A RESIDUAL ECHO SUPPRESSION TECHNIQUE FOR SYSTEMS WITH NONLINEAR ACOUSTIC ECHO PATHS

Kun Shi, Xiaoli Ma, and G. Tong Zhou

School of Electrical and Computer Engineering, Georgia Tech, Atlanta, GA 30332-0250, USA

ABSTRACT

Linear acoustic echo cancellers are widely employed to enhance the quality of telecommunication systems. However, low-cost audio components may generate significant nonlinear distortions in the acoustic echo path, which degrades the performance of adaptive linear acoustic echo cancellers (AECs), thus motivating the nonlinear AEC research. In this paper, we propose to use a postfilter to further attenuate the nonlinear residual echo following a linear AEC. The postfilter design is based on the power spectral density (PSD) of the nonlinear residual echo. By modeling the nonlinearity in the system using a basis expansion form, the PSD estimate becomes independent of the length of the room impulse response. The performance of the proposed algorithm is illustrated by computer simulations.

Index Terms— Acoustic echo cancellation, echo suppression, nonlinear residual echo, postfilter, power spectral density

1. INTRODUCTION

Acoustic echo is an annoying disturbance with its origin in the sound propagation between the loudspeaker and the microphone. Acoustic echo canceller (AEC) is widely used to reduce the echo signal by first identifying the echo path and then subtracting the estimated echo signal from the received signal. The performance of existing AEC approaches strongly relies on the assumption of a linear echo path. However, today's competitive audio consumer market may favor sacrificing linear performance for lower cost of the analog components. The assumption of linearity may not hold anymore, due to the nonlinear distortions introduced by the loudspeakers and/or their amplifiers [1]. In [2], it has been shown that the performance of a linear AEC is limited by nonlinear distortions in the echo path.

One approach to removing nonlinear echo in the loudspeaker-enclosure-microphone system (LEMS) is to use a nonlinear adaptive filter instead of a linear one. In the literature, most of the nonlinear approaches have resorted to Volterra filters, but an adaptive Volterra filter demands high computational complexity [3]. Adaptive realization of cascade structures has been proposed in [1, 4], but it may be difficult to converge to the optimal solution or to guarantee stable adaptation behaviors due to the nonquadratic surface of the objective function.

One way to overcome these drawbacks of the nonlinear AEC is to apply a residual echo suppressor to reduce the nonlinear residual echo that remains after the linear AEC. The structure of this post-processing scheme is shown in Fig. 1; it was first proposed in the context of linear echoes for combined acoustic echo control and noise reduction [5, 6]. However, these methods are all based on the assumption of a linear echo path and, thus, are not applicable when nonlinear distortions are present. The nonlinear residual echo suppressor (NRES) approach proposed in [7] requires a frequency-domain model of the nonlinear residual echo that must be determined in advance. Since this model depends on the hardware components actually included in the echo path, it must be acquired for each hardware set-up separately. Similar to the linear case in [6], the NRES in [8] includes an additional adaptive filter referred to as the residual echo filter to estimate the nonlinear residual echo. In contrast to the existing methods, the NRES proposed in this paper uses basis expansion as a very general model for the nonlinear acoustic echo path. Moreover, the estimation of the nonlinear residual echo power spectral density (PSD) bypasses the estimation of additional filter coefficients. We will show that our proposed method improves the convergence rate and is robust to the length of the acoustic echo path.

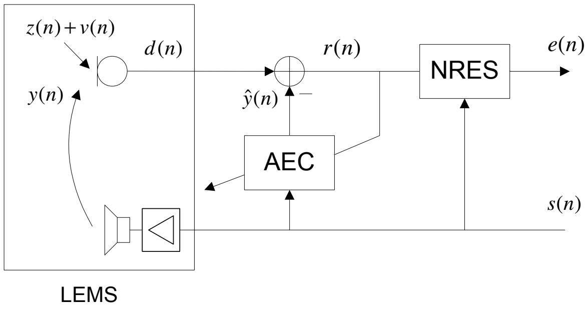

Fig. 1. Nonlinear acoustic echo cancellation with NRES.

2. MODELS OF NONLINEAR ECHO PATH AND RESIDUAL ECHO

The schematic for nonlinear acoustic echo cancellation using NRES is shown in Fig. 1. $s(n)$ denotes the far-end signal and $d(n)$ the microphone received signal, consisting of the near-end speech $z(n)$, the background noise $v(n)$, and the acoustic echo $y(n)$. The adaptive AEC tries to identify the LEMS and produce an estimate of the echo signal denoted by $\hat{y}(n)$. The estimated echo is then subtracted from the microphone signal to produce the residual signal $r(n)$

$$\begin{aligned} r(n) &= d(n) - \hat{y}(n) \\ &= z(n) + v(n) + y(n) - \hat{y}(n). \end{aligned} \tag{1}$$

We define the nonlinear residual echo $p(n)$ as the difference between the true echo signal $y(n)$ and its estimate $\hat{y}(n)$

$$p(n) = y(n) - \hat{y}(n). \tag{2}$$

Suppose that the LEMS consists of a (memoryless) nonlinear amplifier and/or loudspeaker followed by a linear subsystem (the

This work was supported in part by the Texas Instruments DSP Leadership University Program.

1-4244-1484-9/08/$25.00 ©2008 IEEE

room impulse response). We model the nonlinearity $f(\cdot)$ in the amplifier/loudspeaker block as a linear combination of basis functions $b_k(\cdot)$ with corresponding coefficients α_k

$$f(s;\boldsymbol{\alpha}) = \sum_{k=1}^{K} \alpha_k b_k(s), \tag{3}$$

where $b_k(s) = s^k$. If the room impulse response is modeled by a finite impulse response (FIR) filter $h(n)$ with length L_h, the nonlinear acoustic echo can be expressed as

$$y(n) = \sum_{l=0}^{L_h-1} h(l) \sum_{k=1}^{K} \alpha_k b_k(s(n-l)). \tag{4}$$

Assume that the AEC uses an FIR filter $\hat{h}(n)$ (an estimate of $h(n)$) with length L'_h, and then the estimated echo is obtained as

$$\hat{y}(n) = \sum_{l=0}^{L'_h-1} \hat{h}(l) s(n-l). \tag{5}$$

Usually we choose $L'_h < L_h$, since we can decrease the computational complexity by sacrificing some echo cancellation performance when the reverberation time of the LEMS is too long. Combining (4), (5), and (2), we obtain

$$\begin{aligned} p(n) &= \sum_{l=0}^{L_h-1} h(l) \sum_{k=1}^{K} \alpha_k b_k(s(n-l)) - \sum_{l=0}^{L'_h-1} \hat{h}(l) s(n-l) \\ &= \sum_{k=1}^{K} g_k(n) * x_k(n), \end{aligned} \tag{6}$$

where

$$g_k(n) = \begin{cases} \alpha_k h(n) - \hat{h}(n), & k = 1 \\ \alpha_k h(n), & k = 2, ..., K \end{cases} \tag{7}$$

and $x_k(n) = b_k(s(n))$. Therefore, the nonlinear residual echo $p(n)$ can be treated as the output signal of a multiple-input single-output (MISO) system with the input signals being $x_k(n)$. Due to the presence of nonlinearity, there is much energy in the residual echo. Thus, we employ a postfilter to further suppress the echo.

3. NONLINEAR RESIDUAL ECHO SUPPRESSION

Similar to the postfilter commonly used in the linear echo case, the NRES is a frequency-dependent, real-valued gain filter $C(f)$, realized by frequency domain processing on a frame-by-frame basis [5]. Accordingly, for each frame, the NRES output signal $e(n)$ and the residual signal $r(n)$ are related in the frequency domain as

$$E^{(m)}(f) = C^{(m)}(f) R^{(m)}(f), \tag{8}$$

where m is the frame index; $E^{(m)}(f)$ and $R^{(m)}(f)$ are the discrete Fourier transform (DFT) of the m-th frame of $e(n)$ and $r(n)$, respectively, at discrete frequency bins f. The resulting $E^{(m)}(f)$ is transformed back into the time domain by inverse DFT, and the output signal $e(n)$ is then synthesized with the overlap-and-add method. One way to design the gain function $C(f)$ will be described next. For notational simplicity, we will omit the frame index m when feasible from this point on.

The optimal gain $C(f)$ can be derived by minimizing the contribution of the nonlinear residual echo $R(f)$ to the output signal $E(f)$ in the mean square error (MSE) sense. Based on the results obtained from [5, 8], the optimal $C(f)$ is

$$C(f) = \frac{S_r(f) - S_p(f)}{S_r(f)}, \tag{9}$$

where $S_r(f)$ and $S_p(f)$ denote the PSD of $r(n)$ and $p(n)$, respectively. Here, we focus on the suppression of nonlinear residual echo without attenuating the background noise. If noise reduction is considered, the gain function in (9) can be rewritten as

$$C(f) = \frac{S_r(f) - S_p(f) - S_v(f)}{S_r(f)}, \tag{10}$$

where $S_v(f)$ is the PSD of the background noise $v(n)$. Since we assume $v(n)$ is white noise, (9) can be used in place of (10).

In (9), $S_r(f)$ can be estimated easily by recursively smoothing $\left|R^{(m)}(f)\right|^2$ as in

$$\hat{S}_r^{(m)}(f) = \lambda \hat{S}_r^{(m-1)}(f) + (1-\lambda) \left|R^{(m)}(f)\right|^2, \tag{11}$$

where $0 < \lambda < 1$ is the forgetting factor. However, this method can not be directly applied to determine $S_p(f)$ since $y(n)$ is unknown. Next, we propose a method for estimating $S_p(f)$ based on the nonlinear residual echo in (6). The Fourier transform of (6) yields

$$P(f) = \sum_{k=1}^{K} G_k(f) X_k(f), \tag{12}$$

where $G_k(f)$ and $X_k(f)$ are the Fourier transforms of $g_k(n)$ and $x_k(n)$, respectively. Define vectors

$$\boldsymbol{G}(f) = \left[G_1(f), G_2(f), ..., G_K(f)\right]^T, \tag{13}$$

$$\boldsymbol{X}(f) = \left[X_1(f), X_2(f), ..., X_K(f)\right]^T. \tag{14}$$

Using (12), the PSD of $p(n)$ can be expressed as

$$S_p(f) = E\left[|P(f)|^2\right] = \boldsymbol{G}^H(f) \mathbf{S}_{xx}(f) \boldsymbol{G}(f), \tag{15}$$

where $E[\cdot]$ denotes the statistical expectation; ^{H} is Hermitian transpose; and

$$\mathbf{S}_{xx}(f) = E\left[\boldsymbol{X}^*(f)\boldsymbol{X}^T(f)\right] = \begin{bmatrix} S_{11}(f) & \cdots & S_{1K}(f) \\ S_{21}(f) & \cdots & S_{2K}(f) \\ \vdots & & \vdots \\ S_{21}(f) & \cdots & S_{KK}(f) \end{bmatrix} \tag{16}$$

is the autocorrelation matrix of $\boldsymbol{X}(f)$, and its ij-th element is the cross spectral density between signals $x_i(n)$ and $x_j(n)$. Furthermore, the linear MMSE solution for $G_k(f)$ of (12) can be calculated as

$$\boldsymbol{s}_{xp}(f) = \mathbf{S}_{xx}\, \boldsymbol{G}(f), \tag{17}$$

where

$$\boldsymbol{s}_{xp}(f) = E\left[\boldsymbol{X}^*(f) P(f)\right] = \left[S_{1p}(f), ..., S_{Kp}(f)\right]^T \tag{18}$$

is the cross-correlation vector between $\boldsymbol{X}(f)$ and $P(f)$, and its i-th element is the cross spectral density between signals $x_i(n)$ and

$p(n)$. Combining (15) and (17), the PSD of the nonlinear residual echo can be obtained as

$$S_p(f) = \boldsymbol{s}_{xp}^H(f)\mathbf{S}_{xx}^{-1}(f)\boldsymbol{s}_{xp}(f). \tag{19}$$

Note that the nonlinear residual echo $p(n)$ is not accessible, since it is hidden in the microphone signal $r(n)$. Assume that the near-end speech, the background noise, and the far-end speech are mutually independent of each other and their mean has been removed. Thus $\boldsymbol{s}_{xp}(f) = \boldsymbol{s}_{xr}(f)$, and correspondingly (19) can be rewritten as

$$S_p(f) = \boldsymbol{s}_{xr}^H(f)\mathbf{S}_{xx}^{-1}(f)\boldsymbol{s}_{xr}(f). \tag{20}$$

Since the signals $x_i(n)$ and $r(n)$ are known, the recursive estimate of the i-th entry in $\boldsymbol{s}_{xr}^H(f)$ and the ij-th entry in $\mathbf{S}_{xx}$ can be given, respectively as

$$\left[\hat{\boldsymbol{s}}_{xr}^{(m)}(f)\right]_i = \lambda\left[\hat{\boldsymbol{s}}_{xr}^{(m-1)}(f)\right]_i + (1-\lambda)[X_i^{(m)}(f)]^* R(f), \tag{21}$$

$$\left[\hat{\mathbf{S}}_{xx}^{(m)}(f)\right]_{ij} = \lambda\left[\hat{\mathbf{S}}_{xx}^{(m-1)}(f)\right]_{ij} + (1-\lambda)[X_i^{(m)}(f)]^* X_j^{(m)}(f). \tag{22}$$

To avoid the high computational complexity associated with matrix inversion, $S_{xx}^{-1}(f)$ can be calculated recursively according to the Sherman-Morrison-Woodbury matrix inversion lemma [10]

$$(S_{xx}^{(m)}(f))^{-1} = \frac{1}{\lambda}(S_{xx}^{(m-1)}(f))^{-1} - \left(1 - \frac{1}{\lambda}\right) \cdot \frac{(S_{xx}^{(m-1)}(f))^{-1}(\boldsymbol{X}^{(m)}(f))^*(\boldsymbol{X}^{(m)}(f))^T(S_{xx}^{(m-1)}(f))^{-1}}{(\boldsymbol{X}^{(m)}(f))^T(S_{xx}^{(m-1)}(f))^{-1}(\boldsymbol{X}^{(m)}(f))^*}. \tag{23}$$

Therefore, the PSD estimate of the nonlinear residual echo $\hat{S}_p(f)$ can be obtained by substituting (21), (22), and (23) into (20). Correspondingly, the nonlinear gain $C(f)$ can be found using (9).
Remark: We recognize that (20) can be rewritten as

$$\begin{aligned} S_p(f) &= \frac{\boldsymbol{s}_{xr}^H(f)\mathbf{S}_{xx}^{-1}(f)\boldsymbol{s}_{xr}(f)}{S_r(f)} \cdot S_r(f) \\ &= \Gamma_{x_1 \ldots x_K, r}(f) S_r(f), \end{aligned} \tag{24}$$

where $\Gamma_{x_1 \ldots x_K, r}(f)$ is the so-called multiple coherence function [11]. It can be shown that $0 \le \Gamma_{x_1 \ldots x_K, r}(f) \le 1$, $\forall f$; and the multiple cohere function $\Gamma_{x_1 \ldots x_K, r}(f)$ indicates the fraction of the power in the signal $r(n)$ that is attributed to the linear combination of $x_1(n), ..., x_K(n)$. Therefore, (24) extracts the power of the signal which is related to $x_1(n), ..., x_K(n)$ from the signal $r(n)$. This is exactly the PSD of the nonlinear residual echo signal.

4. SIMULATION RESULTS

In this section, the performance of the proposed method is assessed via computer simulations. The nonlinearity of the power amplifier/loudspeaker concatenation is modeled by a $3rd$-order polynomial function

$$f(s) = -0.0325s^3 - 0.0003s^2 + 0.4824s. \tag{25}$$

The room impulse response was generated according to

$$h(n) = \begin{cases} \beta(n)e^{-\alpha n}, & 4 \le n \le L_h \\ 0, & \text{otherwise} \end{cases} \tag{26}$$

where $\beta(n)$ was i.i.d. standard Gaussian distributed; $L_h = 512$ and $\alpha = 0.004$. A white noise $v(n)$ was added and the resulting signal-to-noise ratio (SNR) was 30 dB. Here, the linear AEC is implemented using the frequency domain normalized least mean square (NLMS) algorithm [12], since the subsequent NRES is also performed in the frequency domain. As a performance metric to evaluate nonlinear acoustic echo cancellation, we use the echo return loss enhancement (ERLE) defined as

$$\text{ERLE (dB)} = 10\log_{10}\frac{E\left[d^2(n)\right]}{E\left[e^2(n)\right]}. \tag{27}$$

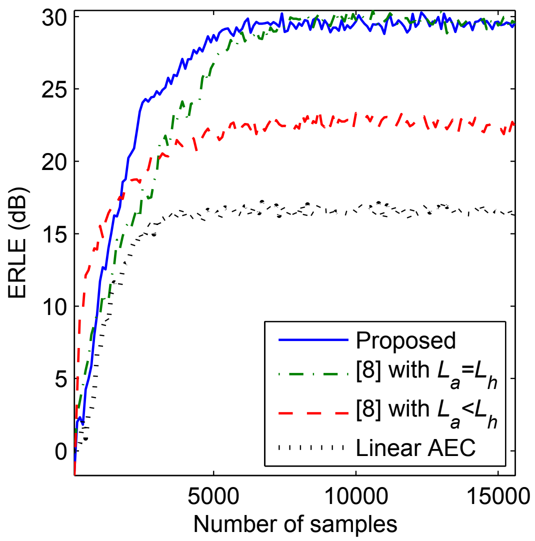

Fig. 2. Performance of linear AEC with and without NRES with white Gaussian noise as the input signal.

In the first experiment, we used white Gaussian noise for the far-end signal. For comparison purposes, we also implemented the method of [8]. The ERLEs obtained for different approaches are shown in Fig. 2. We can see that both nonlinear approaches remarkably improved the echo attenuation performance compared to the purely linear AEC. The proposed method outperforms the method of [8] in terms of convergence rate. This is because the proposed method estimates $S_p(f)$ directly, whereas the estimate of $S_p(f)$ in [8] depends on the convergence of another filter with length L_a. The major advantage of the proposed method is that it bypasses the estimation of the additional filter coefficients, and it requires no knowledge of the room impulse response length L_h. This can also be seen from Fig. 2, where the method in [8] uses $L_a = 350$ ($< L_h$) coefficients for the additional adaptive filter to estimate $S_p(f)$. It is seen that inadequate filter length gives rise to a large bias in the estimate of $S_p(f)$, and correspondingly the performance of the algorithm of [8] is degraded, whereas the proposed method will not be affected.

Next, we evaluated the performance of the proposed method using speech data as the input signal. In Fig. 3, we show the ERLEs obtained with proposed NRES and with a linear RES (LRES) in [5]. We notice that the nonlinear approach provides a consistent increase in echo attenuation throughout the data frame. The PSD estimates $\hat{S}_r(f)$ and $\hat{S}_p(f)$ are shown in Fig. 4. Given the PSD of the background noise, it can be seen that $\hat{S}_r$ and $\hat{S}_p(f) + \hat{S}_v(f)$ almost overlap at each frequency, which indicates the accuracy in the estimate $\hat{S}_p(f)$.

In the last experiment, we evaluated the performance of the proposed method in double-talk situations. The resulting echo signal $y(n)$ is shown in Fig. 5(a). The near-end speech $z(n)$ is depicted in Fig. 5(b). In Fig. 5(c) the NRES output signal $e(n)$ is shown. It can

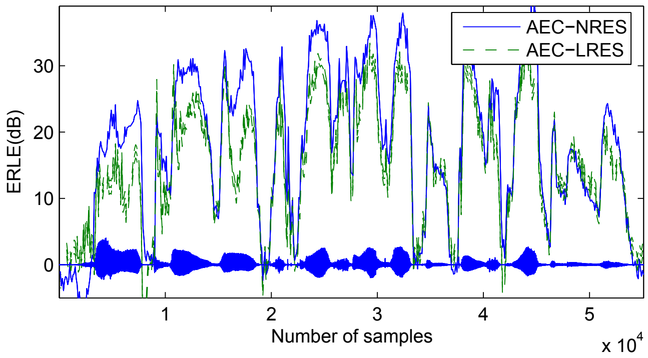

Fig. 3. Performance of the AEC with LRES and NRES using speech as the input signal.

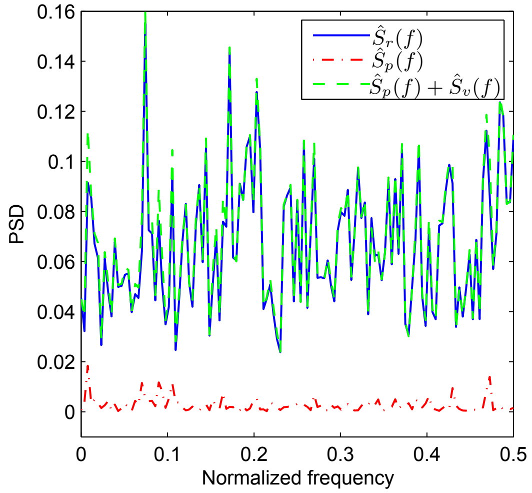

Fig. 4. Estimated PSD of different signals.

be seen the near-end speech is hardly distorted, while the echo signal has been sufficiently suppressed. Note we do not provide ERLE in this experiment, since ERLE can only be calculated during the far-end speech only period.

5. CONCLUSIONS

We investigated the suppression of nonlinear acoustic residual echoes that remain when a linear AEC is applied in the presence of nonlinear distortions. NRES is a spectral weighting approach to attenuate the nonlinear residual echo. The proposed technique relies on the basis expansion model of nonlinearities in the loudspeaker or amplifier. A special feature of the proposed method is the procedure for estimating the PSD of the nonlinear residual echo, which requires no knowledge of the room impulse response length. Simulation results have demonstrated the effectiveness of the proposed method.

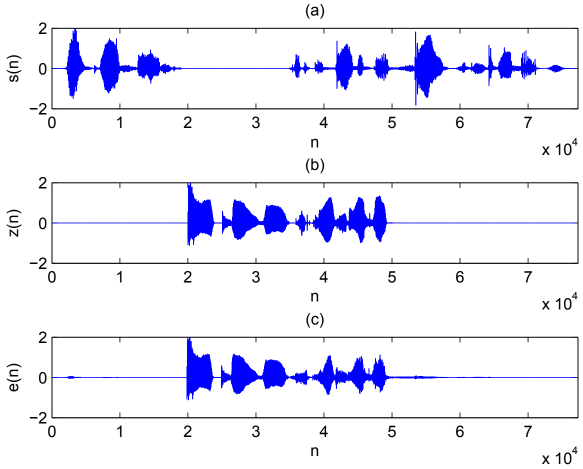

Fig. 5. Different signals for nonlinear acoustic echo cancellation: (1) far-end speech $s(n)$, (2) near-end speech $z(n)$, (3) NRES output $e(n)$.

6. REFERENCES

[1] B. S. Nollett and D. L. Jones, "Nonlinear echo cancellation for hands-free speakerphones," *in Proc. IEEE Workshop on Nonlinear Signal and Image Processing*, Mackinac Island, Michigan, Sept. 1997.

[2] A. N. Birkett and R. A. Goubran, "Limitations of handsfree acoustic echo cancellers due to nonlinear loudspeaker distortion and enclosure vibration effects," *in Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, pp. 103-106, New Paltz, New York, Oct. 1995.

[3] A. Guérin, G. Faucon, and R. Le Bouquin-Jeannes, "Nonlinear acoustic echo cancellation based on Volterra filters," *IEEE Trans. on Speech and Audio Processing*, vol. 11, pp. 672–683, Nov. 2003.

[4] J.-P. Costa, A. Lagrange, and A. Arliaud, "Acoustic echo cancellation using nonlinear cascade filters," *in Proc. IEEE ICASSP*, vol. 5, pp. 389–392, Hong Kong, China, Apr. 2003.

[5] S. Gustafsson, R. Martin, and P. Vary, "Combined acoustic echo control and noise reduction for hands-free telephony," *Signal Processing*, vol. 64, pp. 21-32, 1998.

[6] V. Myllylä, "Residual echo filter for enhanced acoustic echo control," *Signal Processing*, vol. 86, pp. 1193–1205, 2006.

[7] O. Hoshuyama and A. Sugiyama, "An acoustic echo suppressor based on a frequency-domain model of highly nonlinear residual echo," *in Proc. IEEE ICASSP*, pp. 269–272, Toulouse, France, May 2006.

[8] F. Kuech and W. Kellermann, "Nonlinear residual echo suppression using a power filter model of the acoustic echo path," *in Proc. IEEE ICASSP*, pp. 73-76, Honolulu, Hawaii, May 2007.

[9] G. -Y. Jiang and S. -F. Hsieh, "Nonlinear acoustic echo cancellation using orthogonal polynomial," *in Proc. IEEE ICASSP*, pp. 273-276, Toulouse, France, May 2006.

[10] R. A. Horn and C. R. Johnson, *Matrix Analysis*. Melbourne, Australia: Cambridge Univ. Press, 1993.

[11] R. K. Otnes and L. Enochson, *Applied Time Series Analysis*. John Wiley & Sons, 1978.

[12] E. Hänsler and G. Schmidt, *Acoustic Echo and Noise Control: A Practical Approach*, John Wiley & Sons, 2004.