ADAPTIVE DEREVERBERATION METHOD BASED ON COMPLEMENTARY WIENER FILTER AND MODULATION TRANSFER FUNCTION

Kento Ohtani[†], *Tatsuya Komatsu*[†], *Takanori Nishino*[‡], *and Kazuya Takeda*[†]

[†]Graduate School of Information Science, Nagoya University, Nagoya, Japan
[‡]Graduate School of Engineering, Mie University, Tsu, Japan

ABSTRACT

Many methods for dereverberation of speech have been proposed, however most of these methods require large computational resources. In this paper, we evaluate a method based on a complementary Wiener filter which requires very little computation. Because the filter coefficients used with this method have a large impact on dereverberation performance, we propose method to estimate these filter coefficients. Our filter estimation method is based on the observation that modulation transfer functions have inverse characteristics of dereverberation filter. We also use spectral subtraction for noise reduction and an ML estimator to estimate reverberation time. The results of our experimental evaluation show that our proposed method can effectively suppress reverberation with only a slight time delay.

Index Terms— dereverberation, real-time processing, single-channel, Wiener filter, modulation transfer function

1. INTRODUCTION

Speech recognition technology is currently in use in applications such as the voice call function of smart phones, for voice-activated information searches, and in other types of hands-free computer interfaces, and its use is rapidly spreading. These systems are often used in rooms, however, and when we use a microphone indoors, the reflection of sound waves from the walls, called reverberation, reduces the clarity of speech. Therefore, removing acoustic distortion from reverberant speech is an important issue, and many dereverberation techniques have been proposed. When we use dereverberation methods with voice call and teleconferencing systems, it is desirable that there be no delay due to computational processing. Moreover, current dereverberation methods require additional filtering processes, such as noise reduction. However, most of the devices using these systems have limited computational speed and resources. Thus, it is desirable to minimize computation during dereverberation.

Lebart et al. proposed a single channel dereverberation technique using spectral subtraction (SS) [1]. They assumed that the late reverberant component of recorded speech is equal to the target speech, when time delay and amplitude attenuation are taken into consideration [2]. Unoki et al. proposed a dereverberation method based on restoration of the speech power envelope [3]. Using the conventional method, the speech power envelope component is first separated from reverberant speech. Then, the power envelope component is filtered and integrated with the carrier component.

Regarding multi-channel methods, semi-blind MINT (Multiple-input/output INverse filtering Theorem) [4], multichannel multiple-step linear prediction [5] and others have been proposed. These multi-channel methods can perform dereverberation more accurately than single channel methods, however multi-channel methods require multiple microphones and large computational resources.

Kondo et al. proposed a dereverberation method based on a complementary Wiener filter [6]. This method requires much less computation than other proposed methods, and thus has the advantages of little time delay and use of fewer computational resources. The drawback with this method is the care which must be taken when selecting dereverberation parameters. These parameters have a large impact on dereverberation performance [7], thus selecting the appropriate parameters is crucial.

In this paper, we propose a method of dealing with noisy reverberant data. First, we use spectral subtraction for noise reduction. Then, for blind dereverberation, we use a maximum likelihood reverberation time estimator. Next we estimate the parameters of the dereverberation filters. Our parameter estimation method is based on the fact that modulation transfer functions have the inverse characteristics of dereverberation filters. Because our method requires very little computation, it can perform dereverberation very quickly. Finally, we perform an experimental evaluation to confirm the performance of our method.

2. CONVENTIONAL USE OF WIENER FILTER FOR DEREVERBERATION

Kondo et al. proposed a dereverberation method based on a complementary Wiener filter [6]. Dereverberated speech processed using a complementary Wiener filter can be expressed as follows:

$$Z(k,l) = G(k,l)X(k,l) \tag{1}$$

where $Z(k,l)$ represents dereverberated speech, $X(k,l)$ represents the observed speech in the frequency domain, k represents the frequency bin index and l represents the frame index. $G(k,l)$, which is the spectral gain of the dereverberation, is based on a complementary Wiener filter, which can be calculated as:

$$G(k,l) = \begin{cases} 1, & \dfrac{P_X(k,l)}{P_X(k)} \geq 1 \\ \dfrac{P_X(k,l)}{P_X(k)}, & \text{otherwise} \end{cases}, \tag{2}$$

where,

$$P_X(k,l) = \left|X(k,l)\right|^2, \tag{3}$$

$$P_X(k) = E_l[|X(k,l)|^2]. \tag{4}$$

$E_l[\cdot]$ is the expectation of the power spectrum over frame index l. To solve the gain calculation, we approximate expectation $E_l[\cdot]$ using the exponential moving average, as shown in Eq.(5):

$$R_X(k,l) = (1-\alpha)P_X(k,l) + \alpha R_X(k,l-1), \tag{5}$$

where, $\alpha(0 < \alpha < 1)$ is a weighting factor representing how many past values we consider, which is called a smoothing constant. Thus, dereverberation gain can be represented as:

$$G(k,l) = \begin{cases} 1, & \dfrac{P_X(k,l)}{R_X(k,l)} \geq 1 \\ \dfrac{P_X(k,l)}{R_X(k,l)}, & \text{otherwise} \end{cases}. \tag{6}$$

In Eq.(6), we calculated the ratio between instantaneous speech power $P_X(k,l)$ and long term average $R_X(k,l)$. However, when the $P_X(k,l)$ is almost 0, the dereverberation gain also becomes almost 0, which results in a musical tone. Therefore, as an approximate value of $P_X(k,l)$, we use the short term exponential moving average of $P_X(k,l)$. In other words, we use two smoothing constant α_1 and $\alpha_2(\alpha_1 < \alpha_2)$ to calculate short term average $R_1(k,l)$ and long term average $R_2(k,l)$. Finally, dereverberation gain can be represented as:

$$G(k,l) = \begin{cases} 1, & \dfrac{R_1(k,l)}{R_2(k,l)} \geq 1 \\ \dfrac{R_1(k,l)}{R_2(k,l)}, & \text{otherwise} \end{cases}. \tag{7}$$

To calculate this gain, we only use the exponential moving average of speech power. In the discussions below, we will use Eq.(7) for dereverberation.

3. ESTIMATION OF FILTER COEFFICIENTS

When we use a dereverberation method based on a complementary Wiener filter, we must set filter coefficients α_1 and α_2 of the smoothed filter. This is a crucial step, because the performance of this dereverberation method will vary widely in response to changes in the filter coefficients.

In this section, we will discuss filter coefficient estimation methods used to choose the best filter coefficients. Our estimation method is based on the fact that the modulation transfer function has the inverse characteristics of the dereverberation filter.

3.1. Modulation transfer function

In this paper, we assume that the effect of reverberation can be represented by the modulation transfer function (MTF) [3]. If we regard the room where speech occurs as the transduction pathway, MTF reveals the relationship between the input (source sound) and the output (reverberant sound). Input and output power change in the frequency domain can be defined as:

$$\text{Input}: \quad \overline{I_i}(1 + \cos(2\pi f_m t)), \tag{8}$$

$$\text{Output} \quad \overline{I_o}(1 + A(f_m)\cos(2\pi f_m t - \theta)), \tag{9}$$

where $\overline{I_i}$ is input power, $\overline{I_o}$ is output power, and θ is the phase of each modulation frequency. f_m, which represents the modulation frequency, shows the frequency of the speech power envelope. Relationship between modulation frequency f_m and usual freqency f is $f_m = f/R$, where R is the frame shift width. Moreover, frame index l and time t have the same relation, that is $t = l{\cdot}R$.

$A(f_m)$ in the Eq.(9) is the MTF. From Eq.(8) and (9), the MTF under the reverberant condition can be calculated as:

$$A(f_m) = \frac{\left|\int_0^\infty e_h^2(t)\exp(-j2\pi f_m t)dt\right|}{\int_0^\infty e_h^2(t)dt}, \tag{10}$$

where $e_h^2(t)$ is the power envelope of the room impulse response (RIR). Here, we use the exponential decay model of the RIR [8]:

$$e_h^2(t) = \exp\left(-\frac{3\ln 10}{T_{60}}t\right). \tag{11}$$

Therefore, the MTF can be expressed as:

$$A(f_m) = \left\{1 + \left(2\pi f_m \frac{T_{60}}{6\ln 10}\right)^2\right\}^{-1/2}, \tag{12}$$

where T_{60} is the reverberation time. Fig. 1 shows the MTFs of different reverberation times. The vertical axis represents the MTF $A(f_m)$, and the horizontal axis represents the modulation frequency f_m. From this figure we can see that under the reverberant condition, the high frequency component of the speech power envelope is highly suppressed. Moreover, as reverberation time increases, suppression increases.

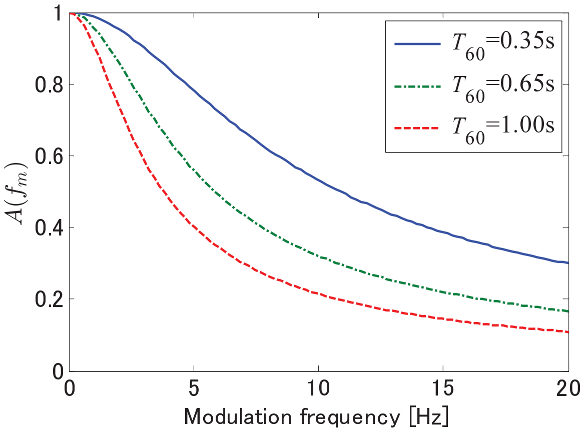

Fig. 1. Relationship between MTFs and modulation frequency.

3.2. Amplitude response of dereverberation filter

When reverberant speech passes through the dereverberation filter, the speech is affected by the amplitude response of the filter. Therefore, we need to consider the amplitude response of the dereverberation filter.

To calculate the amplitude response, we make a z-transform of the exponential moving average shown in Eq.(5):

$$R_X[k,z] = (1-\alpha)P_X[k,z] + \alpha z^{-1}R_X[k,z], \tag{13}$$

where $R_X[k,z]$ and $P_X[k,z]$ represent the z-transformations of $R_X(k,l)$ and $P_X(k,l)$, respectively. Rearranging this equation to obtain $R_X[k,z]$, we get:

$$R_X[k,z] = \frac{1-\alpha}{1-\alpha z^{-1}}P_X[k,z]. \tag{14}$$

Similarly, by making a z-transform of the dereverberation gain from Eq.(2), we get:

$$G[k,z] = \frac{R_1[k,z]}{R_2[k,z]}. \tag{15}$$

Then, we substitute Eq.(14) and $z = \exp(-j2\pi f_m)$ to compute the frequency response $H(f_m)$ of dereverberation gain $G(k,l)$:

$$H(f_m) = \frac{1-\alpha_1}{1-\alpha_2}\frac{1-\alpha_2\mathrm{exp}(-j2\pi f_m)}{1-\alpha_1\mathrm{exp}(-j2\pi f_m)}. \tag{16}$$

Fig. 2 shows the amplitude responses $|H(f_m)|$ of the dereverberation filters for various smoothing constants α_2. In this figure, smoothing constant $\alpha_1 = 0.4$ is fixed. The vertical axis represents amplitude response $|H(f_m)|$, and the horizontal axis represents modulation frequency f_m. From this figure, we can see that the dereverberation filter enhances high frequency components of the speech power envelope, in contrast to the MTF. Moreover, as smoothing constant α_2 becomes larger, enhancement of high frequency components becomes stronger.

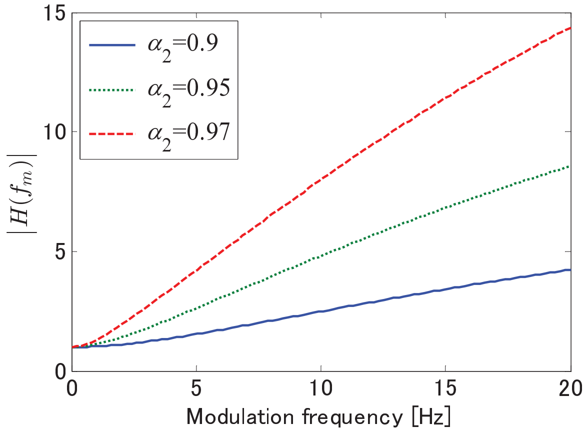

Fig. 2. Amplitude responses of the dereverberation filter.

3.3. Estimation of filter coefficients

An overview of our dereverberation process can be considered as below. First, a sound source produces the speech signal $X(f_m)$. Second, speech signal $X(f_m)$ is affected by RIRs and background noise. Thus, assuming that $A(f_m)$ is an effect of the MTF and background noise, noisy reverberant speech can be considered as $A(f_m)X(f_m)$. Next, the reverberant speech passes through the dereverberation filter, and is affected by the amplitude response of that filter $|H(f_m)|$. Therefore, the dereverberation filter outputs the dereverberated signal $|H(f_m)|\,A(f_m)X(f_m)$. Thus, it follows that if:

$$|H(f_m)|\,A(f_m) = 1 \tag{17}$$

for all of modulation frequency f_m, the dereverberated speech can be considered to be equal to the source speech signal. In this paper, we assume that reverberation time is already given or can be easily estimated. Therefore, the MTF can be found, and we only need to adjust smoothing constants α_1 and α_2.

For filter coefficient estimation, we use the following sum of squares error function:

$$E = \frac{1}{2}\sum_{m=1}^{M}\left\{1-|H(f_m)|A(f_m)\right\}^2, \tag{18}$$

where m denotes the modulation frequency bin index and M is the number of modulation frequency bins. Taking the derivative of E with respect to α_1 and α_2, we get:

$$\frac{\partial E}{\partial \alpha_1} = -\frac{1+\alpha_1}{1-\alpha_1}\sum_{m=1}^{M}\frac{1-\cos(2\pi f_m)B(f_m)}{|1-\alpha_1\mathrm{exp}(-j2\pi f_m)|^2}, \tag{19}$$

$$\frac{\partial E}{\partial \alpha_2} = \frac{1+\alpha_1}{1-\alpha_2}\sum_{m=1}^{M}\frac{1-\cos(2\pi f_m)B(f_m)}{|1-\alpha_2\mathrm{exp}(-j2\pi f_m)|^2}, \tag{20}$$

where $B(f_m)$ is as follows:

$$B(f_m) = |H(f_m)|\,A(f_m)\left\{|H(f_m)|\,A(f_m)-1\right\}. \tag{21}$$

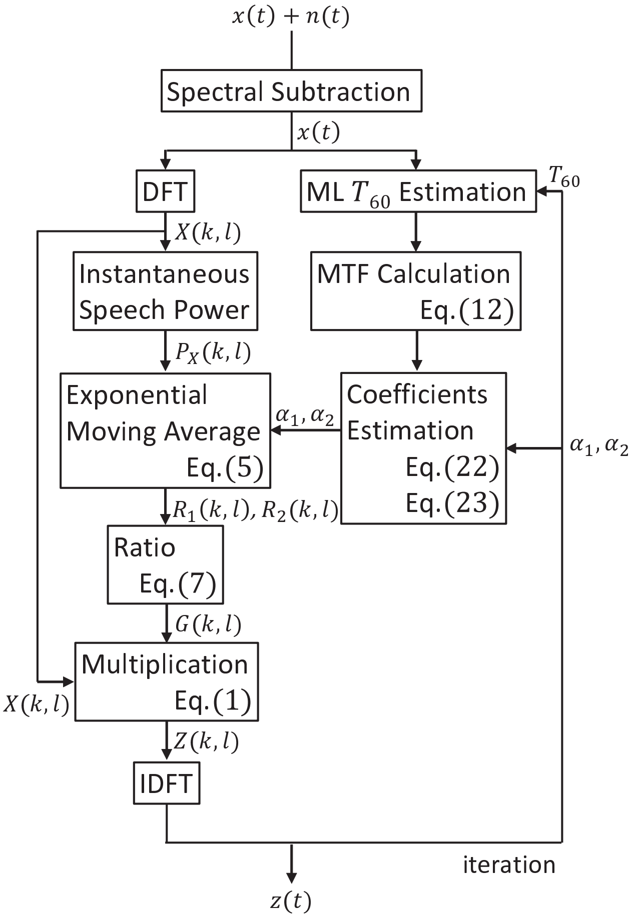

Fig. 3. Overview of a single iteration of our dereverberation method.

To estimate the filter coefficients, we use the steepest descent method and the following iteration equations:

$$\alpha_1^{(i+1)} = \alpha_1^{(i)} - \lambda_1 \left.\frac{\partial E}{\partial \alpha_1}\right|_{\alpha_1=\alpha_1^{(i)}}, \tag{22}$$

$$\alpha_2^{(i+1)} = \alpha_2^{(i)} - \lambda_2 \left.\frac{\partial E}{\partial \alpha_2}\right|_{\alpha_2=\alpha_2^{(i)}}, \tag{23}$$

where i is the iteration index, and λ_1 and λ_2 denote the step width.

4. PROPOSED METHOD

An overview of the dereverberation method described above is shown in Fig. 3. Our method, which uses a complementary Wiener filter, is unable to suppress background noise, however. Therefore, we use ssubmmse() function in Voicebox [9] for noise reduction. Next, we use a maximum likelihood (ML) estimator to estimate reverberation time [10], and then estimate the smoothing constants. Finally, we perform dereverberation using the conventional method. Repeating this algorithm frame by frame, dereverberation can be done adaptively.

In the initialization step, we set the initial values of the reverberation time and smoothing constants α_1 and α_2. In addition, using the initial few frames, we estimate the strength of the background noise for spectral subtraction.

Table 1. Dereverberation conditions

Sampling frequency	16,000 Hz
Frame length	32 ms (512 pt)
Frame shift	8 ms (128 pt)
DFT length	512 pt
Initial reverberation time	0.2 s
Initial value for α_1	0.2
Initial value for α_2	0.8
Noise estimation data length	128 ms (2048 pt)

5. EXPERIMENTS

5.1. Experimental conditions

Dereverberation conditions are shown in Table 1. Reverberation time and smoothing constants α_1 and α_2 are initialized using the values in the table and are updated iteratively using the method described above. REVERB-Challenge data [11] was used for objective evaluation. There are two kinds of speech datasets, namely simulated dataset and real recording dataset.

The simulated dataset consisted of utterances from the WSJCAM0 corpus [12], which are convolved by RIRs measured in different rooms. It simulates six different reverberation conditions: three types of rooms with different volumes (small, medium and large), and two distances between the speaker and the microphone array (near = 50 cm, far = 200 cm). Reverberation times for each room are about 0.25 s, 0.5 s and 0.7 s, respectively. For the simulated dataset, noise is added to the reverberant speech with a signal-to-noise ratio (SNR) of 20 dB.

The real recording dataset consists of utterances from the MC-WSJ-AV corpus [13], which are utterances recorded in a noisy and reverberant room. Reverberation time of the real recording dataset is 0.7 s. Distances between the speaker and the microphone array are either "near" (= 100 cm) or "far" (= 250 cm).

5.2. Experimental results

Experimental results are shown in Table 2. "Reverberant" represents the results for noisy reverberant speech data, and "Enhanced" represents the results after dereverberation.

For objective evaluation, we used four objective measures: cepstrum distance (CD), log likelihood ratio (LLR),

Table 2. Experimental results (lower values of CD and LLR indicate higher sound quality, while higher values of FWSegSNR and SRMR indicate higher sound quality)

Datasets		Simulated Dataset							Real Recording Dataset		
Environments		Small room		Medium room		Large room		Average	Far	Near	Average
		Far	Near	Far	Near	Far	Near				
CD [dB]	Reverberant	2.67	1.99	5.21	4.63	4.96	4.38	3.97	-	-	-
	Enhanced	3.93	3.89	5.02	4.26	4.84	4.08	4.34	-	-	-
LLR	Reverberant	0.38	0.35	0.75	0.49	0.84	0.65	0.58	-	-	-
	Enhanced	0.61	0.60	0.91	0.65	0.94	0.73	0.74	-	-	-
FWSegSNR [dB]	Reverberant	6.68	8.12	1.04	3.35	0.24	2.27	3.62	-	-	-
	Enhanced	7.69	7.63	5.50	7.79	4.78	6.86	6.71	-	-	-
SRMR	Reverberant	4.58	4.50	2.97	3.74	2.73	3.57	3.68	3.19	3.17	3.18
	Enhanced	5.53	5.21	4.32	5.06	4.08	4.92	4.85	5.97	6.13	6.05

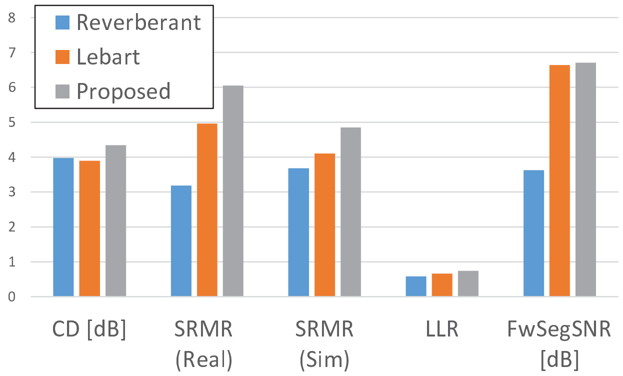

Fig. 4. Results compared with the Lebart method.

frequency-weighted segmental SNR (FWSegSNR) [14] and speech-to-reverberation modulation energy ratio (SRMR) [15]. CD, LLR and FWSegSNR are based on the discrepancy between the target and reference signals. The clean speech signal is used as the reference, thus these measures are used only for the simulated dataset. In contrast, SRMR can be calculated only from target signals, thus SRMR is used for both datasets. The smaller values of CD and LLR mean better dereverberation of speech. On the other hand, greater values of FWSegSNR and SRMR mean better dereverberation of speech.

Compared with the original reverberant speech, CD and LLR became a little lower in quality after dereverberation. In contrast, FWSegSNR and SRMR improved in quality. From these results, we see that our dereverberation method can suppress reverberation with minor processing distortion.

In Fig. 4, we compare the results of the proposed method and Lebart's method [2]. In this figure, reverberant shows the results of reverberant noisy speech. Proposed and Lebart means the results of proposed method and Lebart's method, respectively. From this figure, we can see that results of CD, LLR and FWSegSNR are comparable between the two methods, but that the results of SRMR is higher for the proposed method than for Lebart's method.

Computational complexity is also an important factor in speech dereverberation. We used processing time and real time factor (RTF) to evaluate computational complexity, and the results are shown in Table 3. Our equipment specifications were as follows: we used MATLAB 2012a for dereverberation processing with an Intel® Core^{TM} i5-3470 processor (4 core, up to 3.60 GHz) and a 12 GB memory (DDR3-1333). Compare to the Lebart's method, our proposed method can process much faster.

According to the RTF, we can process speech faster than the speed of the speech. Thus, the processing delay is only 32 ms, which is the frame length for this dereverberation method. For noise estimation, we used the first 128 ms (2048 pt) of speech data. Therefore, the maximum delay using this dereverberation method is estimated to be 160 ms, small enough for real time processing.

6. CONCLUSION

In this paper, we proposed dereverberation method based on the complementary Wiener filter. For filter parameter estimation, we use MTF and the amplitude response of the dereverberation filter. Parameters are selected to neutralize the effect of the MTF. Spectral subtraction is used to suppress noise, and an ML estimator is used to estimate reverberation time. From the results of our experimental evaluation, it is shown that our method can dereverberate speech effectively with limited computational resources and with almost no processing delay.

7. ACKNOWLEDGEMENTS

Special thanks to Kazunobu Kondo of the Yamaha Corporation for his insightful advice and suggestions regarding this research. This work is partly supported by a Grant-in-Aid for Scientific Research (No. 25280060).

Table 3. Processing time

Method	Proposed		Lebart [2]	
Dataset	Simulated dataset	Real recording	Simulated dataset	Real recording
Length of data [s]	17232	2424.1	17232	2424.1
Processing time [s]	1281.8	171.39	6257.9	871.33
RTF	0.0743	0.0707	0.3632	0.3594

8. REFERENCES

[1] S.F. Boll, "Suppression of acoustic noise in speech using spectral subtraction," *IEEE Trans. Acoustics, Speech and Signal Processing*, vol. 27, no. 2, pp. 113–120, Apr. 1979.

[2] K. Lebart and J.M. Boucher, "A new method based on spectral subtraction for speech dereverberation," *Acta Acustica*, vol. 87, pp. 359–366, Apr. 2001.

[3] M. Unoki, K. Sakata, M. Furukawa, and M. Akagi, "A speech dereverberation method based on the mtf concept in power envelope restoration," *Acoustical Science and Technology*, vol. 25, no. 4, pp. 243–254, July 2004.

[4] K. Furuya, "Noise reduction and dereverberation using correlation matrix based on the multipleinput/output inverse-filtering theorem (MINT)," *Proc. of International Workshop on Hands-Free Speech Communication*, pp. 59–62, Apr. 2001.

[5] K. Kinoshita, M. Delcroix, T. Nakatani, and M. Miyoshi, "Suppression of late reverberation effect on speech signal using long-term multiple-step linear prediction," *IEEE Trans. Audio, Speech, and Language Processing*, vol. 17, no. 4, pp. 534–545, May 2009.

[6] K. Kondo, Y. Takahashi, T. Komatsu, T. Nishino, and K. Takeda, "Computationally efficient single channel dereverberation based on complementary Wiener filter," *Proc. of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2013)*, pp. 7452–7456, May 2013.

[7] K. Ohtani, T. Komatsu, K. Kondo, T. Nishino, and K. Takeda, "Objective and subjective evaluation of complementary Wiener filter for speech dereverberation," *Proc. of International Congress on Acoustics 2013 (ICA2013)*, vol. 19, June 2013.

[8] J.D. Polack, "La transmission de l'énergie sonore dans les salles," *Dissertation. Université du Maine*, Dec. 1988.

[9] D. M. Brookes, "Voicebox: Speech processing toolbox for matlab," `http://www.ee.ic.ac.uk/hp/staff/dmb/voicebox/voicebox.html`, 1997, [Online; accessed Aug. 5, 2013].

[10] H. W. Löllmann, E. Yilmaz, M. Jeub, and P. Vary, "An improved algorithm for blind reverberation time estimation," *International Workshop on Acoustic Echo and Noise Control (IWAENC)*, Aug. 2010.

[11] K. Kinoshita, M. Delcroix, T. Yoshioka, T. Nakatani, E. Habets, R. Haeb-Umbach, and V. Leutnant, "The reverb challenge: A common evaluation framework for dereverberation and recognition of reverberant speech," *Proc. of IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA-13)*, Oct. 2013.

[12] T. Robinson, J. Fransen, D. Pye, J. Foote, and S. Renals, "Wsjcam0: A British English speech corpus for large vocabulary continuous speech recognition," *Proc. of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP-95)*, pp. 81–84, 1995.

[13] M. Lincoln, I. McCowan, J. Vepa, and H.K. Maganti, "The multi-channel Wall Street Journal audio visual corpus (mc-wsj-av): Specification and initial experiments," *IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*, pp. 357–362, Nov. 2005.

[14] Y. Hu and P.C. Loizou, "Evaluation of objective quality measures for speech enhancement," *IEEE Trans. Audio, Speech, and Language Processing*, vol. 16, no. 1, pp. 229–238, Jan. 2008.

[15] T. H. Falk, C. Zheng, and W. Chan, "A non-intrusive quality and intelligibility measure of reverberant and dereverberated speech," *IEEE Trans. Audio, Speech, and Language Processing*, vol. 18, no. 7, pp. 1766–1774, Sep. 2010.