On the Enhancement of Dereverberation Algorithms Using Multiple Perceptual-Evaluation Criteria

Rafael Zambrano-López*, Thiago de M. Prego†, Amaro A. de Lima† and Sergio L. Netto*
*Program of Electrical Engineering, Federal University of Rio de Janeiro, Brazil.
†Program of Electrical Engineering, Federal Center for Technological Education, Brazil.
Emails: {rafael.lopez, thiago.prego, amaro.lima, sergioln}@smt.ufrj.br

***Abstract*—This paper describes an enhancement strategy based on several perceptual-assessment criteria for dereverberation algorithms. The complete procedure is applied to an algorithm for reverberant speech enhancement based on single-channel blind spectral subtraction. This enhancement was implemented by combining different quality measures, namely the so-called QAreverb, the speech-to-reverberation modulation energy ratio (SRMR) and the perceptual evaluation of speech quality (PESQ). Experimental results, using a 4211-signal speech database, indicate that the proposed modifications can improve the word error rate (WER) of speech recognition systems an average of 20%.**

I. Introduction

Reverberation can drastically affect the performance of current automatic speech/speaker recognition or hearing-aid systems, motivating the use of appropriate speech enhancement techniques to reduce its effects. Although reverberation degrades speech intelligibility and perceptual quality, in a small amount it makes speech more pleasant to common listeners [1]. This paper analyzes an optimization procedure for choosing the parameter values of a given dereverberation algorithm based simultaneously on several perceptual-assessment measures. In particular, the measures considered here include the QAreverb [2], the speech-to-reverberation modulation energy ratio (SRMR) [3] and the perceptual evaluation of speech quality (PESQ) [4]. The core idea is to combine the ability to quantify the reverberation effect inherent to the QAreverb or SRMR measures with the PESQ ability to evaluate the overall quality of speech signals in the presence of other (coding) artifacts. The method is illustrated with a one-microphone dereverberation algorithm described in [5] but it is suitable to any dereverberation algorithm and any number of sensors. Performance of the final algorithm configuration is then analyzed with the large databases deployed in [6], [7]. Results with the given dereverberation algorithm acting as the front end to a speech recognition system indicate an average improvement of about 20% in the final word error rate (WER) in comparison to the original algorithm performance.

In order to describe the proposed techniques, this paper is organized as follows: In Section II, the main concepts behind the optimization methodology are first introduced. Section III presents the dereverberation algorithm considered in this work, describing its main parameters whose values are investigated later on. This section also details the employed quality-assessment measures, as well as the reverberant-speech databases used when applying the proposed methodology to the dereverberation algorithm. Section IV shows the results of the training experiments in three different scenarios, using the word error rate of an speech recognition system as a comparative measure. Finally, a conclusion concerning the overall performance increase is provided in Section V.

II. Proposed Methodology

There exist many perceptual reverberation-assessing estimators (e.g., [2]–[4], [8]–[12]), both blind (that is, based on the reverberant signal only) and non-blind approaches. The need for such assessments is inherent to modern communications and even practical systems often incorporate quality-assessing tools to evaluate their performance in a reliable manner.

The proposed methodology is intended to optimize simultaneously several complementary quality-assessment measures to improve the performance of dereverberation algorithms. By choosing a training set of speech signals and by selecting a proper set of measures to be optimized, it is possible to fine tune the parameters of the algorithm in question in a more efficient way. The resulting parameter values are the ones that yield a better compromise between the several quality-assessing measures considered.

Figure 1 shows an example of the proposed technique using two objective criteria. The feasible region is formed by all combinations of algorithm parameters, obtained through an exhaustive search. A Pareto frontier, which represents the best trade-offs between all considered objectives, is then found. Finally, a decision maker can decide which of the available trade-offs works best, choosing an operating point (or a set of operating points) that improves the considered measures in a joint manner.

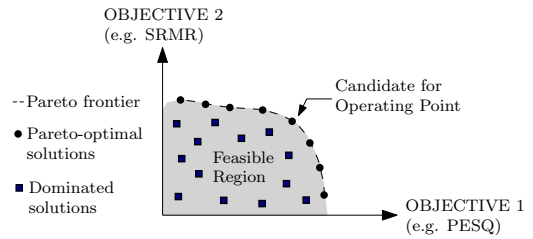

Fig. 1. Illustrative example of proposed 2-objective optimization using PESQ and SRMR as quality-assessment measures: The feasible region represents the SRMR × PESQ values obtained for all combinations of possible parameter values of a certain dereverberation algorithm. This plot allows one to choose an operating point (set of parameter values) of the dereverberation algorithm which provides the best compromise between the two measurements considered in the analysis.

The choice of the quality measures in the proposed optimization process has a significant role. If, for example,

two measures are highly correlated, the joint optimization would be equivalent to the individual optimization, which would entail an unnecessary waste of resources. Therefore, in this proposal the idea is to combine reverberation-based measures, such as the SRMR and QAreverb scores, with a more general purpose (coding) quality measure, such as PESQ, widely used for quality measurement of network transmitted speech. The joint reverberation-coding measure better assesses the algorithm performance, improving the intelligibility of the resulting signal with respect to the results achieved with the individual measures, as verified in Section IV.

III. Practical Considerations

In principle, the proposed optimization strategy may be applied to any dereverberation algorithm, and with distinct quality-assessment criteria. Its effectiveness, however, is illustrated here based on a simple dereverberation scheme [5] described in Subsection III-A and on the QAreverb, SRMR, and PESQ measures discussed in Subsection III-B. In addtion, all experimental data employed in this work are detailed in Subsection III-C.

A. Dereverberation Algorithm

The overall structure of the spectral subtraction dereverberation algorithm devised in [5] (as a simplification of the Wu-Wang algorithm introduced in [13]) is depicted in Fig. 2. In this scheme, the reverberant signal $z(n)$ is modeled as the convolution of the room impulse response (RIR) $h(n)$ and the anechoic (clean) speech signal $s(n)$, that is

$$z(n) = \sum_{l=0}^{N} h(l)s(n-l). \quad (1)$$

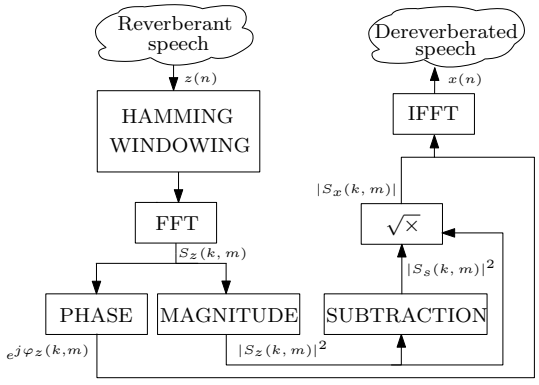

Fig. 2. Diagram of the spectral subtraction algorithm.

Let $S_z(k,m) = |S_z(k,m)|e^{j\varphi_z(k,m)}$ be the FFT of the m-th frame of the windowed version of $z(n)$, where a 32 ms Hamming window with 24 ms overlap between consecutive frames is used, and $w(m)$ be an asymmetrical smoothing window based on the Rayleigh distribution, given by

$$w(m) = \begin{cases} \left(\frac{m+a}{a^2}\right) e^{\left(\frac{-(m+a)^2}{2a^2}\right)} & \text{if } m > -a \\ 0 & \text{otherwise} \end{cases}, \quad (2)$$

where the parameter a controls the overall spread of the function.

The model of the power spectrum of the late reverberation can be described as

$$|S_l(k,m)|^2 = \gamma w(m-\rho) * |S_z(k,m)|^2, \quad (3)$$

where "$*$" represents the convolution operation in the time domain, k is the frequency bin, and m refers to the time frame. The parameter γ is a scaling factor and ρ represents the length of the early reflections.

Considering that the early and late components are mutually uncorrelated [13], the power spectrum of the early impulse components can be estimated by subtracting the power spectrum of the late impulse components from the reverberated speech. The spectrum subtraction scheme performs a weighting in the power spectrum of $z(n)$, and the block SUBTRACTION is given by

$$|S_s(k,m)|^2 = |S_z(k,m)|^2 \max\left[1 - \frac{|S_l(k,m)|^2}{|S_z(k,m)|^2}, \epsilon\right], \quad (4)$$

where ϵ is the floor and corresponds to the maximum attenuation. The power spectrum of the output (dereverberated) spech signal $x(n)$ is given by

$$|S_x(k,m)|^2 = \sqrt{|S_z(k,m)|^2 \times |S_s(k,m)|^2}. \quad (5)$$

Finally, in order to calculate the spectrum of $x(n)$, the phase $\varphi_z(k,m)$ of $S_z(k,m)$ is combined to the magnitude $|S_x(k,m)|$, such that

$$S_x(k,m) = |S_x(k,m)|e^{j\varphi_z(k,m)}, \quad (6)$$

which allows one to estimate the clean signal $x(n)$ as desired.

From this algorithm, four parameters were chosen for the optimization process, as detailed bellow:

- Scaling factor (γ): Specifies the relative strength of the late-impulse components of the reverberant speech signal in Equation (3). Despite many factors contribute to this relative strength (for instance, the reverberation time), the system performance is not very sensitive to specific values of γ [13]. The original value of the scaling factor was $\gamma = 0.35$.
- Attenuation limit (ϵ): Corresponds to the maximum attenuation in Equation (4). The original value of this parameter was $\epsilon = 0.001$, equivalent to an attenuation of 30 dB.
- Early-reflections length (ρ): Indicates the relative delay of the late impulse components in Equation (3). This delay reflects speech properties and is independent of reverberation characteristics. It is commonly considered to correspond to around 50 ms, which implies $\rho = 7$ frames. This value of ρ was set in the original algorithm.
- Spread control (a): This parameter controls the overall spread of the function $w(n)$ from Equation (2). It needs to be smaller than ρ to provide a reasonable match to the equalized impulse-response shape. The original value of this parameter was $a = 6$.

These four parameters were combined within different ranges in order to proceed with the optimization strategy. Table I shows the parameter ranges considered for the optimization of the algorithm, which gave a total of 2475 training setups.

TABLE I
RANGE OF VALUES OF EACH PARAMETER USED IN THE OPTIMIZATION PROCESS OF THE DEREVERBERATION ALGORITHM.

Parameter	Range
γ	$\{0.30, 0.31, 0.32, \ldots, 0.40\}$
ϵ	$\{10^{-5}, 10^{-4}, 10^{-3}, 10^{-2}, 10^{-1}\}$
ρ	$\{1, 2, 3, 4, 5, 6, 7, 8, 9\}$
a	$\{1, 2, 3, 4, 5, 6, 7, 8, 9\}$, with $a \leq \rho$

B. Quality-Assessment Measures

In this work, the perceptual quality of a reverberant speech signal is evaluated according to the following quality measures:

- The QAreverb [2] measure Q is defined as

$$Q = -\frac{\sigma^2 T_{60}}{R^{\gamma}}, \tag{7}$$

  where σ^2 denotes the room spectral variance defined in [15], T_{60} is the reverberation time (time interval required for the sound-pressure to decay 60 dB after its initial stimulus ceases [16]) and R is the direct-to-reverberant energy ratio [17]. In this expression, the constant factor γ sets the importance of R with respect to the other two parameters, and its value was heuristically set to $\gamma = 0.3$, maximizing the correlation between Q and the subjective scores for a reverberant-speech database developed in [2]. In practice, the parameters σ^2, T_{60}, and R can be obtained directly from the RIR, $h(n)$, which is estimated from the deconvolution process between the clean and the reverberant speech signals. In the final stage, the value of Q is mapped onto the 1–5 mean-opinion scale (MOS) using a nonlinear transformation yielding the Q_{MOS} QAreverb measure.
- The speech-to-reverberation modulation energy ratio (SRMR) [3] is a non-intrusive quality and intelligibility measurement of reverberant and dereverberated speech. SRMR is the ratio of the average energy in the low modulation frequencies (4 – 18 Hz) to the high modulation frequencies (29 – 128 Hz). Larger values are assumed to indicate better speech quality.
- The perceptual evaluation of speech quality (PESQ) [4] is an ITU-T standard measure, also suitable for distortions commonly encountered when speech goes through telecommunication channels, such as packet loss, signal delays, and codec distortions. PESQ compares two perceptually-transformed signals and generates a noise disturbance value to estimate the perceived speech quality. Larger values are assumed to indicate better quality.

C. Reverberant-Speech Databases

The main database used in this work was provided by the REVERB (REverberant Voice Enhancement and Recognition Benchmark) Challenge 2014 [18], which divided the data into the so-called development database and evaluation database. Each of these databases were further divided into two datasets:

- SimData: contains speech signals from the WSJCAM0 database [6], artificially convolved with RIRs measured in three different rooms with different volumes (small, medium and large) and two different source-microphone distances (near = 50 cm and far = 200 cm). Background noise was added to each signal at fixed signal-to-noise ratio (SNR) of 20 dB. The reverberation times for these rooms are $\{250, 680, 730\}$ ms.
- RealData: contains a set of real recordings from the MC-WSJ-AV database [7] made in a reverberant and noisy meeting room (which is different from the ones used for SimData) with two different source-microphone distances (near $\approx$ 100 cm and far $\approx$ 250 cm). The reverberation time for this room is about 700 ms.

All utterances considered were captured with single-channel microphones at a sampling frequency of 16 kHz. The development database is composed by 1484 utterances from SimData and 179 utterances from RealData. The wider evaluation database, with similar characteristics to the development database, is composed by 2176 utterances from SimData and 372 utterances from RealData. Both development and evaluation sets are employed in the present work.

IV. EXPERIMENTAL RESULTS

Following the proposal of this work, a small 12-signal training set (composed by one female and one male utterance randomly selected from each reverberation condition within the SimData dataset) was elaborated for the fine-tuning of the spectral-subtraction parameters. Feasible regions were elaborated by combining these parameters within their ranges according to Table I and computing the average of the Q_{MOS}, SRMR, and PESQ scores in the framework of the training set. In that manner, through the SRMR×PESQ and Q_{MOS}×PESQ relations for the obtained regions, a new operating point for the algorithm that jointly maximizes the desired measures was chosen. Among the Pareto-frontier solutions, the overall optimum was determined as the point with lower WER.

In order to show and compare the performance of the proposed method, three distinct scenarios were considered, as detailed below:

Scenario 1 (S1): Corresponds to the original configuration of the two-stage dereverberation algorithm proposed in [13]. In this case, besides the spectral-subtraction block, an inverse-filtering stage is first employed. For the spectral-subtraction stage, the original parameter set was $\{\gamma = 0.32, \epsilon = 10^{-3}, \rho = 7, a = 5\}$.

Scenario 2 (S2): In this scenario, the dereveberation algorithm described in Section III-A was tested. The original set of parameters was $\{\gamma = 0.35, \epsilon = 10^{-3}, \rho = 7, a = 6\}$.

Scenario 3 (S3): This scenario corresponds to the optimized algorithm configuration. After following the proposed method, the operating point of the optimized algorithm was achieved by the set $\{\gamma = 0.39, \epsilon = 0.1, \rho = 9, a = 5\}$.

Figures 3 and 4 illustrate the SRMR×PESQ and $Q_{\mathrm{MOS}}\times$ PESQ plots, respectively, obtained in the training process. Each scattered cross of these plots corresponds to one of the 2475 training setups. These two figures also show the Pareto-optimal solutions, the operating points of the three scenarios S1, S2 and S3, as well as the point corresponding to the unprocessed training signals. The set chosen in S3, represented by an asterisk, was the one with a lower error rate among the Pareto-solutions. The values of the perceptual measures for the three scenarios and for the unprocessed signals within the training process are summarized in Table II.

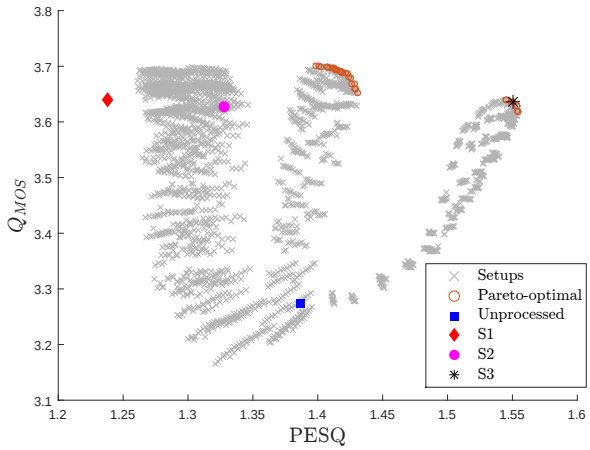

Fig. 3. SRMR×PESQ plot for the training process showing the operating points of the unprocessed signals and scenarios S1, S2, and S3.

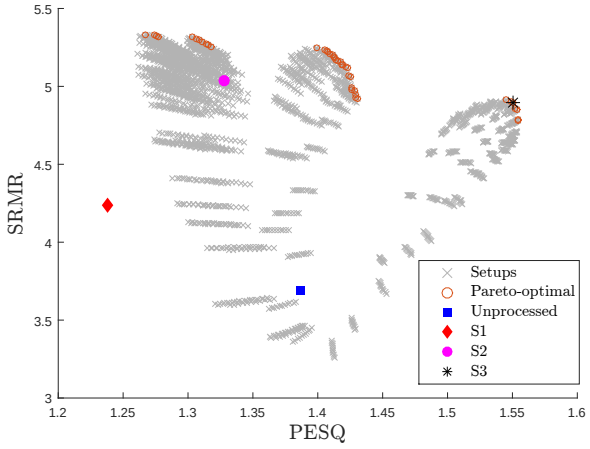

Fig. 4. Q_{MOS}×PESQ plot for the training process showing the operating points of the unprocessed signals and scenarios S1, S2, and S3.

TABLE II
QUALITY MEASURES OBTAINED IN THE TRAINING PROCESS FOR THE UNPROCESSED SIGNALS AND FOR EACH SCENARIO S1, S2, AND S3.

Scenario	SRMR	PESQ	Q_{MOS}
Unprocessed	3.68	1.39	3.28
S1	4.24	1.24	3.64
S2	5.04	1.33	3.63
S3	4.90	1.55	3.64

To evaluate the proposed modifications from the point of view of speech intelligibility, the performance of each algorithm configuration (scenario) was assessed through the word error rate (WER) of a speech recognition system based on HTK and provided by the REVERB challenge.

Tables III and IV show the WER results for each scenario and each subset of development and evaluation databases. From these results, it can be observed that the S3 scenario surpassed all other configurations in almost all cases, thus verifying the effectiveness of the proposed method. The results for SimData Room 1 show a slightly worst performance for the all the processed speech signals. This is because the perceptual quality of the signals from this setup was already considerably high. However, it must be noticed that the S3 scenario stays close to the unprocessed WER results for this room. In comparison to the Wu-Wang algorithm configuration (S1), scenario S3 presents an average improvement of 36% and 37% for SimData, and 21% and 18% for RealData. It also can be seen that for every setup, S1 worsens WER in relation to scenario S2, which reinforces the conclusions drawn in [5], regarding the removal of the inverse-filtering stage from the original two-stage algorithm [13]. Comparing with the unmodified algorithm (S2), scenario S3 improves by 24% for SimData and by 10% for the RealData set.

TABLE III
WORD ERROR RATE (WER) IN % FOR BOTH DEVELOPMENT AND EVALUATION SIMDATA DATASET. BOLD NUMBERS INDICATE BEST RESULTS.

Scenario	Room 1		Room 2		Room 3		Avg.
	Near	Far	Near	Far	Near	Far	–
Development dataset							
Unproc.	**15.3**	**25.3**	43.9	85.8	51.9	88.9	51.8
S1	65.9	76.9	62.4	76.9	72.4	83.8	73.0
S2	54.0	64.6	51.0	66.9	59.9	69.8	61.0
S3	18.7	25.8	**26.8**	**57.5**	**33.1**	**60.6**	**37.1**
Evaluation dataset							
Unproc.	**18.1**	**25.4**	42.9	82.2	53.5	88.0	51.7
S1	82.1	71.9	60.9	72.8	73.7	86.8	74.7
S2	61.3	68.5	49.2	62.1	58.8	77.3	62.0
S3	23.4	28.5	**27.1**	**50.7**	**35.8**	**62.0**	**37.9**

TABLE IV
WORD ERROR RATE (WER) IN % FOR BOTH DEVELOPMENT AND EVALUATION REALDATA DATASET. BOLD NUMBERS INDICATE BEST RESULTS

Scenario	Room 1		Avg.
	Near	Far	–
Development dataset			
Unproc.	88.7	88.3	88.5
S1	83.5	84.9	84.1
S2	73.6	74.2	73.9
S3	**61.6**	**64.6**	**63.1**
Evaluation dataset			
Unproc.	89.7	87.3	88.5
S1	89.2	87.7	88.4
S2	81.5	79.6	80.6
S3	**72.4**	**69.2**	**70.8**

V. CONCLUSION

This work proposes an enhancement strategy for dereverberation algorithms, based on the optimization of several perceptual measures simultaneously. The complete procedure was applied to an algorithm for reverberant speech enhancement based on a single-channel blind spectral-subtraction block. For this algorithm, four of its parameters were finely tuned following a jointly perceptual perspective. Results demonstrate effectiveness of proposed approach as it led to an algorithm scenario that outperformed other configurations in terms of speech intelligibility, as assessed by the lower WER achieved by a speech recognition system.

REFERENCES

[1] R. Appel and J. Beerends, "On the quality of hearing one's own voice," *J. Audio Engineering Society*, vol. 50, no. 4, pp. 237–248, Apr. 2002.

[2] A. A. de Lima, T. de M. Prego, S. L. Netto, B. Lee, A. Said, R. W. Schafer, T. Kalker, and M. Fozunbal, "On the quality-assessment of reverberated speech," *Speech Communication*, vol. 54, no. 3, pp. 393–401, Mar. 2012.

[3] 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 and Language Processing*, vol. 18, no. 7, pp. 1766-1774, 2010.

[4] ITU-T Rec. P.862, "Perceptual Evaluation of Speech Quality (PESQ): An Objective Method for End-to-End Speech Quality Assessment of Narrow-Band Telephone Networks and Speech Codecs," 2001.

[5] T. de M. Prego, A. A. de Lima, and S. L. Netto, " On the enhancement of dereverberation algorithms based on a perceptual evaluation criterion," *Proc. Interspeech*, Lyon, France, pp. 1360-1364, Aug. 2013.

[6] T. Robinson, J. Fransen, D. Pye, J. Foote, and S. Renals, "Wsjcam0: A British English speech corpus for large vocabulary continuous speech recognition," *Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Processing*, pp.81–84, 1995.

[7] M. Lincoln, I. McCowan, J. Vepa, and H. K. Maganti, "The multi-channel Wall Street Journal audio visual corpus (MC-WSJ-AV): Specification and initial experiments," *Proc. IEEE Workshop on Automatic Speech Recognition and Understanding*, 2005.

[8] J. B. Allen, "Effects of small room reverberation on subjective preference," *Journal of the Acoustical Society of America*, vol. 71, 1982.

[9] D. A. Berkley and J. B. Allen, "Normal listening in typical rooms: The physical and psychophysical correlates of reverberation," In *Acoustical Factors Affecting Hearing Aid Performance*, 2nd ed., edited by G. A. Studebaker and I. Hochberg, (Allyn and Bacon, Boston, 1993).

[10] J. Y. C. Wen and P. A. Naylor, "An evaluation measure for reverberant speech using tail decay modeling," *Proc. European Signal Processing Conf.*, Florence, Italy, Sept. 2006.

[11] J. Y. C. Wen, N. D. Gaubitch, E. A. P. Habets, T. Myatt, and P. A. Naylor, "Evaluation of speech dereverberation algorithms using the MARDY database," *Proc. IEEE Int. Workshop on Acoustic Echo and Noise Control*, Paris, France, Sept. 2006.

[12] Y. Hu and P. C. Loizou, "Evaluation of objective quality measures for speech enhancement," *IEEE Trans. Audio, Speech, and Language Processing*, vol. 16, no. 1, pp. 229–238, Jan. 2008.

[13] M. Wu and D. Wang, "A two-stage algorithm for one-microphone reverberant speech enhancement," *IEEE Trans. on Audio, Speech and Lang. Proc.*, vol. 14, May 2006.

[14] J. C. S. Veras, T. de M. Prego, A. A. de Lima, T. N. Ferreira, and S. L. Netto, " Speech quality enhancement based on spectral subtraction," *Proc. Reverberation Challenge*, Florence, Italy, pp. 1-5, May 2014.

[15] J. J. Jetz, "Critical distance measurement of rooms from the sound energy spectral response," *J. Acoustic. Soc. Am.*, vol. 65, pp. 1204–1211, May 1979.

[16] M. Karjalainen, P. Antsalo, A. Mäkivirta, T. Peltonen, and V. Välimäki, "Estimation of modal decay parameters from noisy reponse measurements," *Proc. Conv. Audio Engineering Society*, Amsterdam, Netherlands, pp. 867–878, May 2001.

[17] M. Kuster, "Reliability of estimating the room volume from a single room impulse response," *J. Acoustic. Soc. Am.*, vol. 124, pp. 982–993, Aug. 2008.

[18] K. Kinoshita, et al., "The REVERB Challenge: A Common Evaluation Framework for Dereverberation and Recognition of Reverberant Speech," *Proceedings of the IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA-13)*, 2013.