Speech Enhancement Using the Constrained-Optimization Technique

W. Li and W. C. Siu

***Abstract*—We address a problem of speech enhancement: recovering a speech source from a mixture of its delayed versions and additive noise. By using the constrained-optimization technique, the second order statistics based algorithm is developed. The new proposed algorithm requires no strong limitations to the speech signal and the noise. Simulation results show that our algorithm achieves a better performance as compared to other algorithms.**

I. Introduction

SPEECH enhancement has been active in speech signal processing. The objective is to extract a single speech-source signal from its delayed versions and in noisy environment. Depending upon the amount and the type of noise and the strength of echoes existing in the environment, the resulting speech signals could vary substantially. The quality of the speech may range from being slightly degraded to being annoying to listeners, and in the worst case, it could be totally unintelligible. It is necessary to recover the speech signal from the distortion.

A number of approaches to signal recovery have been proposed [1]–[4]. These approaches make use of second-order statistics [1], [2] or the higher-order statistics [3], [4] of the outputs. They are basically the least-squares solutions. However, if the unknown parameters in an algorithm have inherent, nonlinear relations, they are usually assumed to be independent of each other to apply the linear least-squares technique. A larger estimation error is inevitably generated, although an additional postprocessing step may reduce the error. To accommodate practical applications and achieve accurate estimation, we will design our algorithm with (1) no strong limitations to the source and noise except for all signals being stationary and (2) with consideration to the inherent nonlinear relation among the unknown parameters, while deriving our new algorithm to avoid the additional postprocessing step.

II. Problem and Assumptions

Let us first consider a linear time invariant (LTI) system with the following model:

$$\begin{aligned} y_1(k) &= A(z^{-1})s(k) + T(z^{-1})n(k) \\ y_2(k) &= C(z^{-1})s(k) + n(k) \end{aligned} \tag{1}$$

where

$$\begin{aligned} A(z^{-1}) &= a(0) + a(1)z^{-1} + \cdots + a(p)z^{-p}, \\ C(z^{-1}) &= c(0) + c(1)z^{-1} + \cdots + c(r)z^{-r}, \end{aligned}$$

and

$$T(z^{-1}) = t(0) + t(1)z^{-1} + \cdots + t(l)z^{-1}.$$

z^{-1} is a shift operator (i.e., $z^{-1}s(k) = s(k-1)$.). The received signals $y_1(k)$ and $y_2(k)$ are the outputs of the LTI systems with the same input, and $s(k), n(k)$ is the received noise from a receiver. Because the two receivers are in the same background, the two received noises are highly related. We use a linear, time-invariant operator $T(z^{-1})$ to link the received noises.

To simplify the problem, we assume that 1) all signals are sampled wide-sense stationary random processes with zero-mean and 2) $A(z^{-1})$ and $C(z^{-1})$ are relatively prime and $a(0) = 1$.

III. Algorithm Development

A. Unconstrained LS Solution

To accommodate a practical environment, we will make use of the second-order statistics of the outputs for the parameter estimation. The cross-correlation and auto-correlation of $y_i(k)$ and $y_j(k)$ for lag τ are represented by

$$r_{y_i y_j}(\tau) = E[y_i(k)y_j(k+\tau)] \qquad i, j = 1, 2. \tag{2}$$

Substituting (1) into (2) and applying the Fourier transform, the following results can be obtained:

$$\begin{aligned} &A(\omega)P_{y_2y_1}(\omega) - A(\omega)T^*(\omega)P_{y_2y_2}(\omega) \\ &\quad = C(\omega)P_{y_1y_1}(\omega) - T^*(\omega)C(\omega)P_{y_1y_2}(\omega) \end{aligned} \tag{3}$$

where $P_{y_iy_i}(\omega)$ $(i, j = 1, 2)$ is the power spectrum of the joint random processes y_i and y_j, $P_{ss}(\omega)$ and $P_{nn}(\omega)$ are the power spectrums of the source signal and the additive noise, respectively. Let

$$\begin{aligned} A_t(\omega) &= e^{-j\omega l}A(\omega)T^*(\omega) \\ C_t(\omega) &= e^{-j\omega l}C(\omega)T^*(\omega) \end{aligned} \tag{4}$$

where

$$A_t(\omega) = a_t(0) + a_t(1)e^{-j\omega} + \cdots + a_t(p+l)e^{-j(p+l)\omega}$$

and

$$C_t(\omega) = c_t(0) + c_t(1)e^{-j\omega} + \cdots + c_t(r+l)e^{-j(r+l)\omega}.$$

Manuscript received June 5, 1999. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. K. K. Paliwal.

The authors are with the Department of Electronic and Information Engineering, The Hong Kong Polytechnic University, Hung Hom, Kowloon, Hong Kong (e-mail: enliwei@en.polyu.edu.hk).

Publisher Item Identifier S 1070-9908(00)00933-0.

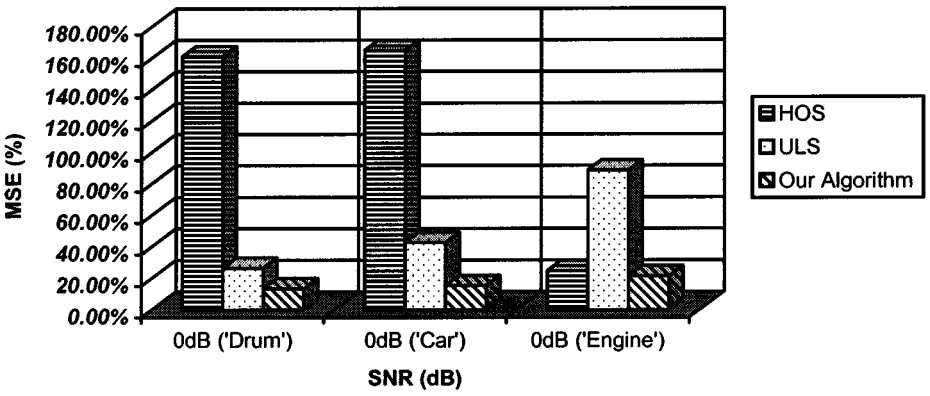

Fig. 1. MSE of the estimated parameters under real noises (data length = 2000).

Substituting (4) into (3) and taking the inverse Fourier transform, we have

$$\sum_{i=1}^{p} a(i)r_{y_2y_1}(k-i) - \sum_{i=0}^{p+l} a_t(i)r_{y_2y_2}(k+l-i) - \sum_{i=0}^{r} c(i)r_{y_1y_1}(k-i) + \sum_{i=0}^{r+l} c_t(i)r_{y_1y_2}(k+l-i) = -r_{y_2y_1}(k) \quad (5)$$

Equation (5) may be viewed as a linear equation with $(2p + 2l + 2r + 3)$ unknowns involving $\{a(k)\ k = 1,2,\cdots,p\}$, $\{c(k)\ k = 0,1,2,\cdots,r\}$,, $\{a_t(k)\ k = 0,1,2,\cdots,p+l\}$, and $\{c_t(k)\ k = 0,1,2,\cdots,r+l\}$ if unknowns $\{a_t(k)\}$ and $\{c_t(k)\}$ are assumed to be independent of $\{a(k)\}$ and $\{c(k)\}$,, respectively. By selecting some values for k in series, a set of overdetermined equations is formed. For instance, let k range from m_1 to m_2, and (5) can be expressed in matrix form

$$R\vec{\theta} = \vec{r} \quad (6)$$

where

$$\vec{\theta} = [a(1)\ \cdots\ a(p)\quad c(0)\ \cdots\ c(r)\quad a_t(0)\ \cdots\ a_t(p+l)\quad c_t(0)\ \cdots\ c_t(r+l)]^T$$
$$\vec{r} = [r_{y_2y_1}(m_1)\quad r_{y_2y_1}(m_1+1)\quad \cdots\quad r_{y_2y_1}(m_2)]^T.$$

R denotes a matrix and can be easily expressed.

The corresponding LS solution is

$$\vec{\theta} = (R^TR)^{-1}R\vec{r}. \quad (7)$$

Note that it is not a final solution for each coefficient. The final solution is usually obtained by synthesising the LS estimates that are related but not are the coefficients in $A(z^{-1})$ and $C(z^{-1})$. This processing is called postprocessing. Obviously, the LS solutions may be far away from the real parameters due to ignoring the inherent nonlinear relations between $\{a(k)\}$ and $\{a_t(k)\}$, and $\{c(k)\}$ and $\{c_t(k)\}$.. To avoid the additional postprocessing step and reduce estimation error, let us consider these relations while deriving our algorithm in the next section.

B. Constrained Optimal Solution

Let $\vec{e} = R\vec{\theta} - \vec{r}$. Set the goal attainment J as

$$J = \vec{e}^T\vec{e}. \quad (8)$$

Minimizing J produces the above unconstrained LS solution, which ignores the inherent nonlinear relations of the unknowns. To obtain a reasonable solution, let us consider these relations together. From (4), it is very easy to attain the following $(p + r + l + 1)$ equations.

$$\sum_{i=0}^{k} a_t(i)c(k-i) = \sum_{i=0}^{k} a(i)c_t(k-i) \quad \text{for } k = 0,1,\cdots,p+r+l. \quad (9)$$

A constrained-optimization problem can then be stated as follows:

$$\text{Min } J$$
$$\text{Subject to } \sum_{i=0}^{k} a_t(i)c(k-i) = \sum_{i=0}^{k} a(i)c_t(k-i) \quad k = 0,1,\cdots,p+r+l.$$

By using the Gauss–Newton method, we can obtain the optimal solutions for $\{a(k), k = 1,2,,p\}$, $\{c(k), k = 0,1,2,,r\}$, $\{a_t(k), k = 0,1,,p+l\}$, and $\{c_t(k), k = 0,1,,r+l\}$. We note that this is not the only algorithm for solving the criterion equations. Iterative gradient-based algorithms such like the steepest-descent or the Newton–Raphson may also be applied here.

C. Signal Reconstruction

Provided that the unknown channel parameters have been identified by the above method denoted by $\hat{A}(z^{-1})$, $\hat{C}(z^{-1})$, and $\hat{T}(z^{-1})$,, respectively, and that all the roots of $\hat{A}(z^{-1}) - \hat{C}(z^{-1})\hat{T}(z^{-1}) = 0$ are inside the unit circle, the estimated signal can be expressed by

$$\hat{s}(k) = \frac{1}{\hat{A}(z^{-1}) - \hat{C}(z^{-1})\hat{T}(z^{-1})}\,[y_1(k) - \hat{T}(z^{-1})y_2(k)]. \quad (10)$$

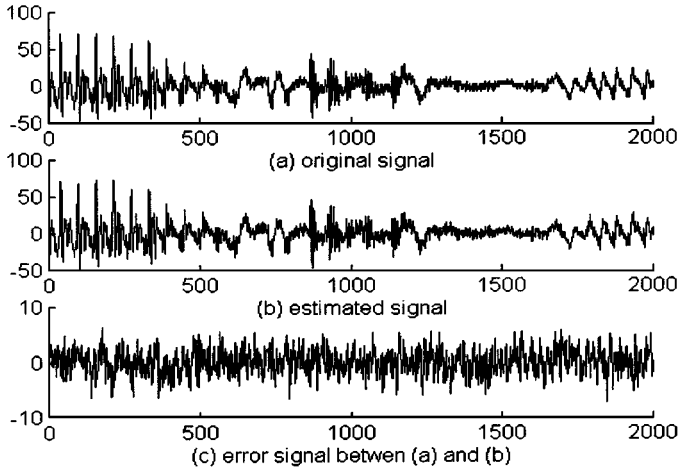

Fig. 2. Speech enhancement with engine noise at SNR 0 dB via our algorithm (a) original speech signal, (b) estimated signal, and (c) error between (a) and (b).

Obviously, if $\hat{A}(z^{-1})$, $\hat{C}(z^{-1})$, and $\hat{T}(z^{-1})$ equal to $A(z^{-1}), C(z^{-1})$ and $T(z^{-1})$, respectively, the estimated source exactly equals the real one.

IV. Simulation Results

Extensive simulations have been carried out to compare the proposed constrained-optimization algorithm (COP) with the higher order statistics-based (HOS) algorithm [3] and the unconstrained least-square algorithm (ULS). For each test case, the coefficient vectors in model (1) are $[a(0)\ \ a(1)\ \ a(2)] = [1\ \ 0.2\ \ 0.1]$, $[c(0)\ \ c(1)] = [0.8\ \ 0.2]$, and $[t(0)\ \ t(1)] = [0.7\ \ 0.15]$. We define the SNR as $SNR = 10\log(\|s(\cdot)\|_2/\|n(\cdot)\|_2)$ (dB) and the mean squared error (MSE) (MSE) as $MSE = [(h-\hat{h})^T(h-\hat{h})/h^Th]^{1/2}$, where $\hat{h}$ is a vector that consists of all estimated coefficients, and h is a vector that consists of real values. We add a speech signal (the data length $= 2000$) as the source signal.

Fig. 1 shows the estimated results for each approach under three noises emitted by a drum, car, and engine at SNR 0 dB. Obviously, only the COP algorithm can adapt these noises with better accuracy. This is because the COP algorithm is not limited to any kind of noise, but the HOS algorithm is available only for the Gaussian noise. Under the noise "engine," the recovered signal by using our algorithm is shown in Fig. 2. It is easy to see that the estimated signal is very close to the original one with very small errors.

V. Conclusions

A new COP algorithm for FIR-channel identification and speech enhancement has been proposed in this paper. Because of using the second-order statistics of outputs, there is no strong limitation on the source signal and additive noise. In addition, by employing the COP technique, our algorithm avoids the additional postprocessing step that usually happens in other algorithms. Simulation results have illustrated the correctness of the coefficients using our new algorithm. By comparing our algorithm with the ULS and the HOS algorithms, we can see that our algorithm performs the parameter estimation and recovers the original speech signal with a better performance under the practical noises even if the SNR is 0 dB.

References

[1] E. Weinstein, M. Feder, and A. V. Oppenheim, "Multi-channel signal separation by decorrelation," *IEEE Trans. Speech Audio Processing*, vol. 1, pp. 405–413, Oct. 1993.

[2] M. Abe *et al.*, "Estimation of the waveform of a sound source by using an iterative technique with many sensor," *IEEE Trans. Speech Audio Processing*, vol. 6, pp. 24–35, Jan. 1998.

[3] D. Yellin and E. Weinstein, "Criteria for multichannel signal separation," *IEEE Trans. Signal Processing*, vol. 42, pp. 2158–2167, Aug. 1994.

[4] C. L. Nikias and A. P. Petropulu, *Higher-Order Spectra Analysis: A Nonlinear Signal Processing Framework*. Englewood Cliffs, NJ: Prentice Hall, 1993.