ADAPTIVE DEREVERBERATION OF SPEECH SIGNALS WITH SPEAKER-POSITION CHANGE DETECTION

Takuya Yoshioka, Hideyuki Tachibana, Tomohiro Nakatani, and Masato Miyoshi

NTT Communication Science Laboratories, NTT Corporation
2-4, Hikari-dai, Seika-cho, Soraku-gun, Kyoto 619-0237, Japan

ABSTRACT

This paper proposes a method for adaptive speech dereverberation and speaker-position change detection, which have not previously been addressed. Signal transmission channels in rooms are modeled as auto-regressive systems in individual frequency bands. The proposed method adaptively estimates the regression coefficients of this model, which are called room regression coefficients (RRCs). The proposed method has two distinguishing features: (1) The method is based on the weighted recursive least squares algorithm, which enables an efficient RRC-estimate update as well as a fast convergence rate; (2) The method detects changes in speaker position and so can quickly catch up with the sudden channel changes that such position changes cause. Detection is realized by finding time frames where the power of dereverberated speech is anomalously amplified. Experimental results showed that the proposed method attained convergence in 5 seconds and successfully detected changes in speaker position.

Index Terms— Dereverberation, speech enhancement, adaptive filters, weighted RLS algorithm,

1. INTRODUCTION

Room reverberation degrades the quality of speech and the automatic speech recognition performance. Therefore, many researchers have studied the speech dereverberation technology.

The existing dereverberation methods may be divided into two classes. One class directly estimates clean speech signals without estimating room impulse responses (RIRs) or their inverse filters [1, 2]. The dereverberation methods in this class are advantageous in that they can work online by employing a small amount of prior knowledge such as the room's reverberation time.

The other class uses inverse filters for RIRs. As long as the speaker does not move during the observation, the methods in this class are able to yield high quality dereverberated speech. Another advantage lies in the fact that inverse filter based dereverberation systems are easily combined with other microphone array systems including beamformers and blind source separation systems [3]. Therefore, we have investigated this class of dereverberation methods in a series of our recent publications (see, for example, [4] and references therin). Most of the existing dereverberation methods in this class work only with batch processing. However, since the RIRs may change during the observation if the speaker moves or if multiple speakers utter alternately, it is essential that we estimate the inverse filters adaptively. Although an adaptive dereverberation method was proposed in [5], the method requires a lot of observation data to attain convergence.

There are three requirements that adaptive dereverberation methods must meet. First of all, the adaptive dereverberation methods need to have a fast convergence rate. Moreover, they need to detect changes in speaker position so as to catch up with sudden RIR changes caused by such speaker-position changes. Finally, they need to mitigate the reverberation effect even at the beginning of utterance. In [6], we addressed the last problem by using prior knowledge about room acoustics. This paper proposes a method for solving the first two problems. The method is derived on the basis of the weighted recursive least squares (RLS) algorithm, which endows the method with a fast convergence rate and an efficient update rule. The method can also detect changes in speaker position.

The remainder of this paper is organized as follows. Section 2 describes the proposed adaptive dereverberation method based on the weighted RLS algorithm. Section 3 proposes the speaker-position change detection method. These sections also report several experimental results. Section 4 concludes the paper.

2. ADAPTIVE SPEECH DEREVERBERATION WITH WEIGHTED RLS ALGORITHM

2.1. Adaptive speech dereverberation task

Throughout this paper, we represent acoustic signals in the short-time Fourier transform (STFT) domain. The STFT domain representation allows us to employ the dereverberation-friendly reverberation model proposed in [4]. Hereafter, we denote the frame and frequency-band indices by t and l, respectively, with $0 \leq l \leq L-1$, where L is the number of frequency bands.

Let $s_{t,l}$ and $x_{t,l}^{(m)}$ denote a clean speech signal and a reverberant speech signal observed by the m-th microphone, respectively. Reference [4] showed that the effect of room reverberation could be well modeled with multi-channel auto-regressive (MCAR) systems in individual frequency bands. Based on MCAR modeling of the room reverberation, the reverberant signal observed by the first microphone, $x_{t,l}^{(1)}$, is given by

$$x_{t,l}^{(1)} = \sum_{m=1}^{M} \sum_{k=1}^{K_l} g_{k,l}^{(m)*} x_{t-k,l}^{(m)} + s_{t,l}, \tag{1}$$

where superscript $*$ stands for the complex conjugate operator and $g_{k,l}^{(m)}$ is the k-th regression coefficient operating on the m-th observed signal in the l-th frequency band. We hereafter call $g_{k,l}^{(m)}$ a room regression coefficient (RRC). We can rewrite (1) as

$$x_{t,l}^{(1)} = \boldsymbol{g}_l^H \boldsymbol{x}_{t-1,l} + s_{t,l}, \tag{2}$$

$$\boldsymbol{g}_l = [g_{1,l}^{(1)}, \cdots, g_{K_l,l}^{(1)}, \cdots, g_{1,l}^{(M)}, \cdots, g_{K_l,l}^{(M)}]^T \tag{3}$$

$$\boldsymbol{x}_{t-1,l} = [x_{t-1,l}^{(1)}, \cdots, x_{t-K_l,l}^{(1)}, \cdots, x_{t-1,l}^{(M)}, \cdots, x_{t-K_l,l}^{(M)}]^T, \tag{4}$$

where superscripts H and T stand for the complex and non-complex conjugate operators, respectively. (2) leads to the following signal recovery model:

$$s_{t,l} = x_{t,l}^{(1)} - \boldsymbol{g}_l^H \boldsymbol{x}_{t-1,l}. \tag{5}$$

Thus, the task to be solved is the adaptive estimation of all RRCs, which are collectively represented as $\Theta = \{\boldsymbol{g}_l\}_{0 \le l \le L-1}$.

The task is more specifically defined as follows. Let us represent the set of all reverberant signals observed at the t-th time frame as $X_t = \{x_{t,l}^{(m)}\}_{0 \le l \le L-1, 1 \le m \le M}$. Furthermore, we represent all the observed signals up to the t-th time frame as $\bar{X}_t = \{X_\tau\}_{1 \le \tau \le t}$. Now, suppose that the observed signals are given frame by frame. Then, the task is defined as updating an estimate of Θ every time a new observation data set X_t is provided.

2.2. Bayesian estimation approach

We approach the above task by using the Bayesian estimation method. With the Bayesian estimation approach, we calculate $p(\Theta|\bar{X}_t)$ sequentially for every $t = 1, 2, \cdots$, where $p(\Theta|\bar{X}_t)$ is the posterior probability density function (PDF) of the RRC set, ${}_g\Theta$, given the observation data set up to the t-th frame, $\bar{X}_t$. The Bayes theorem tells us that $p(\Theta|\bar{X}_t)$ can be calculated from $p(\Theta|\bar{X}_{t-1})$ according to the following update rule:

$$p(\Theta|\bar{X}_t) = \frac{p(X_t|\Theta, \bar{X}_{t-1}) p(\Theta|\bar{X}_{t-1})}{\int p(X_t|\Theta, \bar{X}_{t-1}) p(\Theta|\bar{X}_{t-1}) d\Theta}. \tag{6}$$

The right hand side of (6) includes the observation data PDF for the current frame, $p(X_t|\Theta, \bar{X}_{t-1})$, and the posterior PDF of the RRC set at the immediately preceding frame, $p(\Theta|\bar{X}_{t-1})$. Below, we define these two PDFs and embody the update rule.

Now, let us assume that the clean speech signal, $s_{t,l}$, is drawn from the complex Gaussian distribution with mean 0 and variance ${}_s\lambda_{t,l}$. Note that $\{{}_s\lambda_{t,l}\}_{0 \le l \le L-1}$ corresponds to the short-time power spectrum of clean speech at frame t. Then, the observation data PDF for the current frame is given by

$$p(X_t|\Theta, \bar{X}_{t-1}) = \prod_{l=0}^{L-1} \mathcal{N}_{\mathbb{C}}\Big\{x_{t,l}^{(1)}; \boldsymbol{g}_l^H \boldsymbol{x}_{t-1,l},\ {}_s\lambda_{t,l}\Big\}, \tag{7}$$

where $\mathcal{N}_{\mathbb{C}}\{\boldsymbol{u}; \boldsymbol{\nu}, \Gamma\}$ is the PDF of the (possibly multivariate) complex Gaussian random variable $\boldsymbol{u}$ with mean $\boldsymbol{\nu}$ and covariance matrix Γ.

Next, we assume that the posterior distribution of $\boldsymbol{g}_l$ after observing $\bar{X}_{t-1}$ is a complex Gaussian distribution with mean $\boldsymbol{\mu}_l(t-1)$ and covariance matrix $\Phi_l(t-1)$. Therefore, the RRC posterior PDF at the immediately preceding frame is written as

$$p(\Theta|\bar{X}_{t-1}) = \prod_{l=0}^{L-1} \mathcal{N}_{\mathbb{C}}\{\boldsymbol{g}_l; \boldsymbol{\mu}_l(t-1), \Phi_l(t-1)\}. \tag{8}$$

By reorganizing (6) using (7) and (8), we finally obtain

$$p(\Theta|\bar{X}_t) = \prod_{l=0}^{L-1} \mathcal{N}_{\mathbb{C}}\{\boldsymbol{g}_l; \boldsymbol{\mu}_l(t), \Phi_l(t)\}, \tag{9}$$

where

$$\boldsymbol{\mu}_l(t) = \Phi_l(t)\Big(\frac{\boldsymbol{x}_{t-1,l} x_{t,l}^*}{{}_s\lambda_{t,l}} + \Phi_l(t-1)^{-1} \boldsymbol{\mu}_l(t-1)\Big) \tag{10}$$

$$\Phi_l(t) = \Big(\frac{\boldsymbol{x}_{t-1,l} \boldsymbol{x}_{t-1,l}^H}{{}_s\lambda_{t,l}} + \Phi_l(t-1)^{-1}\Big)^{-1}. \tag{11}$$

Note that the RRC posterior PDF at frame t, given by (9), is in the same form as that at frame $t-1$, given by (8).

Threfore, the adaptive dereverberation algorithm based on the Bayesian estimation method is summarized as follows.

Bayesian adaptive dereverberation algorithm

1. *Initialization*
   Set the initial mean $\boldsymbol{\mu}_l(0)$ and covariance matrix $\Phi_l(0)$ of the RRC posterior distribution for each $l \in \{0, \cdots, L-1\}$.
2. *Clean speech estimation and RRC update*
   Perform the following procedures sequentially for $t = 1, 2, \cdots$.
   (a) Calculate the clean speech signal estimate $\hat{s}_{t,l}$ for all $l \in \{0, \cdots, L-1\}$ as
   $$\hat{s}_{t,l} = x_{t,l} - \boldsymbol{\mu}_l(t-1)^H \boldsymbol{x}_{t-1,l}. \tag{12}$$
   (b) Estimate the clean speech power spectrum $\{{}_s\lambda_{t,l}\}_{0 \le l \le L-1}$. One convenient way to roughly estimate ${}_s\lambda_{t,l}$ is using the power spectrum of one of the M observed signals.
   (c) Update the mean $\boldsymbol{\mu}_l(t-1)$ and covariance matrix $\Phi_l(t-1)$ of the RRC posterior distribution for each $l \in \{0, \cdots, L-1\}$ according to the following update rules:
   $$\boldsymbol{\mu}_l(t) = \Phi_l(t)\Big(\frac{\boldsymbol{x}_{t-1,l} x_{t,l}^*}{{}_s\lambda_{t,l}} + \alpha \Phi_l(t-1)^{-1} \boldsymbol{\mu}_l(t-1)\Big) \tag{13}$$
   $$\Phi_l(t) = \Big(\frac{\boldsymbol{x}_{t-1,l} \boldsymbol{x}_{t-1,l}^H}{{}_s\lambda_{t,l}} + \alpha \Phi_l(t-1)^{-1}\Big)^{-1}. \tag{14}$$
   Note that we have introduced forgetting factor α, where $0 < \alpha < 1$, in order to endow the algorithm with adaptability.

In the experiments described later, the initial mean $\boldsymbol{\mu}_l(0)$ and covariance matrix $\Phi_l(0)$ are set at a zero vector and an identity matrix, respectively, for all l. As an alternative, these parameters may be determined by exploiting prior knowledge about room acoustics as in [6].

2.3. Weighted RLS algorithm

The above algorithm involves the computation of a K_l-dimensional inverse matrix for each l value every time a new observation data set X_t is given. In order to reduce the computational cost, we here propose a new adaptive dereverberation algorithm. This algorithm is based on the RLS algorithm [7, Chapter 9] with time-frequency dependent weights, which is called weighted RLS algorithm.

Now, let us assume that matrix A is defined as

$$A = B^{-1} + CD^{-1}C^H, \tag{15}$$

where B, C, and D are arbitrary matrices. According to Woodbury's identity [7], A^{-1} may be expressed as

$$A^{-1} = B - BC(D + C^H BC)^{-1} C^H B. \tag{16}$$

Setting B, C, and D as

$$B = \frac{\Phi_l(t-1)}{\alpha},\ \ C = \boldsymbol{x}_{t-1,l},\ \ D = {}_s\lambda_{t,l} \tag{17}$$

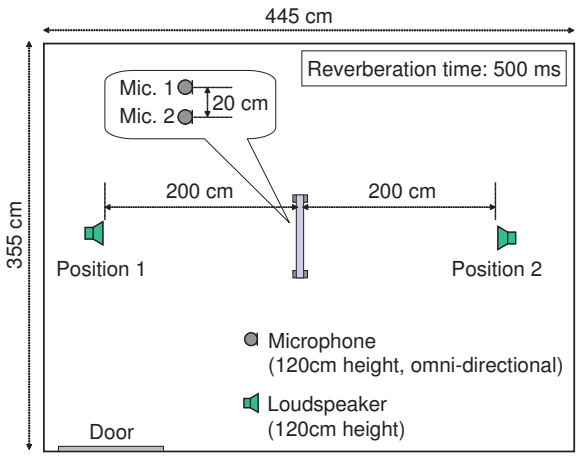

Fig. 1. Room layout.

enables us to rewrite (14) as

$$\Phi_l(t) = \frac{1}{\alpha}\Big\{\Phi_l(t-1) - \Phi_l(t-1) \times \frac{\boldsymbol{x}_{t-1,l}\boldsymbol{x}_{t-1,l}^H}{\alpha\, {}_s\lambda_{t,l} + \boldsymbol{x}_{t-1,l}^H \Phi_l(t-1)\boldsymbol{x}_{t-1,l}} \Phi_l(t-1)\Big\}. \quad (18)$$

Since the denominator on the second line is a scalar, we can avoid the inverse matrix computation.

By reconstructing the above Bayesian adaptive dereverberation algorithm based on (18) along the same lines as [7, Chapter 9], we finally reach the following algorithm. We may see that the time-frequency dependent weights are determined based on clean speech power spectrum $\{ {}_s\lambda_{t,l}\}_{0\le l\le L-1}$.

Weighted RLS based adaptive dereverberation algorithm

The procedures 1 to 2-(b) are common to the above Bayesian adaptive dereverberation algorithm.

2-(c) Update the mean $\boldsymbol{\mu}_l(t-1)$ and covariance matrix $\Phi_l(t-1)$ of the RRC posterior distribution and gain vector $\boldsymbol{k}_l(t)$ for each $l \in \{0, \cdots, L-1\}$ according to the following update rules:

$$\boldsymbol{k}_l(t) = \frac{\Phi_l(t-1)\boldsymbol{x}_{t-1,l}}{\alpha\, {}_s\lambda_{t,l} + \boldsymbol{x}_{t-1}^H \Phi_l(t-1)\boldsymbol{x}_{t-1,l}} \quad (19)$$

$$\boldsymbol{\mu}_l(t) = \boldsymbol{\mu}_l(t-1) + \boldsymbol{k}_l(t)\hat{s}_{t,l}^* \quad (20)$$

$$\Phi_l(t) = \frac{\Phi_l(t-1) - \boldsymbol{k}_l(t)\boldsymbol{x}_{t-1,l}^H \Phi_l(t-1)}{\alpha}. \quad (21)$$

2.4. Experiment on adaptive speech dereverberation

We conducted an experiment to examine the convergence performance of our proposed adaptive dereverberation algorithm. We took Japanece utterances spoken by 20 speakers (10 male and 10 female) from the ASJ-JNAS database. The acoustic signals of the individual utterances were truncated up to 15 seconds. Each signal was convolved with two-channel RIRs measured in the room depicted in Fig. 1, where we used the RIRs for position 1. For each frequency-band index l, the regression order K_l was set so as to cover the room's reverberation time. The forgetting factor was set at $\alpha = 0.99$. The dereverberation results were evaluated in terms of the mel-frequency cepstral coefficient (MFCC) distances between the clean speech signals and test signals, where the test signals were either reverberant or dereverberated signals.

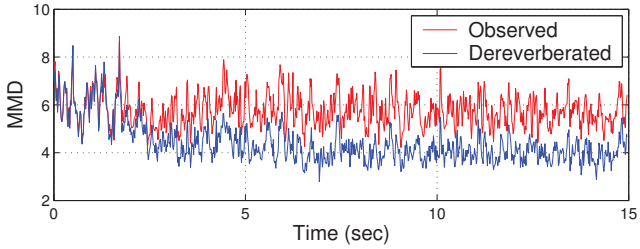

Fig. 2. Mean MFCC distance (MMD).

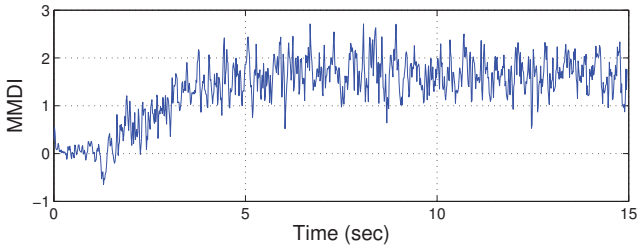

Fig. 3. Mean MFCC distance improvement (MMDI).

Fig. 2 shows the ensemble average of the MFCC distances for all the speakers. The corresponding improvement gained by the dereverberation is plotted in Fig. 3. It can be seen that the proposed method began to improve the mean MFCC distance in 2 seconds, and attained convergence in 5 seconds. On the basis of this result, we may conclude that the proposed algorithm converges relatively fast and greatly reduces the mean MFCC distance in a steady state.

3. SPEAKER-POSITION CHANGE DETECTION

Although the adaptive dereverberation method proposed in Section 2 may be able to track slow changes in RIRs, it cannot keep up with the sudden RIR changes typically caused by speaker movement and change. Fig. 4 exemplifies this problem. Fig. 4 is an example of the MFCC distance improvement gained with the above method for 35-second observed signals, which include a speaker change at 15 seconds. The first 20-second and remaining 30-second parts respectively consist of male speech uttered at position 1 in Fig. 1 and female speech uttered at position 2. 10 seconds was needed to return to the steady state. This long delay in tracking is attributed to the fact that the method estimates RRCs based on both current and previous data. Therefore, in order to catch up with sudden RIR changes more quickly, it is essential to detect changes in speaker position and and reinitialize the RRC estimates.

3.1. Proposed speaker-position change detection method

Let us investigate what happens to signals dereverberated with the method described above when a speaker position changes during the observation. Fig. 5 shows the dereverberated speech waveform for the same observation data as those used in Fig. 4. We can see that

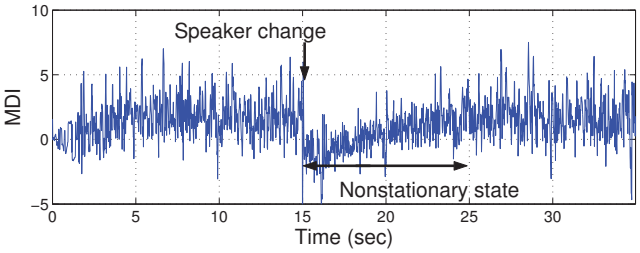

Fig. 4. MFCC distance improvement (MDI) without speaker-position change detection for observation data including speaker change.

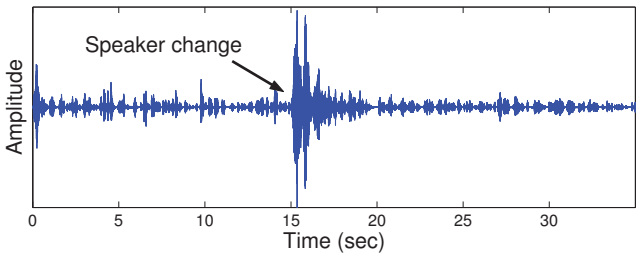

Fig. 5. Dereverberated speech waveform corresponding to the MDI in Fig. 4.

the power of the dereverberated signal was amplified immediately after the speaker changed. This fact may be explained based on the likelihood ratio test, however we omit the details owing to the space limitation. On this basis, we propose the following speaker-position change detection method.

We define the smoothed short-time powers of the observed and dereverberated speech signals respectively as

$$P_x(t) = \beta \sum_{l=0}^{L-1} |x_{t,l}^{(1)}|^2 + (1-\beta) P_x(t-1) \tag{22}$$

$$P_s(t) = \beta \sum_{l=0}^{L-1} |\hat{s}_{t,l}|^2 + (1-\beta) P_s(t-1), \tag{23}$$

where β is a smoothing constant. The proposed method determines that the speaker position has changed at frame t if $P_x(t)/P_s(t) < \delta$, where δ is a threshold.

3.2. Experiment on adaptive speech dereverberation and speaker-position change detection

We conducted an experiment to assess the overall performance of the proposed method. The same clean speech signals were used as in the experiment described in Section 2.4. The proposed method was tested 20 times. In each test, one of the 15-second male (female) speech signals was convolved with the RIRs for position 1 in Fig. 1, while one of the 20-second female (male) speech signals was convolved with those for position 2. These two reverberant signals were then concatenated to synthesize a single 35-second reverberant signal. The system parameters were set at $\beta = 0.99$ and $\delta = 0.2$.

In all of the tests, the proposed method successfully detected the change of speaker. Fig. 6 shows the ensemble average of the MFCC distance improvements. It is clear that the proposed method attained

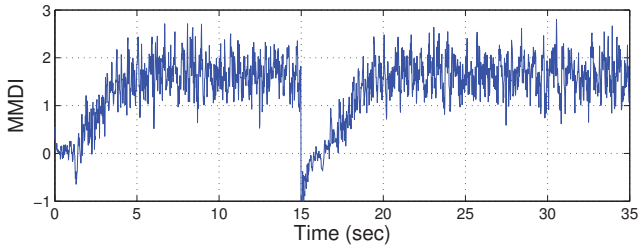

Fig. 6. Mean MFCC distance improvement (MMDI) for observation data including speaker change obtained by the proposed method.

convergence again in 5 seconds after the change of speaker. The drop in the mean MFCC distance improvement immediately after the speaker change would be mitigated by using prior knowledges about room acoustics [6].

4. CONCLUSION

This paper described a method for adaptive speech dereverberation and for detecting changes in speaker position. The proposed method estimates RRCs with the weighted RLS algorithm. Futhermore, if speaker position changes, the proposed method detects these changes and reinitializes the RRC estimates. The effectiveness of the proposed method was confirmed experimentally. Future work will include a comprehensive evaluation of the proposed method. The integration of this approach with adaptive noise suppression and signal separation methods is also a subject for future study.

5. REFERENCES

[1] B. Yegnanarayana and P. S. Murthy, "Enhancement of reverberant speech using LP residual signal," *IEEE Trans. Speech, Audio Process.*, vol. 8, no. 3, pp. 267–281, 2000.

[2] A. Abramson, E. A. P. Habets, S. Gannot, and I. Cohen, "Dual-microphone speech dereverberation using GARCH modeling," in *Proc. Int'l Conf. Acoust., Speech, Signal Process.*, 2008, pp. 4565–4568.

[3] T. Yoshioka, T. Nakatani, and M. Miyoshi, "An integrated method for blind separation and dereverberation of convolutive audio mixtures," in *Proc. Eur. Signal Process. Conf.*, 2008, CD-ROM Proceedings.

[4] T. Nakatani, T. Yoshioka, K. Kinoshita, M. Miyoshi, and B.-H. Juang, "Blind speech dereverberation with multi-channel linear prediction based on short time Fourier transform representation," in *Int'l Conf. Acoust. Speech, Signal Process.*, 2008, pp. 85–88.

[5] B. W. Gillespie, H. S. Malvar, and D. A. F. Florêncio, "Speech dereverberation via maximum-kurtosis subband adaptive filtering," in *Proc. Int'l Conf. Acoust., Speech, Signal Process.*, 2001, vol. VI, pp. 3701–3704.

[6] T. Nakatani, T. Yoshioka, K. Kinoshita, M. Miyoshi, and B.-H. Juang, "Incremental estimation of reverberation with uncertainty using prior knowledge of room acoustics for speech dereverberation," in *Int'l Worksh. Acoust. Echo, Noise Contr.*, 2008, CD-ROM Proceedings.

[7] S. Haykin, *Adaptive Filter Theory*, Prentice Hall, fourth edition, 2001.