BLIND REVERBERATION TIME ESTIMATION BY INTRINSIC MODELING OF REVERBERANT SPEECH

Project Report

by

Divya Prakash (11260)
Swapnil Shwetank Jha (11753)
Rahul Bakolia (11559)
Anirudh Agrawal (11098)
Rohit Agrawal(11612)

Project Supervisor: Prof. Rajesh M. Hegde
TA mentor: Mr. Karan Nathwani

Department of Electrical Engineering
IIT Kanpur

Acknowledgement

We have taken efforts in this project. However, it would not have been possible without the kind support and help of many individuals. We would like to extend our sincere thanks to all of them.

We are highly indebted to Prof Rajesh Hegde for his guidance and constant supervision as well as for providing necessary information regarding the project & also for their support in completing the project.

We would like to express our special gratitude and thanks to Mr. Karan Nathwani for giving us such attention and time.

Our thanks and appreciations also go to our colleagues in developing the project and people who have willingly helped us out with their abilities.

Table of Contents

Topic	Page No.
Abstract	4
Introduction	5
Algorithm	6-7
Problem Formulation	8
Decay Rate Distribution	9
Model for training	10-11
Testing Reverberation time	12-13
Results	14-17
References	18

Abstract

The reverberation time (RT) is a very important measure that quantifies the acoustic properties of a room and provides information about the quality and intelligibility of speech recorded in that room. In a recent study, it has been shown that existing methods for blind estimation of the RT are highly sensitive to noise. In this paper, a novel method is proposed to blindly estimate the RT based on the decay rate distribution. Firstly, a data driven representation of the underline decay rates of several training rooms is obtained via the Eigen Value Decomposition of a specially tailored kernel. Secondly, the representation is extended to a room under test and used to estimate its decay rate and hence it's RT. The presented results show that the proposed method outperforms a competing method and is significantly more robust to noise.

Introduction

The reverberation time (RT) is a very important measure that quantifies the acoustic properties of a room. The RT is defined as the time it takes for the sound to decay by 60dB once the source has been switched off. The RT highly depends on the room geometry and the reflectivity of the surfaces in the room. In contrast to the room impulse response (RIR), the RT is independent of the source-microphone configuration. An estimate of the RT of a room can serve as an indicator of the quality and the intelligibility of speech observed in that room.

Both signal based and channel based methods have been developed to estimate the RT. The channel based method require an estimate of the RIR. Although, this provides accurate estimates of the RT, it may not always be practical or even possible to measure the RIR in a room. Therefore, it is desirable to be able to estimate the RT directly from an observed reverberant speech signal. Several methods have been proposed to blindly estimate the RT. Wen et al. proposed a method that blindly estimates the RT by analyzing the distribution of decay rates of the observed reverberant speech signal. The authors shown that the negative-side variance of the distribution can be related to the RT.

In this paper, a novel method to blindly estimate the RT based on the decay rate distribution has been presented. Instead of using a specific characteristic of the distribution, the proposed method empirically reveals the most significant underlying parameter of the decay rates of the observed reverberant speech signal. It is shown that this parameter is strongly related to the decay rate of the room. Firstly, a data-driven representation of the underlying decay rates of several training rooms is obtained via the eigenvalue decomposition of a kernel. Unlike common kernel methods, this kernel is built based on a specially-tailored distance between the observable decay rate distributions of the reverberant speech and is shown to uncover intrinsic geometric information on the underlying parameter. Secondly, the representation is extended to a room under test and used to estimate its decay rate (and hence it's RT). A major advantage of the proposed method is its robustness to additive noise.

Algorithm

Training Stage

Get several recording of reverberant speech from various training rooms

Compute the decay rates of the reverberant speech signals in short time frames

For each recording, compute a histogram of the obtained decay rates

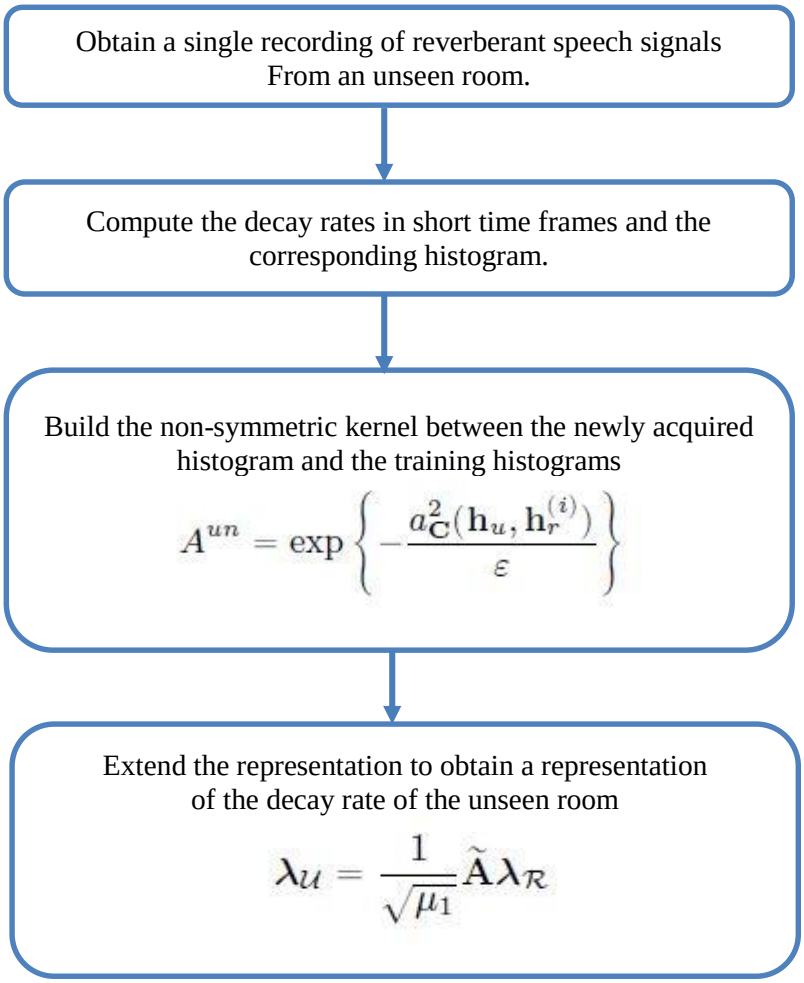
For each training room, compute the covariance matrix of the histograms of different recordings for this room

$$\mathbf{C}_r = \frac{1}{L}\sum_{i=1}^{L}(\mathbf{h}_r^{(i)} - \bar{\mathbf{h}}_r)(\mathbf{h}_r^{(i)} - \bar{\mathbf{h}}_r)^T$$

Build a kernel between the histograms. Apply EVD on the kernel and view the values of its principal eigenvector as data-driven representation of the underlying decay rates of the training rooms.

$$W_{\mathcal{R}}^{nm} = \exp\left\{-\frac{d_{\mathbf{C}}^2(\mathbf{h}_r^{(i)}, \mathbf{h}_\rho^{(j)})}{\varepsilon}\right\}$$

Testing Stage for a single room



Problem Formulation

We assume that each room is characterized by merely a single decay rate value of the energy envelope, which is independent of the frequency. The characteristic decay rate λ_r of a room r is related to RT by

$$\mathrm{RT} = -6\ln(10)/\lambda_r.$$

We rely on the fact that there is a one-to-one mapping between the decay rate of a room and the RT. Estimating the decay rate of a room and estimating the RT are therefore considered equivalent tasks.

Let R be a collection of training rooms with various known characteristic decay rates. In each room $r \in R$ with a characteristic decay rate λ_r, we perform L recordings and collect a set of L reverberant speech signals, denoted by $\{x_r^{(i)}(n)\}_{i=1}^{L}$.

Let λ_s be a random variable that represents the instantaneous decay rate of the energy envelope of an anechoic speech. The decay rates of the room and the anechoic speech are unobservable and may be estimated via the measured reverberant speech. Reverberant speech in an enclosure is usually modeled as the convolution of the anechoic speech signal and the RIR. Thus, the energy envelope of the reverberant speech signal can be viewed as a function of the energy envelopes of the anechoic speech signal and the RIR. Let λ_x denote the *observable* instantaneous decay rate of the energy envelope of the reverberant speech, which can be written as

$$\lambda_x = g(\lambda_r, \lambda_s)$$

where g is an arbitrary (possibly nonlinear) function.

Our objective in this paper is to recover the decay rate (RT) of a room from λ_x without model assumptions. We assume that accurate estimates of λ_x can be obtained from the observable reverberant speech signal. The decay rates can be estimated in the time-frequency domain according to

$$\tilde{H}(t,f) = P(f)e^{\lambda_h(f)t} \quad \text{for} \quad t \geq 0,$$

where, $\tilde{H}(t,f)$ is the energy envelope of RIR at time t and frequency f. $\lambda_h(f)$ is the decay rate at frequency f and $P(f)$ is the initial Power Spectral Density.

Decay Rate Distribution

From each training recording $x_r^{(i)}(n)$ we compute the decay rates in short time frames. Let $h_r^{(i)}$ denote the histogram of the decay rates corresponding to the i-th recording in room r. We compute the empirical covariance matrix of the histograms from the same training room as follows

$$\mathbf{C}_r = \frac{1}{L}\sum_{i=1}^{L}(\mathbf{h}_r^{(i)} - \bar{\mathbf{h}}_r)(\mathbf{h}_r^{(i)} - \bar{\mathbf{h}}_r)^T$$

Where h_r bar is the empirical mean of the histograms in r, for all $r \in R$. The natural variations of the decay rates in different recordings introduce variations of the corresponding histograms in the observable domain (histograms of the decay rates of the measured reverberant speech). We exploit these variations, as manifested in the covariance matrix C_r to empirically invert the function g and reveal the decay rates.

Model for training

We define a symmetric distance function between pairs of training feature vectors (histograms) as

$$d^2_{\mathbf{C}}(\mathbf{h}^{(i)}_r, \mathbf{h}^{(j)}_\rho) = (\mathbf{h}^{(i)}_r - \mathbf{h}^{(j)}_\rho)^T(\mathbf{C}_r^{-1} + \mathbf{C}_\rho^{-1})(\mathbf{h}^{(i)}_r - \mathbf{h}^{(j)}_\rho)$$

for each r; ρ ϵ *R* for all i, j.

This distance is termed as the Mahalanobis distance and has two important properties. The Mahalanobis distance is invariant to linear transformations. Thus, according to the analysis in previous section, in the features (histograms) domain, this distance is invariant to the distortions imposed on the decay rate by the anechoic speech and noise.

Given the pairwise distances between the desired values, we recover the values themselves through the eigenvalue decomposition (EVD) of an appropriate Laplace operator. Let W_R be an affinity matrix (kernel) between pairs of feature vectors, whose (n, m)-th element is given by

$$W^{nm}_{\mathcal{R}} = \exp\left\{-\frac{d^2_{\mathbf{C}}(\mathbf{h}^{(i)}_r, \mathbf{h}^{(j)}_\rho)}{\varepsilon}\right\}$$

Where $\mathcal{E}$ is the kernel scale and $n = rL + i$ and $m = \rho L + j$.

Let D be a diagonal normalization matrix whose diagonal elements are

$$\mathrm{D}^{\mathrm{nn}} = \sum_m W^{\mathrm{nm}}{}_R .$$

Let

$$\hat{\mathrm{W}}_R = D^{-\frac{1}{2}}\,\mathrm{W}_R\,D^{-\frac{1}{2}}$$

be a normalized kernel that shares the eigenvectors with the normalized graph-Laplacian $I - \hat{\mathrm{W}}_R$.

It can be shown that the eigenvectors Φ_k of $\hat{W}_R$ reveal the underlying structure of the data. In the following, we assume that the decay rate of the room is the most significant underlying parameter of the decay rates of the observed reverberant speech signal.

In particular, the n-th coordinate of the principal eigenvector relates to the decay rate as

$$\Phi^{(n)}{}_1 = f(\lambda^{(i)}{}_r)$$

Where,
$n = rL + i$ and f is a monotonic function.

Thus, the principal eigenvector organizes the feature vectors according to the values of the decay rates of the rooms up to a monotonic scaling. Furthermore, since the decay rates of the training rooms are known, we are able to use them for calibrating the values of the eigenvectors to the values of the decay rates.

Estimating the Reverberation Time

Let U denote a collection of "unseen" rooms with unknown RTs. From each such unseen room $u \in U$, we obtain a *single* reverberant speech recording $x_u(n)$. Based on the reverberant speech, we compute the histograms (feature vectors) h_u for $u \in U$ of the decay rates of the energy envelopes of the signal in short time frames. Now, we present the simultaneous estimation of the RTs of all the unseen rooms, which includes the case of a single unseen room as well.

We define a non-symmetric distance function between feature vectors from the unseen rooms and the training rooms as

$$a_C^2(\mathbf{h}_u, \mathbf{h}_r^{(i)}) = (\mathbf{h}_u - \mathbf{h}_r^{(i)})^T \mathbf{C}_r^{-1} (\mathbf{h}_u - \mathbf{h}_r^{(i)})$$

For each $r \in R$, $u \in U$, and $i = 1, \ldots, L$.

Now we define a corresponding non-symmetric affinity matrix A using a Gaussian as-

$$A^{un} = \exp\left\{ -\frac{a_{\mathbf{C}}^2(\mathbf{h}_u, \mathbf{h}_r^{(i)})}{\varepsilon} \right\}$$

Where $n = rL + i$.
We note that the construction of A relies merely on the observed and training data.

Let $\tilde{A} = D_a^{-1} A \, \omega^{-1}$, where D_a is a diagonal matrix whose diagonal elements are the sum of rows of A, and ω is a diagonal matrix whose diagonal elements the sum of columns are of $D_a^{-1} A$.

Also,

$$\hat{W}_R = \tilde{A}^T \tilde{A},$$

Thus, the eigenvectors Φ_k of $\hat{W}_R$, which represent the training rooms, can be obtained from the right singular vectors of $\tilde{A}$.

Define a new affinity matrix between feature vectors of the unseen rooms as,

$$W_U = \tilde{A}\,\tilde{A}^T$$

The principal eigenvector of W_U represents the underlying desired decay rates of the unseen rooms.

By definition of SVD,

$$\psi_k = \left(\frac{1}{\sqrt{\mu k}}\right) \tilde{A}\,\Phi_k \qquad \text{...(1)}$$

where μk is the k-th eigenvalue of W_U and ψ_k is the k-th eigenvector of W_U (and the left singular vector of $\tilde{A}$).
Thus, for k = 1, we obtain the extension of the representation of the underlying desired decay rates of the unseen rooms.

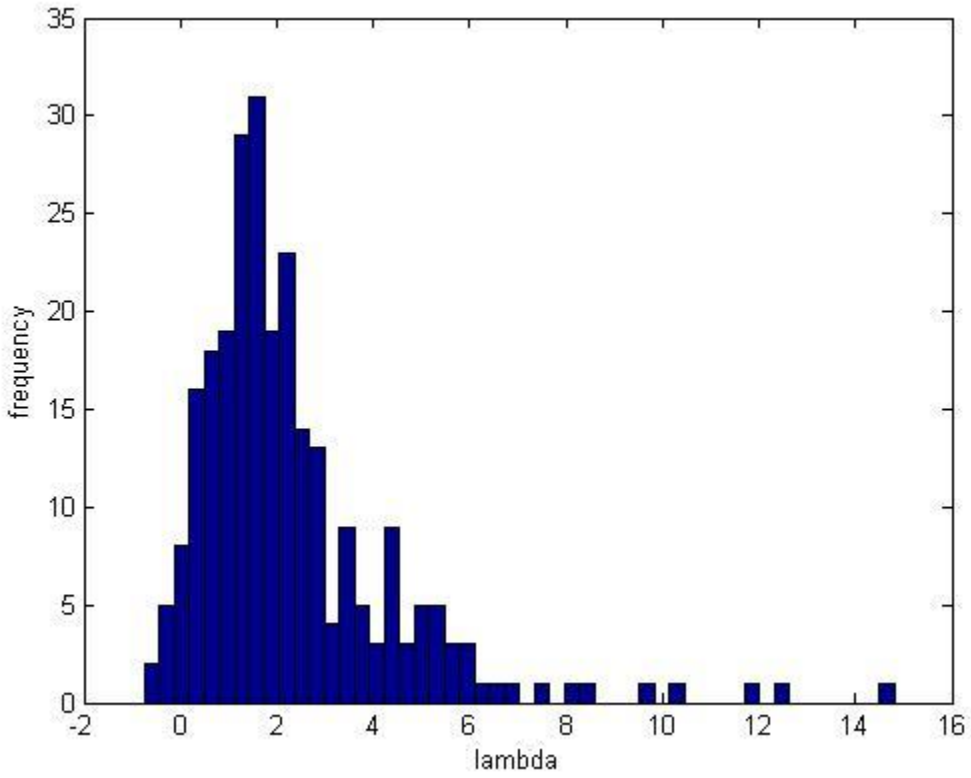
The decay rates (RTs) of all the training rooms are known and can be used to estimate the decay rates (RTs) of the unseen rooms. The SVD expresses the relationship between the representation of the decay rates of the training and unseen rooms. Since the true decay rates of the training rooms are known, we exploit the same relationship to estimate the decay rates of the unseen rooms. Substituting the training decay rates into (1) and setting k = 1 yields

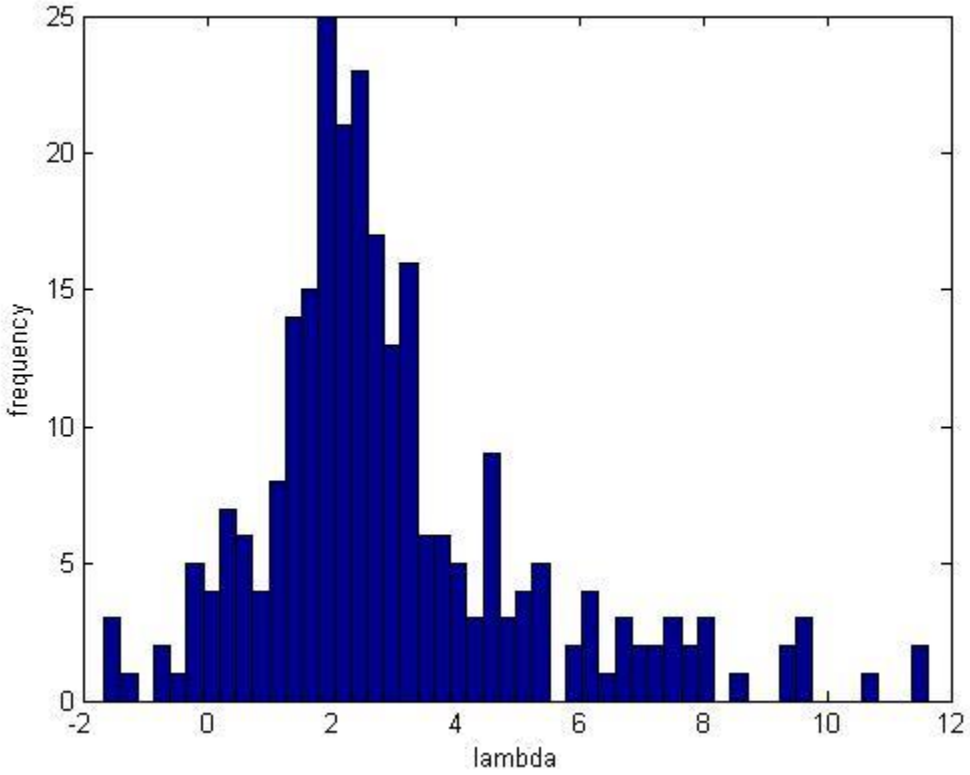
$$\lambda_U = \left(\frac{1}{\mu 1}\right)\tilde{A}\,\lambda_R$$

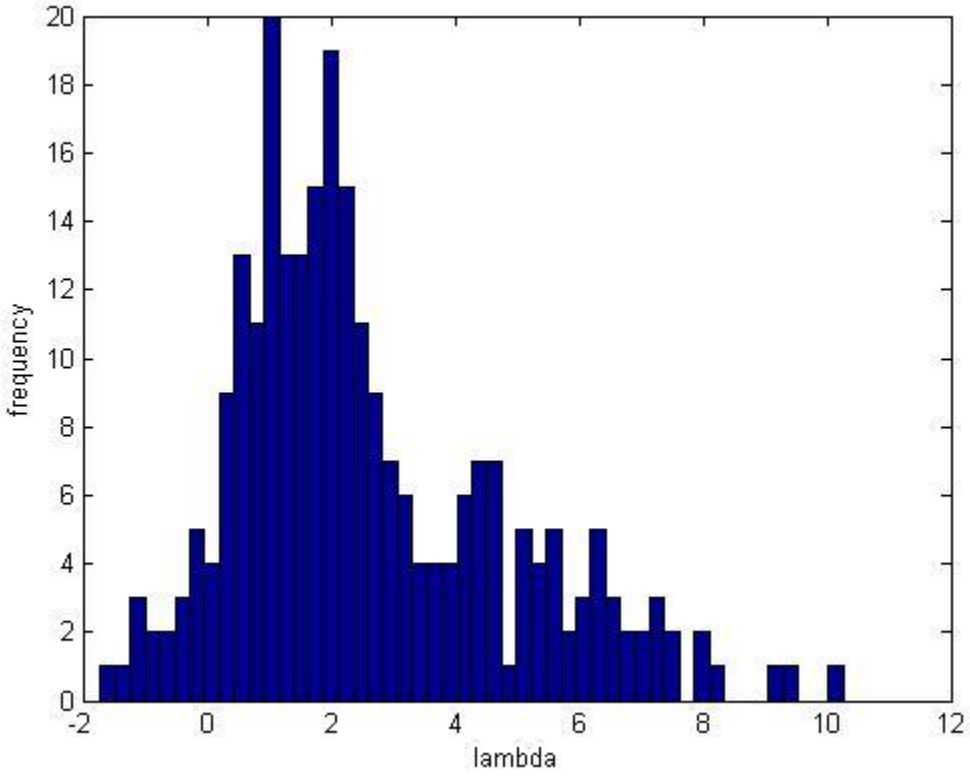
where λ_R and λ_U are vectors consisting of the known decay rates of the training rooms and the obtained estimates of the decay rates of the unseen rooms, respectively.

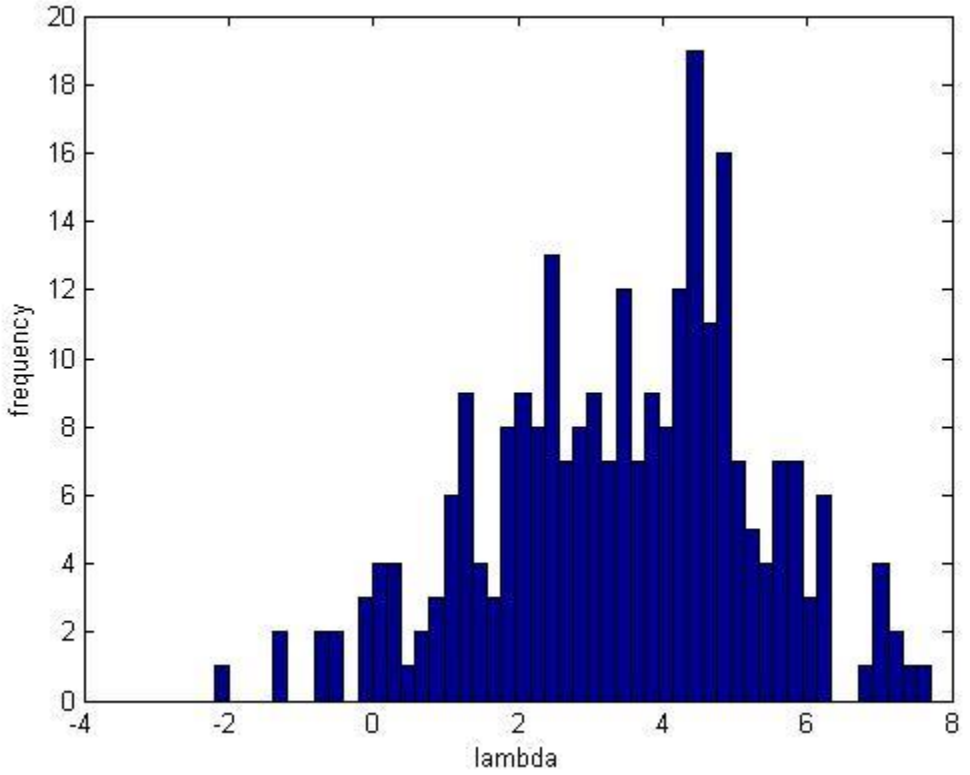
Results

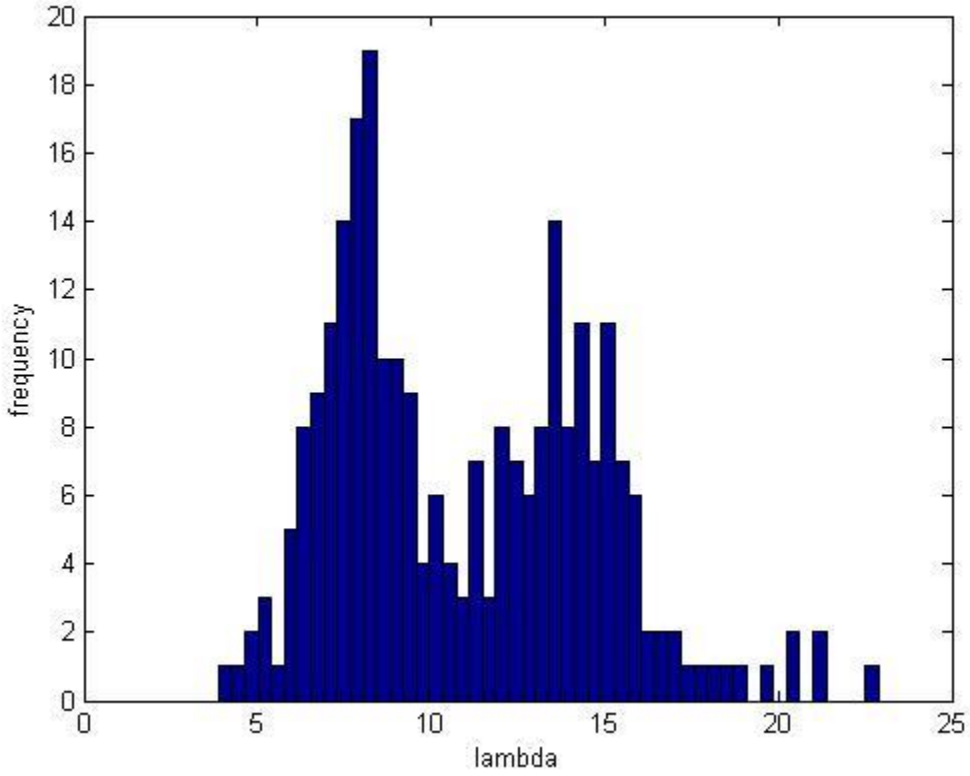
'Decay rate' distributions (**histograms**)

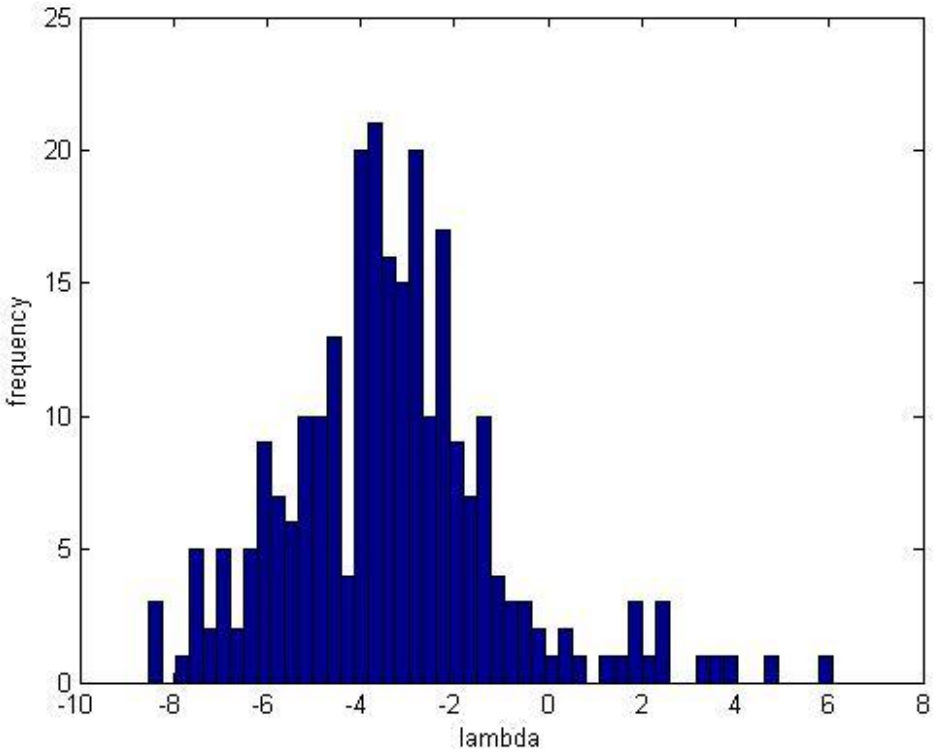



frequency
lambda
25
20
15
10
5
0
-2
0
2
4
6
8
10
12

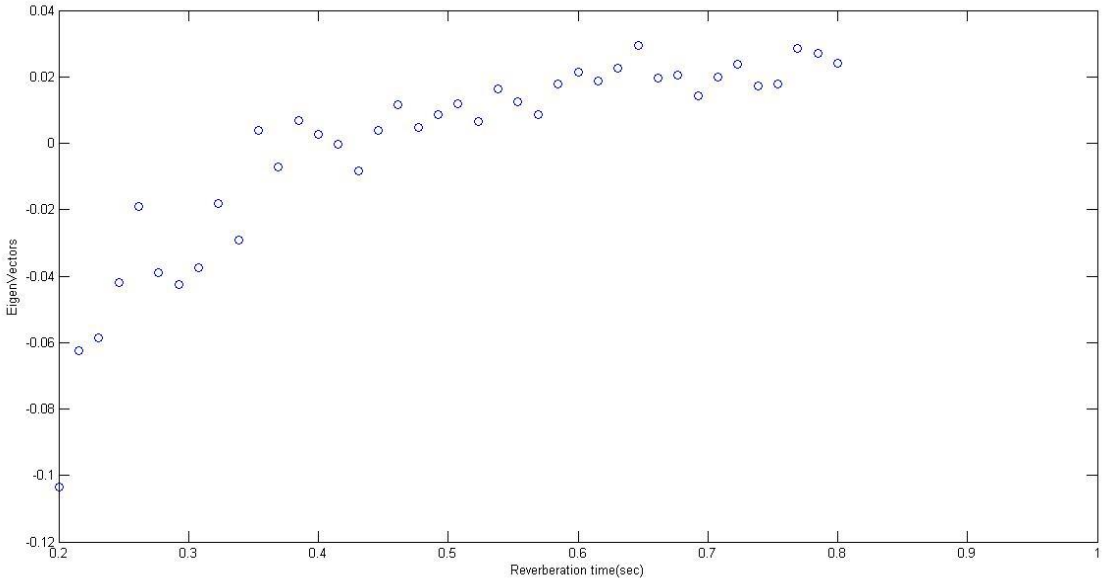

frequency
lambda
20
18
16
14
12
10
8
6
4
2
0
-2
0
2
4
6
8
10
12

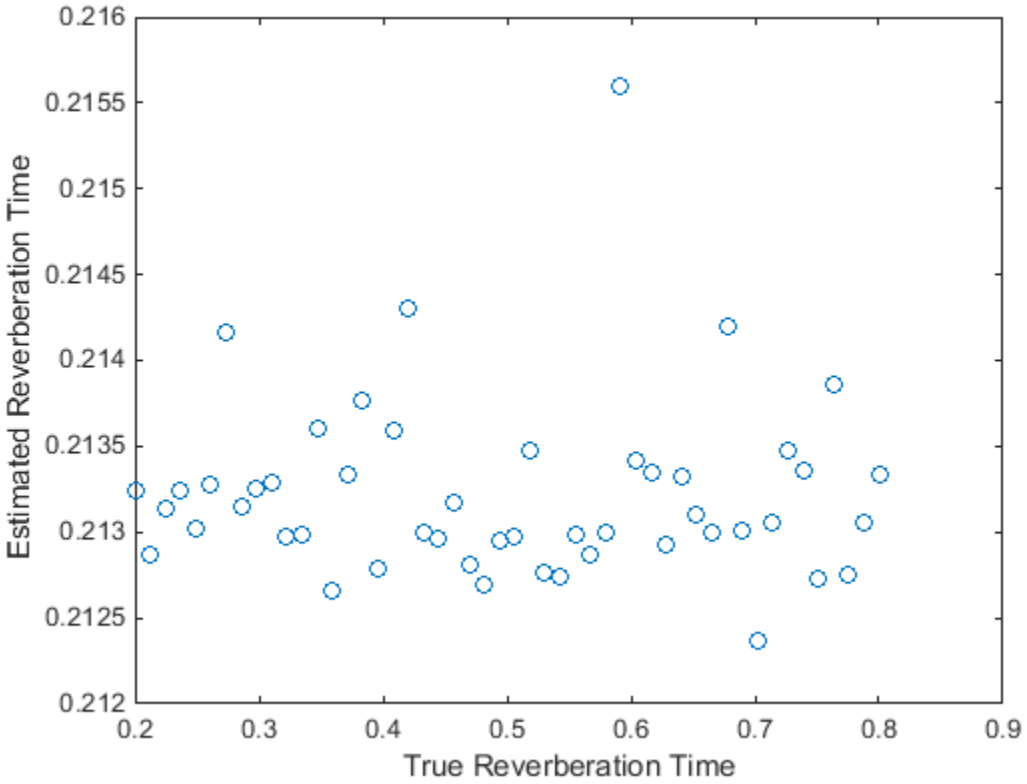

frequency
lambda
0
2
4
6
8
10
12
14
16
18
20
-4
-2
0
2
4
6
8


frequency
lambda
0
2
4
6
8
10
12
14
16
18
20
0
5
10
15
20
25



And, the final primary eigenvector shows monotonic variation of reverberation times:


EigenVectors
Reverberation time(sec)


Estimated Reverberation Time
True Reverberation Time

References

1. J. Y. C. Wen, E. A. P. Habets, and P. A. Naylor, "Blind estimation of reverberation time based on the distribution of signal decay rates," in Proc. IEEE Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP), Las Vegas, USA, Apr. 2008.
2. N. L´opez, Y. Grenier, G. Richard, and I. Bourmeyster, "Low variance blind estimation of the reverberation time," in Proc. Intl. Workshop Acoust. Signal Enhancement (IWAENC), Aachen, Germany, Sept. 2012.
3. M. Hein and J. Y. Audibert, "Intrinsic dimensionality estimation of submanifold in rd," ICML, pp. 289–296, 2005.
4. J. B. Allen and D. A. Berkley, "Image method for efficiently simulating small-room acoustics," J. Acoust. Soc. Am., vol. 65, no. 4, pp. 943–950, Apr. 1979.
5. E. A. P. Habets, "Room impulse response generator," Technische Universiteit Eindhoven, Tech. Rep, 2006.