FEATURE ANALYSIS FOR THE REVERBERATION PERCEPTION IN SPEECH SIGNALS

J. M. F. del Vallado, A. A. de Lima*†‡*, T. de M. Prego*†‡*, S. L. Netto*†

*ETSIT, Universidad Politécnica de Madrid, Spain,
†PEE/COPPE, Federal University of Rio de Janeiro, Brazil,
‡Federal Center Tech. Edu. Celso Suckow da Fonseca, Nova Iguaçu, Brazil.
chema.jmfv@gmail.com and {*amaro,thprego,sergioln*}*@lps.ufrj.br*

ABSTRACT

This paper considers the problem of quantifying the reverberation perception on speech signals. We investigate several combinations of three distinct reverberation-related features (namely, the reverberation time (RT), room spectral variance (RSV), and direct-to-reverberant energy ratio (DRR)), which can be extracted directly from the associated room impulse response. Particular attention is also paid on different post-processing nonlinear mappings in order to provide a more effective quality evaluation algorithm. Results indicate that the RSV feature can be completely disregarded if the RT and DRR estimates are properly weighted. Performance of resulting measure is slightly superior in comparison to previous state-of-the-art method, particularly with respect to the computational cost and robustness (by disregarding the RSV estimation), but also on the statistical correlation level with subjective grades of a large dataset of reverberant speech.

Index Terms— Speech quality assessment, reverberation, MOS, room impulse response

1. INTRODUCTION

Reverberation is one of the most intricate problems in modern acoustical systems, highly affecting the performance, for instance, of a telepresence device or an automatic speech-recognition system. For that matter, in order to ensure appropriate functioning and user satisfaction in these types of systems, reverberation levels have to be regularly monitored and even reduced, if necessary.

This paper considers the human perception of the reverberation effect on high-quality speech signals. Several audio features are considered for that purpose and their optimal combination on a closed-form metric is investigated. Performance of the resulting estimator is compared to the subjective scores of two speech databases with distinct levels of reverberation.

The organization of the paper is as follows: In Section 2, we characterize the reverberation effect, and the three main features to be used in estimating the perceived quality of reverberant speech, namely: reverberation time, room spectral variance, and direct-to-reverberant energy ratio; In Section 3, two established objective measures based on these features are presented, whereas Section 4 details the proposed analysis, which considers a large set of feature combinations and several MOS-mapping functions, to maximize the correlation levels with the subjective scores of two distinct databases; Section 5 shows the analysis results, which can be regarded as a new objective measure for reverberation assessment, along with the complete step-by-step algorithm for its implementations; Finally, Section 6 summarizes the main contributions of the paper.

2. PRIOR WORKS IN THE FIELD

Previous works in reverberation assessment include, for instance, references [1]–[7].

In [1], a pioneering work by Allen describes what is perhaps the first attempt to quantify the reverberation effect, which was later validated experimentally in [2].

In [3], [4], authors present the so-called MARDY database, including 32 speech signals with different reverberation levels, and a new objective reverberation measure. In [5], authors introduce a new blind estimate for reverberation perception, based solely on the reverberated signal, what enables a real-time monitoring without disrupting system's operation.

In [6], authors investigated several speech features for reverberation assessment, which were later combined to generate an extension of Allen's non-blind metric with improved estimation performance, as presented in [7].

The present paper extends even further Allen's work, by considering several alternative combinations of three basic reverberation features (as detailed in Section III below). Results indicate a significant improvement in performance, achieving correlation scores of 92.3% and 95.1% with subjective grades of two independent databases of reverberated speech signals.

3. REVERBERATION CHARACTERIZATION

Reverberation is the effect where the direct sound combined with its attenuated-and-delayed versions (generated from reflections on the room surfaces) are perceived altogether as a single sound signal. Although some reverberation can be acoustically pleasant, excessive levels affects speech intelligibility, thus restricting conversation efficiency [8].

The discrete-time reverberated speech, $s_r(n)$, is commonly modeled as the convolution of a given speech signal, $s(n)$, with a length-N room impulse response (RIR), $h(n)$, which mathematically is given by

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

For analytical purposes, the RIR is often divided into two parts as represented in Fig. 1:

- Early reflections: These are essentially impulsive components, which contain most energy of the RIR and are constrained to the first $80 - 100$ ms portion of $h(n)$. The time, n_d, it takes for the direct sound to reach the listener is also observed in this part of RIR.

- Late reverberation: This includes the remaining portion of the RIR, and presents a diffuse nature, indicating the random nature of the process after a given interval of time.

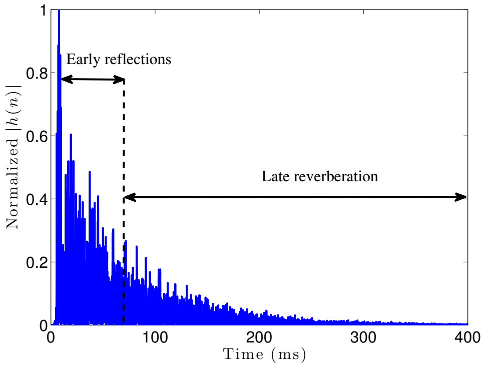

Fig. 1. Example of practical RIR emphasizing its early reflections and late reverberation portions.

As verified in [6], there are three measures highly associated to the perception of the reverberation effect: reverberation time (RT), room spectral variance (RSV), and direct-to-reverberant energy ratio (DRR).

3.1. Reverberation Time

The RT, T_{60}, is defined as the time the sound pressure takes to decrease 60 dB from its maximum energy after the steady-state excitation is instantly terminated. Its perceptual correspondence is the liveness, which is associated to the reflectiveness of the walls. In practice, the higher the value of T_{60}, the more lasting is the perceived reverberation effect on a given sound signal. Typical values of T_{60} range from a few milliseconds, in anechoic rooms; around 100–200 ms, in acoustically-treated recording rooms; and up to a few seconds in large or acoustically non-treated reverberating spaces.

The adopted algorithm to estimate T_{60} [9] combines the models for exponential decay of RIR and the stationary noise floor by using nonlinear optimization, providing a quite robust estimation.

3.2. Room Spectral Variance

The RSV function, σ_I^2, is given by [10]

$$\sigma_I^2 = \overline{(I(k) - \overline{I(k)})^2}, \tag{2}$$

where $\overline{\{\cdot\}}$ denotes the average of a function in the discrete frequency domain k and $I(k)$ is the relative acoustic intensity level, defined as

$$I(k) = 10\log_{10}\left[\frac{|H(k)|^2}{\overline{(|H(k)|^2)}}\right] \; (dB), \tag{3}$$

where $H(k)$ is the discrete Fourier transform of $h(n)$.

In practice [10], the value of σ_I^2 depends on the critical distance d_c, defined as the minimum distance at which the energy of the reverberant sound is no more affected by the source-microphone distance.

3.3. Direct-to-Reverberant Energy Ratio

The DRR, R, is the energy ratio between the direct and reverberant portions of RIR. The direct energy E_d corresponds to the accumulated RIR energy within a discrete time interval equivalent to 2.5 ms around n_d, and the reverberant energy E_r is the RIR energy of the remaining time interval, as given by [11, 12]

$$R = \frac{E_d}{E_r} = \frac{\sum_{(n_d-n_1)}^{(n_d+n_{1.5})} h^2(n)}{\sum_{(n_d+n_{1.5})}^{N} h^2(n)}, \tag{4}$$

where n_1 and $n_{1.5}$ are the discrete time equivalents to 1 ms and 1.5 ms continuous-time intervals. According to [13], only the signal components 20 dB above the noise floor level should be considered in $h(n)$, in order to reduce the noise influence when estimating the DRR.

4. OBJECTIVE MEASURES FOR REVERBERATION ASSESSMENT

Among the several measures for evaluating the perceived quality of reverberant speech signals, the first and yet one of the most efficient is the Allen's score P, defined as [1]

$$P = P_{max} - \sigma_I^2 T_{60}, \tag{5}$$

where P_{max} is upper value of P. Although Allen's score is quite effective, even for today's standards, it was primarily devised for short source-listener distances, when the RSV and DRR become quite correlated [14].

The QAreverb measure Q developed in [7] overcomes that gap by incorporating the DRR measure in its formulation, such that

$$Q = -\frac{T_{60}\sigma_I^2}{R^\gamma}, \tag{6}$$

with $\gamma = 0.3$ obtained empirically during the system's training stage. Taking the DRR measure explicitly into account seems to extend the estimation reliability of the QAreverb measure beyond the critical-distance case and into a wider range of practical reverberation scenarios.

5. PROPOSED ANALYSIS

The proposed analysis generalizes the QAreverb measure by associating a different exponent to each individual feature, T_{60}, σ_I^2, and R, leading to the modified closed-form expression

$$Q_m = -\frac{(T_{60})^\alpha (\sigma_I^2)^\beta}{(R)^\gamma}. \tag{7}$$

In this work, the values of α, β, and γ are chosen in order to maximize Pearson's correlation [15] with the subjective scores of two independent databases described below.

Besides the generalized QAreverb approach, three different nonlinear mapping functions are considered in a similar manner to other objective evaluators:

- A ITU P.563-like mapping of the form [16]

$$Q_m^{(a)} = x_1 Q_m^3 + x_2 Q_m^2 + x_3 Q_m + x_4; \tag{8}$$

- The mapping employed on the ITU recommendation for perceptual evaluation of speech quality (PESQ) [17]

$$Q_m^{(b)} = 1 + \frac{4}{1 + e^{y_1 Q_m + y_2}}; \tag{9}$$

- A modified PESQ-like mapping defined as

$$Q_m^{(c)} = z_3 + \frac{z_4}{1 + e^{z_1 Q_m + z_2}}. \qquad (10)$$

In addition to each of the nonlinear mappings listed above, a subsequent linear mapping of the form

$$Q_{\text{MOS}}^{(x)} = v Q_m^{(x)} + w, \qquad (11)$$

with $x = a, b, c$, was also considered for a proper scaling adjustment [18] of the final estimation score, and the baseline QAreverb mapped score Q_{MOS} is given by Q applied to the mappings of Eqs. (8) and (11).

In this work, two databases were employed to evaluate the performance of the generalized QAreverb system:

- Database A [7]: This database comprises a total of 204 speech signals, sampled at $F_s = 48$ kHz, with distinct levels of reverberation, including: 4 anechoic (2 by a male speaker and 2 by a female speaker) signals; 108 directly degraded signals in real environments comprising 7 different rooms (with RT in the range of $T_{60} = \{140, 390, 570, 650, 700, 890, 920\}$ ms) and several source-microphone distances; 68 degraded signals with estimated RIRs from 4 real environments (with $T_{60} = \{120, 230, 430, 780\}$ ms and distinct source-microphone distances for each case, as detailed in [19]); and 24 degraded signals with 6 artificial RIRs with the early reflections obtained via the image method [20], with a fixed $d = 1.8$ m in a room of dimensions length×widht×height = $4 \times 3 \times 3$ m. As regards the late reverberation, the feedback delay network method [21] was used to emulate $T_{60} = \{200, 300, 400\}$ ms and a modified version of Gardner's method [22], which was devised to emulate reverberation times above 400 ms, was used for $T_{60} = \{500, 600, 700\}$ ms.

  These three scenarios altogether include only 12 signals with a source-microphone distance shorter than the theoretical critical limit, where the RSV and DRR metrics are highly correlated [14].

  All 204 Database A signals were sorted according to the corresponding MOS and then subdivided into two 102-signal databases, Database A$_1$ and Database A$_2$, collecting the odd- and even-index signals, respectively. Database A$_1$ was used for training of all algorithms described above, whereas Database A$_2$ was used only for testing the algorithm performances on untrained data.

- Database B: the so-called MARDY database [4], includes 16 reverberant signals recorded directly from a studio environment, configuring naturally degraded signals, and their dereverberated versions using delay-and-sum algorithm, making a total of 32 speech signals.

The A$_1$ subset of Database A was used in the system training to obtain the best experimental values for the α, β, and γ exponents in Eq. (7) and all coefficients in Eqs. (8)–(11). The remaining A$_2$ subset and Database B were used only on the validation stage of previous results.

6. EXPERIMENTAL RESULTS

6.1. Estimating α, β, and γ

In order to select the best combination for the α, β, and γ exponents in Eq. (7), each of these parameters was varied in the interval $[0, 1]$ with steps of 0.05 in a total of $21^3 = 9261$ indexed combinations. Each setup leads to a different set of Q_m scores, whose correlation factors ρ with the subjective grades of the testing Database A_1 are depicted in Fig. 2, where the crosses indicate a combination with $\beta = 1$ and the scattered dots correspond to a combination with $\beta = 0$. It can be observed that in every region of this figure there is a parameter combination with $\beta = 0$ that yields a higher performance than a similar combination with a different value of β.

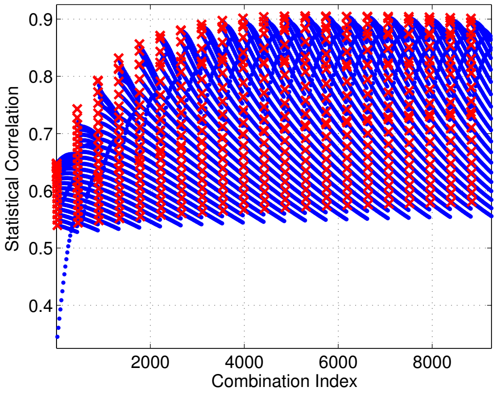

Fig. 2. Statistical correlation between Q_m and the subjective MOS of Database A_1 for all (α, β, γ) combinations, where "×" represents a combination with $\beta = 0$ and "." represents a combination with $\beta = 0.5$ or $\beta = 1$.

Table 1 presents the three best correlation scores for the modified extended QAreverb approach and the subjective scores of Database A_1. From these results, one readily observes that all three setups include $\beta = 0$, indicating that the RSV feature becomes redundant with the RT and DRR features, and can be removed from the final score calculation.

Table 1. Correlation scores ρ (%) for all databases for best (α, β, γ) combinations using training Database A_1.

Parameters			Correlation (Databases)			
α	β	γ	$\rho(A_1)$	$\rho(A_2)$	$\rho(A)$	$\rho(B)$
0.55	0	0.15	90.5	89.5	90.0	94.2
0.60	0	0.15	90.5	89.5	90.0	94.1
0.65	0	0.15	90.5	89.4	89.9	94.0

6.2. Estimating α, β and γ with mappings

The choice of the best (α, β, γ) parameters was also performed by taking the nonlinear mappings detailed in Section 5 into account.

Table 2 shows the three best performances for the three mappings defined in Eqs. (11), (9), and (10). The best combinations for each mapping are somewhat consistent in the sense they all present very small values of β, once again corroborating the conjecture that the RSV is somewhat redundant with the RT and DRR features. From this table, one also notices that PESQ-like mappings seem to work minimally better than the P.563-like mapping, due to the fact that their free parameters adjustment for a good fitting leads to slightly higher correlations for Databases A and B.

Table 2. Correlation scores ρ (%) for all databases and distinct nonlinear mappings for best (α, β, γ) combinations using training Database A_1.

Mapping	Parameter			Correlation (Databases)			
	α	β	γ	$\rho(A_1)$	$\rho(A_2)$	$\rho(A)$	$\rho(B)$
$Q_{\mathrm{MOS}}^{(a)}$	1.00	0.10	0.35	93.0	91.7	92.3	95.0
	1.00	0.15	0.35	93.0	91.7	92.3	95.0
	1.00	0	0.40	93.0	91.7	92.3	95.1
$Q_{\mathrm{MOS}}^{(b)}$	0.70	0	0.2	93.9	91.9	92.9	95.3
	0.90	0.05	0.25	93.9	91.9	92.9	95.3
	0.85	0.05	0.25	93.9	92.0	92.9	95.4
$Q_{\mathrm{MOS}}^{(c)}$	0.85	0	0.25	93.9	92.0	92.9	95.4
	0.70	0	0.20	93.9	91.9	92.9	95.3
	0.90	0.05	0.25	93.9	91.9	92.9	95.3

6.3. Comparison to other systems

In this section, we compare the performance for the modified QAreverb measure (using $\alpha = 0.85$, $\beta = 0$, $\gamma = 0.25$, and the modified PESQ-like mapping described in Eq. (10)) to other objective measures, for reverberation assessment or not, previously presented in the literature. Correlation results from the scores yielded by all methods and the subjective grades for the two databases employed in this work are summarized in Table 3.

Table 3. Statistical correlation coefficient ρ (%) between subjective grades and objective scores by several quality-evaluating algorithms for the Databases A and B.

Algorithm	Databases	
	A	B
W-PESQ	88.6	77.7
P.563	58.8	53.9
R_{DT}	60.6	64.2
SRMR	80.6	76.9
Q_{MOS}	91.2	94.9
$Q_{\mathrm{MOS}}^{(c)}$	92.9	95.4

From this table, one notices how the modified QAreverb does not significantly diminish its performance by disregarding the RSV feature. In fact, the correlation increases in 0.5% for Database B and 1.7% for Database A, which is statistically significant for the amount of signals (204) being considered. Furthermore, the suppression of RSV reduces the computational cost in about 5% compared to the baseline score Q_{MOS}. The estimated scores for the chosen configuration of the modified QAreverb system are shown in Figs. 3 and 4 for Databases A and B, respectively, illustrating its ability to predict such scores in a successful manner.

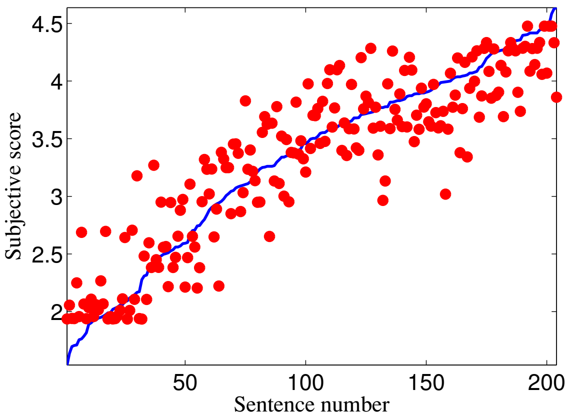

Fig. 3. Objective scores (large dots) yielded by modified QAreverb algorithm and subjective grades (solid line) for all 204 signals in Database A.

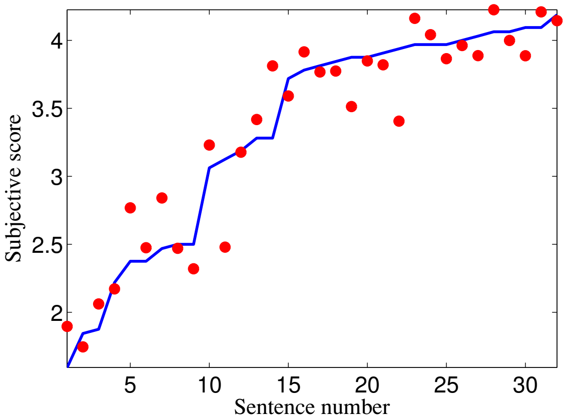

Fig. 4. Objective scores (large dots) yielded by modified QAreverb algorithm and subjective grades (solid line) for all 32 signals in Database B.

7. CONCLUSIONS

This paper analyzed several combination strategies for the RT, RSV, and DRR features to estimate the human perception of the reverberation effect in speech signals. Several mapping functions were also put to the test altogether with all feature combinations previously considered. Performance of the new objective measure was validated for several reverberating scenarios, yielding 92.9% and 95.4% statistical correlations with the subjective MOS of two independent databases and reducing in about 5% the measure computational cost. Results also indicated that the RSV features little contributes to the final estimate if the RT and DRR measures are properly combined with an effective nonlinear mapping. This can be explained from the fact that RSV and DRR features tend to aggregate similar information about the perceived reverberation effects, except from the fact that RSV is region limited with respect to the source-listener positioning. This conclusion can be further exploited on the development of a more reliable and simplified blind counterpart of the QAreverb algorithm.

8. REFERENCES

[1] J. B. Allen, "Effects of small room reverberation on subjective preference," *J. Acoustic. Soc. Am.*, vol. 71, Apr. 1982.

[2] 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, 1993.

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

[4] 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, pp. 1–4, Sept. 2006.

[5] T. H. Falk and W.-Y. Chan, “A non-intrusive quality measure of dereverberated speech,” *Proc. IEEE Int. Workshop on Acoustic Echo and Noise Control*, Seattle, USA, Sept. 2008.

[6] A. A. de Lima, T. de M. Prego, S. L. Netto, B. Lee, A. Said, R. W. Schafer, T. Kalker, and M. Fozunbal, “Feature analysis for quality assessment of reverberated speech,” *Proc. Int. Workshop on Multimedia Signal Proc.*, Rio de Janeiro, Brazil, Oct. 2009.

[7] 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, pp. 393–401, Mar. 2012.

[8] ITU-T Rec. G.191, *Software tools for speech and audio coding standardization*, 1995.

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

[10] 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.

[11] P. Zahorik, “Assessing auditory distance perception using virtual acoustics,” *J. Acoustic. Soc. Am.*, vol. 111, pp. 1832–1846, Apr. 2002.

[12] P. Zahorik, “Direct-to-reverberant energy ratio sensitivity,” *J. Acoustic. Soc. Am.*, vol. 112, pp. 2110–2117, Nov. 2002.

[13] 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.

[14] E. Larsen, N. Iyer, C. R. Lansing and A. S. Feng, “On the minimum audible difference in direct-to-reverberant energy ratio,” *J. Acoustic. Soc. Am.*, vol. 124, pp. 450–461, July 2008.

[15] S. M. Kay, *Fundamentals of Statistical Signal Processing: Estimation Theory*, Prenctice-Hall, Upper Saddle River:NJ, 1993.

[16] ITU-T Rec. P.563, *Single-ended Method for Objective Speech Quality Assessment in Narrow-band Telephony Applications*, 2004.

[17] ITU-T Rec. P.862.1, *Mapping function for transforming P.862 raw result scores to MOS-LQO*, 2003.

[18] S. Zielinski and F. Rumsey, “On some biases encountered in modern audio quality listening test - A review,” *J. Audio Eng. Soc.*, vol. 56, no. 6, pp. 427–451, 2008.

[19] M. Jeub, M. Schäfer, and P. Vary, “A Binaural Room Impulse Response Database for the Evaluation of Dereverberation Algorithms,” *Proc. of 16th Int. Conf. on Digital Signal Processing (DSP)*, Santorini, Greece, 2009.

[20] J. B. Allen and D. A. Berkley, “Image method for efficiently simulating small-room acoustics,” *J. Acoustic. Soc. Am.*, vol. 65, no. 4, pp. 943–950, Apr. 1979.

[21] J.-M. Jot and A. Chaigne, “Digital delay networks for designing artificial reverberators,” *Proc. Conv. Audio Eng. Soc.*, Preprint 3030, Feb. 1991.

[22] W. G. Gardner, *The Virtual Acoustic Room*, MSc Thesis, MIT, Cambridge, Sept. 1992.