SUBSPACE APPROACH FOR THE SUPPRESSION OF THE NONLINEAR ACOUSTIC ECHO INTRODUCED BY LOUDSPEAKERS

Xiaojian Lu and Benoît Champagne

Department of Electrical & Computer Engineering, McGill University
3480 University Street, Montreal, Quebec, H3A 2A7, Canada
{xlu,champagne}@tsp.ece.mcgill.ca

ABSTRACT

In this paper, a new nonlinear acoustic echo canceller (AEC) based on the subspace approach is proposed for the application of hands-free telephony when the loudspeaker nonlinearities cannot be neglected. The new AEC decomposes the microphone signal and the estimated echo produced by an FIR adaptive filter into orthogonal subspaces, where both the acoustic echo and the background noise are attenuated according to Wiener filter scheme. Furthermore, the reconstructed signal is synthesized by the components in the signal subspace, resulting in more suppression of the unwanted signals (i.e. noise and echo). Experiments show that, compared to a conventional AEC, the proposed AEC significantly suppresses the nonlinear echo signal caused by the loudspeaker and the background noise in the near-end.

1. INTRODUCTION

Acoustic echo canceller (AEC) is a very important device for hands-free telephones, and has been extensively studied in the past decades. Most AECs employ a linear adaptive filter to estimate the acoustic echo since the acoustic echo path is assumed to be modelled as a slowly time-varying linear system. The estimated echo is then subtracted from the microphone signal, resulting in an 'echo-cancelled' signal which is sent to the far-end. Indeed, this is a reasonable approximation when the loudspeaker has a good quality (e.g. negligible non-linear distortion) and plays at a moderate volume for the use of the hands-free terminals.

In practice, low-cost loudspeakers are preferably used in hands-free equipments to minimize the system cost. Sometimes the volume of the loudspeaker has to be high in a noisy environment, e.g. a hands-free telephone is used in a running vehicle. In these cases, the non-linearities of the loudspeakers cannot be neglected [1]. Consequently, a conventional AEC that mainly employs a purely linear adaptive filter cannot suppress the acoustic echo to a satisfactory level. Recently, some non-linear adaptive filtering algorithms such as polynomial adaptive filters, neural networks, and specific saturation models have been tried to compensate the non-linearities of loudspeakers in the application of acoustic echo cancellation [2, 3, 4]. However, our simulation shows that these algorithms suffer from slow convergence rate and instability when excited by coloured signal such as speech. Furthermore, the difficulty in accurately modelling the characteristics of loudspeakers which present complex non-linearities [5], limits the attenuation amount of the nonlinear echo in the practical application.

We introduce a new AEC based on the subspace approach in this paper. Our experiments show that the proposed AEC significantly attenuates the acoustic echo, as well as remarkably reduces the near-end background noise.

2. AEC BASED ON SUBSPACE APPROACH

The proposed AEC consists of an echo estimator and a subspace echo processor, as shown in Fig. 1. The estimated acoustic echo $\hat{y}(n)$ is obtained by the echo estimator, then subtracted from the microphone signal $d(n)$ with perceptual suppression gains in the signal subspace, resulting in an echo-suppressed and noise-reduced signal $e(n)$.

2.1. Echo estimator in the new AEC

The purpose of the echo estimator in AEC is to produce a replica of the acoustic echo signal by estimating the echo

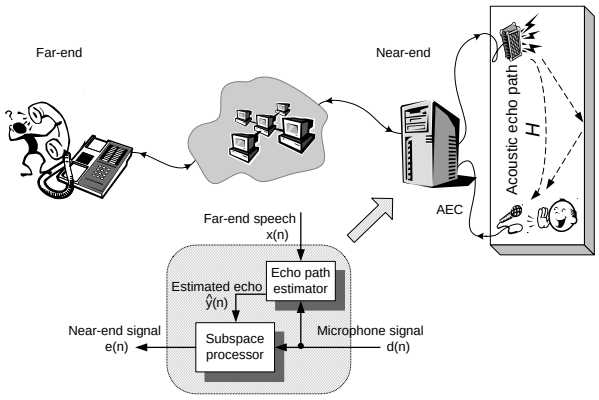

Fig. 1. The acoustic echo cancellation.

path between the far-end speech $x(n)$ and the microphone signal $d(n)$. From the viewpoint of system identification, an adaptive FIR filter can fulfill this task. Actually, this is the case with most conventional AECs.

Due to the nonlinearities of the loudspeaker, the acoustic echo path cannot be precisely estimated by a linear adaptive filter. However, from a practical viewpoint, we may regard the non-linear acoustic echo path as a fast-varying linear system, at least approximately. When a linear adaptive filter is used to identify the loudspeaker-enclosure-microphone system, it tries to track the change of the acoustic echo path by minimizing the mean-square error (MSE) between the acoustic echo and the estimated one. Accordingly, the performance of an adaptive filter in the presence of loudspeaker nonlinearities depends on its tracking behaviour: the better the tracking capability, the lower MSE it achieves.

In the excitation of the speech signal, our experiments suggest that the affine projection (AP) algorithm [6] achieves a lower MSE than some other conventional algorithms such as NLMS and RLS in the presence of loudspeaker nonlinearities. This is consistent with our research result that the former tracks the change of the echo path faster.

Based on this consideration, we choose AP as the estimator to roughly estimate the acoustic echo $\hat{y}(n)$ in the proposed AEC. When the acoustic echo path is linear, the tracking speed of AP is proportional to the projection order, i.e. higher projection order results in a better tracking property, although with higher computational complexity. However, when the non-linearities of the loudspeaker need to be considered, it can be verified that the lower bound of MSE is nearly reached even if the projection order is only set to 2 or 3, and no obvious improvement is observed with higher projection order. Consequently, a low-order AP is suitable to estimate the acoustic echo in terms of the low MSE it achieves and of an acceptable computational requirement.

2.2. Subspace echo suppression

The most important component of the proposed AEC is the subspace echo processor, illustrated in Fig. 2. The subspace approach has been widely used in the signal processing field, including remarkable applications in speech enhancement [7]. Here, we derive an algorithm for subspace echo suppression.

2.2.1. Karhunen-Loeve transform (KLT)

Referring to Fig. 1, the microphone signal is expressed as a K-dimensional vector, defined by

$$\mathbf{d}(n) = [d(n), d(n-1), \ldots, d(n-K+1)]^H, \quad (1)$$

and we have

$$\mathbf{d}(n) = \mathbf{y}(n) + \mathbf{s}(n) + \mathbf{w}(n), \quad (2)$$

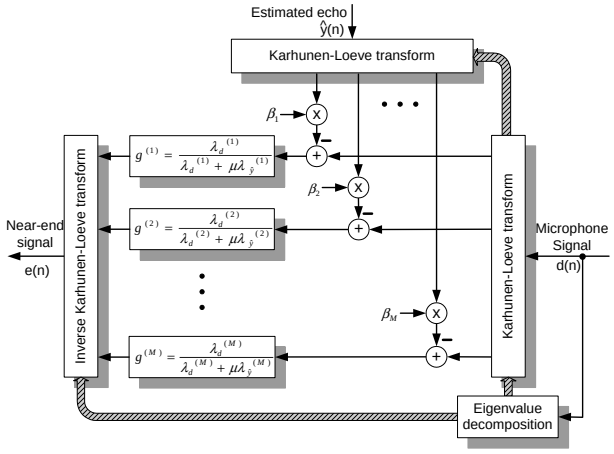

Fig. 2. The subspace processing for AEC.

where, $\mathbf{y}(n)$, $\mathbf{s}(n)$ and $\mathbf{w}(n)$ denote the echo signal, near-end speech and background noise vectors, respectively. Under the assumption that the signals $\mathbf{y}(n)$, $\mathbf{s}(n)$ and $\mathbf{w}(n)$ are mutually uncorrelated, the covariance matrix of $\mathbf{d}(n)$ is obtained as

$$\mathbf{R}_d(n) = \mathbf{R}_y(n) + \mathbf{R}_s(n) + \mathbf{R}_w(n). \quad (3)$$

The eigendecomposition of $\mathbf{R}_d(n)$ can be written as

$$\mathbf{R}_d(n) = \mathbf{Q}(n)\mathbf{\Lambda}_d(n)\mathbf{Q}(n)^H, \quad (4)$$

where

$$\mathbf{Q}(n) = [\mathbf{q}_1(n), \mathbf{q}_2(n), \ldots, \mathbf{q}_K(n)] \quad (5)$$

is an orthonormal matrix of eigenvectors of $\mathbf{R}_d(n)$, and

$$\mathbf{\Lambda}_d(n) = \text{diag}\{\lambda_d^{(1)}(n), \lambda_d^{(2)}(n), \ldots \lambda_d^{(K)}(n)\} \quad (6)$$

is a diagonal matrix of eigenvalues of $\mathbf{R}_d(n)$, where the diagonal elements are in descending order, i.e.

$$\lambda_d^{(1)}(n) > \lambda_d^{(2)}(n) > \ldots > \lambda_d^{(K)}(n) \quad (7)$$

Hence, $\mathbf{Q}(n)^H\mathbf{d}(n)$ is the KLT of $\mathbf{d}(n)$, which projects the microphone signal $\mathbf{d}(n)$ into the noisy signal subspace spanned by $\mathbf{q}_i(n),\ i = 1, 2, \ldots, K$. Similarly, the estimated echo signal $\hat{y}(n)$ can be decomposed into the same subspace by KLT, resulting in $\mathbf{Q}(n)^H\hat{\mathbf{y}}(n)$.

2.2.2. Echo subtraction in KLT domain

The estimated echo and the microphone signal are decomposed into the subspace by KLT, resulting in $\mathbf{Q}(n)^H\hat{\mathbf{y}}(n)$ and $\mathbf{Q}(n)^H\mathbf{d}(n)$. Subtracting the estimated echo from the microphone signal, the residual signal in the transform domain is obtained by subtract the estimated echo from the microphone signal

$$\mathbf{Q}(n)^H\boldsymbol{\epsilon}(n) = \mathbf{Q}(n)^H\mathbf{d}(n) - \boldsymbol{\beta}\mathbf{Q}(n)^H\hat{\mathbf{y}}(n) \quad (8)$$

where the underestimation matrix $\boldsymbol{\beta} = \text{diag}(\beta_1, \beta_2, \ldots, \beta_K)$, $0.8 \leq \beta_m \leq 1,\ m = 1, 2, \ldots, K$, is introduced to reduce the distortion of the near-end speech signal. As mentioned earlier, the $\hat{y}(n)$ is not an ideal estimator of $y(n)$, due to the loudspeaker non-linearities. Thus, the coefficients that are less than 1 can reduce the effect of estimation error which is serious during the double-talk period when the adaptation of the AEC device is stopped.

Hence, the residual signal $\boldsymbol{\epsilon}(n)$ in the time domain is

$$\boldsymbol{\epsilon}(n) = \mathbf{s}(n) + \boldsymbol{\delta}_y(n) + \mathbf{w}(n) \tag{9}$$

where, $\boldsymbol{\delta}_y(n)$ denotes the residual acoustic echo in $\boldsymbol{\epsilon}(n)$, defined by

$$\boldsymbol{\delta}_y(n) = \mathbf{y}(n) - \mathbf{Q}(n)\boldsymbol{\beta}\mathbf{Q}^H(n)\hat{\mathbf{y}}(n). \tag{10}$$

2.2.3. Echo suppression filter

Let $\mathbf{H}(n)$ be a $K \times K$ matrix which is the echo suppression filter, and let $\hat{\mathbf{s}}(n) = \mathbf{H}(n)\boldsymbol{\epsilon}(n)$ be the estimator of the near-end speech signal $\mathbf{s}(n)$, then the estimation error $\mathbf{e}_s(n)$ is written as

$$\begin{aligned} \mathbf{e}_s(n) &= \hat{\mathbf{s}}(n) - \mathbf{s}(n) \\ &= [\mathbf{H}(n) - \mathbf{I}]\mathbf{s}(n) + \mathbf{H}(n)[\boldsymbol{\delta}_y(n) + \mathbf{w}(n)] \end{aligned} \tag{11}$$

where, the first term in (11) is the distortion of the near-end speech, and the second term is further suppression of the echo and reduction of the background noise. The ideal case is $\mathbf{e}_s(n) = \mathbf{0}$, which means both terms in (11) should be $\mathbf{0}$. Because the signals, i.e. $\mathbf{s}(n)$, $\boldsymbol{\delta}_y(n)$ and $\mathbf{w}(n)$, may not be $\mathbf{0}$, to have $\mathbf{e}_s(n) = \mathbf{0}$ requires that $\mathbf{H}(n) - \mathbf{I} = \mathbf{0}$ and $\mathbf{H}(n) = \mathbf{0}$ simultaneously. This is obviously impossible. In other words, we cannot find a way to attenuate the echo without any near-end speech distortion. Hence we minimize the speech distortion in terms of mean-squared error under the constraints of suppressing the acoustic echo and the background noise to a certain level. By employing Kuhn-Tucker necessary conditions [8], we obtain the optimal filter

$$\mathbf{H}_{opt}(n) = \mathbf{Q}(n)\mathbf{G}(n)\mathbf{Q}^H(n) \tag{12}$$

where $G(n) = \text{diag}\{g^{(1)}(n), g^{(2)}(n), \ldots, g^{(K)}(n)\}$ with

$$g^{(m)} = \frac{\lambda_s^{(m)}}{\lambda_s^{(m)} + \mu[\lambda_{\delta_y}^{(m)} + \lambda_w^{(m)}]} \tag{13}$$

where μ is the Lagrange multiplier. In the derivation of (12)-(13), the approximation has been made that the off-diagonal elements in matrix $\mathbf{Q}(n)\mathbf{R}_s(n)\mathbf{Q}^H(n)$ were omitted, and $\lambda_s^{(m)}$ represents the diagonal elements of $\mathbf{Q}(n)\mathbf{R}_s(n)\mathbf{Q}^H(n)$. Similarly, $\lambda_{\delta_y}^{(m)}$ and $\lambda_w^{(m)}$ denote the diagonal elements of $\mathbf{Q}(n)\mathbf{R}_{\delta_y}(n)\mathbf{Q}^H(n)$ and that of $\mathbf{Q}(n)\mathbf{R}_w(n)\mathbf{Q}^H(n)$, respectively.

Unfortunately, it is difficult to apply (13) in practice, because we have to find $\mathbf{R}_s(n)$ and $\mathbf{R}_{\delta_y}(n)$ which can be used to respectively compute $\lambda_s^{(m)}$ and $\lambda_{\delta_y}^{(m)}$, and we also need a voice activity detector to determine $\mathbf{R}_w(n)$ for $\lambda_w^{(m)}$. Let $\mathbf{R}_{\hat{y}}(n)$ be the covariance matrix of $\hat{\mathbf{y}}(n)$, and $\lambda_{\hat{y}}^{(m)}$ a diagonal element of matrix $\mathbf{Q}(n)\mathbf{R}_{\hat{y}}(n)\mathbf{Q}^H(n)$. In order to simplify the structure of the subspace processor and based on a reasonable assumption that $\lambda_{\delta_y}^{(m)}$ is proportional to $\lambda_{\hat{y}}^{(m)}$, we propose a suboptimal acoustic echo suppression filter gain to replace (13)

$$g^{(m)} = \frac{\lambda_d^{(m)}}{\lambda_d^{(m)} + \mu\lambda_{\hat{y}}^{(m)}},\ \ m = 1, 2, \ldots, K \tag{14}$$

where, the Lagrange multiplier μ controls the echo suppression and the near-end speech distortion: larger μ implies higher echo attenuation but more signal distortion.

2.2.4. Dimension of signal subspace

As is known, the K-dimensional microphone signal subspace is the noisy (i.e. signal plus noise) subspace. Assuming that the dimension of the signal subspace is M where $M < K$, one can write $\mathbf{Q}(n)$ as

$$\mathbf{Q}(n) = [\mathbf{Q}_s(n), \mathbf{Q}_n(n)] \tag{15}$$

where $\mathbf{Q}_s(n)$ constitutes the signal subspace

$$\mathbf{Q}_s(n) = [\mathbf{q}_1(n), \mathbf{q}_2(n), \ldots, \mathbf{q}_M(n)] \tag{16}$$

and the noise subspace is spanned by $\mathbf{Q}_n(n)$

$$\mathbf{Q}_n(n) = [\mathbf{q}_{M+1}(n), \ldots, \mathbf{q}_K(n)] \tag{17}$$

Since it is very difficult to find an accurate rank of the signal subspace, M, a fixed value based on empirical data is used as the dimension of signal subspace in our work to simplify the algorithm of the subspace echo processor. Hence, only the signals projected in the signal subspace are considered, illustrated in Fig. 2.

2.2.5. Estimation of the covariance matrix $\mathbf{R}_d(n)$

According to the definition, the covariance matrix of the microphone signal $\mathbf{d}(n)$ is

$$\mathbf{R}_d(n) = \text{E}\{\mathbf{d}(n)\mathbf{d}^H(n)\} \tag{18}$$

where, $\text{E}[\cdot]$ denotes the expectation operator. Practically, empirical data are used to estimate the covariance matrix $\mathbf{R}_d(n)$. Differing from the approach in [7], much more past samples than future samples are used to estimate $\mathbf{R}_d(n)$ in order to reduce the potential delay which is an important issue in acoustic echo cancellation. Referring to (1) and let

n represent the index of the current frame, we estimate the covariance matrix from the samples in the past $T-1$ and current frames

$$\mathbf{R}_d(n) = \frac{1}{TK} \sum_{i=(n-T)K+1}^{nK} \mathbf{d}(i)\mathbf{d}^H(i). \qquad (19)$$

The estimation of covariance matrix $\mathbf{R}_d(n)$ is performed frame by frame with a rectangular window, which maintains the second order statistics of the samples in the window. After the estimation of the covariance matrix, the $\mathbf{Q}(n)$ and $\mathbf{\Lambda}_d(n)$ can be obtained by applying eigendecomposition to $\mathbf{R}_d(n)$.

3. EXPERIMENTAL RESULTS

In the experiments carried out in an office room with 4(L)× 3.5(W)×2.7(H), a 1.3GHz Pentium-IV PC was used as the host, associated with the Delta 1010 Digital Recording System which has 10-input and 10-output full-duplex recording interface, made by Midiman company. A common amplified PC loudspeaker was used to play the far-end speech. The microphone signal was amplified by Tascam MX-80 microphone/line mixer before it was sent to the recording system. The computer fans were the main contributors of background noise.

The third-order AP algorithm was employed to estimate the acoustic echo, where the filter length was 1600 taps, corresponding to 200ms with 8kHz sampling rate. The step-size was set to 0.9.

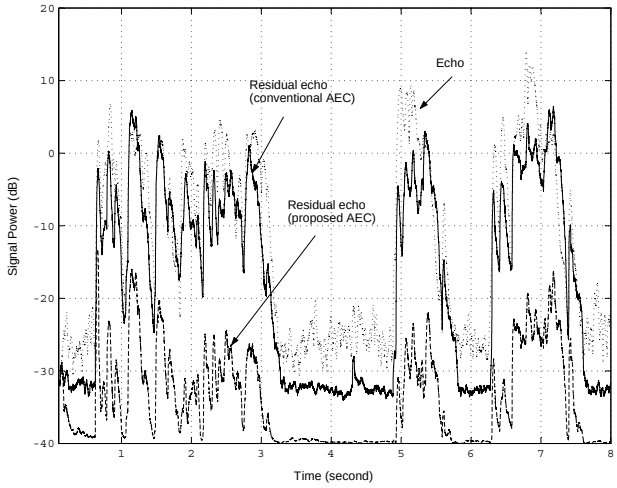

Fig. 3. The signal power versus time: dot – acoustic echo; solid – residual echo of a conventional AEC with AP; dash – residual echo of the proposed AEC.

For the subspace echo processor, the dimensions of the noisy signal subspace and the signal subspace were set $K = 40$ and $M = 32$, respectively. A Hanning window and a rectangular window were respectively used in the synthesis and analysis procedures, with 50% frame overlap. There were $T = 10$ frames used to estimate the covariance matrix $\mathbf{R}_d(n)$. The parameter μ was set to 10 for compromise between echo suppression and signal distortion.

The experimental results are shown in Fig. 3, where a result from a conventional AEC which only employed a linear adaptive filter with AP is also plotted for comparison. The proposed AEC outperforms the conventional AEC in terms of extra 10-20dB echo suppression and about 8dB background noise reduction. Since the frame size is 40 samples, corresponding to 5ms, the delay is acceptable in most acoustic echo cancellation applications.

4. CONCLUSION

A new nonlinear AEC has been proposed in this paper. Compared to conventional AECs, this subspace-based AEC significantly suppresses the nonlinear acoustic echo produced by loudspeakers and remarkably reduces the background noise, which has been verified by our experiments.

Furthermore, the subspace echo processor in the proposed AEC has an open-loop structure, which shows a predictable behaviour. Accordingly, it is much more robust than other nonlinear adaptive algorithm such as Volterra filter and neural networks. Considering that the eigendecomposition operation in the new AEC leads to high computational complexity, an appropriate subspace tracking technique may be used to reduce the computational burden.

5. REFERENCES

[1] W. Klippel, "nonlinear large-signal behavior of electrodynamic loudspeaker at low frequencies," *J. Audio Eng. Soc.*, vol. 40, no. 6, pp. 483–496, Jun. 1992.

[2] A. Stenger and W. Kellermann, "Nonlinear acoustic echo cancellation with fast converging memoryless preprocessor," in *Proc. ICASSP'00*, Jun. 2000, pp. 805–808.

[3] A.N. Birkett and R.A. Goubran, "Nonlinear loudspeaker compensation for hands free acoustic echo cancellation," *Electronics Letters*, vol. 32, no. 12, pp. 1063–1064, Jun. 1996.

[4] B.S. Nollett and D.L. Jones, "Nonlinear echo cancellation for hands-free speakerphones," in *Proc. NSIP'97*, Michigan, USA, Sep. 1997.

[5] A. J. M. Kaizer, "Modeling of the nonlinear response of an electrodynamic loudspeaker by a volterra series expansion," *J. Audio Eng. Soc.*, vol. 35, no. 6, pp. 421–433, Jun. 1987.

[6] K. Ozeki and T. Umeda, "An adaptive filtering algorithm using an orthogonal projection to an affine subspace and its properties," *Elec. and Comm. in Japan*, vol. 67-A, no. 5, pp. 19–27, 1984.

[7] Y. Ephraim and H. L. Van Trees, "A signal subspace approach for speech enhancement," *IEEE Trans. Speech and Audio Processing*, vol. 3, no. 4, pp. 251–266, Jul. 1995.

[8] S. G. Nash and A. Sofer, *Linear and Nonlinear Programming*, McGraw-Hill, New York, 1996.