PERCEPTUAL AND INSTRUMENTAL EVALUATION OF THE PERCEIVED LEVEL OF REVERBERATION

Benjamin Cauchi[1,4], Hamza Javed[2], Timo Gerkmann[3,4], Simon Doclo[1,3,4], Stefan Goetze[1,4], Patrick Naylor[2]

[1]Fraunhofer IDMT, Project Group Hearing, Speech and Audio Technology, Oldenburg, Germany
[2]Imperial College London, Dept. of Electrical Engineering, London, United Kingdom
[3]University of Oldenburg, Dept. of Medical Physics and Acoustics, Oldenburg, Germany
[4]Cluster of Excellence Hearing4all, Oldenburg, Germany

ABSTRACT

Perceptual measures are usually considered more reliable than instrumental measures for evaluating the perceived level of reverberation. However, such measures are costly in both time and money, and, due to variations in stimuli or assessors, the resulting data is not always statistically significant. Therefore, an efficient perceptual measure of the perceived level of reverberation is needed. We compare the use of a multiple stimuli test with the use of pairwise comparison for the evaluation of the perceived level of reverberation. The results suggest that using multiple stimuli is preferable to pairwise comparison as long as the number of conditions to be compared is not too large. Additionally, we use the results from the conducted perceptual measurements to examine the reliability of existing instrumental measures of the perceived level of reverberation. Our observations show which instrumental measures are effective in highlighting differences between RIR characteristics and which ones have to be preferred if one aims at predicting the level of reverberation perceived by a human assessor.

Index Terms— Reverberation, perceptual evaluation, instrumental measures

1. INTRODUCTION

In many speech communication applications, the speech signal uttered by a user is recorded by a distant microphone. In an enclosed space, the recorded signal is thus typically corrupted by both noise and reverberation. The latter can be characterised by the room impulse response (RIR) between the speech source and the microphone [1]. Though early reflections can be beneficial to the speech quality and intelligibility [2], large levels of reverberation are usually detrimental in speech communication applications [3]. Therefore, efforts have been made to design speech enhancement systems able to suppress the reverberation [1] and common evaluation databases have been published in recent challenges [4, 5]. However, the evaluation of the perceived level of reverberation, needed in order to evaluate speech enhancement systems, remains a challenge.

The performance of speech enhancement systems are typically assessed using either instrumental measures, i.e. metrics computed from the recorded signal, or perceptual measurements, i.e. obtained from human assessors. Instrumental measures are convenient to use but many, such as the Bark spectral distortion (BSD) [6], the perceptual objective listening quality assessment (POLQA) [7] or the perceptual evaluation of speech quality (PESQ) [8], are often only correlated with overall speech quality. It has been proposed in [9] to predict the perceived level of late reverberation from the input signal and the RIR. Unfortunately, the RIR is often unavailable, limiting the applicability of this measure. Some instrumental measures, such as the speech to reverberation modulation energy ratio (SRMR) [10, 11], or the reverberation decay tail ($\mathrm{R_{DT}}$) [12, 13], have been designed specifically to evaluate the perceived level of reverberation and their computation does not require knowledge of the RIR. However, such instrumental measures might correlate only poorly with perceptual measurements of the perceived level of reverberation [14], which are generally considered more reliable. Therefore, perceptual measurements are needed for both the evaluation of reverberation suppression systems and the development of reliable instrumental measures.

Unfortunately, perceptual measurements also have some drawbacks. They are expensive and time consuming [15], often restricting researchers to use available instrumental counterparts. Additionally, the most popular standards for the perceptual measurement of dereverberation systems have been developed for different applications and might not always be the best methods to assess reverberation. The 5-point mean opinion score (MOS) [16] used, e.g., for the evaluation of single channel speech dereverberation systems in [3], has originally been developed for the evaluation of noise suppression algorithms. The multi stimuli test with hidden reference and anchor (MUSHRA) [17] was originally developed for the evaluation of audio codecs although it has also been used for the evaluation of single-channel and multichannel speech dereverberation systems, e.g., in [18, 19].

Some perceptual measures, such as MOS or MUSHRA, rely on the simultaneous comparison of several conditions. As different reverberant conditions can be perceived as strongly similar by some assessors, this simultaneous comparison can be difficult and may result in the lack of statistical significance sometimes obtained in such tests [19]. In this paper, we propose to use pairwise comparison for the evaluation of the perceived level of reverberation (PCEPLR), aiming to facilitate the task of the assessors by asking them to compare only two audio signals at a time, similarly as applied to, e.g., the comparison of microphones for the recording of a singing voice [20]. Additionally, we propose modifications to the standard MUSHRA, mostly in the design of the anchor and reference signals, aimed at making this test more efficient in the evaluation of the perceived level of reverberation. This measure will be denoted by multi stimulus test with hidden reference and anchor for reverberant speech

The research leading to these results has received funding from the EU Seventh Framework Programme project DREAMS under grant agreement ITN-GA-2012-316969.

(MUSHRAR) in the remainder of this paper in order to avoid confusion with the standard described in [17].

This paper has two objectives. First, it aims at identifying a suitable perceptual measure for the level of perceived reverberation. We propose and compare two perceptual measurement schemes which are known from the fields of assessment of noise reduction algorithms and audio coding and are adapted. We will refer to these newly adapted scheme as MUSHRAR and PCEPLR and describe them in Section 2. Additionally, this paper examines the relation of the obtained perceptual scores with RIR characteristics and instrumental measures, summarised in Section 3, to compare their reliability. The results, presented in Section 5, provide insight into the choice of perceptual measurements to assess the perceived level of reverberation and on the reliability of the considered instrumental measures.

2. PERCEPTUAL MEASUREMENT SCHEMES

Two perceptual measurement schemes are considered in this paper. During these tests, a total of $N \times M$ audio samples are graded by by each assessor, with N and M denoting the number of anechoic speech utterances and the number of reverberation conditions, respectively. Each grading session begins with a training phase, in which the assessor listens to all files to be graded during the test in order to become familiar with the presented speech material and to adjust the sound volume in the audiological calibrated headphones to a comfortable level. The two testing schemes, which are adopted from known methods are described in the following subsections.

2.1. MUSHRAR

MUSHRA, as described in [17], consists of grading an attribute of audio files on a continuous scale using sliders, as described in [21]. Each assessor is presented with M sliders N times, in order to compare all conditions under test with each other. MUSHRA has been designed for the evaluation of audio codecs and previous studies suggest that some of its elements may not be appropriate for the evaluation of the perceived level of reverberation [19]. Therefore, in this paper, we propose MUSHRAR, which differs from MUSHRA in the following ways.

In our proposed MUSHRAR scheme, the sliders are labeled from "not reverberant" (scoring 0) to "very reverberant" (scoring 100), which is the opposite of what has been used, e.g., in [19], but has been chosen because attributing a higher score to a higher perceived level of reverberation is more intuitive for the assessor. Consequently, the assessors are asked to give a score of 0 to at least one of the samples, which, if identified correctly, should be the hidden reference. In our test, the reference consists of anechoic speech convolved with a RIR having high direct-to-reverberant ratio (DRR) and low T_{60} (cf. Section 4), instead of a clean or anechoic speech signal, in order to make it sound more natural to the assessor. Additionally, the anchor consists of speech convolved with a RIR having low DRR and high T_{60} (cf. Section 4), instead of a low-passed (at 3.5 kHz) speech signal, which is prescribed in the standard MUSHRA. The reason is that, while a 3.5 kHz low pass may be reasonable as an anchor in speech coding, this filter may actually reduce the level of perceived reverberation and is therefore unsuited for our application.

2.2. PCEPLR

During the PCEPLR testing phase, each assessor is presented with two audio samples, labeled "A" and "B". The assessor is asked to listen to both and to indicate if the perceived level of reverberation is larger in "A" or in "B", or if both are perceived as equally reverberant. This selection is made by ticking checkboxes as described in [20]. Each combination of utterance and condition is compared to itself once, no comparison is made between different utterances and no comparison is repeated. This leads to a total of

$$P = N \times \left(\frac{M!}{2(M-2)!} + M \right) \tag{1}$$

pairs to be evaluated by each assessor. In this paper, n denotes the sample index and $y_{i,j}(n)$ denotes the signal consisting of the i-th utterance in the j-th condition. The pairwise score attributed to $y_{i,j}(n)$ by one particular assessor is computed similarly as in [20] and normalized to range between 0 and 100 for easier comparison with MUSHRAR.

3. INSTRUMENTAL MEASURES

The scores resulting from the perceptual measurements are compared with existing instrumental measures of the level of reverberation. The considered measures are the SRMR [10] and its extention $\mathrm{SRMR}_{\mathrm{x}}$ [11] as well as the R_{DT} [12] and its extension $\mathrm{R}_{\mathrm{DT}}\mathrm{x}$ [13]. Additionally, we consider the PESQ score [8], which, though originally developed for the evaluation of audio codecs, has been used in numerous contributions to evaluate the performances of speech dereverberation algorithms, e.g., [18, 19, 22, 23]. It can be noted that the SRMR and the $\mathrm{SRMR}_{\mathrm{x}}$ are non-intrusive measures, whilst the computation of R_{DT}, $\mathrm{R}_{\mathrm{DT}}\mathrm{x}$ and PESQ requires a reference signal.

The computation of the SRMR relies on decomposing the input signal using gamma-tones filter banks from which the temporal envelope is extracted. The SRMR value is computed as the ratio between the modulation energy in the higher and lower frequencies [24]. However, it has been observed that the SRMR is highly correlated with both pitch and speech content, limiting its value as a measure of the perceived level of reverberation. In order to avoid these detrimental effects, the SRMR has been extended in [11]. This extension, denoted by $\mathrm{SRMR}_{\mathrm{x}}$ in this paper, limits the influence of pitch and speech content by using a narrower range of modulation frequency and by introducing upper and lower energy bound values in each band.

The R_{DT} has been proposed in [12] to measure the tail effect of reverberation by jointly characterising the relative energy in the tail of the RIR and the rate of its decay. The computation of the R_{DT} consists of, first, representing both the signal to be evaluated and the corresponding reference signal in the Bark domain. Time periods which would be most affected by reverberation are identified from the Bark representation of the reference signal. Once these periods have been identified, the R_{DT} is then computed as the ratio between the reverberant energy, relative to the direct path energy, and the reverberation rate of decay. The $\mathrm{R}_{\mathrm{DT}}\mathrm{x}$ [13] differs from the R_{DT} most notably in the computation of the Bark spectrum and in the detection of the time periods affected by reverberation.

4. EXPERIMENTAL SETUP

The audio signals presented to the assessors have been generated by convolving anechoic speech with recorded RIRs and adding ambient noise, i.e.

$$y_{i,j}(n) = \left(\sum_{\ell=0}^{L_h-1} s_i(n-\ell) h_j(\ell) \right) + v_j(n), \tag{2}$$

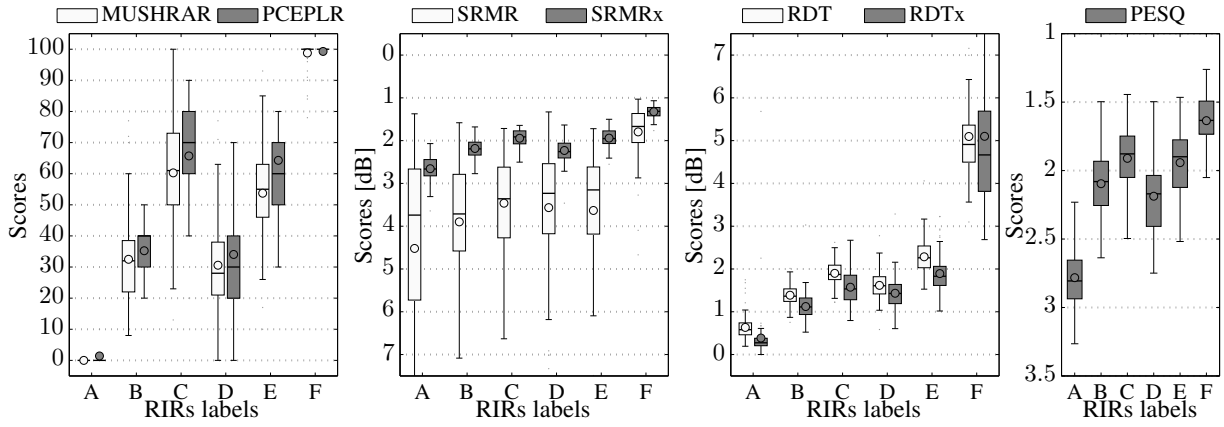

Fig. 1. Scores obtained using the considered perceptual and instrumental measures.

RIR labels	A	B	C	D	E	F
T_{60} [s]	0.18	0.38	0.44	0.62	0.66	1.29
DRR [dB]	11.31	8.38	0.94	12.19	5.09	4.95

Table 1. RIR characteristics

with $s_i(n)$, $v_j(n)$, $y_{i,j}(n)$ and $h_j(n)$ denoting the anechoic speech, the noise, the generated signal and the recorded RIR of length L_h, respectively. All signals are sampled at a sampling frequency of 16 kHz and the stimuli have durations ranging from 14 s to 18 s.

For all assessors, the same $M = 6$ RIRs have been used, $j \in \{1 \cdots 6\}$, as summarised in Table 1. These RIRs are taken from either the SMARD corpus [25] ($j = 1$), or from the ACE challenge corpus [5] ($j > 1$). These RIR databases have been chosen for both their realistic range of RIRs characteristics, and for the fact that each recorded RIR is accompanied by a recording of ambient noise, recorded in the same room and at the same microphone position, from which $v_j(n)$ is extracted. All signals to be assessed have a signal to noise ratio (SNR) of 20 dB, computed using the early reverberant speech, considering the first 50 ms of the RIR, as target signal and computing the speech energy according to [26]. For each assessor, $N = 3$ utterances have been randomly selected from the TIMIT [27] database of anechoic speech $i \in \{1 \cdots 3\}$. The energy of the direct speech has been normalised to be equal over all conditions for each speech utterance.

A total of 28 self-reported normal hearing assessors participated in the perceptual measurements. All assessors conducted the test in a soundproof booth and listened to the diotic signals reproduced using an audio interface (RME: Fireface 800) and closed-back headphones (Senheiser: HDA 300). Each assessor started the grading session with the training phase before evaluating the files using both MUSHRAR and PCEPLR. The order in which the tests appeared was randomised for each assessor in order to limit potential biases and training effects.

5. RESULTS

The scores obtained using both proposed perceptual measures as well as all considered instrumental measures are depicted in Fig. 1, for all tested signals. Considering the perceptual scores, it can be seen in the left panel of Fig. 1 that both anchor and reference have been identified correctly by the assessors. This was not the case in some other studies using MUSHRA, e.g. [19], suggesting that the difference in designing anchor and reference signals are beneficial and that the assessors were reliable. However, it appears that a large interquartile range, i.e. the spread, can be observed in both MUSHRAR and PCEPLR, for RIRs B to E. The fact that the perceptual scores obtained on RIRs B and D as well as C and E show close medians illustrates that the perceived level of reverberation is more largely influence by the DRR than by the T_{60}. PCEPLR yields a lower spread than MUSHRAR in some conditions, e.g. RIR B, but larger in others, e.g. RIR D.

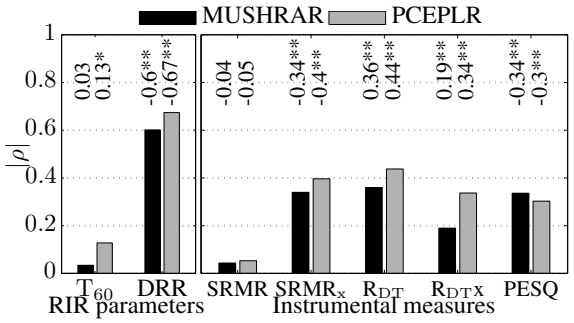

Fig. 2. Pearson correlation coefficients, over all assessed samples, between the perceptual measures and the considered RIRs characteristics and instrumental measures,. Stars and double stars indicate statistical significance at $p < 0.05$ and $p < 0.01$, respectively.

The results show that the scores obtained using the SRMR show the largest spread of all the considered instrumental measures. However, this spread is greatly reduced in the case of SRMR_x, suggesting that the extension of the SRMR measure is beneficial, confirming the results from [11]. On the other hand, R_DT and R_DTx show similar behaviour, except at the largest considered T_{60}, i.e. RIR F, for which the scores obtained using R_DTx show a larger spread than the ones obtained using R_DT.

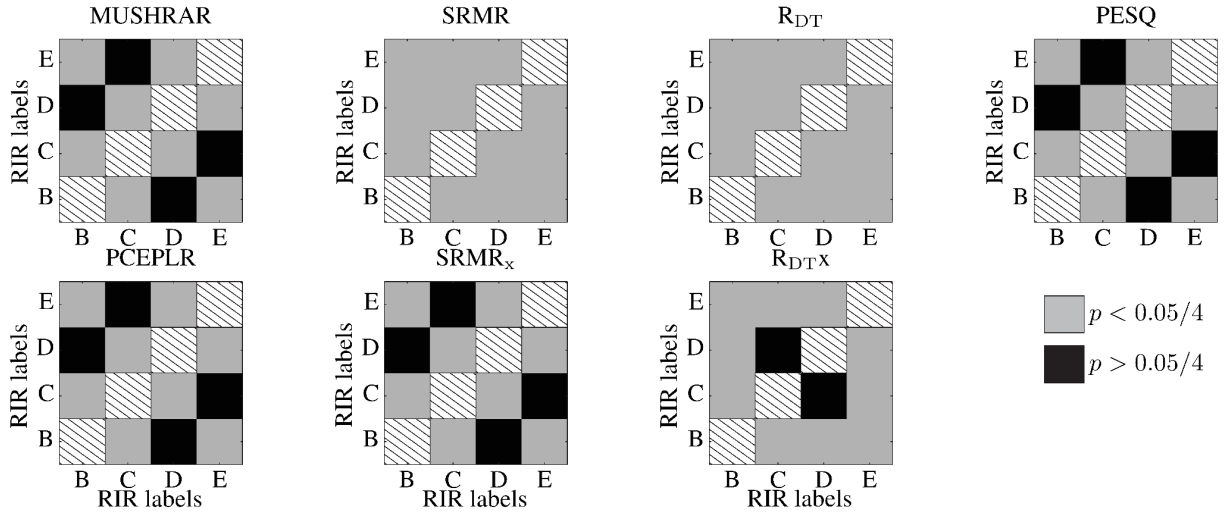

Fig. 3. Results of the Wilcoxon rank-sum test for all considered measures. Dashed areas represent unconsidered self-comparisons. The statistical significance criterion has been set at $p < 0.05/4$ to account for Bonferroni correction. Grey and black areas represent significant and non significant differences between two RIRs, respectively.

5.1. Statistical analysis

This subsection provides further statisitical analysis of the obtained results, excluding the scores of the reference and anchor, i.e, only RIRs B to E are taken into account.

The Pearson correlation coefficients between the perceptual measures and the considered RIRs characteristics and instrumental measures and over all assessed signals are depicted in Fig. 2. It can be observed, in the left panel of Fig. 2, that the perceived level of reverberation measured with either MUSHRAR or PCEPLR is more correlated with the DRR ($|p|$ up to 0.67) than with the T_{60} ($|p|$ up to 0.13). This reinforces the conclusion that the DRR has a larger influence than the T_{60} on the perceived level of reverberation and that these RIR characteristics should be considered in combination to predict the perceived level of reverberation. Among the considered instrumental measures, $\mathrm{SRMR_x}$ and $\mathrm{R_{DT}}$ show the highest correlations with the perceptual scores with $|p|$ up to 0.40 and 0.44, respectively. This illustrates the advantage of instrumental measures specifically designed for the evaluation of the perceived level of reverberation, contrary to PESQ which yields a correlation of $|p|$ up to 0.34. However, as instrumental measures are often used to identify differences between algorithms, further analysis is needed to generalized these results to the evaluation of processed signals.

The Friedman scores [28] for each measure are depicted in Table 2 and show that all considered measures reveal at least one significant difference between all considered conditions. However, in order to examine the pairwise differences between conditions, the results of a Wilcoxon rank sum test [29] are depicted in Fig. 3. Considering the perceptual measures, both MUSHRAR and PCEPLR show significant differences for all pairs of conditions except between B and D and between C and E. The fact that these two perceptual measures achieve similar results suggests that MUSHRAR is preferable for most studies as it requires less time than PCEPLR from each assessor. However, as the number of conditions being compared is limited in MUSHRAR, it is recommended in [17] to not exceed 12 conditions, PCEPLR should be prefered in studies comparing a large number of reverberant conditions.

	MUSHRAR	PCEPLR	SRMR	$\mathrm{SRMR_x}$	$\mathrm{R_{DT}}$	$\mathrm{R_{DT}x}$	PESQ
p	< 0.05	< 0.05	< 0.05	< 0.05	< 0.05	< 0.05	< 0.05
χ^2	149.7	176.4	13.2	123.8	188.9	136.4	115.6

Table 2. Results of the Friedmans test for all considered perceptual and instrumental measures. The value $p < 0.05$ indicates the significance of the results and χ^2 denotes the Friedmans chi square statistic.

Concerning the instrumental measures, our results indicate that SRMR and $\mathrm{R_{DT}}$ show significant differences between all pairs of conditions, suggesting that these instrumental measures are effective in highlighting differences between reverberant conditions but not in predicting the level of reverberation perceived by a human assessor. On the other hand, both $\mathrm{SRMR_x}$ and PESQ exhibit the same behaviour as MUSHRAR and PCEPLR, suggesting that they are more able to predict the level of reverberation that would be perceived by a human assessor.

6. CONCLUSION

We have proposed and compared the use of MUSHRAR and of PCEPLR for the evalaution of the perceived level of reverberation. Our analysis showed that both measures yield similar results and that the proposed MUSHRAR should be prefered to PCEPLR as long as the number of conditions to be compared is small enough, preferably less than 12. The relation between the perceptual scores and RIR characteristics showed that the DRR has a larger influence than the T_{60} on the perceived level of reverberation but that neither the DRR nor the T_{60} alone is sufficient to predict the perceived level of reverberation. The considered instrumental measures showed that some are suitable to highlight differences in reverberation conditions, e.g. $\mathrm{R_{DT}}$, but that others are more suitable for the prediction of the perceived level of reverberation, e.g. $\mathrm{SRMR_x}$ and PESQ.

7. REFERENCES

[1] P. A. Naylor and N. D. Gaubitch, Eds., *Speech Dereverberation*, Springer, 2010.

[2] J. S. Bradley, H. Sato, and M. Picard, "On the importance of early reflections for speech in rooms," *J. Acoust. Soc. Am.*, vol. 113, no. 6, pp. 3233–3244, June 2003.

[3] A. Warzybok, I. Kodrasi, J.O. Jungmann, E.A.P. Habets, T. Gerkmann, A. Mertins, S. Doclo, B. Kollmeier, and S. Goetze, "Subjective speech quality and speech intelligibility evaluation of single-channel dereverberation algorithms," in *Proc. Intl. Workshop Acoust. Signal Enhancement (IWAENC)*, Sept 2014, pp. 333–337.

[4] K. Kinoshita, M. Delcroix, T. Yoshioka, T. Nakatani, A. Sehr, W. Kellermann, and R. Maas, "The reverb challenge: Acommon evaluation framework for dereverberation and recognition of reverberant speech," in *Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, New Paltz, NY, USA, oct 2013, pp. 1–4.

[5] J. Eaton, N. D. Gaubitch, A. H. Moore, and P. A. Naylor, "The ACE challenge - corpus description and performance evaluation," in *Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, New Paltz, NY, USA, 2015.

[6] S. Wang, A. Sekey, and A. Gersho, "An objective measure for predicting subjective quality of speech coders," *IEEE J. Sel. Areas Commun.*, vol. 10, no. 5, pp. 819–829, June 1992.

[7] ITU-T, "Perceptual objective listening quality assessment: An advanced objective perceptual method for end-to-end listening speech quality evaluation of fixed, mobile, and IP-based networks and speech codecs covering narrowband, wideband, and super-wideband signals," Standard P.863, International Telecommunications Union (ITU-T), Jan. 2011.

[8] 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," Recommendation P.862, International Telecommunications Union (ITU-T), Feb. 2001.

[9] C. Uhle, J. Paulus, and J. Herre, "Predicting the perceived level of late reverberation using computational models of loudness," in *Proc. IEEE Intl. Conf. Digital Signal Processing (DSP)*, Corfu, Greece, 2011, pp. 1–7.

[10] T. H. Falk, C. Zheng, and W.-Y. Chan, "A non-intrusive quality and intelligibility measure of reverberant and dereverberated speech," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 18, no. 7, pp. 1766–1774, Sept. 2010.

[11] J. F. Santos, M. Senoussaoui, and T. H. Falk, "An improved non-intrusive intelligibility metric for noisy and reverberant speech," in *Proc. Intl. Workshop Acoust. Signal Enhancement (IWAENC)*, Sept. 2014, pp. 55–59.

[12] J. Y. C. Wen and P.A. Naylor, "An evaluation measure for reverberant speech using tail decay modelling," in *Proc. European Signal Processing Conf. (EUSIPCO)*, Florence, Italy, sep 2006, pp. 1–4.

[13] H. Javed and P. A. Naylor, "An extended reverberation decay tail metric as a measure of perceived late reverberation," in *Proc. European Signal Processing Conf. (EUSIPCO)*, Nice, France, sep 2015, pp. 1–4.

[14] S. Goetze, A. Warzybok, I. Kodrasi, J.O. Jungmann, B. Cauchi, J. Rennies, E.A.P. Habets, A. Mertins, T. Gerkmann, S. Doclo, and B. Kollmeier, "A study on speech quality and speech intelligibility measures for quality assessment of single-channel dereverberation algorithms," in *Proc. Intl. Workshop Acoust. Signal Enhancement (IWAENC)*, Sept 2014, pp. 233–237.

[15] B. Bech and N. Zacharov, "Fundamental of experimentation," in *Perceptual Audio Evaluation, Theory, Method and Application*, chapter 3. Wiley, 2006.

[16] ITU-T, "Subjective test methodology for evaluating speech communication systems that include noise suppression algorithms," Standard P.835, International Telecommunications Union (ITU-T), Nov. 2003.

[17] ITU-T, "Method for the subjective assessment of intermediate quality levels of coding systems," Standard BS.1534–3, International Telecommunications Union (ITU-T), Nov. 2003.

[18] K. Kinoshita, M. Delcroix, S. Gannot, E.A.P. Habets, R. Haeb-Umbach, W. Kellermann, V. Leutnant, R. Mass, T. Nakatani, B. Raj, A. Sehr, and T. Yoshioka, "A summary of the REVERB challenge: state-of-the-art and remaining challenges in reverberant speech processing research," *EURASIP Journal on Advances in Signal Processing*, vol. 2016, pp. 1–19, Jan. 2015.

[19] B. Cauchi, I. Kodrasi, R. Rehr, S. Gerlach, A. Jukić, T. Gerkmann, S. Doclo, and S. Goetze, "Combination of MVDR beamforming and single-channel spectral processing for enhancing noisy and reverberant speech," *EURASIP Journal on Advances in Signal Processing*, vol. 2015, pp. 1–12, July 2015.

[20] B. De Man and J.D. Reiss, "A pairwise and multiple stimuli approach to perceptual evaluation of microphones types," in *Proc. Audio Eng. Soc. (AES) Convention*, Rome, Italy, May 2013.

[21] E. Vincent, M. G. Jafari, and M. D. Plumbley, "Preliminary guidelines for subjective evalutation of audio source separation algorithms," in *UK ICA Research Network Workshop*, Southampton, United Kingdom, Sept. 2006.

[22] B. Cauchi, P. A. Naylor, T. Gerkmann, D. Doclo, and S. Goetze, "Late reverberant spectral variance estimation using acoustic channel equalization," in *Proc. European Signal Processing Conf. (EUSIPCO)*, Nice, France, sep 2015, pp. 1–4.

[23] I. Kodrasi, S. Goetze, and S. Doclo, "Regularization for partial multichannel equalization for speech dereverberation," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 21, no. 9, pp. 1879–1890, Sept. 2013.

[24] T. H. Falk and W.-Y. Chan, "Temporal dynamics for blind measurement of room acoustical parameters," *IEEE Trans. Instrum. Meas.*, vol. 59, no. 4, pp. 978–989, Apr. 2010.

[25] J. K. Nielsen, J. R. Jensen, S. H. Jensen, and M. G. Christensen, "The single- and multichannel audio recordings database (SMARD)," in *Proc. Intl. Workshop Acoust. Signal Enhancement (IWAENC)*, Antibes, France, Sept. 2014.

[26] ITU-T, "Objective measurement of active speech level," Recommendation P.56, International Telecommunications Union (ITU-T), Dec. 2011.

[27] J.S. Garofolo, L.F. Lamel, W.M. Fisher, J.G. Fiscus, D.S. Pallett, and Dahlgren N.L., "Darpa timit acoustic-phonetic continuous speech corpus cd-rom TIMIT," NIST Interagency/Internal Report (NISTIR) 4930, National Institute of Standards and Technology, Feb. 1993.

[28] M. Friedman, "The use of ranks to avoid the assumption of normality implicit in the analysis of variance," *Journal of the American Statistical Association*, vol. 32, no. 200, pp. 675–701, 1937.

[29] J.D. Gibbons and S. Chakraborti, *Nonparametric statistical inference*, Springer, Berlin, 2011.