BLIND ESTIMATION OF REVERBERATION TIME IN OCCUPIED ROOMS

Yonggang Zhang[†], *Jonathon A. Chambers*[†], *Francis F. Li*[*], *Paul Kendrick*[‡], *Trevor J. Cox*[‡]

[†]The Centre of Digital Signal Processing, Cardiff School of Engineering
Cardiff University, Cardiff CF24 0YF, UK. email: zhangy15@cf.ac.uk, chambersj@cf.ac.uk
[*]Department of Computing and Mathematics
Manchester Metropolitan University, Manchester M1 5GD, UK. email: f.li@mmu.ac.uk
[‡]School of Acoustics and Electronic Engineering
University of Salford, Salford M5 4WT, UK. email: P.kendrick@salford.ac.uk, T.J.Cox@salford.ac.uk

ABSTRACT

A new framework is proposed in this paper to solve the reverberation time (RT) estimation problem in occupied rooms. In this framework, blind source separation (BSS) is combined with an adaptive noise canceller (ANC) to remove the noise from the passively received reverberant speech signal. A polyfit preprocessing step is then used to extract the free decay segments of the speech signal. RT is extracted from these segments with a maximum-likelihood (ML) based method. An easy, fast and consistent method to calculate the RT via the ML estimation method is also described. This framework provides a novel method for blind RT estimation with robustness to ambient noises within an occupied room and extends the ML method for RT estimation from noise-free cases to more realistic situations. Simulation results show that the proposed framework can provide a good estimation of RT in simulated low RT occupied rooms.

1. INTRODUCTION

Room reverberation time is a very important parameter that qualifies the room acoustic quality [1]. This parameter is defined as the time taken by a sound to decay 60 dB below its initial level after it has been switched off. Many methods have been proposed to estimate the RT during recent years [2][3][4][5]. The maximum-likelihood (ML) estimation method proposed in [5] which utilizes a passively received speech signal has received a lot of attention due to its simplicity and efficiency. In this method, an exponentially damped Gaussian white noise model is used to describe the reverberation diffusive tail signal. An ML estimation method is then performed on segments of the speech signal to measure the time-constant of the decay. The most likely RT is identified from a series of estimates by using an order-statistic filter. As shown by the authors, it provides reliable RT estimates in a noise free environment. To estimate the RT in noisy environments, such as occupied rooms, where many noises are generated by the occupants, this method potentially only considers the signal decay range between the initial maximum of the decay curve and the point where the decay curve intersects the background noise. When the noise is large, for example comparable with the excited speech signal, the results will be contaminated or even incorrect. Therefore this method is limited by the noise level and not suitable for occupied rooms.

To make the ML RT estimation method more robust and accurate, an intuitive way is to remove the unknown noise signal from the received speech signal as much as possible before RT estimation. A powerful tool for extracting some noise interference signal from a mixture of signals is the convolutive BSS method [6]. Naturally, given two spatially distinct observations, BSS can separate the mixed signals to yield two independent signals. One of these two signals mainly consists of the excitation speech signal plus residue of noise and the other signal contains mostly the noise signal. Using this estimated noise signal as a reference signal the noise contained in the received speech signal can then be removed by an ANC. Our new framework is motivated by BSS and ANC. Different stages of this framework in an occupied room are shown in Fig.1. The signal $s_1(n)$, which is assumed to be the noise signal in this work, is independent with the excitation speech signal $s_2(n)$. The passively received signals $x_1(n)$ and $x_2(n)$ are modelled as convolutive mixtures of $s_1(n)$ and $s_2(n)$. The room impulse response $h_{ji}(n)$ is the impulse response from source i to microphone j. BSS is used firstly to obtain the estimated excitation speech signal $\hat{s}_2(n)$ and the estimated noise signal $\hat{s}_1(n)$. The estimated noise signal $\hat{s}_1(n)$ then serves as the reference signal for the ANC to remove the noise component from $x_1(n)$. The output of the ANC $\hat{y}_{12}(n)$ is an estimation of the noise free reverberant speech signal $y_{12}(n)$. As compared with $x_1(n)$, it **crucially** retains the reverberant structure of the speech signal and has a low level of noise, therefore it is more suitable to estimate the RT of the occupied room. To remove the guess work of the window length selection in the ML method and reduce the variance of the RT estimates, we use an overlap polyfit method as a preprocessing step. The decay segments in $z(n)$ which contain most of the free decay samples of the reverberant speech signal are extracted by this preprocessing. Then

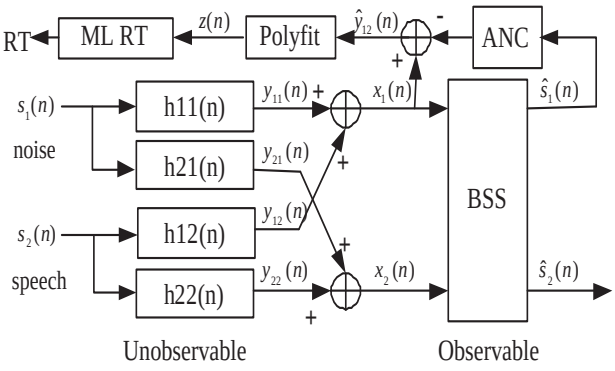

Figure 1: Proposed blind RT estimate frame work for occupied rooms.

the ML estimation method is performed only on these decay segments. Based on the idea of bisection [5], a new method to calculate the RT is also provided in the ML RT estimation method. Compared with other calculation methods this method has some advantages, as will be discussed later.

The following section introduces the BSS process. The ANC is described in Section 3. The polyfit preprocessing is described in Section 4. Section 5 describes the ML method. A bisection algorithm for the ML estimation method is also introduced. Simulation results are given in Section 6. Section 7 summarizes the paper.

2. BLIND SOURCE SEPARATION

As shown by Fig.1, the goal of BSS is to extract the estimated noise signal $\hat{s}_1(n)$ from received mixture signals $x_1(n)$ and $x_2(n)$. If we assume that the room environment is time invariant, the received mixtures $x_1(n)$ and $x_2(n)$ can be modeled as weighted sums of convolutions of the source signals $s_1(n)$ and $s_2(n)$. Assume that N sources are recorded by M microphones (here M=N=2) the equation that describes this convolved mixing process is:

$$x_j(n) = \sum_{i=1}^{N} \sum_{p=0}^{P-1} s_i(n-p) h_{ji}(p) \tag{1}$$

where $s_i(n)$ is the source signal from a source i, $x_j(n)$ is the received signal by a microphone j, and $h_{ji}(n)$ is the P-point response from source i to microphone j. Using a T-point windowed discrete Fourier transformation (DFT), time domain signal $x_j(n)$ can be converted into the time-frequency domain signal $X_j(\omega,n)$ where ω is a frequency index and n is a time index. For each frequency bin we have

$$\mathbf{X}(\omega,n) = \mathbf{H}(\omega)\mathbf{S}(\omega,n) \tag{2}$$

where $\mathbf{S}(\omega,n) = [s_1(\omega,n),\cdots,s_N(\omega,n)]^T$ and $\mathbf{X}(\omega,n) = [x_1(\omega,n),\cdots,x_M(\omega,n)]^T$ are the time-frequency representations of the source signals and the observed signals respectively and $(\cdot)^T$ denotes vector transpose. The separation can be completed by the unmixing matrix $\mathbf{W}(\omega)$ in a frequency bin ω

$$\hat{\mathbf{S}}(\omega,n) = \mathbf{W}(\omega)\mathbf{X}(\omega,n) \tag{3}$$

where $\hat{\mathbf{s}}(\omega,n) = [\hat{s}_1(\omega,n),\cdots,\hat{s}_N(\omega,n)]^T$ is the time-frequency representations of the estimated source signals and $\mathbf{W}(\omega)$ is the frequency representation of the unmixing matrix. $\mathbf{W}(\omega)$ is determined so that $\hat{s}_1(\omega,n),...,\hat{s}_N(\omega,n)$ become mutually independent. Exploiting the nonstationary of the speech signal we define the cost function as follows:

$$J(\mathbf{W}(\omega)) = \arg\min \sum_{w=1}^{T} \sum_{k=1}^{K} F(\mathbf{W})(\omega,k), \tag{4}$$

where K is the number of signal segments and $F(\mathbf{W})(\omega,k)$ is defined as

$$F(\mathbf{W})(\omega,k) = \|R_{\hat{S}}(\omega,k) - diag[R_{\hat{S}}(\omega,k)]\|_F^2 \tag{5}$$

where $R_{\hat{S}}(\omega,k)$ is the autocorrelation matrix of the separated signals and $\|\cdot\|_F^2$ denotes the squared Frobenius norm, k is the block index. The separation problem is then converted into a joint diagonalization problem. Obviously, the solution $\mathbf{W}(\omega) = 0$ will lead to the minimization of $F(\mathbf{W})(w,k)$. To avoid this some constraints should be added to the unmixing matrix. In [6] a penalty function is added to convert the constrained optimization problem into an unconstrained optimization problem. The cost function of penalty function based joint diagonalization is as follows:

$$J(\mathbf{W}(\omega)) = \arg\min \sum_{w=1}^{T} \sum_{k=1}^{K} F(\mathbf{W})(\omega,k) + \lambda g(\mathbf{W})(\omega,k) \tag{6}$$

where λ is the penalty weight factor and $g(\mathbf{W})(\omega,k)$ is a form of penalty function based on a constraint of the unmixing matrix. With a gradient-based descent method we can calculate the unmixing matrix after several iterations from equation (6). The separated signals $\hat{s}_1(n)$ and $\hat{s}_2(n)$ can then be obtained from (3) after applying an inverse DFT.

3. ADAPTIVE NOISE CANCELLER

After BSS we obtain the estimated noise signal $\hat{s}_1(n)$. This signal is then used as a reference in the ANC stage signal to remove the noise component from the received signal $x_1(n)$. A new variable step size LMS algorithm which is suitable for speech processing is used in the ANC. The updates of the step size can be formulated as follows:

$$e(n) = x_1(n) - \hat{\mathbf{s}}_1^T(n)\mathbf{w}(n) \tag{7}$$

$$\mathbf{g}(n) = \frac{e(n)\hat{\mathbf{s}}_1(n)}{\sqrt{L[\hat{\sigma}_e^2(n) + \hat{\sigma}_s^2(n)]}} \tag{8}$$

$$\mathbf{p}(n) = \beta\mathbf{p}(n-1) + (1-\beta)\mathbf{g}(n) \tag{9}$$

$$\mu(n+1) = \alpha\mu(n) + \gamma\|\mathbf{p}(n)\|_F^2 \tag{10}$$

where $\mu(n)$ is the variable step size, $\hat{\mathbf{s}}_1(n) = [\hat{s}_1(n),\cdots,\hat{s}_1(n-L+1)]^T$, $\mathbf{w}(n)$ is the weight vector of the adaptive filter, L is the filter length, $\hat{\sigma}_e^2(n)$ and $\hat{\sigma}_s^2(n)$ are estimations of the temporal error energy and the temporal input energy, $0<\alpha<1$, $0<\beta<1$, $\gamma>0$, $\mathbf{g}(n)$ is the square root normalized gradient vector, $\mathbf{p}(n)$ is a smoothed version of $\mathbf{g}(n)$. The recursion of the filter weight vector is as follows

$$\mathbf{w}(n+1) = \mathbf{w}(n) + \mu(n)\frac{e(n)\hat{\mathbf{s}}_1(n)}{L[\hat{\sigma}_e^2(n) + \hat{\sigma}_s^2(n)]} \tag{11}$$

The square root normalized gradient vector $\mathbf{g}(n)$ in (8) is used to obtain a robust measure of the adaptive process. The first-order filter based averaging operation in (9) removes the disturbance brought by the target signal. The variable step size $\mu(n)$ in (10) is adapted to obtain a fast convergence rate during the early adaptive process and a small misadjustment after the algorithm converges. The adaptation of the weight vector in (11) is based on the sum method in [7] which is designed to minimize the steady state mean square error. Equations (7)(8)(9)(10)(11) provide a new variable step size LMS algorithm for the ANC stage. The output signal of the ANC $\hat{y}_{12}(n)$ should then be a good estimation of the noise free reverberant speech signal $y_{12}(n)$. Next, estimating of the RT from $\hat{y}_{12}(n)$ must be considered.

4. POLYFIT PREPROCESSING

In this stage, the input signal is the estimated noise free reverberant speech signal $\hat{y}_{12}(n)$. The overlap polyfit method is used to extract the decay segments of this signal. The output of this stage is the signal $z(n)$ which contains the decay segments of $\hat{y}_{12}(n)$. In accordance with the ML estimation of RT, we use the same exponentially damped Gaussian white noise model which has been used in [5]. The mathematical formulation is as follows:

$$\hat{y}_{12}(n) = a(n)v(n) \tag{12}$$

where $v(n)$ is an i.i.d. term with normal distribution $N(0,\sigma)$ and $a(n)$ is a time-varying envelope term. Let a single decay rate τ describe the damping of the sound envelope during free decay, then the sequence $a(n)$ is uniquely determined by

$$a(n) = \exp(-n/\tau) = a^n \tag{13}$$

where

$$a = \exp(-1/\tau) \tag{14}$$

In this stage, we first use a moving window with an appropriate length and shift to obtain overlap speech frames. From the model of the reverberant speech tail in (12), which is assumed to hold in each frame, the logarithm of the envelope of the free decay segment is a line with negative slope. Because in reality the RT should have a reasonable span, for example, 0s to 3s, such a slope should have a corresponding range. A polyfit operation is then performed on each frame to extract the slope. By discarding the frames whose slopes are outside such a range, the speech signal is divided into several continuous decay segments. The longest segments contained in $z(n)$ should contain the most likely free decay segments of the speech signal.

As a preprocessing stage for the ML RT estimation method it has several advantages. At first it provides the window length for the ML RT estimation method automatically. Although in [5] it has been found that increasing window length reduces the variability in the estimates, the window length is limited by the duration and occurrence of the gaps between sound segments. The choice of the window length is a trade off between the accuracy and variance of the estimated RTs. After the polyfit preprocessing the window length of the ML estimation must be less than the length of the extracted signal segment. It is then chosen automatically according to the segment length. Simulation results show that half length of the segment length will be a good choice of the window length. Secondly, the variance of RT estimates is reduced because most samples of these segments are in agreement with the Gaussian damped model, as will be confirmed in later simulations.

5. ML RT ESTIMATION METHOD

The ML estimation method is then performed on the chosen segments $z(n)$. From the definition of RT and the signal model, the relationship between RT and the decay rate τ is as follows [5]:

$$T_{60} = \frac{-3\tau}{\log_{10}(\exp(-1))} = 6.91\tau \tag{15}$$

The decay rate τ is extracted by the ML estimation method. Denote the N-dimensional vectors of $z(n)$ and $a(n)$ (the same as that in (13) and (14)) by $\mathbf{z}(n)$ and $\mathbf{a}(n)$ and N is the estimation window length, we can obtain the logarithm likelihood function

$$E\{L(\mathbf{z};a,\sigma)\} = -\frac{N(N-1)}{2}\ln(a) - \frac{N}{2}\ln(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum_{n=1}^{N} a^{-2n}z^2(n) \tag{16}$$

where σ is the initial power of the signal. With this function the parameters a and σ can be estimated using an ML approach. From each segment we can obtain a series of estimates of RT. All estimates are used to identify the most likely RT of the room.

By considering the relationship between a and decay rate τ, we propose a new bisection method with respect to the RT rather than with respect to a in [5]. The range of the RT is set between 3s and 0.1s. As the time-constant is not required to be arbitrarily precise, the accuracy is limited to 10ms in our method. The update of our bisection method is as follows:

i). Initialization

$$T60\_min = 0.1; T60\_max = 3; accuracy = 0.01;$$

$$iter = \log_2((T60\_max - T60\_min)/accuracy))$$

where *accuracy* is the accuracy of the estimation of RT and *iter* is the iteration number.

2). Iteration

$$\begin{aligned} T(i) &= (T60\_min + T60\_max)/2 \\ a(i) &= exp(-6.91/T(i)) \\ g(i) &= \frac{\partial L(\mathbf{y};a,\sigma)}{\partial a} \\ g(i) &> 0 \quad then \quad T60\_min = T(i) \\ g(i) &< 0 \quad then \quad T60\_max = T(i) \end{aligned}$$

As the authors point out in [5], the disadvantage of the bisection method is that it works poorly in regions near the true value of a. From (14) and (15) we know that a is not a linear transform of RT. Our bisection on RT is actually a non-equivalent bisection with respect to a. Compared with the fast block algorithm proposed in [8] our algorithm has a number of advantages:

1. No step size needs to be selected.
2. No initial value of a is needed.
3. It always converges and converges quickly within a fixed number of steps.

6. SIMULATION

In this section we examine the performance of the proposed framework. The flow chart of the simulations is shown in Fig.1. The occupied room and its impulse response h_{ji} between source i and microphone j are simulated by an image room model [9]. The room size is set to be 10*10*5 meter^3 and the reflection coefficient is set to be 0.7 in rough correspondence with the actual room. The RT of this room measured by Schroeder's method [2] is 0.27s. The excitation speech signal and the noise signal are two anechoic 40 seconds male speech signals with a sampling frequency of 8kHz, and scaled to make the signal to noise ratio (SNR) to be 0dB over the whole observation. The position of these two

sources are set to be [1m 3m 1.5m] and [3.5m 2m 1.5m]. The positions of the two microphones are set to be [2.45m 4.5m 1.5m] and [2.55m 4.5m 1.5m] respectively. As shown by Fig.1, BSS is performed firstly to extract the estimated noise signal $\hat{s}_1$. This signal contains mostly the noise signal and a low level of the desired speech signal. To evaluate the BSS performance we use a noise to signal ratio (NSR) which is the energy ratio defined between the component of the noise signal and the component of the speech signal contained in $\hat{s}_1$. The NSR of $\hat{s}_1$ in this simulation is 38dB, therefore it has a strong correlation with the noise signal s_1 and a slight correlation with the speech signal. This signal is then used in the ANC model as a reference signal. The filter length of the ANC is set to be 500 and the parameters α, β, γ are set to be 0.99, 0.9999, 200 respectively. The last 1000 samples of the filter coefficients are used to measure the steady-state performance. The output signal of the ANC contains two components: the reverberant speech signal and the residue of the noise signal. The signal to noise ratio (SNR) between these two components is 43dB. The first approximately 10s of this signal will be used to estimate the RT. We plot the first approximately 10s of the received signal x_1 and the output signal of ANC $\hat{y}_{12}$ in Fig.2(a) and Fig.2(b) respectively. It is easy to see that after BSS and ANC the noise contained in x_1 is reduced greatly.

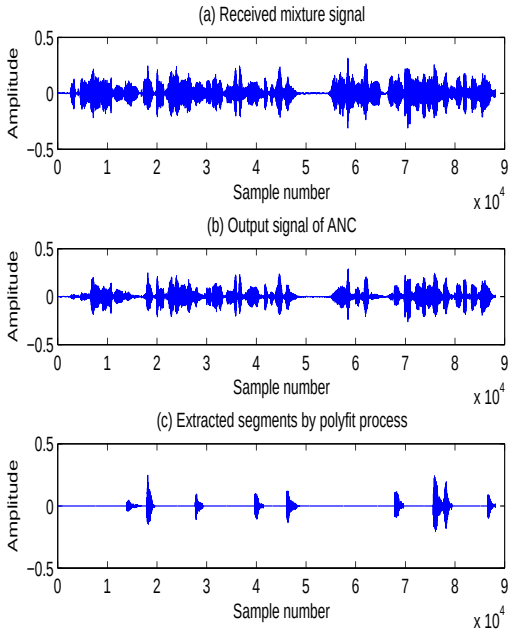

Figure 2: The received mixture signal, the output of ANC and the extracted signal by polyfit process

At first we estimate the RT by using the ML RT estimation method [5] with the whole output signal of ANC first. According to the analysis and simulations in [5], the window length is set to be 1200, which is approximately equal to 4τ, to provide a good choice of the window length. The results are shown in Fig.3(a).

Then the polyfit process is performed to extract the free decay segments. The window length of our polyfit method is set to be 400 samples (0.05s) and the shift is set to be 10 samples. Ten segments extracted by the polyfit stage are shown

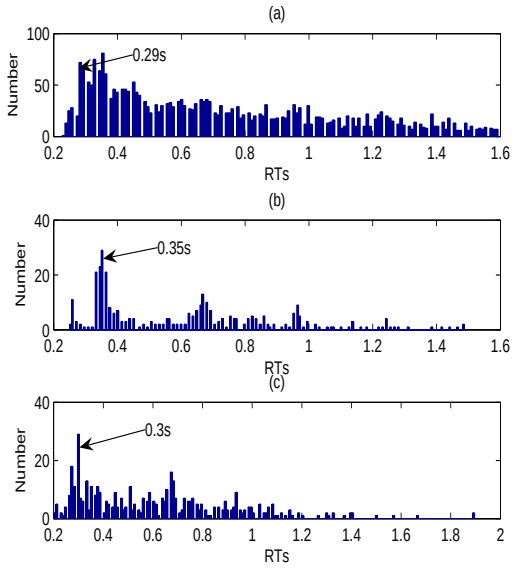

Figure 3: (a) The estimates of RTs by ML method with the whole output signal of ANC and a window length of 1200. (b)The estimates of RTs by ML method with the output signal of polyfit process and a window length of 1200. (c)The estimates of RTs by ML method with the output signal of polyfit process and automatical decided window length.

in Fig.2(c). Note that three segments are connected in the figure in the interval 75,000 to 80,000.

Finally, two experiments are performed to show the two advantages of the polyfit process which we analyzed in section 4. In the first experiment, the extracted signal which is the output of the polyfit process is used to estimate the RT by using the ML method with a window length of 1200. The results are shown in Fig.3(b). The second experiment is the same as the first experiment except the window length of the ML method is decided automatically, which is half of the segment length. The results of this experiment are shown in Fig.3(c).

Compare Fig.3(a) with Fig.3(b) we can see that the variance of the estimates of the RT is reduced greatly by using the polyfit process, where most decay samples are extracted. The first peak in Fig.3(a) is 0.29s, but it is not clear and the variance of the RT estimates is very large. In Fig.3(b) the variance of the RT estimates is reduced greatly and the first peak is 0.35s. Although both results in these two figures are larger than the theoretical RT of 0.27s due to the lack of sharp transients in the clean speech, the bias of the model in ML method and the influence of the interference, they are reasonable and acceptable in most applications.

Compare Fig.3(c) with Fig.3(b) we can see that the variance of the RT estimates in both figures are comparable. The first peak in Fig.3(c) is 0.3s, which is also a reasonable and acceptable result. Thus we can conclude that performance of the ML method with automatical decided window length is comparable, if not better, with the performance of the ML method with a good choice of the window length.

From all the simulations above we can see that the combination of BSS and ANC can remove the noise signal greatly whilst retaining the key reverberant structure to make the high-noise environment RT estimation possible. Further

more, the polyfit process has been added before the ML RT estimation method to reduce the variance of the results and remove the 'guess' work of the window length of the ML RT estimation method.

We have performed other experiments where one of the speech signals in the previous simulations is replaced by a white noise signal, as a simulated interference in the occupied room. Similar estimation results are also obtained. However limited by the room model and the performance of frequency domain BSS, this framework is designed to estimate RT in the occupied room whose RT is less than 0.3s. As shown by our simulations above, nonetheless, reliable RT can be extracted using this framework within a highly noisy occupied room, something that has not previously been possible.

7. CONCLUSION

This paper proposes a new framework for blind RT estimation in occupied rooms. In this framework, BSS is combined with an ANC to remove the noise of the received speech signal. A polyfit stage is added to improve the performance of the ML RT estimation method. A bisection method is used in the ML method which provides many advantages over the previous calculation method. Simulation results show that the noise is removed greatly from the reverberant speech signal and the performance of this frame work is good in a simulated low RT occupied room environment. Due to the motivation of our framework BSS and ANC can be potentially used in many reverberation time estimation methods as a preprocessing. Although the mixing model used in this paper is not suitable for many applications, this framework provides a new way to overcome the noise disturbance in RT estimation. However, limited by the performance of convolutive BSS, this framework is only appropriate for the low RT estimation case. Future work will focus on the theoretic analysis of this blind RT estimate framework and the improvement of its stages, especially the improvement of convolutive BSS under long reverberation environments.

REFERENCES

[1] H. Kuttruff, *Room Acoustics 4th ed.*, Spon Press, London, 2000.

[2] M. R. Schroeder, "New method for measuring reverberation time," *J. Acoust. Soc. Am*, vol. 37, pp. 409–412, 1965.

[3] ISO 3382, "Acoustics-measurement of the reverberation time of rooms with reference to other acoustical parameters," *International Organization for Standardization*, 1997.

[4] T. J. Cox, F. Li and P. Darlington, "Extracting room reverberation time from speech using artificial neural networks," *J. Audio. Eng. Soc*, vol. 49, pp. 219–230, 2001.

[5] R. Ratnam, D. L. Jones, B. C. Wheeler, W. D. O'Brien Jr., C. R. Lansing and A. S. Feng, "Blind estimation of reverberation time," *J. Acoust. Soc. Amer.*, vol. 114, no. 5, pp. 2877–2892, Nov. 2003.

[6] W. Wang, S. Sanei and J. A. Chambers, "Penalty function-based joint diagonalization approach for convolutive blind separation of nonstationary sources," *IEEE Tran. Signal Processing*, vol. 53, no. 5, pp. 1654–1669, May 2005.

[7] J. E. Greenberg, "Modified LMS algorithm for speech processing with an adaptive noise canceller," *IEEE Trans. Signal Processing*, vol. 6, no. 4, pp. 338–351, July 1998.

[8] R. Ratnam, D. L. Jones and W. D. O'Brien Jr., "Fast algorithms for blind estimation of reverberation time," *J. Acoust. Soc. Amer.*, vol. 11, no. 6, pp. 537–540, June 2004.

[9] J. B. Allen and D. A. Berkley, "Image method for efficiently simulating small-room acoustics," *J. Acoust. Soc. Amer.*, vol. 65, pp. 943–950, Apr. 1979.