Frequency Dependent Statistical Model for the Suppression of Late Reverberations

Tariqullah Jan [#1], Wenwu Wang [*2]

[#] *Department of Electrical Engineering*
University of Engineering & Technology Peshawar, Pakistan
[*] *Centre for Vision, Speech and Signal Processing*
University of Surrey, Guildford, United Kingdom
[1] tariqullahjan@yahoo.com
[2] w.wang@surrey.ac.uk

***Abstract*—Suppression of late reverberations is a challenging problem in reverberant speech enhancement. A promising recent approach to this problem is to apply a spectral subtraction mask to the spectrum of the reverberant speech, where the spectral variance of the late reverberations was estimated based on a frequency independent statistical model of the decay rate of the late reverberations. In this paper, we develop a dereverberation algorithm by following a similar process. Instead of using the frequency independent model, however, we estimate the frequency dependent reverberation time and decay rate, and use them for the estimation of the spectral subtraction mask. In order to remove the processing artifacts, the mask is further filtered by a smoothing function, and then applied to reduce the late reverberations from the reverberant speech. The performance of the proposed algorithm, measured by the segmental signal to reverberation ratio (SegSRR) and the signal to distortion ratio (SDR), is evaluated for both simulated and real data. As compared with the related frequency indepenent algorithm, the proposed algorithm offers considerable performance improvement.**

I. Introduction

Room reverberation is one of the main causes of performance degradation in automatic speech recognition systems, digital hearing aids, binaural telephone headsets, hands free communication devices, audio analytic sensors, and acoustic surveillance systems. The reverberant speech consists of three components: the direct response, the early reflections and the late reverberations [7]. The direct response is the direct signal that passes from the source to the receiver, (i.e., a microphone or a listener). It is affected by the distance between the source and the receiver. The early reflections are considered harmless since they can reinforce the direct signal. However, in the frequency domain, the spectrum of the speech will be distorted because the frequency response of the early reflections is non-flat (known as coloration). Late reverberation refers to the signal component that has a long delay as compared to the direct signal, which is the main cause of the distortion of speech intelligibility. This is due to the two masking effects introduced by the late reverberations, namely self masking where the speech spectrum is smeared by the late reverberations, and overlap masking where the energy of the preceding phoneme overlaps with that of the subsequent phonemes.

Several algorithms have been proposed to deal with the reverberant speech signal, e.g. [2], [5]-[7], [11], [14]. These methods may be broadly classified into three categories, namely, methods for dealing with the early reflections, the late reverberation, or both (in two stages). For example, inverse filtering, despite targetting the full room impulse response (RIR) by convolving the reverberant signal with an inverse filter derived from the RIR [9], has shown to be effective in mitigating the early reflections. Inverse filtering has been applied in either temporal or spectral domain. Another category of methods attempt to suppress the late reverberations using e.g. the spectral subtraction technique [2], [7], where the variance of the late reverberations is estimated and then subtracted from the reverberant speech. The third category of methods consider both the early reflections and late reverberations, e.g. in [14], where inverse filtering is combined with spectral subtraction in a two-stage process.

Recently, Lebart *et al.* [7] proposed a statistical model for late reverberations. With this model, the spectral variance of the late reverberations can be estimated from the reverberant speech [7], which was further used by Jeub *et al.* for the suppression of late reverberations [5]. This original model was developed as frequency independent where a fix reverberant time (T_{60}) was used for all the frequency channels in the estimation of the decay rate of room reverberations. In this paper, with a frequency dependent model due to Habets *et al.* [3], we develop an improved version of the dereverberation algorithm presented in [5]. Section II formulates the problem and its model. Section III describes the proposed approach which includes the estimation of frequency dependent T_{60}, the estimation of spectral subtraction mask, and the filtering (smoothing) of the mask. Section IV presents the evaluation results, followed by a conclusion in Section V.

II. Problem Formulation and Modelling

The reverberant speech signal $x(n)$ can be modelled as the convolution of the anechoic speech signal $s(n)$ and the RIRs

$h(n)$, [11]

$$x(n) = \sum_{l=0}^{\infty} h(l)s(n-l) \tag{1}$$

where n is the discrete time index. The RIR of length T_r in seconds can be modelled as [7]

$$h(n) = \begin{cases} h_{early}(n) & \text{for } 0 \leq n < T_{le} \cdot f_s, \\ h_{late}(n) & \text{for } T_{le} \cdot f_s \leq n \leq T_r \cdot f_s, \\ 0 & \text{otherwise} \end{cases} \tag{2}$$

where $h_{early}(n)$ denotes the direct and early path, $h_{late}(n)$ is the late reflection path, f_s is the sampling frequency, and T_{le} is the time after which we assume that the late reverberation starts. The range of T_{le} usually lies in 50 to 100 ms. The reverberant speech signal can now be represented as the combination of two main parts, i.e., $x_{early}(n)$ and $x_{late}(n)$,

$$x(n) = \underbrace{\sum_{l=0}^{T_{le}f_s-1} s(n-l)h(l)}_{x_{early}(n)} + \underbrace{\sum_{l=T_{le}f_s}^{T_r f_s} s(n-l)h(l)}_{x_{late}(n)} \tag{3}$$

In order to reduce the effects of early reflections ($x_{early}(n)$), inverse filtering may be used as in [14]. For the suppression of late reverberations ($x_{late}(n)$), spectral subtraction technique such as [7], [14], [2] is usually employed, where the spectral variance of the late reverberations is estimated from the reverberant speech. A recent technique for the spectral variance estimation was proposed by Lebart *et al.* [7] in which the late impulse responses are statistically modelled as

$$h_{late}(n) = \begin{cases} \beta(n)e^{-\alpha n} & \text{for } n \geq 0, \\ 0 & \text{otherwise} \end{cases} \tag{4}$$

where $\beta(n)$ is a sequence of zero-mean mutually independent and identically distributed (i.i.d.) Gaussian random variables, and α denotes the decay rate given as

$$\alpha = \frac{3ln(10)}{T_{60}f_s} \tag{5}$$

where ln is the natural logarithm. Using the above model originally proposed by Lebart *et al.* in [7], Jeub *et al.* [4], [5] have recently presented a dereverberation algorithm with a frequency independent α. However, it was shown in [3] that a frequency dependent α may provide more accurate estimation of the spectral variance of the late reverberations. We present in the next section a new dereverberation algorithm using this frequency-dependent model.

III. The Proposed Method

A. Frequency Dependent RIR Model

Applying the short-time Fourier transform (STFT), we can rewrite equation (2) in the T-F domain as

$$H(m,k) = \begin{cases} H_{early}(m,k) & \text{for } 0 \leq m < N_{le}, \\ H_{late}(m,k) & \text{for } N_{le} \leq m \leq N_r, \\ 0 & \text{otherwise} \end{cases} \tag{6}$$

where N_{le} and N_r are the number of frames corresponding to T_{le} and T_r respectively. With the statistical model (4) and a frequency-dependent α, $H_{late}(m,k)$ can also be written as [3],

$$H_{late}(m,k) = \begin{cases} \beta(m,k)e^{-\alpha(k)mR} & \text{for } m \geq 1, \\ 0 & \text{otherwise} \end{cases} \tag{7}$$

where $\beta(m,k)$ is a sequence of zero-mean mutually i.i.d. Gaussian random variables, m is the time frame index, k is the frequency bin index, R denotes the hop size, and $\alpha(k)$ denotes the decay rate which can be obtained from the frequency dependent reverberation time $T_{60}(k)$ as below

$$\alpha(k) \triangleq \frac{3ln(10)}{T_{60}(k)f_s} \tag{8}$$

B. Estimation of $\mathbf{T_{60}}(k)$

We estimate $T_{60}(k)$ from the RIRs using a method similar to the one defined in ISO standard (ISO 3382-1:2009). First, we pass $h(n)$ through a Gammatone filter-bank to get sub-band signals $h(p,n)$, where p is the sub-band index. Subsequently, $h(p,n)$ were analyzed using Schroeder's method [13] to estimate the reverberation time $\check{T}_{60}(p)$ in each sub-band p. Since this filterbank (indexed by p) is different from the one used in the above section (indexed by k), the $\check{T}_{60}(p)$ values need to be inter- and extra-polated to obtain the estimate of $T_{60}(k)$ in each frequency bin k.

First we apply interpolation to $\check{T}_{60}(p)$ so that $\check{T}_{60}(p)$ from each sub-band p is mapped to $\tilde{T}_{60}(f)$, where $f \in [f_c - \frac{bw}{2}, f_c + \frac{bw}{2}]$ denotes the frequency range (in Hz) of sub-band p, f_c and bw are the centre frequency and the bandwidth of this sub-band respectively. Then we apply smoothing across the overlapped regions between the neighbouring sub-bands

$$\tilde{T}_{60}(f) = \tilde{T}_{60}(f_1) + \frac{\tilde{T}_{60}(f_2) - \tilde{T}_{60}(f_1)}{f_2 - f_1}(f - f_1) \tag{9}$$

where f_1 and f_2 are the frequency points of the neighbouring sub-bands at which their overlap begins and ends respectively. $\tilde{T}_{60}(f_1)$ and $\tilde{T}_{60}(f_2)$ are the reverberation times at frequency points f_1 and f_2 respectively. For non-overlapped regions, no such interpolation as (9) is required for $\tilde{T}_{60}(f)$. Finally, $\tilde{T}_{60}(f)$ is then mapped to the STFT sub-bands by an extrapolation method as

$$T_{60}(k) = \sum_{f=(k-1)\frac{F}{K}+1}^{k\cdot\frac{F}{K}} \tilde{T}_{60}(f)/(F/K-1) \tag{10}$$

Note that, $f = 1, 2, ..., F$, where F is the whole frequency range and K denotes the number of frequency bins (indexed by k). An alternative method without using the inter- and extrapolation process is to set the hop size as a single sample when calculating the STFT, and then calculate $T_{60}(k)$ directly for each frequency band k. Note that $T_{60}(k)$ can also be estimated from the reverberant signal [12] using the similar procedure described here.

C. Spectral Subtraction Mask Estimation

The statistical model discussed above in equation (7) is valid when the energy of the direct signal is low in comparison to that of all the given reflections. As a result the spectral variance of the late reverberant speech can be estimated as [3]

$$\sigma^2_{x_{late}}(m,k) = e^{-2\alpha(k)RN_{le}} \cdot \sigma^2_x(m - N_{le}, k) \quad (11)$$

where $\sigma^2_x(m,k)$ is the variance of the reverberant speech which can be estimated by recursive averaging

$$\sigma^2_x(m,k) = e^{-2\alpha(k)R}[\tau \cdot \sigma^2_x(m-1,k) + (1-\tau) \cdot \mid X(m,k) \mid^2] \quad (12)$$

where $\tau \in [0,1]$ is a forgetting factor and $X(m,k)$ is the T-F representation of $x(n)$ in (3). Note that N_{le} is the number of samples after which the late reverberation begins and $e^{-2\alpha(k)R}$ measures the reverberation decay rate. We can then estimate the *posteriori* signal-to-distortion ratio (SDR) [5] as follows

$$\varphi(m,k) = \frac{\mid X(m,k) \mid^2}{\sigma^2_{x_{late}}(m,k)} \quad (13)$$

To reduce the late reverberations, we apply the following spectral subtraction mask [5] to $X(m,k)$

$$\tilde{G}_{late}(m,k) = 1 - \frac{1}{\sqrt{\varphi(m,k)}} \quad (14)$$

In order to avoid over-estimation of $\sigma^2_{x_{late}}(m,k)$, a lower bound $\tilde{G}^{min}_{late}$ is applied to all the weighting gains in the mask.

D. Spectral Gain Smoothing

A common problem with spectral masking is the processing artifacts, i.e. the so-called musical noise. Therefore, similar to [5], we apply a moving average operation to $\tilde{G}_{late}(m,k)$. To this end, the power ratio between the enhanced signal and the reverberant signal is calculated. However, different from [5], we compute this power ratio at each frequency bin k and each time frame m

$$\rho(m,k) = \frac{\mid \tilde{G}_{late}(m,k) \cdot X(m,k) \mid^2}{\mid X(m,k) \mid^2} \quad (15)$$

We then generate a moving average window, as follows:

$$E_s(m,k) = \begin{cases} 1, & \text{if } \rho(m,k) \geq C, \\ 2 \cdot \lfloor (1 - \frac{\rho(m,k)}{C}) \cdot \psi \rfloor + 1, & \text{otherwise.} \end{cases} \quad (16)$$

where C is a constant controlling the trade off between the speech distortion and reduction of musical noise, ψ is a scaling factor for determining the level of smoothing, and $\lfloor \cdot \rfloor$ rounds the argument to its nearest integer. This window function can now be used to create a smoothing filter as

$$F_s(m,k) = \begin{cases} \frac{1}{E_s(m,k)}, & \text{if } k < E_s(m,k), \\ \frac{1}{2k}, & \text{otherwise} \end{cases} \quad (17)$$

By convolving $\tilde{G}_{late}(m,k)$ with $F_s(m,k)$, we obtain a smoothed mask as follows:

$$G_{late}(m,k) = \tilde{G}_{late}(m,k) * F_s(m,k) \quad (18)$$

E. Signal Reconstruction

Finally, the smoothed mask is applied to the T-F representation of the reverberant signal as follows:

$$\hat{S}(m,k) = X(m,k) \cdot G_{late}(m,k) \quad (19)$$

After transforming $\hat{S}(m,k)$ back to the time domain using the inverse STFT, we can obtain the dereverberated signal $\hat{s}(n)$.

IV. Experimental Results and Discussion

In this section, we evaluate the performance of our proposed method using the simulated RIRs from the image model [1] and the real RIRs from the Aachen Impulse Response (AIR) database [4]. 10 different anechoic speech signals from the TIMIT database, uttered by 5 males and 5 females all sampled at 16 KHz, are convolved with the RIRs to generate the reverberant speech files. The size of the room used in the case of simulated RIRs is 10 x 10 x 10 (m^3). The Hanning window of 256 samples is used with an overlap factor set to 50%. The STFT length is 256. The rest of the parameters are set as: τ= 0.1, C= 2.5, N_{le}= 13, R= 128, ψ= 25, and $\tilde{G}^{min}_{late}$= 2.22×10^{-16}. Performance indices used in the evaluations are the mean segmental signal to reverberation ratio (SegSRR) [6] [10] and the SDR [8] [10]. We use the first method in [5] (called for short Jeub *et al.* method hereafter) as the baseline which represents the state-of-the-art and uses the frequency-independent model for decay rate estimation. Note that, the second method in [5] was proposed for dealing with early reflections, and thus not considered here.

First, we present a dereverberation example for the real data recorded in a lecture room [4], where the T_{60} is approximately 900 ms and the source-microphone distance is 2.25 m. The spectrograms of the signals are shown in Figure 1. For comparison we highlight 3 different regions which are marked as A_i, B_i and C_i, where $i = 1$ is for the clean signal, $i = 2$ for the dereverberated signal by Jeub *et al.* method and $i = 3$ for the dereverberated signal from the proposed method. From the highlighted regions we can observe that the signal obtained by our proposed method is closer to the clean one as compared to the Jeub *et al.* method.

We further evaluate the performance of our proposed method in comparison to the Jeub *et al.* method using SDR and the mean SegSRR of the output. First we used the simulated RIRs to generate the reverberant signals from the anechoic speech signals at three different reverberation times, i.e., T_{60} = {300, 500, 600} ms, and two different source-microphone distances, i.e., 0.5 and 2.5 m respectively. For each T_{60} and source-microphone distance, 5 different source-microphone positions and the 10 anechoic signals from the TIMIT database, resulting in 50 different reverberant signals, were used for testing the algorithms. In total, 300 independent tests were run for the simulated data. Table I shows for each T_{60} and source-microphone distance the results (mean values $\pm$ standard deviations) averaged over the 50 tests. The results indicate that our proposed method gives consistently higher

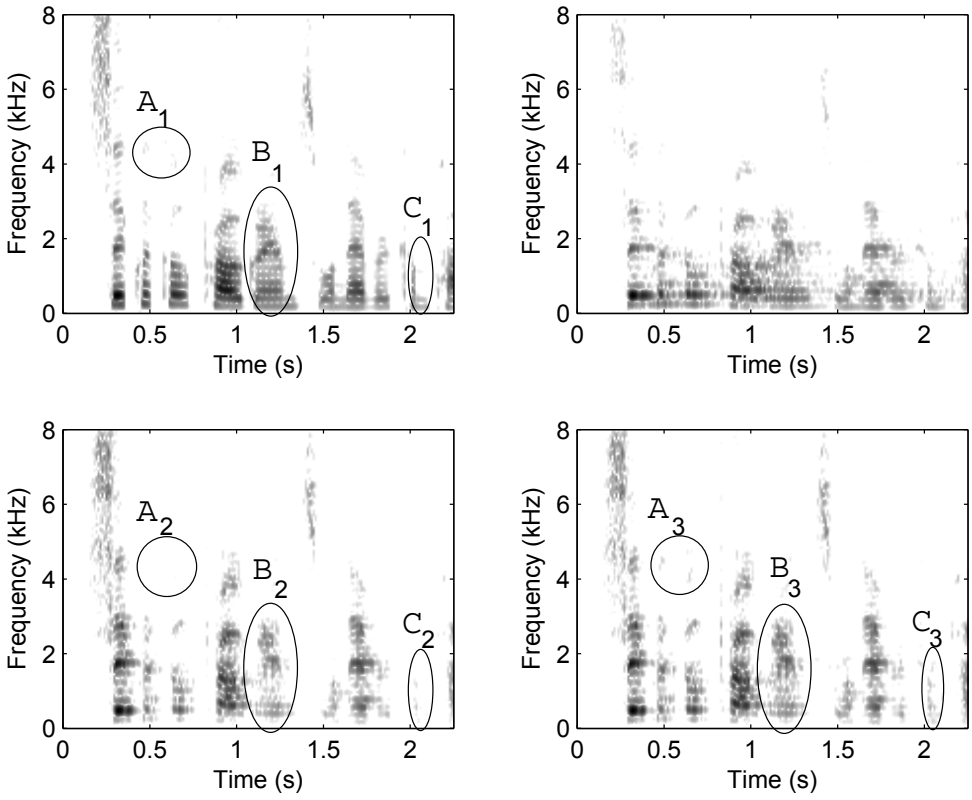

Fig. 1. Comparison of the spectrograms of the clean signal (top left) with the enhanced signals obtained by the proposed method (bottom right) and the Jeub *et al.* method (bottom left). The top right plot shows the reverberant signal.

SDRs and SegSRRs than Jeub *et al.* method for various source-microphone distances and reverberation times.

In another set of experiments, we used the real RIRs from the AIR database [4] which contains five different types of RIRs, recorded in five different room environments, namely booth, office, meeting, lecture, and stairway. For each room environment, we selected a pair of source-microphone distances $\{D_1, D_2\}$ m, respectively $\{0.5,\ 1.5\}$, $\{1,\ 3\}$, $\{1.45,\ 2.8\}$, $\{2.25,\ 7.1\}$, and $\{1,\ 3\}$. The 10 anechoic signals from the TIMIT database are then convolved with each of these RIRs, resulting in 100 reverberant signals in total. For each room type and source-microphone distance, the average results of SDR and SegSRR over the 10 different signals, are given in Figure 2. The proposed method performs significantly better than Jeub *et al.* method for shorter source-microphone distances. For example, for the booth and D_1 = 0.5 m, both SDR and SegSRR obtained by the proposed method are about 10 dB higher than those by Jeub *et al.* method. Such an improvement, observed for nearly all the testing cases, decreases when the source-microphone distance increases. Averaged over all the 100 tests, the SDR and SegSRR of the proposed method are respectively 2.3 dB and 2.5 dB higher than those of the Jeub *et al.* method. These results demonstrate the advantage of the frequency dependent model in particular for shorter source-microphone distances.

TABLE I
SDR AND SEGSRR FOR SIMULATED DATA UNDER DIFFERENT $T_{60}s$ AND SOURCE-MICROPHONE DISTANCE

D (m)	T_{60} (ms)	SDR (dB)		SegSRR (dB)	
		Proposed method	Jeub *et al.* method	Proposed method	Jeub *et al.* method
0.5	300	17.44 ± 0.98	15.68±1.40	12.92±0.71	10.57±0.74
	500	13.12±1.12	11.53±1.20	8.83±0.70	7.27±0.66
	600	11.98±1.14	10.38±1.17	7.74±0.69	6.41±0.63
2.5	300	8.67±0.69	8.20±0.74	5.80±0.55	5.21±0.57
	500	5.06±0.72	4.37±0.75	2.17±0.58	1.94±0.53
	600	3.98±0.74	3.23±0.76	1.07±0.60	1.01±0.53

V. CONCLUSION

We have presented a dereverberation algorithm based on a frequency dependent statistical model of the reverberation time. The algorithm is composed of the estimation of the decay rate of the late reverberations based on this model, the estimation of the mask containing spectral subtraction gains, and the smoothing of the spectral mask by a frequency dependent filter. We have shown that the proposed algorithm offers considerably higher performance as compared with a related recent approach using the frequency independent model.

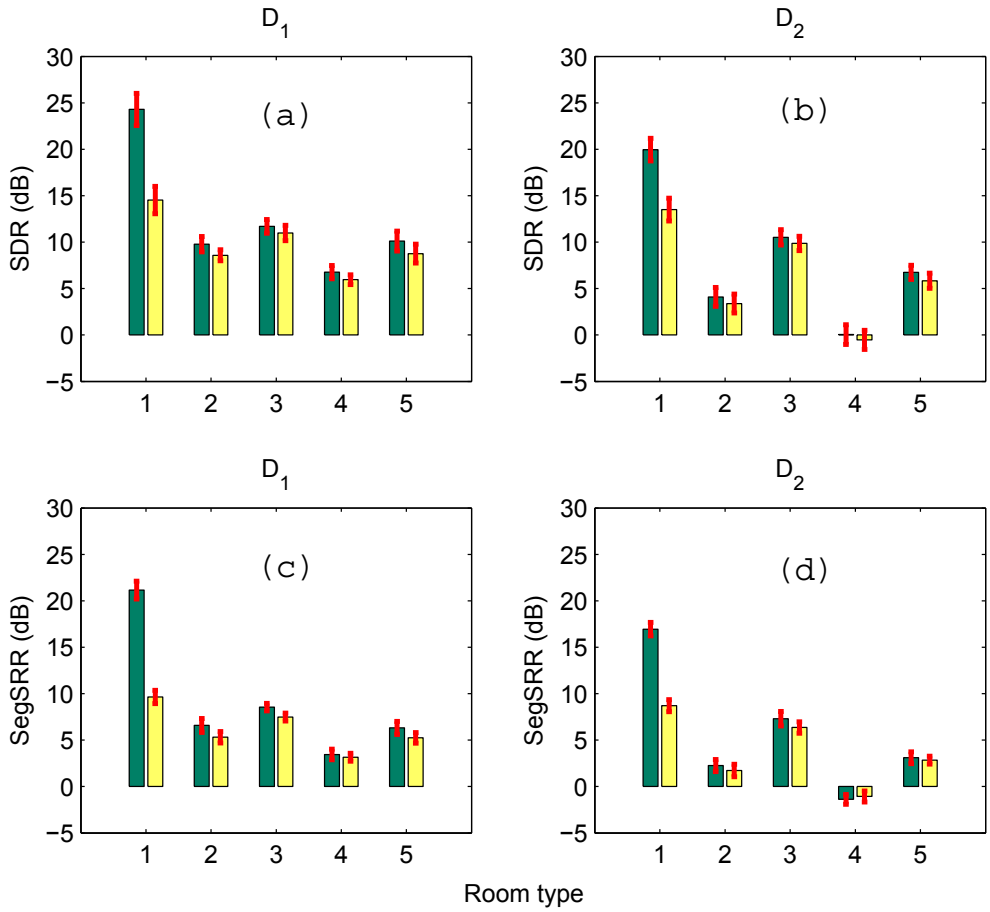

Fig. 2. SDR and SegSRR for the AIR database of the proposed method (green bars) and Jeub *et al.* method (yellow bars). The labels on the horizontal axis represent different room types, namely, 1 - booth, 2 - office, 3 - meeting, 4 - lecture, 5 - stairway. For each of the five rooms, two different source-microphone distances were tested, respectively D_1 = {0.5, 1, 1.45, 2.25, 1} m and D_2 = {1.5, 3, 2.8, 7.1, 3} m. The standard deviations are also plotted as short lines on top of the bars.

ACKNOWLEDGMENT

We are grateful to E. A. P. Habets for discussion on spectral variance estimation and for proofreading the manuscript. T. Jan was supported by the UET Peshawar Pakistan, and W. Wang was supported in part by the EPSRC of the UK (Grant EP/H050000/1 and EP/H012842/1).

REFERENCES

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

[2] E. A. P. Habets, "Single and multi-microphone speech dereverberation using spectral enhancement," Ph.D. dissertation, Technische Univ. Eindhoven, Eindhoven, The Netherlands, Jun. 2007.

[3] E. A. P. Habets, S. Gannot, and I. Cohen, "Late reverberant spectral variance estimation based on a statistical model," *IEEE Signal Process. Letters*, vol. 16, no. 9, pp. 770–773, Sep. 2009.

[4] M. Jeub, M. Schafer, and P. Vary, "A binaural room impulse response database for the evaluation of dereverberation algorithms," in *Proc. Int. Conf. Digital Signal Process. (DSP)*, Santorini, Greece, 2009.

[5] M. Jeub, M. Schafer, T. Esch, and P. Vary "Model-based dereverberation preserving binaural cues," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 18, pp. 1732–1745, Sep 2010.

[6] P. Krishnamoorthy and S. R. Mahadeva Prasanna, "Reverberant speech enhancement by temporal and spectral processing," *IEEE Trans. Audio, Speech, and Language Process.*, vol. 17, no. 2, pp. 253–266, 2009.

[7] K. Lebart, J. M. Boucher, and P. N, Denbigh, "A new method based on spectral subtraction for speech dereverberation", *Acta Acust. United With Acust.*, vol. 87, no. 3, pp. 359–366, 2001.

[8] M. I. Mandel, R. J. Weiss, and P. W.Ellis, "Model-based expectation maximization source separation and localization," *IEEE Trans. Audio, Speech Lang. Process.*, vol. 18, pp. 382–394, Feb 2010.

[9] M. Miyoshi and Y. Kaneda, "Inverse filtering of room acoustics," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 36, no. 2, pp. 145–152, Feb. 1988.

[10] P. A. Naylor, N. D. Gaubitch and E. A. P. Habets, "Signal-based performance evaluation of dereverberation algorithms," *Journal of Electrical and Computer Engineering*, Vol. 2010, Article ID 127513, 5 pages, doi:10.1155/2010/127513, 2010.

[11] S. Neely and J. Allen "Invertibility of a room impulse response," *J. Acoust. Soc. Amer.*, vol. 66, pp. 165–169, 1979.

[12] R. Ratnam, D. L. Jones, B. C. Wheeler, W. D. O´Brien, C. R. Lansing, and A. S. Feng, "Blind estimation of reverberation time," *J. Acoust. Soc. Am.*, pp. 2877–2892, 2003.

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

[14] M. Wu and D. L. Wang, "A two-stage algorithm for one-microphone reverberant speech enhancement," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 14, pp. 774–784, May 2006.