# Feedback to Reviewers' Comments on Manuscript JAES-D-16-00064: Instrumental and Perceptual Evaluation of Dereverberation Techniques based on Robust Acoustic Multichannel Equalization

Ina Kodrasi, Benjamin Cauchi, Stefan Goetze, Simon Doclo

June 29, 2016

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 **several paragraphs in the revised manuscript have been reformulated/added**, hence, the numbering of equations has changed.

## 1 Reviewer 1

- *The m/s uses the proportional Gaussian noise model proposed by Zhang and Naylor which clearly is a very practical choice for modelling perturbations. However, there are many other types of errors possible in the measurement chain and the acoustical system (i.e. the room): 1) temperature gradients that would cause temporal modulations in the RIR, 2) diffuse additive noise at microphones that are partly correlated, 3) microphone directivity and transfer function mismatches (phase mismatches being especially hard to equalise), 4) uncorrelated component noise at the microphones, and 5) jitter and quantisation noise in the recording chain (see the references below). I would ideally like to see a thorough investigation including the simulation and the comparisons for these different second-order perturbations as well. However, I also understand that this may require some additional work (including a rerun of the subjective experiment which is not desirable). Therefore, I suggest that the authors merely discuss the other causes of RIR perturbations. This can be done in a new section after shortening the introduction and the rationale for using the Zhang/Naylor model should also be clarified in that section.*

  We agree with the reviewer that in typical acoustic scenarios there are many other sources of interference which we have not considered, such as additive noise, RIR perturbations arising due to temperature variations, or microphone directivity and transfer function mismatches. Although noise is typically present in many acoustic scenarios, the aim of this paper is to evaluate only the dereverberation performance of acoustic multichannel equalization techniques (as is clearly mentioned in the last paragraph of Page 2), hence the presence of noise has been disregarded. However, in order to address this comment, the description of the acoustic system setup in Section 2.1 **has been reformulated in the revised manuscript** to:

  *"The mth microphone signal $y_m(n)$, $m = 1, \ldots, M$, at time index $n$ is given by*

$$y_m(n) = \underbrace{\sum_{l=0}^{L_h-1} h_m(l)s(n-l)}_{x_m(n)} + v_m(n) = x_m(n) + v_m(n), \tag{1}$$

  *where $h_m(l)$, $l = 0, \ldots, L_h - 1$, are the coefficients of the time-invariant RIR between the speech source and the mth microphone, $s(n)$ is the clean speech signal, $x_m(n)$ is the reverberant speech*

*component, and $v_m(n)$ is the additive noise component. Since this paper aims to investigate the dereverberation performance of acoustic multichannel equalization techniques, in the following it is assumed that $v_m(n) = 0$, hence $y_m(n) = x_m(n)$."*

Accordingly, **Fig. 2 has been changed in the revised manuscript** such that it also includes the noise component.

In order to discuss other sources of RIR perturbations, **we have reformulated and extended the first paragraph of Section 2.2 in the revised manuscript** to:

*"Since the true RIRs are typically not available in practice, the reshaping filter is designed using the perturbed multichannel convolution matrix $\hat{\mathbf{H}}$ constructed from the available RIRs $\hat{\mathbf{h}}_m$. This matrix is equal to $\hat{\mathbf{H}} = \mathbf{H} + \mathbf{E}$, where $\mathbf{E}$ represents the convolution matrix of the RIR perturbations arising due to, e.g., temperature fluctuations [15], source-microphone geometry mismatches [16, 17], RIR estimation errors from blind and supervised system identification methods [18, 19], or microphone transfer function mismatches. It should be noted that microphone transfer function mismatches result in convolutive RIR perturbations instead of additive perturbations. However, the techniques discussed in the remainder of this paper are independent of the type of RIR perturbations present in the system, as long as a model is available to characterize these perturbations."*

# 2 Reviewer 2

1) *Please find a better formulation for "For all considered techniques, the conventionally used reshaping filter length is ...". Explain why you think this exact filter length is optimal.*

   **Rephrased in the revised manuscript** as:

   *"For the R-RMCLS, S-RMCLS, R-PMINT, and S-PMINT techniques the reshaping filter length is set to $L_w = \left\lceil \frac{L_h - 1}{M - 1} \right\rceil$, i.e., $L_w = 1200$ for the system $S_1$ and $L_w = 1627$ for the system $S_2$. As shown in [11], this filter length is the minimum length required for perfect dereverberation performance."*

2) *Some techniques perform almost equally well (thanks for the online listening samples!). In that case, one would choose the algorithm with lowest complexity. Can you add a few remarks about complexity of the various algorithms (MIPS, memory)?*

   **We have added the following paragraph in the revised manuscript** at the end of Section 2.3:

   *"**Computational complexity considerations:** The computational complexity of all considered methods is at most cubic[1], since matrix multiplications and matrix inversions account for the dominant operations in all reshaping filter computations, cf. (8), (9), (14)-(16). The complexity of using a shorter reshaping filter length is $O(n_r^3)$, where $n_r$ denotes the number of rows of the matrix $\mathbf{W}\hat{\mathbf{H}}$ when using $L_w < \left\lceil \frac{L_h - 1}{M - 1} \right\rceil$. The complexity of using regularization is $O(n_c^3)$, where $n_c$ denotes the number of columns of the matrix $\mathbf{W}\hat{\mathbf{H}}$ when $L_w = \left\lceil \frac{L_h - 1}{M - 1} \right\rceil$. Finally, the complexity of using a sparsity-promoting penalty function is $O(L_z^3)$, where $L_z$ denotes the length of the output signal vector. Since typically $n_r < n_c \ll L_z$, decreasing the reshaping filter length results in the lowest computational complexity, whereas incorporating a sparsity-promoting penalty function results in the highest computational complexity. In addition, the execution of the sparsity-promoting method takes a significantly longer time than the execution of the other methods due to the multiple number of iterations."*

# 3 Reviewer 3

1. *Page 2, column 1, line 25: "the objective is to relax the constraints on the reshaping filter design by suppressing only the late reflections": however, PMINT doesn't seem to relax the constraints compared to MINT? It changes the target equalized impulse response but the number of constraints stay the same.*

[1] This upper bound may be tightened when exploiting the fact that the matrices involved are symmetric or Toeplitz.

It is indeed true that the number of constraints is the same for MINT and PMINT and the difference between these techniques consists in the suppression or preservation of the early reflections. In order to avoid any confusion, **we have rephrased this sentence in the revised manuscript to**:

*"Since early reflections tend to improve speech intelligibility [21] and late reflections are the major cause of speech intelligibility degradation, the objective of these techniques is to suppress only the late reflections."*

2. *Eq (4): The vectors s and $x_m$ is indexed with n, but not the vectors h or w just before eq (2). This suggests that 'n' is a time index, and that the source signals are time-varying but the RIRs and equalizing filters are not. Yet, the scalars h and w in eq (1) are also indexed with n. Can this confusion be resolved?*

   We would like to thank the reviewer for pointing out this mistake. It is true that the source signal is time-varying whereas the RIRs and the filters are assumed to be time-invariant. **We have resolved this confusion in the revised manuscript** by explicitly expressing the convolution operation as a summation in (1) and (2).

3. *There is no background additive noise considered in the acoustic system setup in Eq (1). This is fine as the work here deals only with dereverberation, but can a comment be added about this?*

   In order to address this comment, the description of the acoustic system setup **is reformulated in the revised manuscript** to:

   *"The mth microphone signal $y_m(n)$, $m = 1, \ldots, M$, at time index n is given by*

$$y_m(n) = \underbrace{\sum_{l=0}^{L_h-1} h_m(l)s(n-l)}_{x_m(n)} + v_m(n) = x_m(n) + v_m(n), \tag{2}$$

   *where $h_m(l)$, $l = 0, \ldots, L_h - 1$, are the coefficients of the time-invariant RIR between the speech source and the mth microphone, $s(n)$ is the clean speech signal, $x_m(n)$ is the reverberant speech component, and $v_m(n)$ is the additive noise component. Since this paper aims to investigate the dereverberation performance of acoustic multichannel equalization techniques, in the following it is assumed that $v_m(n) = 0$, hence $y_m(n) = x_m(n)$."*

   Accordingly, **Fig. 2 has been changed in the revised manuscript** such that it also includes the noise component.

4. *Page 3, column 1, line 59: Do you have any insights into how to determine if a desired EIR is perceptually close to the true RIRs?*

   To the best of our knowledge, there is currently no reliable measure to evaluate the perceptual similarity of RIRs. Our informal listening tests suggest that, e.g., when the desired EIR exhibits an exponential decay determined by the reverberation time of the true RIRs, the coloration of the desired EIR is perceived as being similar to the coloration of the true RIRs. For this reason, the desired EIR in PMINT is selected as the first taps of one of the estimated RIRs. Although these first taps are not exactly the same as the first taps of the true RIRs due to estimation errors, they still exhibit the desired exponential decay.

5. *Section 4, bullet 4: It looks like R-PMINTs DRR is greater than S-PMINT except in S1-NPM1, where they are approximately equal.*

   **Fixed in the revised manuscript** by reformulating the addressed bullet point to:

   - *"The R-PMINT technique typically yields a higher DRR than the S-PMINT technique (except for the scenario $S_1$-$NPM_1$, where the R-PMINT and S-PMINT techniques yield a similar DRR)."*

6. *Section 5, page 8: What are the implications of two algorithms having results with no statistically significant difference? Does this mean that it does not matter which algorithm is used for dereverberation in practice?*

Indeed, when there are no statistically significant differences in the perceived speech quality, it does not matter which algorithm is used for dereverberation in practice.

7. *Section 5, page 9: There are some surprising results here. E.g. in Table 6, L-RMCLS only has two ticks vs R-RMCLS and R-PMINT, yet I would have expected more ticks based on the boxplots in Fig. 3, where L-RMCLS has a much smaller range and median compared to all other results there. On the other hand, L-RMCLS has all ticks in Table 8, despite it having a larger range in Fig. 3 for S2-NPM2 compared to S2-NPM1. Additionally, in Table 7 and again for L-RMCLS, I would have expected some crosses based its similar range and median in Fig. 3 to, e.g. L-PMINT. Can you please clarify this, and check the other results if necessary?*

   Although these results may seem surprising at first, it should be realized that due to the repeated measures study design (i.e., the same subjects are tested for all conditions), a repeated-measures analysis of variance (rm-ANOVA) is conducted. For such a study design the significant effect can be found in the "within-subject" variance, which is not depicted in the boxplots in Fig. 3. The total variance depicted in the boxplots in Fig. 3 is not an indicator of this "within-subjects" variance, and hence cannot be used to judge upon significant or non-significant differences.

## Minor comments:

1. *1. Page 1, column 2: ' ... enable to exploit both ...' → 'enable the exploitation of both'*

   In order to address the comment of Reviewer 4, **this sentence has been changed in the revised manuscript to**:

   *"In the last decades, many single as well as multichannel dereverberation techniques have been proposed [6], with multichannel techniques being generally preferred since they exploit both the spectro-temporal and the spatial characteristics of the received microphone signals."*

2. *Page 2, column 1, line 52: 'techniques is extensively' → 'are'*

   **Fixed in the revised manuscript.**

3. *Page 3, column 1, line 58: 'generality, also other desired EIRs could be used' → 'generality, other desired EIRs could also be used'*

   **Fixed in the revised manuscript.**

4. *Page 4, column 1, line 11: $\delta$ 'is' a regularization ...*

   **Fixed in the revised manuscript.**

5. *Page 4, column 1, line 31: for completion, please add how $\mathbf{X}$ is constructed.*

   **Added in the revised manuscript** in (12).

6. *Having three $L_*$ seems to be complicating things unnecessarily. My suggestion is that you can just continue using*

   a) *$L_w$ instead of introducing $L_t$ (page 5, col 1, line 28)*

   b) *$L_w$ instead of introducing $L_s$ in eq (17) and replacing $L_t$ with $\lceil \frac{L_h-1}{M-1} \rceil$ explicitly.*

   *Since $L_w$ is understood to be a variable, I don't believe this introduces ambiguity or confusion in having different actual values used for different algorithms and scenarios.*

   **Fixed in the revised manuscript.**

7. *Page 5, column 1, line 46-51: the latter two of 'considered' could be dropped instead of repeating it 3 times.*

   **Fixed in the revised manuscript.**

8. *Page 8, column 2, line 62: 'Furthermore, also the auditory...' → 'Furthermore, the auditory'*

   **Fixed in the revised manuscript.**

# 4 Reviewer 4

- *Abstract should be self-contained. So don't define an acronym in the abstract unless that acronym is then used in the abstract.*

  **Fixed in the revised manuscript** by removing the acronym RIR.

- *Make it clear in the abstract that this only concerns the case of a single source with many microphones. That is not made clear and so this reviewer assumed it was many sources, many microphones.*

  **Fixed in the revised manuscript** by rephrasing the third sentence in the abstract to:

  *"This paper focuses on evaluating the performance of these methods for single-source multi-microphone scenarios, both using instrumental performance measures as well as using subjective listening tests."*

- *Replace all occurences of multi-channel with multichannel. Only hyphenate if there is genuine confusion without it, or if the compound word or combining form is not accepted without it.*

  **Fixed in the revised manuscript.**

- *Avoid repetition. There are at least 5 occurences of 'it has been proposed to', the first three of which occur in immediate succession. On some of these, you could replace 'In [X] it has been proposed to (do something)' with just '[X] proposed to do something' or '[X] did something'.*

  We believe that the current formulation makes it easier to follow the paper. Since we prefer the current formulation, **we have not made any changes in the revised manuscript** and leave it to the editorial office to make changes if necessary.

- *"In the last decades, many single- as well as multi-channel dereverberation techniques have been proposed [6], with" - Either remove this completely, or find a better reference than [6], which is almost 10 years old and doesn't actually review the many single and multichannel techniques. If you remove it, then the beginning of this paragraph should be, 'multi-channel dereverberation techniques are generally preferred"*

  **Fixed in the revised manuscript** by changing the reference to "P. A. Naylor and N. D. Gaubitch, Eds., Speech dereverberation. London, UK: Springer, 2010", which provides an extensive overview of state-of-the-art single and multichannel dereverberation techniques.

- *Change 'they enable to exploit' to either 'they exploit' or 'they enable one to exploit'.*

  **Fixed in the revised manuscript** using "'they exploit".

- *"Acoustic multi-channel equalization ... aim ... can ... comprising ..." is actually three sentences. Break it into at least two. So: "Acoustic multi-channel equalization techniques aim to reshape the available room impulse responses (RIRs) between the speaker and the microphone array. They can in theory achieve perfect dereverberation performance [11], and hence comprise an attractive approach to speech dereverberation."*

  **Fixed in the revised manuscript.**

- *Also, the 'perfect dereverberation' claim is far too strong, but I guess it is acceptable here since the claim is made by the reference [11], not directly by the authors.*

  The "perfect dereverberation" claim has been theoretically proven in [11] under the assumptions of perfect knowledge of the RIRs and no common zeros between the RIRs. It is indeed true that these assumptions almost never hold in practice. For this reason we write *"They can in theory achieve perfect dereverberation performance"*, which is the correct statement.

- *"Since the available [13, 14]." Is another run-on sentence. Break it in two, with text such as 'MINT fails to invert the true RIRs since the available RIRs typically differ from the true RIRs due to, e.g., temperature or position variations [15, 16] or due to the sensitivity of blind and supervised system identification methods to near-common zeros or background noise [13, 17-20]. This may lead to perceptually severe distortions in the output signal [13, 14].'*

  **Fixed in the revised manuscript** by breaking the sentence into:

*"Since the available RIRs typically differ from the true RIRs due to, e.g., temperature or position variations [15-17] or due to the sensitivity of blind and supervised system identification methods to near-common zeros or background noise [13, 18-21], MINT fails to invert the true RIRs. This may lead to perceptually severe distortions in the output signal [13,14]."*

- *Line 42 of page 2 mentions instrumental performance measures for the first time. The reader doesn't yet know what this is, and it should be explained when first mentioned. It is also confusing since a lot of dereverberation is applied to musical signals, where performance can depend on (musical) instrument. For instrumental and perceptual, do you mean objective and subjective?*

  The words instrumental and perceptual are indeed used to refer to objective and subjective results. "Instrumental measures" is a well-established term used to describe objective performance measures in the context of speech enhancement techniques, e.g., see Hendriks et. al[2]. Furthermore, this manuscript clearly deals with speech signals (e.g., the first sentence of the introduction starts with "Speech signals..."), hence, we believe that there is no confusion of the term "instrumental performance measures" to "musical instruments". The term "perceptual results" is also commonly used to refer to subjective results, since the term "subjective" is widely criticized within the quality assessment research community. Therefore **we have not implemented any changes** in the revised manuscript regarding this issue.

- *Line 52, page 2 - remove 'extensively' from 'extensively compared'*

  **Fixed in the revised manuscript.**

- *Section 2.1 - m-th should be written as mth. Similarly, i-th should be written as ith after Eq. 13.*

  **Fixed in the revised manuscript.**

- *Are there any assumptions about placement of the microphones? Is there prior knowledge about source and microphone locations? Does one assume that the source remains fixed? Is there a minimum number of microphones? The problem formulation, including assumptions should be made clear here. And these questions have a large bearing on the ability to manipulate the signals and solve multichannel signal processing problems, see for instance; L. Wang, et al, 'Self-Localization of Ad-hoc Arrays Using Time Difference of Arrivals,' IEEE Transactions on Signal Processing, 64 (4), Feb., 2016. T-K. Hon, et al, 'Fine landmark-based synchronization of ad-hoc microphone arrays,' 23rd European Signal Processing Conference (EUSIPCO), p. 1341-1345, Nice, France, 2015.*

  Acoustic multichannel equalization techniques do not rely on assumptions or prior knowledge about the source-microphone geometry, hence, the suggested references are not relevant. Furthermore, there is no minimum number of microphone required (having more than one microphone is enough for theoretical perfect dereverberation performance). The assumptions made in acoustic multichannel equalization techniques are:

  1) The RIRs are time-invariant.
  2) Measurements or estimates of the RIRs are provided.

  In order to clarify 1), **we have added the following sentence in the revised manuscript** after (1):

  *"... where $h_m(l)$, $l = 0, \ldots, L_h - 1$, are the coefficients of the time-invariant RIR between the speech source and the mth microphone ..."*

  In order to clarify 2), **we have added the following sentence in the revised manuscript** at the beginning of Section 2.2:

  *"Acoustic multichannel equalization techniques assume that measurements or estimates of the RIRs are available. Such techniques aim at speech dereverberation by designing a reshaping filter* $\mathbf{w}$ *such that the (weighted) EIR in (3) is equal to a (weighted) dereverberated target EIR."*

[2] R. C. Hendriks, T. Gerkmann, and J. Jensen, DFT-domain based single microphone noise reduction for speech enhancement - a survey of the state of the art, Synthesis Lectures on Speech and Audio Processing, Morgan & Claypool Publishers, Jan. 2013.

- *Section 2.3 - change 'In [22] it has been shown' to 'In [22] it was shown' or '[22] showed'. Similarly change 'In [14] it has been proposed' to '[14] proposed'. There may be other occurrences of this which should be changed too.*

  **Fixed in the revised manuscript.**

- *Remove the footnote before Eq. 11 and just write 'the following update rules [23, 24] until ...'*

  **Fixed in the revised manuscript.**

- *Section 3, 2nd paragraph - change 'We have considered' to 'We considered'. Similarly, for 'have been measured'. You are describing what was done and completed within a finite duration (past tense), not what was done up to the current time or without duration (present perfect or present perfect continuous).*

  **Fixed in the revised manuscript.**

- *From Eq.s 1 to 16, a very large amount of notation was introduced, to the point where it became quite difficult to follow, especially when comparing the proposed improvements of section 2.3 against each other, and in seeing how they relate to the problem formulation of section 2.1. I encourage the authors to make every effort to make the maths clear and to avoid unnecessary additional notations.*

  In order to address this comment, **we have reformulated Section 2.3** in the revised manuscript by adding (11) and (12) and by providing a step-by-step definition of all new variables. We believe the readability of this section has significantly improved. However, although we would like the manuscript to be self-contained, it is simply not possible to thoroughly describe 6 different techniques since we would rather like to concentrate on the instrumental and perceptual results. Hence, the interested reader is referred to the appropriate references where these techniques have been proposed and described in full extent.

- *Section 5- 'multi-stimulus' or 'multistimulus', not 'multi stimulus'*

  The method is called "Multi stimulus test with hidden reference and anchor" in the specifications in [36]. Therefore **we have not implemented any changes** in the revised manuscript regarding this issue.

- *Fig. 3 - It looks like the hidden reference was always rated 100, for all 21 participants in all scenarios. Really? That seems too perfect.*

  It is indeed true that the hidden reference was always rated 100. This is due to the fact that in the MUSHRA specifications [38] it is required that at least one of the signals is rated 100 and the subjects have access to the reference signal and can compare to it. According to the MUSHRA specifications, the hidden reference is included as a "sanity check", such that if it is rated lower than 90, the subject needs to be excluded from the aggregated responses.

- *Section 6 - I quite like the fact that the authors looked into this. However, I think they should look at and cite the work of E. Vincent and co-authors: E. Vincent, Improved perceptual metrics for the evaluation of audio source separation. 10th Int. Conf. on Latent Variable Analysis and Signal Separation (LVA/ICA), Mar 2012, Tel Aviv, Israel. pp.430-437, 2012. Emiya, V. et al, Subjective and objective quality assessment of audio source separation. IEEE Trans. Audio Speech Lang. Process. 19(7), 2046-2057 (Sep 2011) (there may be better references from them). They compared subjective measures of source separation quality with existing objective measures, found them lacking, and proposed new measures that should show better agreement with perception and preference. Im not sure that their new measures achieve this, but the concept and approach is a very good one, and closely related to the work in this submission for dereverberation measures.*

  The measures proposed by Vincent and Emiya were developed for the evaluation of audio source separation, and to the best of our knowledge, have never been validated for dereverberation performance evaluation. While the development of appropriate measures for evaluating dereverberation algorithms is currently an active research field, see e.g., [2], [3], [34], [36], it is not the objective of this manuscript. Since the objective of Section 6 is to analyze the correlation between subjective results and commonly used objective dereverberation measures, we only included some measures that were developed or validated for dereverberation performance evaluation, i.e., DRR, SRMR,

LLR, and PESQ. As mentioned in the manuscript, the channel-based DRR measure was shown to correlate well with the perceived amount of reverberation for unprocessed signals [35], while the signal-based SRMR, LLR, and PESQ measures were shown to correlate well with the perceived overall quality of signals processed by speech enhancement algorithms for dereverberation and noise reduction [36]. While the work suggested by the reviewer is definitely interesting and worth looking into when designing new instrumental performance measures for speech dereverberation, this is beyond the scope of this manuscript. Hence, **we have not implemented any changes** in the revised manuscript regarding this issue.

- *Section 7. This conclusion would benefit from some discussion and critique. At present, it is mainly a summary. But what are the limitations of the work? Where does the approach fail or perform worse than expected? What was surprising or unusual? What was left unexplored? What new directions of research did it open up? Perhaps the Conclusion can give some ideas to researchers who might build on and cite this work.*

  As the reviewer points out, we intend for the conclusion section to be a summary, highlighting the most important conclusions and we believe that it should not include any additional information. A discussion of the limitations, expected results, or unexpected results is provided in Sections 4 and 5 when relevant, e.g.,

  - Page 5, Column 2: *"It should be noted that the computation of the PESQ score for selecting the optimal parameters is an intrusive procedure that is not applicable in practice, since knowledge of the true RIRs is required in order to compute the reference signal and the resulting EIR."*
  - Page 6, Column 1: *"The robust extensions of the RMCLS technique generally yield a similar or higher DRR than the robust extensions of the PMINT technique. This is to be expected since the robust extensions of the RMCLS technique relax the constraints on the filter design and aim only at suppressing the late reverberation, whereas the robust extensions of the PMINT technique also aim at preserving the perceptual speech quality (which is not reflected by the DRR measure)."*
  - Page 6, Column 1: *"The L-RMCLS and L-PMINT techniques yield the lowest DRR out of all considered robust extensions. This is not surprising since these techniques simply use a shorter reshaping filter length, without explicitly taking into account the structure of the RIR perturbations or the characteristics of the output speech signal."*
  - Page 6, Column 2: *"The robust extensions of the PMINT technique generally yield a similar or better SRMR and LLR than the robust extensions of the RMCLS technique. Surprisingly, the robust extensions of the RMCLS technique generally yield a similar or better PESQ score than the robust extensions of the PMINT technique, implying that PESQ does not appear to reflect the better preservation of the early reflections achieved by the robust extensions of the PMINT technique but puts more emphasis instead on the better reverberant tail suppression achieved by the robust extensions of the RMCLS technique."*

  In addition, in the last sentence of the conclusion section we highlight an important new research direction, i.e.:

  *"Furthermore, the provided correlation analysis highlights the need to develop more accurate instrumental performance measures, reliably reflecting the distortions introduced by acoustic multichannel equalization techniques."*

- *Also, though the paper is clearly about speech dereverb, I don't think the problem formulation and multichannel equalization approaches require that the signal must be speech. And I think only PESQ and SRMR are speech-specific measures. Even then, they can still be valid with non-speech signals. So given that dereverberation and multichannel equalization are useful for nonspeech dereverberation, it would be useful for the authors to weaken the assumption that signals are only speech, such as in 'with s(n) the clean speech signal', and stress in the conclusion and elsewhere the relevance of the work for advances in the general problem of (speech and nonspeech) dereverberation.*

  It is indeed true that the problem formulation does not require the signal to be speech and the reviewer brings up an interesting avenue for further research. However, we have only validated

multichannel equalization techniques for speech signals and have no experience on the performance and perceptual effects these techniques have on music signals. Since we would not like to make any statements that we have not tested, **we have not implemented any changes** in the revised manuscript regarding this issue.