ALGORITHMS AND EVALUATION ON BLIND ESTIMATION OF REVERBERATION TIME

Jens-Alrik Adrian and Joerg Bitzer

Jade University of Applied Sciences, Institute for Hearing Technology and Audiology (IHA)
Ofener Str. 16/19, D - 26121 Oldenburg

ABSTRACT

In this contribution, we propose an algorithm to analyze early and late reverberation in monaural recordings in an off-line processing framework with emphasis on live recordings. This algorithm is evaluated against known state-of-the-art solutions. Our baseline method uses cepstral mean along signal blocks to acquire an estimation of the reverberation's impulse response which is analyzed with respect to its decay characteristics. Further improvements are a cepstral lifter to increase the method's performance by removing nonrelevant cepstral coefficients and a polynomial of second order to map the results onto final estimates. Results indicate larger deviations in the estimated decay times of late reverberations, while estimates for the early decay times are within the just noticable difference (JND) and deviate only slightly from the true values. State-of-the-art algorithms show small correlation with the true reverberation times.

***Index Terms*—** Blind estimation, reverberation time, cepstral analysis

1. INTRODUCTION

The blind estimation of reverberation time (RT) is a current research issue since it provides valuable information in applications like mobile telephony and hearing aids. Often, those technologies apply dereverberation algorithms to enhance the speech quality and rely therefore on information about the RT at the speaker's position. Three state-of-the-art algorithms for speech are compared by Gaubitch et al. [1] and incorporate methods using spectral decay rate distributions [2], modulation energy ratios [3] and maximum likelihood estimations of the prevailing decay constant of the room impulse response during sound decays [4, 5]. All those methods are optimized for speech and imply signal properties of speech like commonly occurring pauses and/or modulation spectra with most of the energy in low bands.

With a focus on audio restoration, we try to find a robust method to estimate the RT independently from the type and characteristics of the audio signal. Therefore, also music recordings have to be considered as possible audio signal. Some publications address this problem. Hansen [6] uses the auto correlation function to evaluate energy decays within the signal by deploying an ordinary backwards integration proposed by Schroeder [7] but the method lacks accuracy [8]. Kendrick et al. [9] compute an envelope spectrum of the bandpass-filtered signal in the 1 kHz-octave to train an artificial neural network with respect to the true RT. Baskind and Warusfel [10] applied a method utilizing a complex cepstral mean to reconstruct the early reflections within the room impulse response (RIR) governing the signal. The late decay time (RT) has been estimated by analyzing sound decays in the music signal.

Except for parts of [10], every method relies in some way on the evaluation of sound decays to estimate the amount of late reverberation. This constraint does not hold for complex music signals since proper sound decays, i.e. signal pauses, are not necessarily present in every recording. For this reason, an approach is proposed which does not require specific properties of the audio signal.

The remainder of the paper is organized as follows. To introduce the context of the algorithm, a working hypothesis is presented in section 2. In section 3, the algorithm is described and evaluated in section 4. The findings are finally concluded in section 5.

2. WORKING HYPOTHESIS

Since the application for the blind estimation proposed in this work is audio restoration, the need of a real-time implementation does not exist and the method can perform off-line. Furthermore, the audio material is expected to be monaural. Therefore, a benefit of multichannel recordings via microphone arrays or binaural recordings does not exist. The impulse responses (IR) which are associated with the reverberation are treated as stationary over the period of one audio track. This also corresponds to the practice of applying artificial reverberation in music production. In those situations the amount of reverberation time or the direct-to-reverberant ratio (DRR) is held nearly constant over one track. Furthermore, the location of the sources in the audio material are supposed to be stationary as well, so musicians or speakers do not move significantly over the progress of the audio signal. At last, the audio material which is considered will be degraded by stationary additive noise caused e.g. by the analog storage media or the recording devices such as microphones. As a constraint, the range of the considered reverberation times spans from $T_{30,\text{min}} = 0.3\,\text{s}$ up to $T_{30,\text{max}} = 3.0\,\text{s}$ since this covers common values in the field of musical recordings. To introduce a measure for the perceptive quality of the estimation the just noticable difference (JND) as stated in [11] is considered. Thus, differences in RTs of less than 100 ms can be neglected.

3. METHOD

The reverberated signal $x(n)$ is divided into L frames with block length N, block index ℓ and frame shift N_Δ, yielding

$$x(n,\ell) = w(n + \ell N_\Delta)x(n), \quad \text{where } \ell = 1,2,\dots,L \quad (1)$$

denotes the block index and $w(n)$ being a window function (Hamming window) in time domain. The block length N is chosen with $N/f_\mathrm{s} \gtrapprox 6\,\mathrm{s}$ to capture the RIR's full length where f_s is the sampling frequency. This follows from the chosen maximum reverberation time to be considered of $T_{30} = 3.0\,\mathrm{s}$. Due to performance reasons, N corresponds to the next power of two so the actual block length might be greater than 6 s. The frame shift N_Δ was set to $N_\Delta/f_\mathrm{s} \approx 2\,\mathrm{s}$ to ensure enough signal innovation within each subsequent cepstral frame. Every signal block ℓ is transformed into the cepstral domain by the discrete real cepstrum transformation

$$c(\kappa,\ell) = \frac{1}{N}\sum_{k=0}^{N-1} \log\left(|X(k,\ell)|\right) \mathrm{e}^{j2\pi\kappa k/N} \quad (2)$$

with $\kappa = 0,1,\dots,N-1$ and

$$X(k,\ell) = \sum_{n=0}^{N-1} x(n,\ell)\mathrm{e}^{-j2\pi nk/N} \quad (3)$$

with $k = 0,1,\dots,N-1$ and κ being the cepstral variable ($[\kappa/f_\mathrm{s}] = \mathrm{s}$). The resulting short time cepstrum $c(\kappa,\ell)$ can be averaged along the blocks ℓ to obtain $\bar{c}_L(\kappa)$, an estimate of the stationary convolutive part's cepstrum prevailing the input signal $x(n)$.

Subsequently, a lifter $\eta(\kappa)$ is applied to remove cepstral coefficients which do not correspond to the actual RIR's cepstrum (c.f. section 3.1). It is defined by

$$\eta(\kappa) = \begin{cases} 1 & \text{if} \quad \kappa_1 < |\kappa| < \kappa_2 \\ 0 & \text{else} \end{cases} \quad (4)$$

with κ_1 and κ_2 being the edge quefrencies of the rectangular lifter. The liftered cepstral mean $\bar{c}_{L,\mathrm{lift}}(\kappa)$ is transformed back into time domain using a minimum-phase reconstruction-window $\xi_\mathrm{min}(\kappa)$. A minimum phase has been chosen although, strictly, this does not hold for a real RIR. Nevertheless, experiments with real and artificial RIRs have shown that the early and late decay times are less distorted by evaluating the minimum phase part than using the zero phase part.

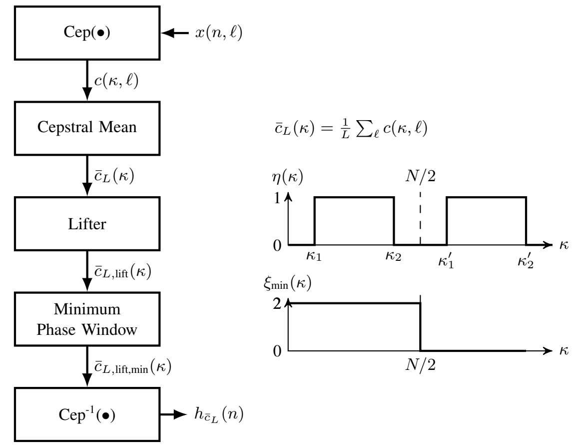

Fig. 1. Algorithm to estimate the stationary impulse response within the audio signal.

The latter would result when both sides of the cepstrum are transformed into time domain. $\xi_\mathrm{min}(\kappa)$ is defined as

$$\xi_\mathrm{min}(\kappa) = \begin{cases} 1 & \text{if} \quad \kappa = 0 \\ 2 & \text{if} \quad \kappa > 0 \\ 0 & \text{if} \quad \kappa < 0 \end{cases}. \quad (5)$$

Applying the two lifters $\xi_\mathrm{min}(\kappa)$ and $\eta(\kappa)$ in the cepstral domain, the mean cepstrum becomes

$$\bar{c}_{L,\mathrm{lift,min}}(\kappa) = \bar{c}_L(\kappa)\cdot\xi_\mathrm{min}(\kappa)\cdot\eta(\kappa) \quad (6)$$

and, thus, the minimum phase part of the impulse response $h_{\bar{c}_L}(n)$ is reconstructed by applying the inverse discrete, real cepstrum.

The described steps to obtain the impulse response corresponding to the reverberation are illustrated in fig. 1 where the Cep(•) operator symbolizes the transformation from equation (2) and Cep^{-1}(•) the inverse transformation, respectively. On the right pane the lifter windows are depicted which were used in the corresponding step on the left pane.

The IR $h_{\bar{c}_L}(n)$ is analyzed by using the Schroeder backwards integration [7] for an estimate of the early decay time ($\widetilde{EDT}$) describing the decay within the first 10 dB of the energy decay curve. For a better comparability with the RT, the EDT is extrapolated to a decay of 60 dB. Furthermore, the late reverberation time ($\widetilde{T}_{30}$) is estimated by applying the method according to Xiang [12], utilizing a non-linear regression approach to handle the occuring noise in the time domain introduced by the averaging process in cepstral domain.

In order to compensate for systematical bias in the results, the estimates are mapped using a polynomial of second order whose coefficients are computed using a robust regression approach (c.f. section 3.2).

3.1. Liftering in cepstral domain

So called liftering in cepstral domain is known from speech processing in order to, for instance, separate the pitch excitation signal from the filter response of the vocal tract [13].

In our approach, the cepstral mean along the frames of short time cepstra show strong components around the origin, as depicted in fig. 2. The shown impulse response computed by Dahl's and Jot's method [14] has been used to reverberate an anechoic orchestral recording of 30 s duration and the cepstral mean $\bar{c}_L(\kappa)$ (c.f. fig. 1) in the lower part of fig. 2 is computed by an arithmetical average of each cepstral bin along the short-time cepstral frames. The strong components at the origin do not correspond to actual cepstral coefficients of the underlying room impulse response which is also plotted in fig. 2 but seem to originate from the audio signal's cepstra at those lower quefrencies. The actual cepstrum of the used Jot impulse response has no decisive peaks below approximately $\kappa = 15$ ms. For this reason, the introduction of a lifter is proposed which cancels those coefficients near the origin that do not correspond to the underlying room impulse response. In order to find an appropriate edge quefrency one has to be aware of the effect of canceling the cepstral coefficients near the origin. Schafer and Oppenheim [15, 16] could show that the cepstrum of a simple IR consisting of a few echoes is zero up to the quefrency which corresponds to the shortest echo delay time. A room impulse response possesses a large number of reflections that can be interpreted as echoes. To find an edge quefrency an assumption of the pre-delay (or initial time delay gap) within a typical RIR would be necessary. Since this is a difficult task, an edge quefrency has been found empirically with $\kappa_1 = 15$ ms. This would cancel the lower quefrencies in the cepstral mean $\bar{c}_L(\kappa)$ in fig. 2 which do not correspond to the RIR's cepstrum and yet leaves the higher coefficients unaltered. Additionally to κ_1, an upper edge quefrency κ_2 has been found beneficial to avoid unnecessary convolutive noise in the reconstructed IR $h_{\bar{c}_L}(n)$ and has been chosen empirically as $\kappa_2 = 600$ ms. κ_1 and κ_2 were found using all available IRs as stated in section 4.

3.2. Mapping of the estimates

In order to map the raw estimates x of the early and late decay times onto the final estimates y_{map}, a polynomial of the form

$$y_{\text{map}}(x) = \sum_{m=0}^{M} \beta_m x^m, \quad \text{with } M = 2 \tag{7}$$

is trained. The order is chosen since additional coefficients do not lead to a better coefficient of determination (R^2). The coefficients β_m were computed using a robust regression by applying an iteratively reweighted least squares algorithm with a bisquare weighting function [17]. This leads to a more robust regression against outliers in the raw estimations. The coefficients and the corresponding R^2 are shown in table 1.

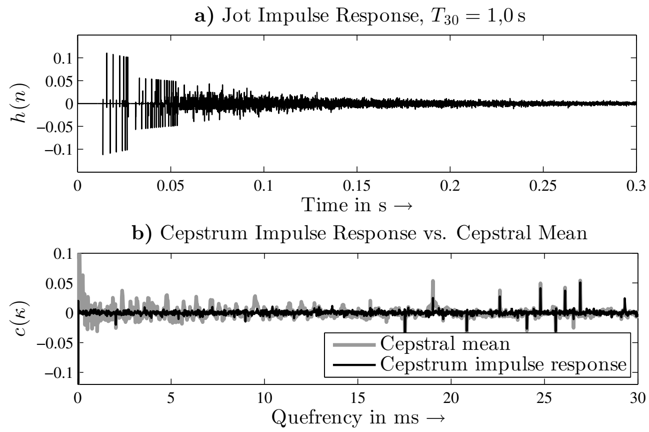

Fig. 2. **a)** Time signal of an impulse response generated using Jot's reverberator [14]. **b)** Cepstral mean computed from an anechoic signal reverberated with the impulse response from a) and the actual cepstrum of the impulse response from a).

Table 1. Coefficients of the polynomial in (7) used to map the estimates onto the final results.

Estimate	β_0	β_1	β_2	R^2
$\widetilde{T}_{30}$	−0.1090	0.8843	0.1773	0.82
$\widetilde{EDT}$	−0.1732	1.0302	0.0881	0.92

For the training, 70% of all available IRs have been used to reverberate the training set. The remaining 30% have been used to reverberate the test set which is subsequently mapped to the final estimates. The anechoic sound files consisted of 18 files with durations between two and five minutes and have been used for both the training and test set.

4. EXPERIMENTS AND RESULTS

For evaluation, we convolved anechoic signals with impulse responses that span the T_{30} range from 0.3 s up to 3.0 s. A sampling rate of $f_s = 16$ kHz has been used throughout all experiments. In this work, the test signals have not been degraded by additive noise and the DRR of the IRs has been adjusted to be exactly 0 dB. The latter follows from the fact that the cepstral influence of the IR is very low for DRR > 0 dB resulting in a cepstral mean which does not correspond to the room's IR. The chosen value of DRR = 0 dB therefore resembles a "worst-case" with respect to a robust estimation.

Since the method focuses mainly on reverberated music signals, in a first experiment a database of anechoic music recordings which includes orchestral music provided by Jukka Pätynen et al. [18] and [19] and anechoic choral recordings provided by Ron Freiheit et al. [20] was formed. Those recordings had durations between 2 min and 5 min and were reverberated with a set of IRs from public sources [21–25] and [14]. All IRs were selected on their true RT using

Schroeder's method [7] prior to estimation to suit the T_{30} value range as stated above. As mentioned in section 3.2, 70% of the total 111 IRs were used as a training set and the remaining 30%, i.e. 33 IRs, formed the test set. The 18 anechoic audio signals were the same for both sets and for all algorithms.

As reference, two algorithms from literature have been evaluated with the same audio material. First, the method by Jeub [5] which was inspired by Löllmann et al. [4] was chosen and is called in the following "ML". The approach focuses on speech signals and incorporates a maximum likelihood estimation (ML) of decay constants in detected signal decays. Since the method works on signal frame basis, a final estimate for the total signal is obtained by averaging the framewise results disregarding values at the beginning and end of the complete signal. Secondly, the approach of Wen et al. [2] was applied on the test signals and is called "SDD" in the following. The latter exploits the spectral decay distribution in time-frequency domain and maps the negative-side variance of the distribution to a final estimate. The parameters used for the proposed method are stated in table 2, both the ML and the SDD algorithm were used with default parameters.

The following measures have been utilized to evaluate the algorithm's performances. The *mean interquartile range* (MIQR) describes the arithmetical average distribution width of the estimates over all J true T_{30} groups (i.e. T_{30} = 0.3 s up to T_{30} = 3.0 s). It is defined by

$$\text{MIQR} = \frac{1}{J}\sum_{j=1}^{J}\left(Q_{j,0.75} - Q_{j,0.25}\right) \tag{8}$$

with $Q_{j,0.75}$ being the 75% quantile of the jth RT group and $Q_{j,0.25}$ the 25% quantile, respectively. The average relative deviation of the medians over the J groups are measured by the *mean relative deviation* (MRD) which is calculated by

$$\text{MRD} = \left(\left\{\prod_{j=1}^{J}\left[1 + \frac{\left|\text{med}\left\{\widetilde{T}_{j,30}\right\} - T_{j,30}\right|}{T_{j,30}}\right]\right\}^{1/J} - 1\right) \cdot 100\% \tag{9}$$

using the geometric mean since we are dealing with an average of ratios. The overall *correlation* (Corr) between estimation medians and ground truth is measured by an ordinary Pearson correlation coefficient ϱ

$$\varrho = \frac{\text{Cov}\left(\text{med}\left\{\widetilde{T}_{30}\right\}, T_{30}\right)}{\sigma\left(\text{med}\left\{\widetilde{T}_{30}\right\}\right)\cdot\sigma\left(T_{30}\right)}. \tag{10}$$

The results for the estimates using music recordings are stated in table 3. Both the reference algorithm ML and SDD show large MIQR of 0.698 s for ML and 1.087 s for SDD respectively which indicate little consistency of the methods for the chosen audio material. Also, the MRD of the reference algorithms show a large average deviation of the medians of about 53% for ML and 46% for SDD indicating a rather large bias. The correlation between true values and medians of the estimates confirm the average deviation of the medians for ML and SDD. Since both reference algorithms were proposed for speech signals and therefore signals with inherent decay parts, the results for music signals and the large set of evaluated RTs is expected as it was assumed in section 1. The proposed method performs in the chosen setup with an MIQR of 0.373 s and MRD of 22%, leading to a correlation of $\varrho = 0.897$ for the estimation of the late decay time $\widetilde{T}_{30}$. The average IQR for the $\widetilde{EDT}$ estimates is even lower with 0.127 s and the medians deviate in average with about 11% leading to a correlation of $\varrho = 0.981$. The $\widetilde{EDT}$ estimates remain therefore almost within the considered JND limits. These findings lead to the assumption that the stronger cepstral coefficients of the early reflections result in more consistent estimates of the underlying impulse response.

In a second experiment, the same setup was used for speech signals of German and English language which were recorded in a rather dry recording booth. Those signals had durations between 2 min and 5 min as well and were convolved with the same training and test set of IRs. The results are shown in table 3. With this chosen test signals containing speech, the reference algorithms perform better in comparison to music signals. The MIQR dropped to each 0.108 s for ML and 0.231 s for SDD but the medians deviate in average with about 49% for ML and 54% for SDD. The correlations show poor values, respectively. This can be explained by a strong dependency on the signal's properties and the large range of chosen RTs for the evaluation. The proposed method performs for the $\widetilde{T}_{30}$ estimates with an MIQR of 0.37 s and MRD of about 29% leading to a correlation of $\varrho = 0.891$. The $\widetilde{EDT}$ estimates show an MIQR of 95 ms and and MRD of about 9% leading to a correlation of $\varrho = 0.981$. Again, the $\widetilde{EDT}$ estimates lie within the considered JND limits.

5. CONCLUSIONS

The proposed method exploiting an enhanced cepstral mean performs well for estimating the early decay time for music and speech signals in an off-line processing framework. This can be explained by the prominent cepstral coefficients of the early reflections. The reference algorithms show poorer results for speech than expected and fail to estimate RTs in music accurately. This leads to the conclusion that an evaluation with a large set of test signals is essential to gain meaningful results. Further investigations are planned to evaluate the proposed method's robustness to test signals degraded by additive noise and the capability of estimating frequency dependent results. In addition, the effects of a DDR > 0 dB and the assumption of stationary IRs must be further evaluated in detail.

Table 2. Parameters used for the proposed algorithm.

N	N_Δ	$w(n)$	$\xi(\kappa)$	κ_1	κ_2
$\geq 6\,\mathrm{s}$	$\approx 2\,\mathrm{s}$	Hamm.	min. phase	$15\,\mathrm{ms}$	$600\,\mathrm{ms}$

Table 3. Performance of the estimates computed by the analyzed algorithms using *music* and *speech* signals. Best performing results are highlighted in bold. The ref. algorithms do not provide $\widetilde{EDT}$ estimates.

Signal	Algorithm	Type	MIQR in s	MRD in %	Corr. ϱ
Music	ML	$\widetilde{T}_{30}$	0.698	53.2	0.592
	SDD	$\widetilde{T}_{30}$	1.087	45.9	0.421
	Proposed	$\widetilde{T}_{30}$	**0.373**	**22.3**	**0.897**
		$\widetilde{EDT}$	**0.127**	**11.2**	**0.981**
Speech	ML	$\widetilde{T}_{30}$	**0.108**	49.3	0.650
	SDD	$\widetilde{T}_{30}$	0.231	53.7	0.489
	Proposed	$\widetilde{T}_{30}$	0.370	**29.4**	**0.891**
		$\widetilde{EDT}$	**0.095**	**09.1**	**0.981**

6. ACKNOWLEDGEMENTS

This research was (partly) funded by grant 03FH030PX2 of the German Federal Ministry of Education and Research (BMBF). The views and conclusions contained in this document, however, are those of the authors. We would like to thank Ron Freiheit from Wenger Corp. and Jukka Pätynen from University of Aalto for providing anechoic recordings and Nikolay Gaubitch for providing code for the reference algorithms.

REFERENCES

[1] N. D. Gaubitch, H. W. Löllmann, M. Jeub, T. H. Falk, P. A. Naylor, P. Vary, and M. Brookes, "Performance comparison of algorithms for blind reverberation time estimation from speech," in *Acoustic Signal Enhancement; Proceedings of IWAENC 2012; International Workshop on*, 2012, pp. 1–4.

[2] Jimi YC Wen, Emanuël AP Habets, and Patrick A. Naylor, "Blind estimation of reverberation time based on the distribution of signal decay rates," in *Acoustics, Speech and Signal Processing, 2008. ICASSP 2008. IEEE International Conference on*, 2008, p. 329–332.

[3] Tiago H. Falk and Wai-Yip Chan, "A non-intrusive quality measure of dereverberated speech," in *Proc. Int. Workshop on Acoustic Echo and Noise Control (IWAENC)*, 2008.

[4] Heiner Löllmann, Emre Yilmaz, Marco Jeub, and Peter Vary, "An improved algorithm for blind reverberation time estimation," in *Proceedings of International Workshop on Acoustic Echo and Noise Control (IWAENC)*, 2010, p. 1–4.

[5] Marco Jeub, *Joint dereverberation and noise reduction for binaural hearing aids and mobile phones*, Ph.D. thesis, RWTH Aachen, 2012, http://darwin.bth.rwth-aachen.de/opus/volltexte/2012/4235/.

[6] Martin Hansen, "A method for calculating reverberation time from musical signals," Technical report, Technical University of Denmark, 1995.

[7] M. R. Schroeder, "New method of measuring reverberation time," *The Journal of the Acoustical Society of America*, vol. 37, no. 3, pp. 409–412, 1965.

[8] Jens Schröder, Thomas Rohdenburg, Volker Hohmann, and Stephan D. Ewert, "Classification of reverberant acoustic situations," in *Proceedings of the International Conference on Acoustics NAG/DAGA*, 2009, p. 606–609.

[9] Paul Kendrick, Trevor J. Cox, Francis F. Li, Yonggang Zhang, and Jonathon A. Chambers, "Monaural room acoustic parameters from music and speech," *The Journal of the Acoustical Society of America*, vol. 124, no. 1, pp. 278–287, 2008.

[10] Alexis Baskind and Olivier Warusfel, "Methods for blind computational estimation of perceptual attributes of room acoustics," *Watermark*, vol. 1, 2012.

[11] Heinrich Kuttruff, *Room acoustics*, Applied Science Publishers, London :, 2d ed. edition, 1979.

[12] Ning Xiang, "Evaluation of reverberation times using a nonlinear regression approach," *The Journal of the Acoustical Society of America*, vol. 98, pp. 2112, 1995.

[13] Jacob Benesty, M. Mohan Sondhi, and Yiteng (Arden) Huang, *Springer Handbook of Speech Processing*, Springer-Verlag New York, Inc., Secaucus, NJ, USA, 2007.

[14] Luke Dahl and Jean-Marc Jot, "A reverberator based on absorbent all-pass filters," in *DAFx-00: Proceedings*, 2000.

[15] Ronald W Schafer, "Echo removal by discrete generalized linear filtering," Tech. Rep., DTIC Document, 1969.

[16] A.V. Oppenheim and R.W. Schafer, *Digital Signal Processing*, Prentice Hall International Editions. Prentice Hall, 1975.

[17] John Fox, "Robust regression," *An R and S-Plus companion to applied regression*, 2002.

[18] Jukka Patynen, Ville Pulkki, and Tapio Lokki, "Anechoic recording system for symphony orchestra," *Acta Acustica united with Acustica*, vol. 94, no. 6, pp. 856–865, 2008.

[19] Denon, "Anechoic orchestral recordings," 1989, CD.

[20] Ron Freiheit, John Alexander, and John Ferguson, "Making an anechoic choral recording," *The Journal of the Acoustical Society of America*, vol. 118, no. 3, pp. 2020–2021, 2005.

[21] Angelo Farina, "Room impulse responses," *Available online at: http://pcfarina.eng.unipr.it/Public/IMP-RESP/ [Last viewed 29-Okt-2013]*, 2013.

[22] University of York, "Open air library," *Available online at: http://www.openairlib.net/ [Last viewed 29-Oct-2013]*, Oct. 2013.

[23] M. Jeub, M. Schafer, and P. Vary, "A binaural room impulse response database for the evaluation of dereverberation algorithms," in *Digital Signal Processing, 2009 16th International Conference on*, July 2009, pp. 1–5.

[24] Juha Merimaa, Timo Peltonen, Tapio Lokki, and Akukon Oy Consulting Engineers, "Concert hall impulse responses-pori, finland: Analysis results," *Available online at: http://www. acoustics.hut.fi/projects/poririrs [Last viewed 26-Jun-2013]*, 2005.

[25] Fokke van Saane, "Impulse responses," *Available online at: http://fokkie.home.xs4all.nl/IR.htm [Last viewed 29-Oct-2013]*, 2004.