# Feature Analysis for Quality Assessment of Reverberated Speech

Amaro A. de Lima[†], Thiago de M. Prego[†], Sergio L. Netto[†],
Bowon Lee[‡], Amir Said[‡], Ronald W. Schafer[‡], Ton Kalker[‡], Majid Fozunbal[‡]

[†] *PEE/COPPE, Federal University of Rio de Janeiro - Rio de Janeiro, Brazil*
{amaro,thprego,sergioln}@lps.ufrj.br

[‡] *HP Labs - Palo Alto, USA*
{bowon.lee,amir_said,ron.schafer,ton.kalker,majid.fozunbal}@hp.com

***Abstract*—This paper analyzes the ability of several measurements to quantify the reverberation effect in speech signals. We consider an intrusive scheme, in which the clean and reverberated signals are available, allowing one to estimate the corresponding room impulse response (RIR) signal. An artificial neural network (ANN) is trained for all features and used in a regression approach to estimate the human perceptual evaluation in a mean opinion score (MOS) 1–5 scale. Dimensionality reduction approaches are applied to generate a simpler ANN regression, establishing the most representative features for the problem at hand. A correlation level of 85% with subjective test scores was achieved by reducing the input-vector dimension from 10 to 3, including only the features of reverberation time, room spectral variance, and direct-to-reverberant energy ratio.**

## I. Introduction

Reverberation is an intricate acoustic effect in which reflections of a given signal are perceived altogether as a single modified signal. This paper considers the problem of evaluating reverberation levels in high-quality speech signals. We follow a similar scheme to the one employed in ITU-R BS.1387-1 PEAQ [1] (perceptual evaluation of audio quality) recommendation: Several numerical features are extracted from corrupted speech signals in an attempt of quantifying the reverberation effect. These measurements are fed into an artificial neural network (ANN) [2] that maps the input feature space to the one-dimensional quality score. A database of subjective scores is used as reference to train the ANN mapping. Resulting scores are then correlated to the reference ones to evaluate ANN mapping efficiency.

Three dimensionality reduction approaches are tested in order to simplify the ANN structure, determining the most representative features of the initial input space. In all three approaches, feature importance was evaluated based on the correlation factor between resulting scores and reference grades.

This paper is organized as follows: In Section II, the perceived reverberation effect is characterized and the ten features considered in this work are described in detail; Section III presents several experiments analyzing the individual and collective performances of all features in evaluating the reverberation perception in speech signals; Section IV emphasizes the main contributions of the paper.

*MMSP'09, October 5-7, 2009, Rio de Janeiro, Brazil.*
*978-1-4244-4464-9/09/$25.00 ©2009 IEEE.*

## II. Reverberation Characteristics

Reverberation is commonly modeled as the result of a linear convolution operation with the room impulse response (RIR), $h(t)$, that represents the acoustical characteristics of a room. In practice, one distinguishes two portions of the RIR: early reflections, which comprises several reflections and contains most of the RIR energy; late reverberation, which constitutes the remaining RIR portion and presents a diffusive nature, as indicated in Figure 1. In this context, one defines the direct-sound $t_d$ and first-reflection $t_r$ instants as the time values of the first and second impulsive components of the RIR early reflection portion.

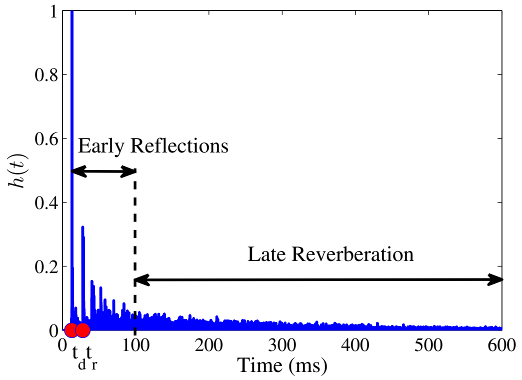

Fig. 1. RIR example indicating the direct-sound ($t_d$) and first-reflection ($t_r$) time instants.

The features considered in this work are all related to the associated RIR. In that context, we consider the so-called intrusive scheme, in which the clean and reverberated signals are available, allowing one to estimate $h(t)$ using a deconvolution operation. The analysis performed in this work could act as a starting point for the development of a non-intrusive method, which would consider a first stage of blind RIR estimation.

### A. Reverberation Time

The reverberation time $T_{60}$ is defined as the period of time required for the sound-pressure to decay 60 dB when a steadystate excitation is abruptly interrupted. In practice, higher $T_{60}$ indicates a more reverberant room. The value of $T_{60}$ may be estimated using, for instance, one of the following schemes:

**Algorithm T1:** By performing a linear fitting of the energy decay curve (EDC) defined as

$$EDC(t) = \int_{t}^{\infty} h^2(\tau)\mathrm{d}\tau, \tag{1}$$

one may then estimate the parameter $T_{60}$ as the time interval required for this approximation to decay from 0 to $-60$ dB [3].

**Algorithm T2:** In [4], Lundeby et al. propose a linear fitting directly on the RIR signal (in the dB scale), down to the time instant in which this signal is significantly affected by noise. Once again, the linear approximation is then employed to estimate $T_{60}$ based on its original definition.

**Algorithm T3:** The scheme proposed in [5] determines a curve fitting of the EDC function, as in Algorithm T1, but also takes into account the noise-floor level, similarly to what is done in Algorithm T2.

Figure 2 illustrates the general mechanism behind all $T_{60}$ estimating algorithms. In this figure, the EDC function is approximated by a linear fitting, and the slope of this line determines the required $T_{60}$ for it to fall from 0 to $-60$ dB.

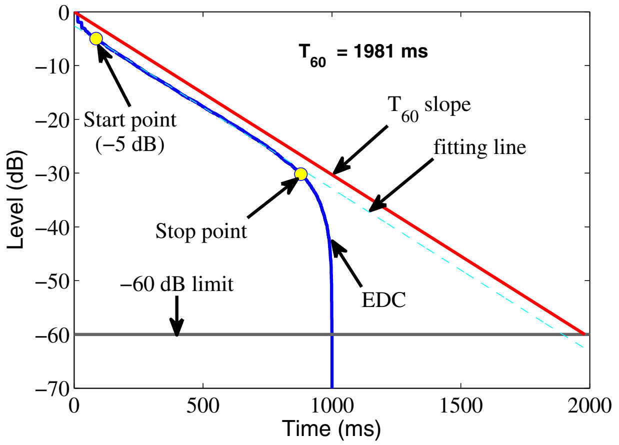

Fig. 2. Reverberation time estimation: The solid curve generates a dashed linear fitting, the slope of which determines the decaying time $T_{60}$ estimate from 0 to $-60$ dB levels.

### B. RIR Total Energy

The RIR is commonly normalized such that its maximum value is set to 1. The resulting total energy is determined as

$$E = \int_{0}^{\infty} h^2(t)\mathrm{d}t. \tag{2}$$

### C. Direct-to-Reverberant Energy Ratio

This feature [6], [7], [8] is determined as the ratio between the direct $E_d$ (within an interval around $t_d$) and reverberant $E_r$ (remaining) energy levels, as given by

$$\frac{E_d}{E_r} = \frac{\displaystyle\int_{(t_d-1)\ \text{ms}}^{(t_d+1.5)\ \text{ms}} h^2(t)\mathrm{d}t}{\displaystyle\int_{(t_d+1.5)\ \text{ms}}^{\infty} h^2(t)\mathrm{d}t}. \tag{3}$$

To reduce noise influence, one may consider only the signal components 20 dB above the noise floor level and stop the energy accumulation at the stop point employed by $T_{60}$ algorithm, as suggested in [8].

### D. Definition

For speech signals, this parameter is determined as a ratio between the energy of first the 50 ms and the total RIR energies [9], that is

$$D_{50} = \frac{\displaystyle\int_{0}^{50\ \text{ms}} h^2(t)\mathrm{d}t}{E}. \tag{4}$$

In practice, this feature depends on the reflection pattern of the room walls, source-microphone distance, and room dimensions.

### E. Clarity

This feature is determined by the dB ratio between the energy of the first 50 ms of speech and the remaining RIR energy [9], that is

$$C_{50} = 10\log_{10}\left(\frac{\displaystyle\int_{0}^{50\ \text{ms}} h^2(t)\mathrm{d}t}{\displaystyle\int_{50\ \text{ms}}^{\infty} h^2(t)\mathrm{d}t}\right), \tag{5}$$

or, equivalently,

$$C_{50} = 10\log_{10}\left(\frac{D_{50}}{1-D_{50}}\right). \tag{6}$$

### F. Center Time

This measurement represents the RIR center of mass, which may be determined as

$$CT = \frac{\displaystyle\int_{0}^{\infty} th^2(t)\mathrm{d}t}{E}. \tag{7}$$

Lower $CT$ indicates that there is more energy at the beginning, which means higher clarity. On the other hand, larger $CT$ values indicate a more reverberant sound [10].

### G. Initial Time Delay Gap

This feature is determined by the time difference between the direct-sound $t_d$ and first-reflection $t_r$ instants [9], that is,

$$ITDG = t_r - t_d, \tag{8}$$

which, in practice, represents "intimacy": Lower $ITDG$ values are perceived as smaller rooms.

### H. Source-Microphone Distance

This feature can be estimated from the RIR signal as

$$d = ct_d \tag{9}$$

where $c$ is the sound speed and $t_d$ is the first-reflection time instant, as illustrated in Figure 1. In our context, parameters $d$ and $t_d$ are completely equivalent, as the constant $c$ does not bring any new information. We have opted for feature $d$, however, since it can be more easily related to a reverberation effect in a given room.

### *I. Room Spectral Variance*

This feature is defined as [11]

$$RSV = \mathbb{E}_f \left[ (I(f) - \mathbb{E}_f[I(f)])^2 \right], \tag{10}$$

where $\mathbb{E}_f[.]$ denotes the expectation value operator in the frequency domain and $I(f)$ is the relative acoustic intensity level, defined as

$$I(f) = 10 \log_{10} \left( \frac{H^2(f)}{\mathbb{E}_f[H^2(f)]} \right) \ (dB), \tag{11}$$

with $H(f)$ is the Fourier transform of $h(t)$. The $RSV$ appears in Allen's measure to assess the reverberation effect [12]. Its value depends on the ratio $d/d_c$, where $d_c$ is the critical distance, defined as the minimum distance at which the energy of the reverberant sound is no more affected by the source-microphone distance, $d$.

### *J. Room Volume*

The number and energy of all sound reflections that compose the reverberation effect are directly related to the room geometry. This aspect can be represented, in a simplified manner, by the room volume, $V$, which can be estimated from the RIR as given by:
**Algorithm V1:** Assuming spherical geometry, one may write [13]

$$V(\text{V1}) = \frac{4\pi}{3N}(cT)^3, \tag{12}$$

where once again $c$ is the sound speed and $N$ is the number of sound reflections within the RIR over a time interval $T$.
**Algorithm V2:** A more elaborate estimate for the room volume is given by [8]

$$V(\text{V2}) = \frac{E_d}{E_r} \left( \frac{4\pi d^2 c T_{60}}{6 \ln(10)} \right). \tag{13}$$

**Algorithm V3:** In practice, Algorithms V1 and V2 tend to under- and over-estimate the room volume, respectively. A third volume estimator can then be devised as a convex sum of these two estimations, that is,

$$V(\text{V3}) = \alpha V(\text{V1}) + (1 - \alpha) V(\text{V2}), \tag{14}$$

with $0 \leq \alpha \leq 1$. Parameter $\alpha$ can be determined experimentally by minimizing the energy of estimation error between theoretical $V$ and $V(\text{V3})$, as performed in Subsection III-B.

## III. Numerical Experiments

The results presented in this section are divided into four parts:

1) Experiment 1: Evaluation of $T_{60}$ estimators described in Subsection II-A;
2) Experiment 2: Evaluation of volume estimators described in Subsection II-J;
3) Experiment 3: Correlation of all individual features against subjective scores for databases using artificial [14] and real [15] RIRs;
4) Experiment 4: Correlation of all combined features through an ANN against the subjective scores.

### *A. Experiment 1*

To validate the algorithms described in Subsection II-A, the reverberation effect was emulated through artificial RIRs generated by the image method [16] with $T_{60} = \{300, 400, 500, 600, 700\}$ ms and $V = \{36, 140, 240, 324, 450\}$ m$^3$. Performances were evaluated through the mean $\mu_T$ estimate, relative standard deviation $\gamma_T$ (which is the standard deviation normalized by the mean) and total squared relative error, defined as

$$E_{rel} = \sum_V \left( \frac{T_{60} - T_i(V)}{T_{60}} \right)^2, \tag{15}$$

where $T_{60}$ corresponds to the theoretical reverberation time and $T_i(V)$ denotes the estimation yielded by Algorithm T$_i$, with $i = 1, 2, 3$, for a given volume $V$. From Table I, one concludes that, for this artificial-RIR database, Algorithm T2 outperformed Algorithms T1 and T3 in the contexts of minimum $E_{rel}$ and highest statistical correlation with theoretical $T_{60}$. From these results, Algorithm T2 will serve as the basis for the room volume estimators in the following experiment, as also suggested by [8].

TABLE I
Comparison of estimation Algorithms T1, T2, and T3 for reverberation time $T_{60}$ (ms).

| $T_{60}$ | T1 ($\mu_T$, $\gamma_T$) | T2 ($\mu_T$, $\gamma_T$) | T3 ($\mu_T$, $\gamma_T$) |
|---|---|---|---|
| 300 | 151, 0.52 | **278**, 0.01 | 129, 0.21 |
| 400 | 291, 0.29 | **368**, 0.02 | 241, 0.27 |
| 500 | 441, 0.06 | **526**, 0.02 | 429, 0.15 |
| 600 | **548**, 0.02 | 700, 0.04 | 518, 0.14 |
| 700 | **715**, 0.20 | 858, 0.07 | 608, 0.19 |
| $Erel$ | 2.36 | **0.51** | 3.06 |
| $Corr(\%)$ | 93 | **99** | 93 |

### *B. Experiment 2*

The algorithms described in Subsection II-J were employed to estimate the room volume in the artificial RIRs determined in Experiment 1. Estimation results are given in Table II, in which Algorithm V3, as defined in Eq. (14) with $\alpha = 0.4319$, yields the best estimates in all cases except when $V = 36$ m$^3$, which, by the way, gets poor results by all three algorithms.

TABLE II
Comparison of estimation Algorithms V1, V2, and V3 for room volume $V$ (m$^3$).

| $V$ | V1 ($\mu_V$, $\gamma_V$) | V2 ($\mu_V$, $\gamma_V$) | V3 ($\mu_V$, $\gamma_V$) |
|---|---|---|---|
| 36 | **65**, 0.08 | 114, 0.31 | 93, 0.24 |
| 140 | 90, 0.06 | 216, 0.36 | **162**, 0.29 |
| 240 | 131, 0.12 | 285, 0.51 | **219**, 0.40 |
| 324 | 192, 0.63 | 448, 0.31 | **337**, 0.27 |
| 450 | 262, 0.27 | 559, 0.39 | **431**, 0.34 |

### C. *Experiment 3*

This experiment uses a combination of two speech databases to analyze the effectiveness of each individual feature described in Section II in assessing the reverberation effect in speech signals.

Database 1 [14] comprises a total of 72 Brazilian-Portuguese speech signals from 4 speakers (2 male and 2 female), recorded with $F_s = 48$ kHz sampling rate, and including 3 repetitions of each theoretical $T_{60}$. Reverberation effect was enforced through artificial RIRs, in which the early reflections were obtained via the image method [16], using a room of dimensions 4m-length, 3m-width, and 3m-height and assuming a fixed distance source-microphone $d = 1.8$ m. For the late reverberation, the FDN method [17] was used to simulate $T_{60} = \{200, 300, 400\}$ ms, and a modified version of Gardner's method [18], which was conceived for higher reverberation times, was used for $T_{60} = \{500, 600, 700\}$ ms.

Database 2, the so-called MARDY database [15], includes 32 speech signals with $F_s = 16$ kHz obtained from 2 different speakers (1 male and 1 female); 4 different source-microphone distances; 2 types of wall panels (1 reflexive and 1 absorbent); 2 different reverberation levels, including speech degraded by reverberation and dereverberated signals by a delay-and-sum beamformer.

Both databases were divided into two parts – one for Training and other for Test – each containing a total of 52 signals, 36 from Database 1 and 16 from the Database 2. Table III includes correlation scores between each individual feature and the subjective MOS scores provided for the so-called Training and Test databases. Results indicate that the reverberation time seems the most important individual measure, followed by $RSV$, $E$, $CT$, and $E_d/E_r$.

TABLE III
STATISTICAL CORRELATION (%) BETWEEN INDIVIDUAL REVERBERATION FEATURES AND SUBJECTIVE SCORES FOR THE TRAINING AND TEST DATABASES WITH ARTIFICIALLY AND NATURALLY GENERATED RIRS.

| Features | Training | Test |
|---|---|---|
| $T_{60}$ | **-81** | **-86** |
| $RSV$ | -72 | -67 |
| $E_d/E_r$ | 46 | 42 |
| $d$ | -30 | 04 |
| $V$(V3) | -31 | 19 |
| $C_{50}$ | -19 | 27 |
| $D_{50}$ | 24 | 12 |
| $CT$ | -57 | -65 |
| $ITDG$ | -33 | -05 |
| $E$ | -68 | -63 |

### D. *Experiment 4*

This subsection reports the nonlinear combination of all features through an ANN to predict the direct subjective MOS for the degraded speech signal. The designed ANN is a typical 2-layer perceptron with linear activation function for the neurons, where the input layer has as much neurons as the number of features and the output layer has only one neuron to generate the desired score.

Initially, the ANN input vector comprised the 10 reverberation features considered in Table III. Three dimensionality reduction approaches were employed to determine the most representative features for reverberation evaluation. Approach 1 uses a full-dimensional ANN and nullifies one input feature at a time. Approach 2 considers several reduced-order ANNs, each one ignoring one input feature at a time. Approach 3 is based on standard principal component analysis (PCA) [19]. In all three approaches, feature importance was evaluated according to the decrease on the correlation factor between resulting scores and reference grades in each case: A more significant decrease indicates greater feature importance.

Approach 1 indicated the importance sequence (from the most to the least important feature):

$$RSV, T_{60}, E_d/E_r, d, C_{50}, CT, ITDG, D_{50}, E, V(\text{V3}), \quad (16)$$

whereas Approach 2 yielded the sequence

$$ITDG, C_{50}, E_d/E_r, d, V(\text{V3}), CT, E, RSV, T_{60}, D_{50}. \quad (17)$$

Both approaches coincided in having the direct-to-reverberant energy ratio and the distance source-to-microphone among the four most important measures but they generate sort of contradictory results with respect to other features and to Table III.

Applying Approach 3, the accumulated eigenvalues of the whole training data covariance matrix is given by $[98, 99, 100, 100, 100, 100, 100, 100, 100, 100]$, which means that only 3 coordinates should contain all information provided by the full 10-dimensional data representation.

Table IV includes the statistical correlation between the ANN results and the subjective scores provided for both databases. The ANN setups included the full 10-dimensional case and the results from all three approaches to the 4-, 3-, and 2-dimensional cases.

TABLE IV
STATISTICAL CORRELATION (%) BETWEEN ANN OUTPUTS AND SUBJECTIVE SCORES FOR THE TRAINING AND TEST DATABASES.

| Input Dimension | Training | Test |
|---|---|---|
| 10 | **93** | **75** |
| 4 (Approach 1) | **92** | 81 |
| 4 (Approach 2) | 47 | 52 |
| 4 (Approach 3) | 88 | **85** |
| 3 (Approach 1) | **92** | **85** |
| 3 (Approach 2) | 46 | 47 |
| 3 (Approach 3) | 84 | **85** |
| 2 (Approach 1) | **89** | **87** |
| 2 (Approach 2) | 35 | 05 |
| 2 (Approach 3) | 31 | 15 |

From Table IV, the lowest ANN dimension that sustained a high correlation factor was 3, as predicted by the PCA method. In this case, Approach 1 yielded 92% for the training database, employing features $RSV$, $T_{60}$, and $E_d/E_r$, as indicated in Eq. (16). Comparing these results with the ones in Table IV, it can be observed that $T_{60}$ alone could not deal with the problem of estimating reverberation quality, justifying the use of an ANN to combine several features for this purpose. By reducing the ANN input dimension from 10 to 3, non-essential information is discarded, causing the correlation increase observed in the testing stage.

The training stage for the 3-dimensional input ANN, as determined by Approach 1, comprised 11 epochs and yielded a final mean squared error (MSE) between resulting and reference scores equal to MSE=0.0637, as depicted in Figure 3. The test scores for this ANN setup are seen in Figure 4, whose irregular behavior is also observed in standard methods such as PESQ [20]. Proposed ANN system compares favorably well with state-of-the-art method which achieves 80% correlation with subjective scores provided for the MARDY database [21].

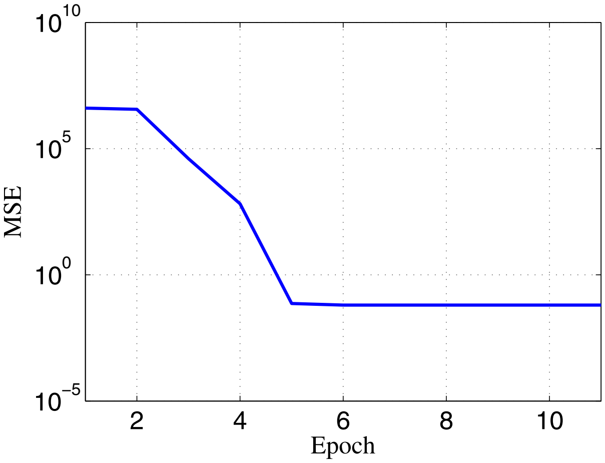

Fig. 3. ANN training performance for 3-dimensional input determined by Approach 1 reaching MSE=0.0637.

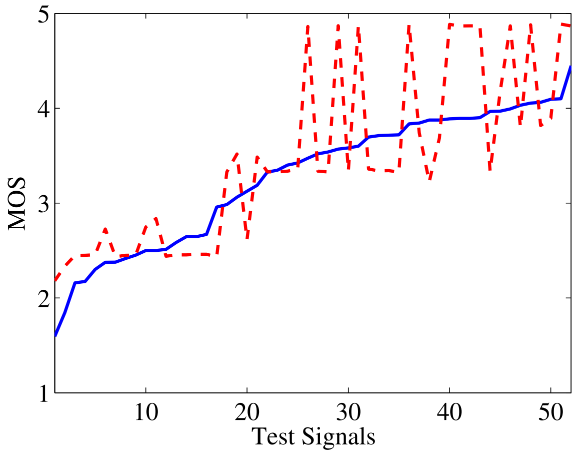

Fig. 4. Subjective (solid line) and estimated (dashed line) scores for test database using 3-dimensional input determined by Approach 1 achieving 85% correlation factor.

## IV. Conclusion

This work addressed the estimation of the perceived reverberation effect using several characteristics extracted from the associated RIR. For this matter, a set of 10 features were analyzed individually and in combination through an ANN. Several experiments comprising artificial and real reverberation effects indicated that the ANN could efficiently map the input features onto a reverberation MOS-like scale. Input relevance tests indicated reverberation time, room spectral variance and direct-to-reverberant energy ratio as the three main features for reverberation assessment.

## Acknowledgments

Authors would like to thank Dr. Luiz W. P. Biscainho, for a suggestion that originated the volume estimator V3, and Rodrigo C. Torres, MSc, for his valuable assistance on the ANN training stage.

## References

[1] ITU-R Rec. BS.1387-1, *Method for Objective Measurements of Perceived Audio Quality*, 2001.

[2] S. Haykin, *Neural Networks, A Comprehensive Foundation*, 2nd ed., Upper Saddle River, NJ: Prentice-Hall, 1999.

[3] M. R. Schroeder, “Integrated-impulse method measuring sound decay without using impulses,” *J. Acoustic. Soc. Am.*, vol. 66, pp. 497–500, Aug. 1979.

[4] A. Lundeby, T. E. Vigran, H. Bietz, and M. Vorlander, “Uncertainties of measurements in room acoustics,” *Acustica*, vol. 81, pp. 344–355, 1995.

[5] 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.

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

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

[8] 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.

[9] F. L. Figueiredo and F. Iazzetta, “Comparative study of measured acoustic parameters in concert halls in the city of Sao Paulo.” *Proc. Int. Congress and Exposition on Noise Control Engineering,* Rio de Janeiro, Brazil, Aug. 2005.

[10] N. Toma, M. D. Topa, V. Popescu and E. Szopos, “Comparative performance analysis of artificial reverberation algorithms,” *Proc. IEEE Int. Conf. Automation, Quality and Testing Robotics,* Cluj-Napoca, Romania, pp. 138–142, May 2006.

[11] 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.

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

[13] E. Larsen, A. Feng, and D. L. Jones, “Room volume and room dimension estimation,” US Patent 2006/0126858 A1, June 2006.

[14] A. A. de Lima, F. P. Freeland, P. A. A. Esquef, L. W. P. Biscainho, B. C. Bispo, R. A. de Jesus, S. L. Netto, R. Schafer, A. Said, B. Lee, and A. Kalker, “Reverberation assessment in audioband speech signals for telepresence systems,” *Proc. Int. Conf. Signal Processing in Multimedia Applications*, Porto, Portugal, pp. 257–262, July 2008.

[15] 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.

[16] 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.

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

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

[19] I. T. Jolliffe, *Principal Component Analysis*, 2nd ed., Springer, 2002.

[20] 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.

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