Acoustic Echo Control with Frequency-Domain Stage-Wise Regression

Kaiyu Jiang, Chao Wu, Yanmeng Guo, Qiang Fu, and Yonghong Yan

***Abstract*—This letter introduces frequency domain stage-wise regression to acoustic echo control. By approximating the echo path as concatenate short segments in frequency domain, simple regression in the frequency domain can be carried out to estimate the echo contributed by consecutive far-end signal blocks stage by stage. A non-stationarity controlled smoothing factor is proposed alongside the regression procedure to mitigate the increasing variance of estimation when no significant echo but only near-end background noise is present. Experiments are carried out to demonstrate the superiority of the proposed approach, especially in unstable environment.**

***Index Terms*—Acoustic echo cancellation, acoustic echo suppression, stage-wise regression.**

I. Introduction

ACOUSTIC echo often arises in full-duplex speech communication and speech recognition based interactive systems due to the acoustic coupling between the loudspeaker and the microphone. It can deteriorate the quality and reliability of the systems substantially. Modeling the acoustic coupling by a linear system, acoustic echo canceller (AEC) [1] is usually exploited to eliminate the echo from the microphone signal.

AEC attempts to identify the linear system either in time [1] or frequency domain [2]. However, the quality of the parameter estimation is usually limited by several facts. First, the closed form least squares solution of this estimation needs an inversion of the far-end signal covariance matrix, which is computationally expensive and numerically unstable in case of ill-conditioned covariance matrix. Stochastic gradient decent method is frequently applied avoiding this matrix inversion. But it endures slow convergence rate when the eigenvalue spread of the far-end signal covariance matrix is large, which is common for speech signal. Other methods compromising between this two kinds of method exist such as affine projection [3], but it remains hard to achieve low computational complexity, robustness against near-end disturbances and fast convergence with colored far-end signal simultaneously. Second, in practice, the echo path is usually longer than several tens of milliseconds on which most audio signal can be assumed nearly stationary, and is constantly changing due to body movements, temperature fluctuations [4] or even change abruptly due to unpredictable disturbances [5]. Third, near-end speech can disturb the linear relationship between the far-end and microphone signals which is known as "double-talk". Besides, near-end background noise can enlarge the parameter estimation variance. As a consequence, a conventional AEC is often combined with a double-talk detector to prevent the AEC from divergence and an echo suppressor to reduce the residual echo or work independently. Obviously, a fast converging method insensitive to such changes and disturbances would be of considerable privileges in many practical applications.

Coherence estimation has been exploited in both the echo suppressor [6][7] and the double-talk detector modules [8]–[11] for the consideration of fast convergence and relative robustness to non-persistent disturbance like near-end speech. As this technique is based on the short-time spectral estimation, a relatively large modeling error is expected when the echo path is long as in many speech applications. Efforts have been made to extend it to longer system by multiple-frame coherence analysis to suppress the residual echo after an AEC [12]. However, as the residual error corresponding to each partition of the far-end signal is estimated independently, the overall residual echo tends to be overestimated. In this paper, we extend short-time spectral estimation to longer system by stage-wise regression [13] in the complex frequency domain. A simple regression is performed in each stage to estimate the echo contributed by a certain block of the far-end signal in a pre-specified order, thus short-time spectral estimation exploiting weighted least squares (WLS) estimate can be exploited and fast convergence with speech excitation can be achieved. Moreover, when enough independent samples are included for estimation, inherent robustness to non-persistent disturbance as near-end speech can be achieved in a similar way like coherence based double-talk detector [8]. In summary, a compact structure for acoustic echo control is obtained without assuming the stability of the echo path [14] or additional double-talk detector [6][7]. To further improve the robustness against near-end background noise when the echo is absent, an adaptive smoothing factor based on non-stationarity of the near-end signal is presented as well.

The rest of the paper is organized as follows. In Section II, frequency-domain stage-wise regression is introduced for acoustic echo control along with a time-frequency varying smoothing coefficient. In Section III, performance of the proposed method is evaluated under various practical conditions and compared with an implementation of partitioned block frequency domain

Manuscript received February 22, 2014; revised April 24, 2014; accepted June 10, 2014. Date of publication June 17, 2014; date of current version June 26, 2014. This work was supported by the NSFC under Grant 11161140319, the Strategic Priority Research Program of the CAS under Grants XDA06030100 and XDA06030500, the National 863 Program under Grant 2012AA012503, and by the CAS Priority Deployment Project under Grant KGZD-EW-103-2. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Muhammad Zubair Ikram.

The authors are with the Key Laboratory of Speech Acoustics and Content Understanding, Institute of Acoustics, Chinese Academy of Sciences, Beijing 100190, China (e-mail: jiangkaiyu@hccl.ioa.ac.cn; wuchao@hccl.ioa.ac.cn; guoyanmeng@hccl.ioa.ac.cn; qfu@hccl.ioa.ac.cn; yyan@hccl.ioa.ac.cn).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/LSP.2014.2331108

adaptive filter (PBFDAF) based AEC. Conclusions are given in Section IV.

II. Problem Formulation and Stage-wise Regression

A. Model Approximation

Let the microphone signal be represented by

$$y(t) = \mathbf{h}^T(t)\mathbf{s}(t) + n(t) \tag{1}$$

where $\mathbf{h}(t) = [h_1(t) \cdots h_M(t)]^T$ is the echo path, $\mathbf{s}(t)$ is the far-end signal vector, $n(t)$ is the near-end signal, $(\bullet)^T$ denotes transpose. Note that $n(t)$ is composed of near-end speech and background noise, thus may be approximated by an identical independent distributed(i.i.d.) random vector with heavier tailed distribution than Gaussian. Partition $\mathbf{h}(t)$ and $\mathbf{s}(t)$ into L blocks of length B, thus $M = LB$ and

$$\mathbf{h}_l(t) = [h_{(l-1)B+1}(t) \cdots h_{lB}(t)]^T \tag{2}$$

$$\mathbf{s}_l(t) = [s(t - lB + 1) \cdots s(t - lB + B)]^T \tag{3}$$

for $l = 1, 2, \cdots, L$. Let $\mathbf{F}$ and $\mathbf{F}^{-1}$ represents $2B \times 2B$ DFT and IDFT matrix. Define

$$\mathbf{H}_l(k) = \mathbf{F}[\mathbf{h}_l^T(kB)\ 0 \cdots 0]^T \tag{4}$$

$$\mathbf{S}_l(k) = \mathbf{F}[\mathbf{s}_l^T((k-1)B)\mathbf{s}_l^T(kB)]^T \tag{5}$$

where k is the block index. Substitute it into (1), we get [2]

$$\mathbf{y}(k) = [\mathbf{0}_B\ \mathbf{I}_B] \sum_{l=1}^{L} \mathbf{F}^{-1}\mathbf{H}_l(k) \odot \mathbf{S}_l(k) + \mathbf{n}(k) \tag{6}$$

where $\odot$ denotes Hadamard product, $\mathbf{0}_B$ and $\mathbf{I}_B$ denote zero and identity matrix of $B \times B$, and $\mathbf{y}(k) = [y(kB - B + 1) \cdots y(kB)]^T$, $\mathbf{n}(k) = [n(kB - B + 1) \cdots n(kB)]^T$. This formulation can lead to the popular structure of PBFDAF. If $\mathbf{s}(t)$ is periodic with period B, we have

$$\mathbf{Y}(k) = \sum_{l=1}^{L} \mathbf{H}_l(k) \odot \mathbf{S}_l(k) + \mathbf{N}(k) \tag{7}$$

where $\mathbf{Y}(k) = \mathbf{F}[\begin{matrix}\mathbf{y}(k-1)\\ \mathbf{y}(k)\end{matrix}]$, $\mathbf{N}(k) = \mathbf{F}[\begin{matrix}\mathbf{n}(k-1)\\ \mathbf{n}(k)\end{matrix}]$.

Although this assumption is not always valid, Avendano showed the approximation error is usually small [6]. No parameter reduction is achieved by this formulation, but overlap and add synthesis used in spectral modification framework can be applied instead of circular to linear convolution conversion. Viewing $\mathbf{S}_l(k)$ as regressors, regression techniques can be applied directly.

B. Stage-wise Regression

The stage-wise regression procedure is described as below. Let the ith element of $\mathbf{Y}(k)$, $\mathbf{S}_l(k)$, $\mathbf{N}(k)$, $i = 1 \cdots 2B$ be denoted as $Y(i,k)$, $S_l(i,k)$, $N(i,k)$ respectively. Denotes the residual after the mth stage of regression as $V_m(i,k), m = 0 \cdots L$, where $V_0(i,k) = Y(i,k)$. Assumption can be made that $S_l(i,k)$ is more correlated with $V_m(i,k)$ than $S_{l-1}(i,k)$ when $l > Q$, where Q corresponds to the known delay between the two channels. Here, for simplicity, we assume the far-end signal and microphone signal are well aligned within one block beforehand, then the order for the regressors to enter the stage-wise regression should be specified as $S_1(i,k) \cdots S_L(i,k)$.

To perform stage-wise regression, the output of the mth stage is used as the input of the $(m+1)$th stage. Let $\hat{\Phi}_{S_{m+1}S_{m+1}}(i,k)$ denotes the auto spectrum of $S_{m+1}(i,k)$ and $\hat{\Phi}_{V_mS_{m+1}}(i,k)$ denotes the cross correlation between $S_{m+1}(i,k)$ and $V_m(i,k)$ in the $(m+1)$th stage. The simple regression coefficient of the $(m+1)$th regression ($m = 0 \cdots L-1$) is given by the least squares solution:

$$\hat{H}_{m+1}(i,k) = \frac{\hat{\Phi}_{V_mS_{m+1}}(i,k)}{\hat{\Phi}_{S_{m+1}S_{m+1}}(i,k)} \tag{8}$$

Accordingly, the residual is

$$V_{m+1}(i,k) = V_m(i,k) - \hat{H}_{m+1}(i,k)S_{m+1}(i,k) \tag{9}$$

Eq. (8) indicates the inherent double-talk robustness of the proposed approach due to the single coefficient WLS estimation. In fact, when the interference is modeled by an uncorrelated sequence with constant variance, even not Gaussian, the Gauss-Markov theorem claims the least squares estimation to be the best linear unbiased estimator.

After regression of L stages, the final estimate of the echo signal can be obtained as

$$\hat{E}_L(i,k) = \sum_{m=0}^{L-1} \hat{H}_{m+1}(i,k)S_{m+1}(i,k) \tag{10}$$

and the final output $V_L(i,k)$ represents the near-end signal.

The auto spectrum $\hat{\Phi}_{S_{m+1}S_{m+1}}(i,k)$ and the cross correlation $\hat{\Phi}_{V_mS_{m+1}}(i,k)$ can be recursively estimated using exponential weighting,

$$\hat{\Phi}_{S_{m+1}S_{m+1}}(i,k) = \alpha\hat{\Phi}_{S_{m+1}S_{m+1}}(i,k-1) + (1-\alpha)S_{m+1}(i,k)S_{m+1}^*(i,k) \tag{11}$$

$$\hat{\Phi}_{V_mS_{m+1}}(i,k) = \alpha\hat{\Phi}_{V_mS_{m+1}}(i,k-1) + (1-\alpha)V_m(i,k)S_{m+1}^*(i,k),\ m = 0 \cdots L-1 \tag{12}$$

where $(\bullet)^*$ denotes complex conjugate, and α is a smoothing factor related to the number of independent frames used for averaging. For speech applications, in this paper, $\alpha = 0.98^{8000B/(128fs)}$, where fs is the sampling frequency. In addition, since $S_{m+1}(i,k) = S_m(i,k-1), m = 2 \cdots L$, we have

$$\hat{\Phi}_{S_{m+1}S_{m+1}}(i,k) = \hat{\Phi}_{S_mS_m}(i,k-1), m = 2 \cdots L \tag{13}$$

In this way, we extend short-time spectral estimation to a long system by stage-wise regression in the complex spectral domain. Although stage-wise regression generally gives a biased estimate from the estimation of the multiple linear regression [13], it achieves faster convergence and inherent robustness to double-talk. Actually, it may not always be possible to obtain an unbiased estimate of the echo path in unstable reverberated environment with satisfactory variance as discussed in Section I.

Besides, incorporating a prior on the rough envelope of the echo path blocks, the stage-wise order is pre-specified and the estimated regression coefficients can be further reasonably constrained. Subsequently, the enlarged mean square error due to the bias can be kept relatively low (see Appendix A).

C. Non-stationarity Controlled Smoothing

The effect introduced by non-persistent disturbance as near-end speech can be mitigated by a large enough α, however, persistent disturbance (near-end background noise) may lead to large variance of $\hat{H}_l(i,k)$ when no significant echo but only background noise presents for some time. As a result, when echo reappears, not enough samples are used to estimate $\hat{H}_l(i,k)$. This will lead to over-fitting of the estimation and distortion on the near-end speech. Consequently, it would further enhance the robustness of the procedure by taking a rough estimate of the echo to near-end ratio (ENR) into consideration. In this paper, a non-stationarity controlled smoothing factor $\alpha(i,k)$ is proposed and incorporated to preserve the estimated $\hat{H}_l(i,k)$ through periods when the microphone signal is stationary, assuming the far-end signal is non-stationary.

$\alpha(i,k)$ should be set to the normal value when $Y(i,k)$ is determined to be non-stationary, but close to 1 when $Y(i,k)$ is quasi-stationary. A simple method proposed in [15] to estimate the mean of the periodogram of the quasi-stationary background noise is exploited. And the variance of the periodogram of the quasi-stationary background noise is estimated in a similar way. Denotes the estimated mean and variance by $\hat{\mu}_{\tilde{N}}(i,k)$ and $\hat{\sigma}^2_{\tilde{N}}(i,k)$, $\alpha(i,k)$ is calculated by

$$\alpha(i,k) = \alpha^{min\left\{max\left\{\frac{Y(i,k)-\hat{\mu}_{\tilde{N}}(i,k)+b\hat{\sigma}_{\tilde{N}}(i,k)}{(a+b)\hat{\sigma}_{\tilde{N}}(i,k)},0\right\},1\right\}} \tag{14}$$

where a and b are parameters related to the upper and lower quantiles of the distribution of the periodogram of the quasi-stationary background noise. Larger a and b can better preserve the estimated $\hat{H}_l(i,k)$ when only noise is present, but may reduce the convergence rate if ENR is too low. Further parameter optimization is out of the scope of this letter. Typically, they are set as 6 and 3 respectively in this paper.

D. Signal Reconstruction

Reconstruction can be carried out using overlap and add on $V_L(i,k)$ as in ordinary spectral modification framework. As complex subtraction rather than magnitude subtraction is exploited in the stage-wise regression procedure, when the estimated $\hat{H}_l(i,k)$ is far from the true value, complex subtraction may lead to enlarged microphone signal. This problem can be circumvented by posing a magnitude constraint on $V_L(i,k)$.

$$\tilde{V}_L(i,k) = min\{V_L(i,k), cY(i,k)\} \tag{15}$$

where c is a constant around 1.

E. Computational Complexity

Assume a $2B$-point Fast Fourier Transform (FFT) needs $2B\log_2 B$ real multiplications, and a complex division requires 8 real multiplications plus 1 division. Then, to produce an output block of length B, PBFDAF adapted with stochastic gradient method needs $(4M+6B)\log_2 B + 8M - (4M+6B)\log_2 L$ real multiplications [2], and the proposed procedure with 50% frame-shift and fixed α needs $4B\log_2 B + 24M + 3B$ real multiplications plus M divisions which is comparable with PBFDAF for typical settings.

III. Experiments

As analyzed above, the proposed estimator is expected to achieve faster convergence rate and relatively better robustness against unstable echo path than traditional methods. Experiments are carried out to evaluate the proposed scheme considering the effect of under-modeling and near-end noise which is usually inevitable in practical scenarios, in comparison with an implementation of PBFDAF with adaptive controller named Speex [2], [16], [17].

The simulations are carried out with 16 kHz sampling frequency. The echo path used for simulation is measured in an ordinary conference room sized $5 \times 5 \times 3m^3$ and truncated to 2048 points. Far-end signal $s(t)$ for each test case is a 10 s segment drawn from TIMIT. A piece of 10 s Chinese speech signal is used as near-end speech for the convenience of informal listening evaluation. The near-end background noise is the output of AR(1) system $1/(1-0.9z^{-1})$ excited by a white Gaussian sequence of constant variance. For all the experiments, the block length of Speex and the B of the proposed procedure are set to 256. Hanning window with 50% frame-shift are used for the proposed procedure.

Echo return loss enhancement (ERLE) defined in short time as $ERLE(k) = 10\lg\frac{\mathbf{y}^T(k)\mathbf{y}(k)}{\hat{\mathbf{n}}^T(k)\hat{\mathbf{n}}(k)}$ is used to measure the echo reduction. Log spectral distance (LSD) defined in [18] and short-time objective intelligibility (STOI) [19] are used as objective indexes related to the near-end speech distortion.

A. Non-stationarity Controlled Smoothing

To evaluate the robustness against near-end background noise as discussed in Section II-C, the echo signal is present for 10 s and absent for 10 s then reappears for 10 s again and the near-end speech is only active for the last 10 s. Background noise is added to make ENR $=$ 10 dB. The filter length of Speex and M of the proposed procedure are set 2048 to fully model the echo path. The Euclidean norm of the estimated system $\|H\| = \sum_{i=1}^{B}\sum_{m=0}^{L-1}|\hat{H}_{m+1}(i,k)|^2$ and the true and estimated near-end waveform are depicted in Fig. 1. Compared to the fixed α, the estimation with the proposed adaptive $\alpha(i,k)$ generally shows lower variance and does not significantly distort the near-end signal when echo reappears at the beginning of 20 s, while the residual echo level is much the same, both lower than Speex. As the difference between the fixed and adaptive $\alpha(i,k)$ is insignificant when the echo keeps being active, the adaptive $\alpha(i,k)$ is used hereafter for the evaluation of the proposed stage-wise regression procedure.

B. Transient Evaluation in Single and Double-talk

ERLE improvement of the proposed method over Speex defined as subtraction of the two ERLE curves is evaluated with ENR=20 dB as depicted in Fig. 2 along with the waveforms. The filter length of Speex and M of the proposed procedure are set 1024 to simulate the under-modeled echo path. The first 10 s contains echo only, and the last contains double-talk. The echo path changes by reversing the sign at the end of 10 s. The overall signal to echo ratio (SER) is about 0 dB in the last 10 s.

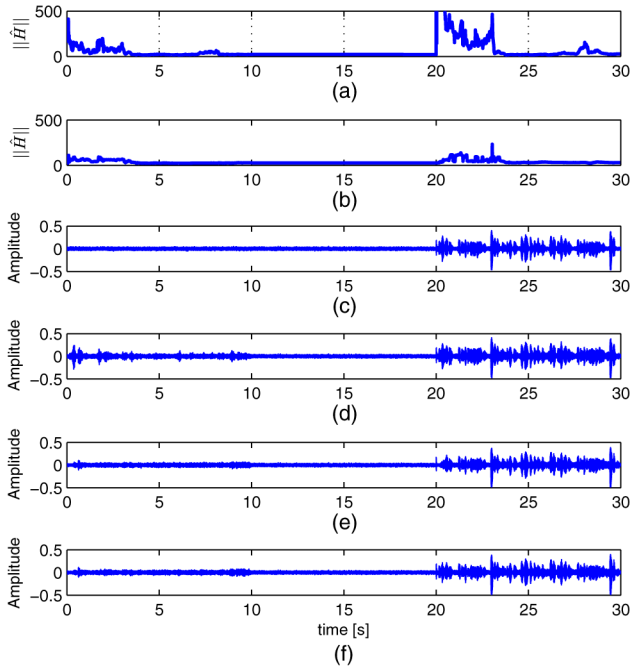

Fig. 1. Evaluation of the non-stationarity controlled smoothing with ENR = 10 dB. (a) $\|H\|$ of the proposed method with fixed α; (b) $\|H\|$ of the proposed method with non-stationarity controlled $\alpha(i,k)$; (c) pure near-end signal; (d) residual of Speex; (e) residual of fixed α; (f) residual of non-stationarity controlled $\alpha(i,k)$.

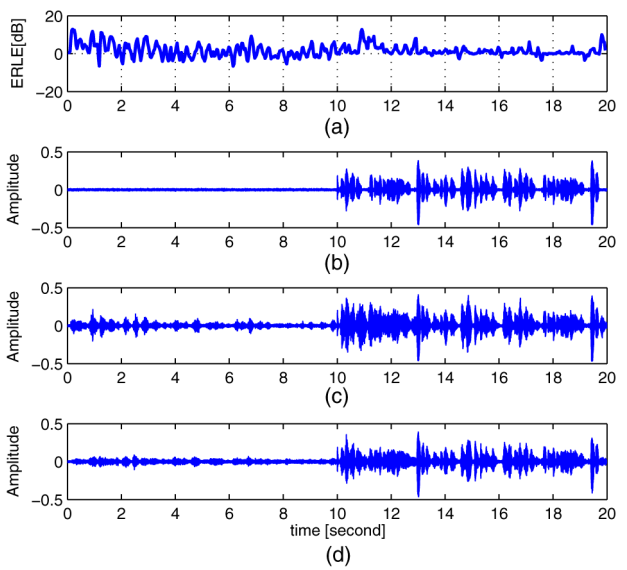

Fig. 2. Evaluation under single and double talk. (a) ERLE improvement of the proposed method over Speex; (b) the near-end signal; (c) the estimated near-end signal by Speex; (d) the estimated near-end signal by the proposed method.

The proposed method achieves significantly higher ERLE in the 0 s-3 s and 11 s-13 s and comparable with Speex in the rest part of the time. The proposed method introduces slightly higher distortion on the near-end speech, but insignificant as indicated by the waveforms in Fig. 2 and more detailed analysis presented in the next subsection with averaged evaluation.

C. Averaged Evaluation with Unstable Echo Path

To simulate unstable echo path, the echo path is changed by reversing the sign every 10 s. The filter length of Speex and M of the proposed procedure are set 2048 to simulate fully-modelled echo path. The ERLE is calculated on the whole time interval. Table I shows the results averaged on a total length of 100 s by randomly align the echo and near-end signal 10times. Although it is observed that the steady state ERLE of the proposed method can be lower when the ENR is higher than 20 dB due to the bias. It is evident that the proposed procedure achieves better overall echo reduction while the speech intelligibility is not affected as indicated by the LSD and STOI.

TABLE I
EVALUATION ON DIFFERENT ECHO-TO-NOISE-RATIO

Echo-to-Noise Ratio		-5dB	0dB	10dB	20dB	30dB
ERLE[dB] (double-talk)	Speex	0.22	0.67	1.31	1.42	1.43
	Proposed	0.98	1.80	2.90	3.15	3.20
LSD[dB] (double-talk)	Speex	26.94	22.70	15.40	10.74	9.39
	Proposed	26.70	22.41	14.97	10.43	9.15
STOI (double-talk)	Speex	0.54	0.64	0.82	0.88	0.89
	Proposed	0.53	0.64	0.83	0.89	0.90
ERLE[dB] (single-talk)	Speex	0.50	1.44	4.03	4.82	4.94
	Proposed	1.09	2.56	7.33	9.89	10.48

IV. CONCLUSION

A frequency-domain stage-wise regression procedure has been proposed for acoustic echo control. The WLS estimation in each stage ensures fast convergence and inherent robustness against near-end speech. The superiority compared to a PBFDAF implementation is demonstrated in practical scenarios considering the effect of under-modeling and near-end background noise. A non-stationarity controlled smoothing technique is proposed and showed to further improve the robustness of the proposed procedure against near-end background noise.

APPENDIX A
BIAS OF STAGE-WISE REGRESSION

For the ith row in (7), define

$$\begin{aligned}\vec{S}_l(i,k) &= [S_l(i,k-K)\cdots S_l(i,k)]^T, l = 1\cdots L\\ \vec{H}(i,k) &= [H_1(i,k)\cdots H_L(i,k)]^T,\\ \vec{Y}(i,k) &= [Y(i,k-K)\cdots Y(i,k)]^T\end{aligned} \tag{16}$$

where K is the number of frames used for estimation. Further define $\vec{\mathbf{S}}(i,k) = [\vec{S}_1(i,k)\cdots\vec{S}_L(i,k)]$. Omit the indexes for simplicity. The least squares estimate of $\vec{H}(i,k)$ is given by:

$$LS\{\vec{H}\} = (\vec{\mathbf{S}}^T\vec{\mathbf{S}})^{-1}\vec{\mathbf{S}}^T\vec{Y} \tag{17}$$

And the stage-wise estimation of $H_1(i,k)$ is given by

$$\begin{aligned}SW\{H_1\} &= (S_1^TS_1)^{-1}S_1^T(S_1H_1 + \vec{\mathbf{S}}_{\{2,L\}}\vec{H}_{\{2,L\}})\\ &= H_1 + (S_1^TS_1)^{-1}S_1^T\vec{\mathbf{S}}_{\{2,L\}}\vec{H}_{\{2,L\}}\end{aligned} \tag{18}$$

where $(\bullet)_{\{2,L\}}$ denotes the vector formed by elements from the 2nd to the Lth place of $(\bullet)$. The bias is given by the 2nd term, which indicates that larger S_1 and less correlation between S_1 and $\vec{\mathbf{S}}_{\{2,L\}}$ will result in lower bias.

REFERENCES

[1] M. Sondhi, "An adaptive echo canceller," *Bell Syst. Techn. J.*, vol. 46, no. 3, pp. 497–511, 1967.

[2] J.-S. Soo and K. K. Pang, "Multidelay block frequency domain adaptive filter," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 38, no. 2, pp. 373–376, 1990.

[3] K. Ozeki, "An adaptive filtering algorithm using an orthogonal projection to an affine subspace and its properties," *IEICE Trans.*, vol. 67, no. 5, pp. 126–132, 1984.

[4] G. W. Elko, E. Diethorn, and T. Gaensler, "Room impulse response variation due to temperature fluctuations and its impact on acoustic echo cancellation," Tech. Rep., 2002 [Online]. Available: http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.14.7040

[5] C. Breining, P. Dreiscitel, E. Hansler, A. Mader, B. Nitsch, H. Puder, T. Schertler, G. Schmidt, and J. Tilp, "Acoustic echo control. an application of very-high-order adaptive filters," *IEEE, Signal Process. Mag.*, vol. 16, no. 4, pp. 42–69, 1999.

[6] C. Avendano, "Acoustic echo suppression in the stft domain," in *2001 IEEE Workshop Applications of Signal Processing to Audio and Acoustics, IEEE*, 2001, pp. 175–178.

[7] C. Faller and J. Chen, "Suppressing acoustic echo in a spectral envelope space," *IEEE Trans. Speech Audio Process.*, vol. 13, no. 5, pp. 1048–1062, 2005.

[8] T. Gansler, M. Hansson, C.-J. Ivarsson, and G. Salomonsson, "A double-talk detector based on coherence," *IEEE Trans. Commun.*, vol. 44, no. 11, pp. 1421–1427, 1996.

[9] J. Benesty, D. R. Morgan, and J. H. Cho, "A new class of doubletalk detectors based on cross-correlation," *IEEE Trans. Speech Audio Process.*, vol. 8, no. 2, pp. 168–172, 2000.

[10] Y. Zhou and X. Li, "A variable step-size for frequency-domain acoustic echo cancellation," in *2007 IEEE Workshop on. Applications of Signal Processing to Audio and Acoustics, IEEE*, 2007, pp. 303–306.

[11] I. J. Tashev, "Coherence based double talk detector with soft decision," in *2012 IEEE Int. Conf. Acoustics, Speech and Signal Processing (ICASSP), IEEE*, 2012, pp. 165–168.

[12] G. Enzner, R. Martin, P. Vary, G. Enzner, R. Martin, and P. Vary, "Partitioned residual echo power estimation for frequency-domain acoustic echo cancellation and postfiltering," *Eur. Trans. Telecommun.*, vol. 13, no. 2, pp. 103–114, 2002.

[13] *Applied Regression Analysis*, ser. Wiley Series in Probability and Mathematical Statistics. New York, NY, USA: Wiley, 1981.

[14] G. Enzner and P. Vary, "Frequency-domain adaptive kalman filter for acoustic echo control in hands-free telephones," *Signal Process.*, vol. 86, no. 6, pp. 1140–1156, 2006.

[15] G. Doblinger, "Computationally efficient speech enhancement by spectral minima tracking in subbands," *Power*, vol. 1, p. 2, 1995.

[16] Speex-aec-matlab [Online]. Available: https://github.com/wavesaudio/Speex-AEC-matlab

[17] J.-M. Valin, "On adjusting the learning rate in frequency domain echo cancellation with double-talk," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 15, no. 3, pp. 1030–1034, 2007.

[18] I. Cohen, "Analysis of two-channel generalized sidelobe canceller (gsc) with post-filtering," *IEEE Trans. Speech Audio Process.*, vol. 11, no. 6, pp. 684–699, 2003.

[19] C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen, "An algorithm for intelligibility prediction of time–frequency weighted noisy speech," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 19, no. 7, pp. 2125–2136, 2011.