Feedback to Reviewer's Comments on Manuscript T-ASL-03898-2012 Regularization for Partial Multichannel Equalization for Speech Dereverberation

Ina Kodrasi, Stefan Goetze, Simon Doclo

December 20, 2012

General Comments to all Reviewers

We would like to thank the reviewers for their careful and detailed review and for their useful suggestions and comments. We believe that we have addressed most concerns raised by the reviewers by implementing the changes described in the following sections.

Please note that the **simulation parameters and hence, the presented results (Tables I-II and Figs. 3-8) have been changed in the revised paper** in order to address Comment 4 by Reviewer 1.

Further, **minor paragraphs in the text have been shortened/reformulated in the revised paper**, in order to allow the incorporation of a new experimental section (Section V-D. Robustness in the Presence of Channel Estimation Errors and Additive Noise) which addresses Comment 1 by Reviewer 1 and Comment 1 by Reviewer 3.

However, **the conclusions drawn in the paper remain valid**, i.e.

- **Regularization yields a significant performance improvement in P-MINT, whereas a smaller improvement is observed for RMCLS and CS.**
- **The intrusively regularized P-MINT technique is robust and yields the highest perceptual speech quality.**
- **The automatic selection procedure for the regularization parameter in P-MINT yields a very similar performance as the intrusive selection procedure.**

1 Reviewer 1

1. *So far, the manuscript neglects background noise, which is usually recorded by the microphones in addition to reverberation. Since the robustness of channel equalization to background noise is extremely important in practical scenarios, background noise should be included in the paper at the following points:*

   a) In equation (1)

   b) In the discussion of regularization, equations (40) - (42)

   c) There should be at least one experiment including realistic background noise (recorded noise, not white noise)

   We agree with the reviewer that it is important to analyze the performance of the considered acoustic multichannel equalization techniques in the presence of background noise, hence **we have made the following changes in the revised paper:**

   - **Equation (1) has been changed to**

     $$y_m(n) = \underbrace{s(n) * h_m(n)}_{x_m(n)} + v_m(n) = x_m(n) + v_m(n), \quad (1)$$

     with $v_m(n)$ being the additive noise signal.

   - Since for the design of reshaping filters in acoustic multichannel equalization the presence of background noise is typically disregarded, we have also assumed that $v_m(n) = 0$ in the provided derivations/discussions. However, **we have added the following sentence in the revised paper** on page 2.

     "*Since acoustic multichannel equalization techniques generally design reshaping filters disregarding the presence of noise, in the following it is assumed that $v_m(n) = 0$, hence $y_m(n) = x_m(n)$.*"

     For the sake of consistency and to avoid confusion for the reader, we believe that we should not reintroduce noise in the discussion of equations (40)-(42). Hence, **we have not changed equations (40)-(42) in the revised paper**. However, we refer to these equations again in the newly introduced experimental section (Section V-D).

     [Please note that to allow the incorporation of a new experimental section, some of the labeled equations in the first version of the paper have been incorporated within the text in the revised version. Hence, the labeling of the equations in the revised paper has been changed, meaning that equations (40)-(42) in the first submitted version correspond to equations (35)-(37) in the revised paper.]

   - **In the revised paper, we have added simulation results describing the performance of acoustic multichannel equalization techniques in the presence of both channel estimation errors and recorded noise (Section V-D).**

     Simulation results show that when equalization techniques such as P-MINT are used without any regularization, the additive noise at the microphones is severely amplified as expected. Furthermore, it is shown that incorporating regularization is effective not only for increasing the robustness of P-MINT to channel estimation errors, but also to avoid the amplification of the additive noise.

2. *The experimental results indicate that the performance gain obtained by the channel equalization technique strongly depend on the channel estimation error (as could be expected).*

For the experiments, scaled white noise is added to the true RIRs as suggested in [32] at a normalized channel mismatch E_m of -33 dB and -15 dB. However it is not clear how realistic these conditions are in practical scenarios. To give the reader an idea of how the channel equalization techniques work in real-world scenarios either experiments with RIRs obtained by state-of-the-art channel identification techniques should be added or at least it should be discussed whether the conditions described above(scaled white noise at E_m of -33 dB or -15 dB) can be achieved by current channel identification algorithms.

It is indeed true that the performance of equalization techniques in practice critically depends on the performance of blind system identification (BSI) methods. However, to the best of our knowledge, the performance of BSI methods is highly dependent on the acoustic scenario being considered. Furthermore, no model has been established that allows to systematically describe estimation errors generated by BSI methods.

Since investigating the performance and the effect of different BSI methods on the presented channel equalization techniques is beyond the scope of this paper, **we have not added any experiments in the revised paper** with RIRs estimated by BSI methods.

However, we also believe it is important to comment on this issue, hence, **we have added the following paragraph in the revised paper** on page 6:

"*In practice, BSI methods [18], [23] should be used to directly estimate the acoustic system. However, to the best of our knowledge the performance of state-of-the-art BSI methods highly depends on the considered acoustic system and no model has been established to systematically describe the estimation errors caused by them. Therefore, (52) and (53) are used to generate the considered estimation errors in the following simulations.*"

3. *Page 2, paragraph 3 of the abstract:*
"... outperforms all other intrusively regularized ... techniques"
should be replaced by
"... outperforms all other considered intrusively regularized ... techniques"
Similar in the second paragraph of the conclusion (page 20).

   Replaced in the revised paper.

4. $L_h = 2000$ *appears to be very short for a reverberation time of* 600 *ms. To cover at least* 50*% of the reverberation time would require* $L_h = 4800$ *at* $f_s = 16$ *kHz for a reverberation time of* 600 *ms. Therefore, the simulations should be performed with RIRs of length* $L_h \geq 4800$ *and corresponding* L_g. *If such long RIRs are a problem to any of the algorithms, then you should at least use* $L_h \geq 4800$ *for generating (some of) the microphone signals and investigate the difference to* $L_h = 2000$.

   We completely agree with the reviewer that a longer RIR length should have been used. Hence, we have re-measured RIRs in a room with reverberation time $T_{60} \approx 550$ ms and all simulations have been run using $L_h = 4400$ and $L_g = 4399$ at the sampling frequency $f_s = 16$ kHz. **We have changed all presented simulation results in the revised paper**, to account for this change in the considered acoustic scenario and simulation parameters.

   Please note that changing the acoustic scenario and simulation parameters has obviously resulted in a change in the obtained performance values, i.e. the values presented in Tables I-II and Figs. 3-8. However, the main conclusions still remain the same as in the first version of the paper.

5. *Figure* 4 *and Figure* 6*: The best performance with regularized P-MINT is achieved with the longest considered desired window length of* $L_d = 50$ *ms. Therefore please add tests*

with longer windows to see were the maximum performance is obtained. Furthermore, please provide color diagrams.

Perceptually motivated objective measures (e.g. the well-known D_{50} measure[1]) suggest that speech intelligibility is guaranteed only when the RIR is shortened so that the energy is concentrated within 50 ms after the first impulse of the RIR. Since the main focus of this article is to derive a perceptually advantageous equalization technique, we have limited the maximum investigated desired window length to $L_d = 50$ ms. Therefore **we have not implemented any changes in the revised paper** regarding this issue.

6. *Figure* 8*: Please add the PESQ score for the unprocessed micophone signal $x_1(n)$ to the plots.*

   Added in the revised paper.

7. *Figure* 5*: From the reviewers point of view, there is no need for the vertical axis to start at -60 dB. The plots would be better readable if the vertical axis started at -30 dB.*

   We agree with the reviewer that a reverberant tail below -30 dB has an insignificant perceptual relevance. However, for a moderate channel mismatch of -33 dB, all regularized partial equalization techniques achieve a very similar level of reverberant tail suppression, with their reverberant tails being below -30 dB (cf. Fig. 3b). If the vertical axis started at -30 dB, it would not be possible to compare these techniques to each other.

   However, to gain more readability while still being able to compare the considered techniques **we have changed the vertical axis for the EDCs in the revised paper** to start from -50 dB instead of -60 dB.

[1] ISO Norm 3382: Acoustics-Measurement of the Reverberation Time of Rooms With Reference to Other Acoustical Parameters, SO Norm 3382, Int. Org. Standardization (ISO)

2 Reviewer 2

1. *It is not clear why the inverse filter becomes robust if it equalizes only the late reverberation part of RIR. Please show clearly based on which equations/operations/nature we can expect it becomes robust to RIR estimation error. It is generally required to have such theoretical validation in addition to the experimental results to support the claim of the paper.*

   We agree with the reviewer that it would be very interesting to provide a theoretical derivation showing why partial equalization is more robust than complete equalization. However, it should be realized that there is a fundamental problem when attempting to tackle this issue from a theoretical perspective.

   - We are aiming at increasing the perceptual speech quality in multichannel equalization techniques, hence the robustness to channel estimation errors is evaluated in terms of perceptual performance measures such as PESQ (as is typically done in the literature on acoustic multichannel equalization). Unfortunately, such measures are not directly related to any of the cost functions being optimized by acoustic multichannel equalization techniques, and hence can not be written as analytical mathematical expressions. Therefore the effect of estimation errors on such measures can not be analytically quantified.

   In addition, we are sure that the reviewer is also aware that such a theoretical derivation is also missing in all previous works where partial equalization techniques such as RMCLS[2] and CS[3] have been proposed.

   We believe that the sentence on page 1 of the paper

   "*It has been experimentally validated that partial equalization techniques lead to a significant increase in robustness in the presence of RIR estimation errors as compared to complete equalization.*"

   is not misleading for the reader who is familiar with the fundamental problem mentioned above. Hence, **we have not added any further remarks about this issue in the revised paper**.

2. *In the experimental section, they compared MINT, Relaxed multi-channel least squares (RMCLS), channel shortening (CS), P-MINT and their regularized versions, and showed the superior performance of RMCLS in terms of energy decay curve (EDC). It is not clear to me why RMCLS is superior to the others. Please add more technical discussions to the experimental section that can explain these differences in performance. It is also curious to see how the performance of the regularized RMCLS and CS may change if the regularization parameters are selected not from the view point of perceptual quality, but from the view of reverberation suppression rate.*

   Unfortunately, for the same fundamental reason as mentioned in the feedback to Comment 1, a theoretical/analytical derivation why certain techniques are more robust than others remains an open research question.

[2] W. Zhang, E. A. P. Habets, and P. A. Naylor, *On the use of channel shortening in multichannel acoustic system equalization*, in Proc. International Workshop on Acoustic Echo and Noise Control (IWAENC),Tel Aviv, Israel, Sep. 2010.

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

As for the selection of the optimal regularization parameter, we have followed the approach proposed by the reviewer in our previous work[4], where it has been shown that selecting the regularization parameter in P-MINT such that the reverberant tail suppression is optimized leads to a higher performance as compared to state-of-the-art equalization techniques in terms of reverberant tail suppression. However, since the aim in this journal paper is to derive a perceptually advantageous equalization technique, optimizing the performance of regularized equalization techniques in terms of reverberant tail suppression is not extremely important. Hence, **we have not implemented any changes in the revised paper**.

3. *Please add the PESQ score of the ideally dereverberated signal for the reference purpose. It can be realized for example as the convolution of clean signal and the first part of RIR* $\mathbf{h}_p^d$*. By doing so, we can see how close the performances of the discussed dereverberation methods are to ideal one.*

   The reviewer should realize that the reference signal we used for calculating the PESQ score was indeed $s(n) * h_p^{\rm d}(n)$. However, since this might not have been clear in the first version of the paper, **we have reformulated the following paragraph in the revised paper** on page 7.

   "*The perceptual speech quality of the output signal* $\hat{s}(n)$ *is evaluated using the objective speech quality measure PESQ [33] which generates a similarity score between the reference and output signals in the range of* 1 *to* 4.5*. The reference signal employed in PESQ is* $s(n) * h_1^{\rm d}(n)$*, i.e. the clean speech signal convolved with the first part of the true first RIR for each value of the desired window length* L_d*.*"

4. *Please remove or modify the last paragraph of the section V-B (the one starting with "Summarizing the simulation results..."). Although it is mentioned as if the regularized P-MINT is the best among the compared methods, we can see that it is just not true, for example, by taking a look at Fig.3, where we can see that RMCLS suppresses the late reverberation more effectively.*

   We agree with the reviewer and **have modified this paragraph in the revised paper** to:

   "*Summarizing the simulation results, we conclude that regularized P-MINT is a robust and perceptually advantageous equalization technique, outperforming all other considered equalization techniques in terms of perceptual speech quality.*"

5. *Please add some comments on the estimation accuracy for the regularization parameter. We can see that it is working reasonably well by observing PESQ score (Fig. 7). In addition to the PESQ score, it may be helpful for readers to understand the effectiveness of the proposed scheme, if the paper includes how close the value itself is to the intrusively selected parameter.*

   We would like to thank the reviewer for this useful comment. In order to evaluate the estimation accuracy for the regularization parameter, we have computed the normalized error between the automatic regularization parameter $\delta_{\rm auto}$ and the optimal regularization parameter $\delta_{\rm opt}$ as

[4] I. Kodrasi and S. Doclo, *Robust partial multichannel equalization techniques for speech dereverberation*, in Proc. International Conference on Acoustics, Speech, and Signal Processing (ICASSP), Kyoto, Japan, Mar. 2012, pp. 537–540.

$$\frac{\log_{10}\delta_{\text{opt}} - \log_{10}\delta_{\text{auto}}}{\log_{10}\delta_{\text{opt}}}, \tag{2}$$

and the normalized mean square error values over all considered scenarios are computed. Hence, **we have added the following sentences in the revised paper** on page 10:

"*Furthermore, the normalized mean square error between the optimal and automatic regularization parameter over all considered* L_d *is* 0.03*, where the normalized error is defined as* $(\log_{10}\delta_{\text{opt}} - \log_{10}\delta_{\text{auto}})/\log_{10}\delta_{\text{opt}}$."

"*Furthermore, the normalized mean square error between the optimal and automatic regularization parameter over all channel mismatch values is* 0.64 *(*$T_{60} \approx 450$ *ms),* 0.05 *(*$T_{60} \approx 550$ *ms), and* 0.05 *(*$T_{60} \approx 750$ *ms).*"

3 Reviewer 3

1. *The authors have considered the estimation error in the RIRs but ignored the additive noise (with the clean speech signal) that causes the estimation error (see Eq. 42). Even if we are given the true RIR, the output signal obtained from the MINT equalization filters is severely distorted due to the additive noise. The regularization term may be robust against the additive noise but it should be demonstrated in the paper.*

   We agree with the reviewer that it is important to analyze the performance of acoustic multichannel equalization techniques in the presence of background noise, and more importantly, to analyze the effect of incorporating regularization.

   In the revised paper, we have added simulation results describing the performance of acoustic multichannel equalization techniques in the presence of both channel estimation errors and recorded noise (Section V-D).

   Simulation results show that when equalization techniques such as P-MINT are used without any regularization, the additive noise at the microphones is severely amplified as expected. Furthermore, it is shown that incorporating regularization is effective not only for increasing the robustness of P-MINT to channel estimation errors, but also to avoid the amplification of the additive noise.

2. *In order to simulate estimation errors, the measured RIRs have been perturbed by adding scaled white noise as proposed in [32]. However, noise robust RIR estimation technique from the noisy speech signal has been already proposed in the literature (see [18] and the ref. therein). These techniques can be used to simulate more realistic scenario.*

   It is indeed true that for a more realistic experimental setting, blind system identification (BSI) methods could be used to estimate the RIRs. However, to the best of our knowledge, the performance of BSI methods is highly dependent on the acoustic scenario being considered. Furthermore, no model has been established that allows to systematically describe estimation errors generated by BSI methods.

   Since investigating the performance and the effect of different BSI methods on the presented channel equalization techniques is beyond the scope of this paper, **we have not added any experiments in the revised paper** with RIRs estimated by BSI methods.

   However, we also believe it is important to comment on this issue, hence, **we have added the following paragraph in the revised paper** on page 6:

   "*In practice, BSI methods [18], [23] should be used to directly estimate the acoustic system. However, to the best of our knowledge the performance of state-of-the-art BSI methods highly depends on the considered acoustic system and no model has been established to systematically describe the estimation errors caused by them. Therefore, (52) and (53) are used to generate the considered estimation errors in the following simulations.*"

3. *The RMCLS algorithm produces better PESQ score than the P-MINT without the regularization term. The performance of RMCLS does not change even if the regularization term is incorporated (Fig. 4). However, the performance of regularized P-MINT is significantly better than the P-MINT. Will you please explain that?*

   Since the presented results have been changed in the revised paper (in order to address Comment 4 by Reviewer 1), we would first like to clarify the following:

   - It is indeed true that in the first version of the paper, the perceptual speech quality of RMCLS when incorporating regularization did not change for the normalized channel mismatch $E_m = -33$ dB (Fig. 4). However, an improvement could be noticed for the normalized channel mismatch $E_m = -15$ dB (Fig. 6). Hence, the only conclusion drawn in this respect was that the performance improvement for RMCLS when incorporating regularization is smaller than the performance improvement for P-MINT.

- Please note that after having changed the experimental conditions and simulation parameters in the revised paper (in order to address Comment 4 by Reviewer 1), a performance improvement can now be noticed for RMCLS for both considered mismatches $E_m = -33$ dB (Fig. 4) and $E_m = -15$ dB (Fig. 6). Nevertheless, the conclusion remains the same, i.e. the performance improvement for RMCLS when incorporating regularization is smaller than the performance improvement for P-MINT.

The incorporation of regularization is unfortunately not directly related to the perceptual speech quality. By incorporating a regularization parameter, the energy of the reshaping filter is decreased, implying that there will likely be less distortions caused by RIR estimation errors in the output signal. This might lead to a similar or slightly better reverberant tail suppression in an already robust technique such as RMCLS (which is important for a high perceptual speech quality), however, it does not in any way control the remaining early reflections in the equalized impulse response (which are also very important for a high perceptual speech quality).

On the other hand, P-MINT unfortunately inherits the sensitivity of MINT, yielding no reverberant tail suppression and a rather low perceptual speech quality in the presence of RIR estimation errors (as can be observed in Figs. 3a and 5a). The incorporation of regularization in P-MINT significantly improves the achieved reverberant tail suppression, resulting in a similar reverberant tail suppression as regularized RMCLS (as can be observed in Figs. 3b and 5b). This contributes to a large improvement in the perceptual speech quality for P-MINT. The remaining advantage in regularized P-MINT which leads to a higher perceptual speech quality than regularized RMCLS is the direct control of the early reflections in the equalized impulse response.

In order to clarify these differences, **we have added the following paragraph in the revised paper** on page 10:

"*The large performance improvement obtained for P-MINT when regularization is incorporated can be explained by the significantly higher reverberant tail suppression that is achieved. The remaining advantage that leads to regularized P-MINT outperforming state-of-the-art techniques lies in the direct control of the early reflections.* "

4. *Can you please demonstrate how the performance of regularized P-MINT varies with selection of different acoustic channels in Eq.* 29*?*

   [Please note that to allow the incorporation of a new experimental section, some of the labeled equations in the first version of the paper have been incorporated within the text in the revised version. Hence, the labeling of the equations in the revised paper has been changed, meaning that equation (29) in the first submitted version corresponds to equation (24) in the revised paper.]

   In order to address this comment, we ran simulations for the regularized P-MINT technique using acoustic channel 2 as the desired equalized impulse response (EIR) for the normalized channel mismatch $E_m = -33$ dB.

The figure presented below depicts the PESQ scores obtained by regularized P-MINT for several desired window lengths L_d with $\hat{\mathbf{h}}_1^{\mathrm{d}}$ and $\hat{\mathbf{h}}_2^{\mathrm{d}}$ as the desired EIR, i.e. with $p = 1$ and $p = 2$ respectively in equation (24). The obtained PESQ scores are slightly different and depend on the acoustic channel being used as the desired EIR. However, the mean square difference over all considered L_d is 0.004, which is an insignificant value. Hence, we believe that the sentence on page 4 of the paper

> "*Without loss of generality, also other desired EIRs could be used instead of* (24)*, as long as they are perceptually close to the true RIRs.*"

is sufficient and **we have not added any further remarks about this issue in the revised paper.**

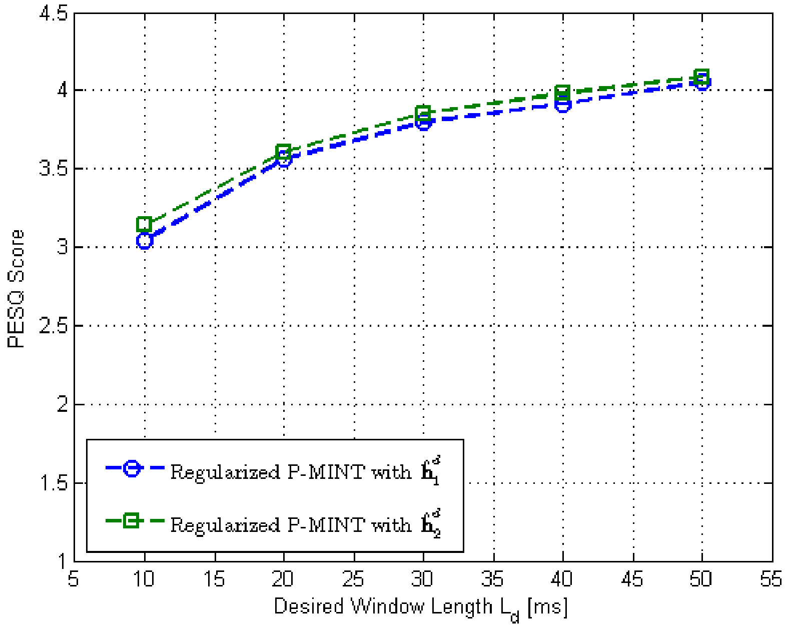

Figure 1: PESQ score of the system's output $\hat{s}(n)$ obtained using regularized P-MINT with $\hat{\mathbf{h}}_1^{\mathrm{d}}$ and regularized P-MINT with $\hat{\mathbf{h}}_2^{\mathrm{d}}$ as the desired equalized impulse response ($E_m = -33$ dB)