Feedback to Reviewers' Comments on Manuscript T-ASL-05976-2016 Signal-Dependent Penalty Functions for Robust Acoustic Multi-Channel Equalization

Ina Kodrasi and Simon Doclo

February 21, 2017

We would like to thank the reviewers for their review and for their useful suggestions and comments. We believe that we have addressed most concerns raised by the reviewers by implementing the changes (denoted in green in the revised manuscript) described in the following sections.

1 Reviewer 1

1. *Page 2, line 13, left column: the beamformer in the cited paper [34] utilizes the non-Gaussianity of the speech signal, not the sparsity.*

   We agree with the reviewer that the authors in [34] exploit the non-Gaussianity of the speech signal for deriving the adaptive beamformer. More precisely, they model the speech signal with a generalized Gaussian distribution. Since the family of generalized Gaussian distributions is a common model of sparse distributions, we believe it is correct to state that the authors in [34] exploit the sparsity of speech signals and **have not made any changes in the revised manuscript**.

2. *Energy decay curves in equations (43) and (44) are known as Schroeder formula, you may want to cite the original 1965 paper "New method for measuring reverberation time" in JASA.*

   We have added this reference in the revised manuscript.

3. *Recommendation to add a fourth column in Tables IV, V, and VI with the average numbers. You may even consider leaving only the average numbers and reducing the size of sections VI.C and VI.D.*

   We agree with the reviewer that providing the average performance for each technique facilitates the comparison between them, hence, **we have added the average performance values in all tables in the revised manuscript**. However, since the performance of the techniques depends on the acoustic system, we have decided to leave the individual results as well such that the advantages and shortcomings of the considered equalization techniques can be studied more carefully for each acoustic system.

2 Reviewer 2

1. *First paragraph of Section III: The authors should ensure that the error being an unknown quantity is made clear in the text.*

   We have reformulated the following sentence in the revised manuscript (page 3, left column):

   "However, it should be realized that in practice only the perturbed RIRs $\hat{h}_m$ are available, i.e., $\hat{h}_m = h_m + e_m$, where e_m represents the unknown RIR perturbations arising due to fluctuations (e.g., temperature or position fluctuations [22]) or due to the sensitivity of BSI and SSI methods to near-common zeros or interfering noise [23]-[25]."

2. *The algorithm assumes that the error correlation matrix is an identity matrix. This may not be true in some cases and the authors should elaborate on when such case can happen and show results on such cases (where the assumption is not valid). They should also show, in the simulation results, how deviation from such assumption will adversely affect the performance of the algorithm.*

   It should be realized that the proposed sparsity-promoting algorithms actually do not require any knowledge of the error correlation matrix and that only existing signal-independent regularized algorithms require knowledge of the error correlation matrix. The performance of regularized algorithms has been extensively investigated when proposed in [19]. Since the aim of this manuscript is to propose and investigate sparsity-promoting algorithms, we believe that a detailed analysis and comparison with existing regularized techniques for different error correlation matrices is out of scope and **have not added additional simulations in the revised manuscript**.

   Nevertheless, the reviewer is completely right in pointing out that for the regularized algorithms this assumption is not always true. We believe that the following paragraph (already included in the original manuscript) clarifies that when errors arise due to microphone position fluctuations or BSI/SSI methods, the error correlation matrix is not equal to an identity matrix but follows e.g., the structure derived in [17, 40] (page 4, left column):

   "When knowledge is available about the type of RIR perturbations (e.g., arising due to microphone position fluctuations or arising from BSI or SSI methods), the matrix $\mathbf{R}_e$ can be constructed based on an appropriate perturbation model [17], [40]. When such knowledge is not available, the RIR perturbations are often assumed to be spatially and temporally white, i.e., $\mathbf{R}_e = \mathbf{I}_w$, with $\mathbf{I}_w$ denoting the $ML_w \times ML_w$–dimensional identity matrix [19], [26]. This assumption has been used for the regularized techniques in Section VI-E."

3. *In (29), the weight is made to be inversely proportional to the energy of the received signal. What would happen if the energy of the signal is low (such as a silent period)? Would the algorithm generate a hissing/howling effect?*

   We have ensured that the algorithm is stable and does not generate such effects in speech absence by including the small positive scalar ζ in computing the weights in (29) (already included in the original manuscript). Using ζ in speech absence (e.g., $\zeta = 10^{-8}$) ensures that the computed weights are large but finite (e.g., $\approx 10^8$), such that the reconstructed signal has (nearly) zero energy. In order to clarify this point, **we have rephrased the following sentence in the revised manuscript** (page 5, right column):

   "However, since the clean speech signal is obviously not available, we propose to use one of the reverberant microphone signals and compute the weights as

   $$u(q) = \frac{1}{|\tilde{x}_p(q)| + \zeta}, \quad q = 0,\ 1,\ \ldots,\ NL-1,$$

   with $\tilde{x}_p(q)$ the STFT coefficients of the p-th microphone signal computed similarly as in (21) and $\zeta > 0$ a small positive scalar included to avoid division by 0 and to ensure stability of the weighted l_1-norm sparsity-promoting techniques. When $|\tilde{x}_p(q)| \approx 0$, e.g., during speech absence, using ζ ensures that $u(q)$ is large but finite, such that the magnitude of the STFT coefficients of the reconstructed signal is also (nearly) 0. "

4. *The authors should also consider more cases where the reverberation is varied. They could for example use synthetically generated RIRs so that the results are re-producible.*

   We agree with the reviewer that the performance for other reverberation times could be analyzed to gain more insight on the proposed techniques. However, in our opinion synthetically generated RIRs (e.g., using the image source model) are not really realistic and unfortunately databases with a large variety of realistic RIRs are not commonly available. Furthermore, we believe that the considered reverberation times ranging from 300 ms to 600 ms span a realistic range of reverberation times found in small to medium size rooms. Hence **we have not added any additional simulations in the revised manuscript**.

5. *How would the performance of the algorithm change with ζ, ρ and η?*

   We would like to thank the reviewer for this useful comment. Since preliminary simulations showed that the performance of the proposed techniques is hardly influenced by the choice of ζ (as long as it is small enough), we have not included additional simulations for different values of this parameter. In order to evaluate the sensitivity of the proposed techniques to the weighting and penalty parameters η and ρ, **we have included Section VI-D in the revised manuscript**, where the performance of the weighted l_1-norm sparsity-promoting PMINT technique is investigated for different values of η and ρ. The presented results show that the performance of the proposed techniques is unfortunately rather sensitive to the choice of these parameters. However, as shown in Section VI-E, once a set of optimal parameters has been determined, the same parameters can be used for different acoustic systems and NPMs to obtain a near-to-optimal performance.

3 Reviewer 3

1. *In the paragraph before equation (17), two assumptions are made and three papers are cited to support these assumptions. I checked with those papers but found none of those explained why the two assumptions are needed for the solution given in (17). Please cite the proper reference or explain clearly why the two assumptions are needed.*

   We would like to thank the reviewer for pointing out this typo. Since the second and third references are wrong, we have removed them from the revised manuscript. The first assumption is derived in [15], whereas the second assumption is derived in [39]. **We have included the correct references in the revised manuscript.**

2. *In your simulations, the acoustic impulse responses are measured by the swept-sine method, which is not very practical for use in real applications. It would be helpful to the reader to comprehend the usefulness of the method if a set of results are presented using channel impulse responses identified using the state-of-the-art acoustic channel identification algorithms.*

   We agree with the reviewer that using state-of-the-art blind system identification (BSI) methods to identify the RIRs would yield a more realistic evaluation of the performance of the proposed techniques. However, the addition of Gaussian distributed errors to perturb measured RIRs to a certain NPM is a widely used technique to systematically simulate RIR perturbations (e.g., in [18, 19, 30, 31]). By considering NPM levels in the range achieved by state-of-the-art BSI methods [25], i.e., between -10 and -20 dB, we believe that the presented results depict the usefulness of the methods, and hence, **we have not added new simulations in the revised manuscript**. However, we believe it is important to comment on this issue, hence, **we have reformulated the following paragraph in the revised manuscript** (page 6, right column):

   "In order to simulate RIR perturbations, the measured RIRs are perturbed by proportional Gaussian distributed errors as proposed in [55], such that a desired level of normalized projection misalignment (NPM), i.e.,

$$NPM = 20\log_{10}\frac{\left\|\mathbf{h}-\frac{\mathbf{h}^T\hat{\mathbf{h}}}{\hat{\mathbf{h}}^T\hat{\mathbf{h}}}\hat{\mathbf{h}}\right\|_2}{\|\mathbf{h}\|_2}, \tag{1}$$

   is generated. Introducing proportional Gaussian distributed errors is a widely used technique to systematically simulate RIR perturbations. The considered NPMs for each acoustic system are

$$NPM \in \{-33\ \text{dB},\ -27\ \text{dB},\ \ldots,\ -3\ \text{dB}\}, \tag{2}$$

   with -33 dB a moderate perturbation level and -3 dB a rather large perturbation level. The reported NPM values achieved by state-of-the-art BSI methods (for relatively short RIRs) range between -10 dB and -20 dB [25], hence, the values in (2) can be considered as rather realistic."

3. *Signal-dependent sparsity promoting method has been investigated for acoustic channel identification. Please explain the difference between the method in this paper and those used for acoustic channel identification. After all, both channel identification and equalization are two closed related problems.*

   To the best of our knowledge, sparsity-promoting methods for acoustic channel identification [43, 44] rely on a sparse model of the room impulse responses in the time-domain and not of the clean signal in the short-time Fourier transform domain. Hence, the difference between the methods consists in the fact that while sparsity-promoting methods for identification aim at enforcing sparsity on the estimated room impulse responses, the method proposed in this manuscript aims at enforcing sparsity on the estimated clean speech signal. In order to clarify this point, **we have added the following paragraph in the revised manuscript** (page 4, right column):

   "It should be noted that sparsity-promoting techniques have also been proposed to increase the robustness of BSI methods to additive noise [43], [44]. However, unlike the sparsity-promoting equalization techniques proposed in the following which aim at enforcing sparsity on the output speech signal, sparsity-promoting techniques for BSI aim at enforcing sparsity on the estimated RIRs."

4. *In Fig. 2 (c), what is the reason that the PESQ improvement is not a decreasing function with respect to NPM?*

   Since measures such as PESQ are actually not originally developed to assess the perceptual quality of signals processed by equalization techniques, they do not necessarily always reflect the distortions introduced by such techniques. Our informal listening tests suggest that particularly when distortions are large (i.e., for large NPMs and non-robust techniques) PESQ does not correlate well with the perceived signal quality. Hence, the fact that PESQ is a non-decreasing function of NPM does not mean that the quality of the signal is not deteriorating with increasing NPM, but that PESQ does not completely capture this deterioration. As shown in the recently organized REVERB Challenge, reliable performance measures to evaluate the quality of signals processed by dereverberation algorithms are unfortunately still lacking.