Proceedings of Meetings on Acoustics

Volume 19, 2013 http://acousticalsociety.org/

ICA 2013 Montreal
Montreal, Canada
2 - 7 June 2013

Signal Processing in Acoustics
Session 2pSP: Acoustic Signal Processing for Various Applications

2pSP2. Towards blind reverberation time estimation for non-speech signals

Joao F. Santos*, Nils Peters and Tiago H. Falk

***Corresponding author's address: INRS-EMT, Institute National de la Recherche Scientifique, Montréal, H5A-1K6, Québec, Canada, jfsantos@emt.inrs.ca**

Reverberation time (RT) is an important parameter for room acoustics characterization, intelligibility and quality assessment of reverberant speech, and for dereverberation. Commonly, RT is estimated from the room impulse response (RIR). In practice, however, RIRs are often unavailable or continuously changing. As such, blind estimation of RT based only on the recorded reverberant signals is of great interest. To date, blind RT estimation has focused on reverberant speech signals. Here, we propose to blindly estimate RT from non-speech signals, such as solo instrument recordings and music ensembles. To estimate the RT of non-speech signals, we propose a blind estimator based on an auditory-inspired modulation spectrum signal representation, which measures the modulation frequency of temporal envelopes computed from a 23-channel gammatone filterbank. We show that the higher modulation frequency bands are more sensitive to reverberation than the modulation bands below 20 Hz. When tested on a database of non-speech sounds under 23 different reverberation conditions with reverberation time (T40) ranging from 0.18 to 15.62 s, a blind estimator based on the ratio of high-to-low modulation frequencies outperformed two state-of-the-art methods and achieved correlations with EDT as high as 0.92 for solo instruments and 0.87 for ensembles.

Published by the Acoustical Society of America through the American Institute of Physics

 [DOI: 10.1121/1.4800598]
Received 21 Jan 2013; published 2 Jun 2013
Proceedings of Meetings on Acoustics, Vol. 19, 055059 (2013)

INTRODUCTION

Reverberation plays an important role on the quality of a sound signal produced on an enclosed environment. Speech signal intelligibility, for example, is severely degraded in highly reverberant environments, as perceptual artifacts such as coloration and echoes are added to the direct sound signal. For non-speech signals such as music, however, the effect is not always prejudicial: totally anechoic music signals are usually perceived as not natural by listeners but, on the other hand, high reverberation times do degrade music quality and intelligibility [1].

Quantifying reverberation is usually done by using the reverberation time (RT), which is the time it takes for a sound to decay by a given amount (typical values are between 10-60 dB) after its source has become inactive. Many standardized methods exist for measuring the reverberation time and related features from the room impulse response (RIR) of a given environment. However, there are cases where such information is not available and estimating the reverberation time directly from an observed reverberant audio signal is necessary. Such methods are called blind, as they are able to estimate the RT without access to the RIR.

Many blind reverberation time estimation methods exist, but most of them are tailored for speech signals. A few works target explicitly blind reverberation time estimation from non-speech signals such as music signals but present some shortcomings: the method described on [2], for example, finds only a rough RT estimate for using as a parameter for a deverberation algorithm. The method shown on [3] requires recording the reverberant signal using multiple sensors, as it is based on the spatial coherence of the signal.

In this work, we have evaluated the performance of two blind RT estimation methods which were originally designed for speech signals and an intrusive quality metric when estimating RT related features from reverberant musical signals. These metrics, which are described in Section 2.2, were compared to reference estimates obtained directly from RIRs (described on 2.1). In Sections 2.3 and 2.4, the databases and experimental setup are presented, and the results are reported and discussed in Section 3. Finally, we present our conclusions in Section 4.

MATERIALS AND METHODS

Reverberation Metrics Based on the Room Impulse Response

Many reverberation-related measurements can be calculated by using the impulse response of a given environment [4]. These measurements may be related to perceptual qualities of sounds produced under such environment. Some characteristics that may be estimated based on the energy of the RIR, where the impulse response is noted as $g(t)$ and its energy as $[g(t)]^2$, are:

Clarity (C) is the logarithmic ratio between the initial energy of the impulse response (t_0 equal to 50 or 80 ms) and the remainder energy, given by:

$$C = 10 \log_{10} \left(\frac{\int_0^{t_0} [g(t)]^2 \, \mathrm{dt}}{\int_{t_0}^{\infty} [g(t)]^2 \, \mathrm{dt}} \right) \tag{1}$$

This value is intended to characterize the clarity and transparency of music in a concert hall and its typical range is from about -5 to +3 dB.

Definition (D) the ratio between the initial energy of the impulse response (again, with t_0 equal to 50 or 80 ms) and the total energy:

$$D = \frac{\int_0^{t_0} [g(t)]^2 \, \mathrm{dt}}{\int_0^{\infty} [g(t)]^2 \, \mathrm{dt}} \times 100\% \tag{2}$$

This measure is related to distinctness of sound which may be subjectively assessed as, for example, the intelligibility of speech on a given room.

Central Time (CT) the gravity center of the whole impulse response in milliseconds, given by:

$$\mathrm{CT} = \frac{\int_0^\infty [g(t)]^2 t \, \mathrm{dt}}{\int_0^\infty [g(t)]^2 \, \mathrm{dt}} \tag{3}$$

A higher CT means that the energy is spread over time, which means the perceived effects of the reverberation last longer.

Early Decay Time (EDT) and Reverberation Time (T) EDT is the decay time for the first 10 dB of the RIR energy. Times for other decays (such as 20 dB, 30 dB, 40 dB, and 60 dB) are also frequently measured. As in some cases the measured RIR does not have the necessary dynamic range for computing some of this measures, the standard procedure is to extrapolate the decay curve to find the reverberation times.

Blind Reverberation Time Estimation Methods

RSMR

The Speech to Reverberation Modulation Ratio, or SRMR, is a non-intrusive objective measure of speech quality and intelligibility proposed by Falk et al. [5]. It is based on the estimation of spectral modulation energy shifts, across frequency, caused by early and late reflections. The measure considers eight modulation frequency bands, from 4 to 128 Hz, for each of the 23 acoustic frequency bands of a gammatone filterbank. From this representation, it measures the energy shift from lower to higher spectral modulation bands. The inverse of the SRMR measure, that is the ratio of the total energy on the higher bands to the lower bands), is called here RSMR. It correlates well with the reverberation time, as shown in [6]; there, the authors have employed a training stage, using speech signals with known reverberation time, to find a 2^{nd} order polynomial mapping function between the ratio values and the RT_{60} values.

Löllmann's method

In [7], the authors have proposed a maximum likelihood approach to reverberation time estimation. This approach, as the one previously proposed in [8], is based on a statistical model of sound decay of reverberant speech, and considers also a statistical model for the acoustic impulse response. In order to reduce computational complexity, the speech signals are downsampled prior to RT estimation (as downsampling does not affect the energy decay behavior). The dowmsampled signal is divided into frames, which are then divided again in subframes. On these subframes, the energy, maximum and minimum value are checked to detect possible sound decay regions. ML estimates are then computed for each sound decay region and a smoothed histogram of these estimates is then used to obtain the final RT estimate. This method was also evaluated in [6], where its performance was shown to be similar to RSMR for short reverberation times (0.2-1.0s).

PEAQ

PEAQ [9] is the ITU standard for objective measurement of perceived audio quality. PEAQ is based on a psychoacoustic model, which takes into account both the peripheral ear and higher level processing stages. Two peripheral ear models, one based on Fast Fourier Transform representations and the other on filterbanks, transform the input into excitation patterns.

These patterns are then processed by the cognitive model, which calculates several features (model output values) that are then used by an artificial neural network to compute a distortion index and an objective difference grade between the clean and distorted signals. While PEAQ is not a reverberation time estimation metric, it is related to the amount of distortion on the target signal. PEAQ is not a blind measure, as it requires access to the clean (anechoic) signal in order to estimate the amount of distortion.

Non-Speech Sound and Room Impulse Response Databases

Anechoic sound recordings from musical ensembles and solo instruments were used to evaluate the performance of the RT estimation methods. A total of 24 musical ensemble recordings was used. Fifteen files come from [10], and another four were generated by mixing anechoic recordings of solo instruments provided in [11]. Each solo instrument channel was pre-processed using a noise gate prior to the mixing, in order to reduce background noise. For the solo instrument tests, 248 recordings were used; the files in [11] included the following orchestral instruments in various playing styles: bassoon, clarinet, flute, French horn, timpani, trumpet, viola, violin, cello, double bass, and trombone. One of the recordings is from a soprano singing voice. Additional databases [12, 13] were used, including samples from acoustic guitar, various percussion instruments and bagpipes.

For the room impulse responses, the OpenAIR library was used [13]. OpenAIR is a public database of acoustic impulse responses containing RIR obtained from different environments. While the database includes both real and simulated RIRs, only the real RIRs were used in our experiments, to a total of 23 different responses. The database RIRs range from short (T40 = 0.18 s) to very long reverberation times (T40 = 15.62 s).

Experiment setup and evaluation metrics

The experiment setup consisted on the following steps. First, we obtained direct measurements of several RT related metrics directly from the RIRs. We consider these measurements as references, since they had access to the impulse responses. The AcMus toolbox [14] was used for computing all the measurements described in Section 2.1. Then, we convolved all the RIRs with the ensemble and solo recordings to generated reverberant versions of the source signals. For this step, all RIR and source signals were downsampled to 16 kHz. While some of the source signals were stereo, only the left channel was used as SRMR and Löllmann's method operate on single channel signals only. The resulting files were then processed using the three methods described on the previous section. PEAQ's basic configuration was used, and files had to be upsampled to 48 kHz prior to processing as the implementation did not support other sampling rates [15].

For the metrics evaluation, we computed the mean values for all reverberant files under the same RIR and used them to find the Pearson correlation coefficient for each method. Correlation coefficients were computed for ensemble files and solo instruments separately. Since PEAQ's maximum value is zero (for no distortion), the absolute value was used to compute the correlations so the correlation coefficients would have the same signal as for the other methods.

Results and discussion

Figures 1 and 2 show, respectively, the heat maps for the Pearson correlation coefficients between the estimated values from the sound files and the features obtained directly from the RIRs. Each of the values has three versions, being "A" and "C" the values obtained by using the A- and C-weighting filters from the IEC standard [16] and "Linear" the values obtained without

weighting. Clarity and definition are inversely proportional to the reverberation time, while all other measurements are directly proportional.

RSMR obtained the highest correlations of the three evaluated methods, with results as high as 0.87 for the early decay time in ensemble recordings, while Löllmann's method and PEAQ had correlations of 0.60 and 0.33 respectively. For solo instruments, RSMR obtained similar results; however, correlations decreased for the Löllmann method and increased for PEAQ, both being around 0.50 in this case.

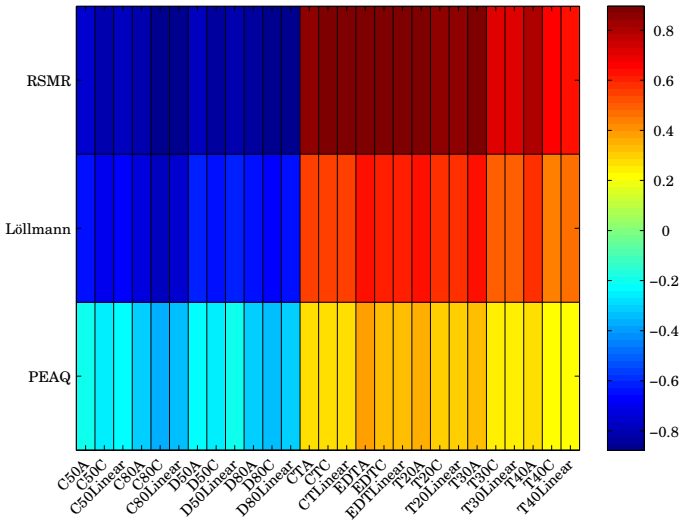


FIGURE 1: Heat map for the ensemble correlations

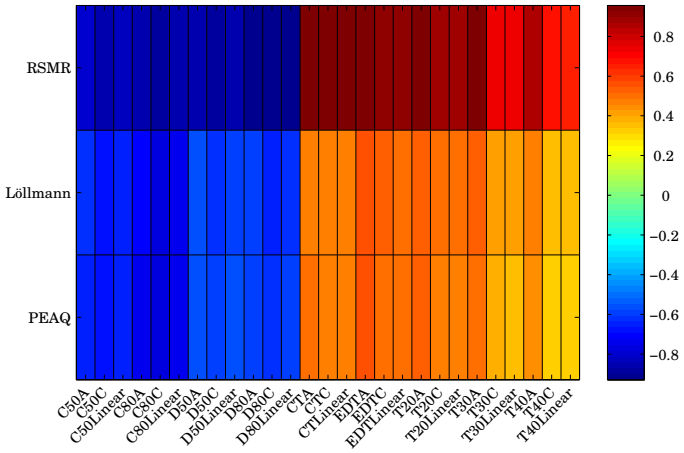
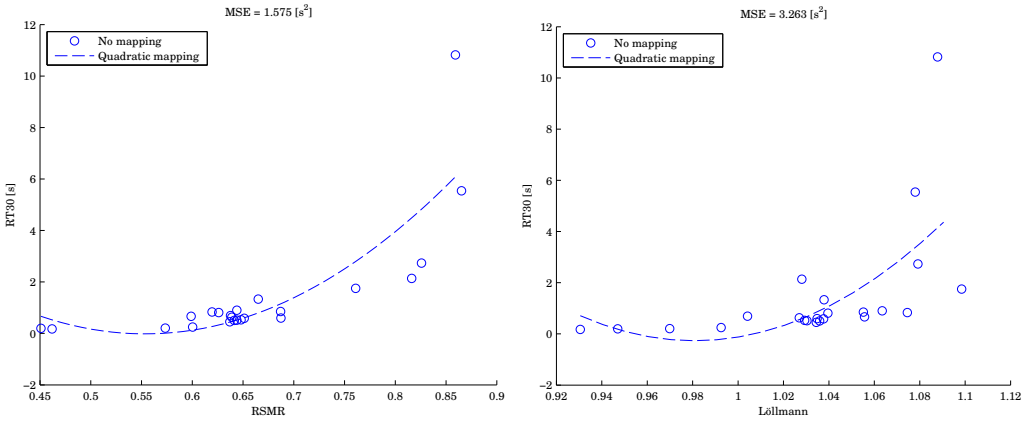


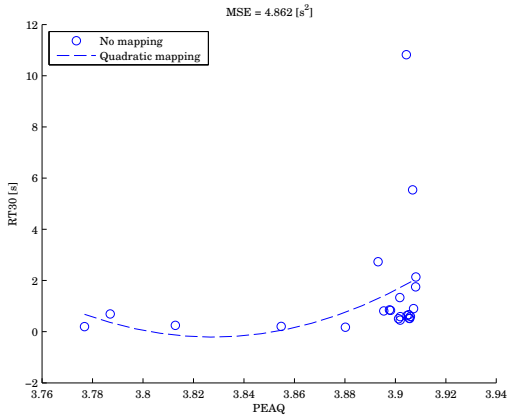
FIGURE 2: Heat map for the solo instrument correlations

We also examined the mean square error of the T30Linear predictions for each method. A quadratic mapping function was used to map the measure values into T30Linear values, in seconds. This mapping step was also performed for Löllmann's method. While the original method results are values in seconds for the T60 RT metric, we noted that the mapping was not appropriate for non-speech signals, as the value range was between 0.9 and 1.1 and our reverberation time range for T30 values (which should be lower than T60) were between 0.1 and 11 s approximately. The mean square error was computed between the same points used for computing the correlation and the "true" T30Linear value (which was estimated directly from

the RIR). The results for RSMR, Löllmann's and the PEAQ method can be seen on Figures 3a, 3b, and 3c respectively. RSMR shows the best MSE score, of 1.575 s^2, against 3.263 s^2 for Löllmann's method and 4.862 s^2 for PEAQ. It can also be noted that all the evaluated methods show higher errors for high T30 values, however, the mapping found for the measured RSMR values predicts values up to 6 s, while the other methods have a more limited range.



(A) Quadratic mapping for the RSMR method **(B)** Quadratic mapping for Löllmann's method



(C) Quadratic mapping for the PEAQ method

FIGURE 3: Mapping functions and mean square errors for the three RT estimation methods

CONCLUSIONS

We compared the performance of two state of the art blind reverberation time estimation methods for non-speech signals. These methods have previously been demonstrated to have similar performances for speech signals [6]. However, this is not the case for non-speech signals, as RSMR has shown higher correlation coefficients than Löllmann's method: correlations as high as 0.87 and 0.92 were found for early decay time estimation for ensembles and solo instruments respectively. Both RSMR and Löllmann's method were tested using the same parameters as employed for speech signals. We also included PEAQ, an intrusive quality metric for audio signals, in our evaluation; still, it has shown poor performance both for musical ensemble and solo instrument signals.

Acknowledgements

The authors wish to acknowledge the Natural Sciences and Engineering Research Council of Canada for their financial support.

References

[1] N. Valverde and M. Fernandez, "Musical perception within a highly reverberant room", Journal of the Acoustical Society of America **123**, 3797 (2008).

[2] A. Tsilfidis and J. Mourjopoulos, "Blind single-channel dereverberation for music post-processing", in *Audio Engineering Society 130th Convention* (London, UK) (2012).

[3] R. Scharrer and M. Vorländer, "Blind Reverberation Time Estimation", in *International Congress on Acoustics*, August, 1–6 (2010).

[4] H. Kuttruff, *Room Acoustics* (Spon Press) (2000).

[5] T. H. Falk, C. Zheng, and W.-Y. Chan, "A Non-Intrusive Quality and Intelligibility Measure of Reverberant and Dereverberated Speech", IEEE Transactions on Audio, Speech, and Language Processing **18**, 1766–1774 (2010).

[6] N. D. Gaubitch, M. Jeub, T. H. Falk, P. A. Naylor, P. Vary, M. Brookes, and H. Löllmann, "Performance comparison of algorithms for blind reverberation time estimation from speech", in *Proceedings of the IWAENC 2012*, 2–5 (2012).

[7] H. Löllmann and E. Yilmaz, "An improved algorithm for blind reverberation time estimation", in *International Workshop on Acoustic Signal Enhancement*, 2, 1–4 (2010).

[8] R. Ratnam, D. L. Jones, B. C. Wheeler, W. D. O'Brien, C. R. Lansing, and A. S. Feng, "Blind estimation of reverberation time", The Journal of the Acoustical Society of America **114**, 2877 (2003).

[9] T. Thiede, W. Treurniet, and R. Bitto, "PEAQ—The ITU standard for objective measurement of perceived audio quality", Journal of the Audio Engineering Society **48**, 3–29 (2000).

[10] Denon, "Anechoic orchestral music recording", Audio CD (1995).

[11] T. Lokki, J. Pätynen, and V. Pulkki, "Recording of anechoic symphony music", The Journal of the Acoustical Society of America **123**, 3936 (2008).

[12] B. . Olufsen, "Music for archimedes", Audio CD (1992).

[13] S. Murphy, Damian T.; Shelley, "OpenAIR: An Interactive Auralization Web Resource and Database", in *Audio Engineering Society Convention 129* (2010).

[14] M. Queiroz, F. Iazzetta, F. Kon, M. H. a. Gomes, F. L. Figueiredo, B. Masiero, L. K. Ueda, L. Dias, M. H. C. Torres, and L. F. Thomaz, "AcMus: an open, integrated platform for room acoustics research", Journal of the Brazilian Computer Society **14**, 87–103 (2008).

[15] P. Kabal, "An Examination and Interpretation of ITU-R BS.1387: Perceptual Evaluation of Audio Quality", Technical Report, McGill University (2002).

[16] International Electrotechnical Commission, "International Standard IEC 61672-1, Electroacoustics - Sound level meters, Part 1: Specifications", (2002).