

[16] M. Kowalski, E. Vincent, and R. Gribonval, "Beyond the narrowband approximation: Wideband convex methods for underdetermined reverberant audio source separation," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 18, no. 7, pp. 1818–1829, Sept. 2010.

[17] B. K. Natarajan, "Sparse approximate solutions to linear systems," *SIAM Journal on Computing*, vol. 24, no. 2, pp. 227–234, Apr. 1995.

[18] S. P. Boyd and L. Vandenberghe, *Convex Optimization*, Cambridge University Press, Cambridge, United Kingdom, 2004.

[19] R. Chartrand, "Shrinkage mappings and their induced penalty functions," in *Proc. International Conference on Acoustics, Speech, and Signal Processing*, Florence, Italy, May 2014, pp. 1026–1029.

[20] N. Parikh and S. P. Boyd, "Proximal algorithms," *Foundations and Trends in Optimization*, vol. 1, no. 3, 2014.

[21] E. Hadad, F. Heese, P. Vary, and S. Gannot, "Multichannel audio database in various acoustic environments," in *Proc. International Workshop on Acoustic Echo and Noise Control*, Antibes, France, Sept. 2014, pp. 313–317.

[22] M. Nilsson, S. D. Soli, and A. Sullivan, "Development of the Hearing in Noise Test for the measurement of speech reception thresholds in quiet and in noise," *Journal of the Acoustical Society of America*, vol. 95, no. 2, pp. 1085–1099, Feb. 1994.

[23] W. Zhang and P. A. Naylor, "An algorithm to generate representations of system identification errors," *Research Letters in Signal Processing*, vol. 2008, Jan. 2008.

[24] P. A. Naylor and N. D. Gaubitch, Eds., *Speech dereverberation*, Springer, London, UK, 2010.

[25] ITU-T, *Perceptual evaluation of speech quality (PESQ), an objective method for end-to-end speech quality assessment of narrowband telephone networks and speech codecs P.862*, International Telecommunications Union (ITU-T) Recommendation, Feb. 2001.

[26] S. Quackenbush, T. Barnwell, and M. Clements, *Objective measures of speech quality*, Prentice-Hall, New Jersey, USA, 1988.

[27] 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, Oct. 2008.