ADAPTIVE REDUCTION OF INTERFERING SPEAKER NOISE USING THE LEAST
MEAN SQUARES ALGORITHM

S. T. Alexander

CENTER FOR COMMUNICATIONS AND SIGNAL PROCESSING

North Carolina State University
Raleigh, N.C. 27650

CCSP-TR-83/6

September 7, 1983

Abstract

The phenomenon of interfering speakers is a significant problem in the design of effective compression algorithms for speech. The desired signal (i.e., the main speaker) and the interfering signal (i.e., other speakers in the background) have similar statistical and spectral characteristics which is quite different from the environment of a main speaker plus white noise. In this paper, the adaptive Least Mean Squares (LMS) algorithm is used to separate speaker-produced "information" from interferer-produced "noise" on the basis of the difference in power levels associated with the two phenomena. This method exploits the property of LMS that it rapidly adapts for the dominant excitation modes while simultaneously adapting very slowly for the weaker modes of excitation. As applied to speech, the LMS quickly adapts to the strongest signals (main speaker) in a composite waveform, while converging very slowly for the weaker signals (the background interferers). This selective convergence property of LMS is next analyzed using an eigenvalue-eigenvector approach which easily displays the signal separation property. Lastly, computer simulations are presented which verify the analysis above for representative speech waveforms.

1.0 Introduction

The design of efficient coding (or compression) algorithms for the main speaker-plus-interfering speaker environment is a significant problem in speech communication systems. This situation commonly occurs for telephone communication in crowded rooms, voice-entry on workshop floors, and others. The statistical and spectral similarity between the desired signal (main speaker) and the interfering signal (background speakers)* often prohibit an improvement using only spectral filtering techniques.

However, the interferers are usually reduced in power level compared to the speaker. Whereas this is of no consequence in spectral filtering methods (since the reduced level interferers are still in the spectral band of the speaker), it allows one to employ the class of adaptive methods known as gradient-search techniques to remove much of the interferer power, while preserving much of the speaker characteristics. A very simple, computationally-efficient member of this class of algorithms is the Least Mean Squares (LMS) algorithm [1], which has been successfully applied to spectral estimation [2], beamforming [3], image compression [4] and other areas. Sambur [5] has applied the LMS algorithm in a speech enhancement application to cancel sinusoidal noise components, but the present paper represents the first application toward reducing the effects of interfering speakers upon the speech encoder. In the present application, the LMS algorithm is structured as a pre-processor which reduces the interferer power levels prior to source and channel encoding. Such a configuration is shown in Figure 1. In this configuration, the "output" of the LMS algorithm is the filter prediction, $\hat{x}(k)$, which is an estimate of the main speaker waveform, $x(k)$. It will be developed that this estimate is more nearly an estimate of $s(k)$, the speaker waveform.

*In this paper, the main speaker will be referred to as the "speaker" and the interfering speakers will simply be called "interferers" or "interference".

This estimate $\hat{x}(k)$ could then be passed to a suitable source encoder, as in DPCM [6], which would extract the redundancy from $\hat{x}(k)$ and transmit the residual over the channel. Since the two problems of interferer extraction and source encoding are somewhat independent, we only treat the former in this paper. For an in-depth survey of source encoding, the reader is referred to the excellent work by Gibson [7].

2.0 The LMS Prediction Filter

The LMS algorithm [1] is used as an adaptive linear prediction filter to make a prediction of the current sample value based upon a knowledge of previous samples. The structure of the filter for this speech application is shown in Figure 1, where the blocks τ_1 denote delays of τ_1 samples and the w_1 are the adaptive filter coefficients. In general, for speech applications a quantizer would be placed in the forward path; however, since the scope of this paper is an analysis and simulation of interfering speaker reduction, the quantizer consideration will not be considered at this time.

As seen from Figure 1, a prediction $\hat{x}(k)$ is made of the current sample based on a knowledge of previous samples:

$$\hat{x}(k) = \sum_{i=1}^N w_i(k)x(k-\tau_i) = \underline{w}^T(k)\underline{X}(k) \quad (2-1)$$

where the $w_i(k)$ are the weighting or prediction filter coefficients. An error value is then formed between the predicted current sample and the actual current sample:

$$e(k) = x(k) - \hat{x}(k) \quad (2-2)$$

This error is then used in the LMS algorithm to update the weighting vector coefficients,

$$\underline{w}(k+1) = \underline{w}(k) + \alpha e(k)\underline{X}(k) \quad (2-3)$$

Figure 1. Configuration of LMS Pre-processor.

where α is a feedback gain parameter controlling convergence of the algorithm.

It is desired to solve (2-3) for the expected weight value $E\{w(k)\}$ which will display the convergency properties of LMS. However, (2-3) represents a set of coupled difference equations. The expected weight value may equivalently be obtained in terms of an uncoupled weight vector $\underline{v}(k)$, where

$$\tilde{\underline{w}}(k) = A\underline{v}(k) \quad (2-4)$$

and the tilde notation signifies the expected value (or ensemble average) at iteration k . In (2-4), the matrix A is the matrix of eigenvectors of the data autocorrelation matrix, R_{xx} , where

$$R_{xx} = E[\underline{X}(k)\underline{X}^T(k)] \quad (2-5)$$

Further, even though R_{xx} may eventually be time varying, it is still symmetrical [8] and may be expanded in terms of its eigenvalues and eigenvectors by the transformation

$$R_{xx} = ADA^T \quad (2-6)$$

where D is the diagonal matrix of eigenvalues of R_{xx} . This matrix is a constant for stationary data. Although in a true sense, speech is a nonstationary process, we may examine a portion of a speech segment as a short-term stationary process (i.e., constant pitch frequency, constant formant frequencies, and power constant levels) as done by Rabiner and Schafer [9]. It will also be useful later to think of A as partitioned into its eigenvectors, $\underline{a}_i$:

$$A = [\underline{a}_1, \underline{a}_2, \dots, \underline{a}_N] \quad (2-8)$$

Solutions for the uncoupled expected weight vector have been obtained by various authors for specific cases [1, 8, 10]. The form obtained for an image compression application by Alexander and Rajala [4] is especially applicable to the speech problem and is given by

$$v_i(k) = \gamma_i^k v_i(0) + \frac{c_i}{\lambda_i} [1 - \gamma_i^k] \quad (2-8)$$

where $v_i(k)$ is the i th uncoupled weight at iteration k and

$$\lambda_i = i\text{th eigenvalue of } R_{xx} \quad (2-9)$$

$$c_i = \underline{a}_i^T \underline{p} \quad (2-10)$$

$$\underline{p} = E[x(k) \underline{x}(k)] \quad (2-11)$$

$$\gamma_i = 1 - \alpha \lambda_i \quad (2-12)$$

One may also use (2-8) to derive a solution for the vector, $\underline{v}(k)$, of uncoupled weights. This derivation is straightforward, but somewhat lengthy, and is omitted to preserve continuity. The result is given by

$$\underline{v}(k) = G^k \underline{v}(0) + D^{-1} [I - G^k] A^T \underline{p} \quad (2-13)$$

where

$$G = [I - \alpha D] \quad (2-14)$$

is the diagonal matrix of the γ_i defined as in (2-12).

In this paper, we will examine the consequences of the weight vector solution from (2-4) and (2-13) as it impacts the interference reduction problem. The weight solution (2-4) will be examined for the simple, but important, case of $N=2$ weights and this will display the signal separation property of LMS. Due to the disparity in power levels between the speaker and interferer (which translates into an eigenvalue disparity in the data), the LMS weights will quickly converge toward the same values which would result from a speaker-only case. Thus, the speaker component will be enhanced and the interferer will be processed as a noise sequence.

An important distinction exists, however, between the properties of the filter output as the convergence process proceeds and the filter output after steady-state has been reached. During convergence, the LMS filter will converge relatively rapidly to the dominant input component (speaker) and, thereafter, more slowly to the secondary components (interferers). Thus, there will be a span of time in which the filter output will pass the main speaker sequence, but will not have begun to converge appreciably for the interferer. During this time span, the interferer sequence will be substantially reduced at the filter output, resulting in a greatly enhanced main speaker sequence.

Although the analysis in this paper is done for stationary signals, it is the ultimate nonstationarity of the speech process which allows for long-term processing gain (i.e., reduction of interferer power) using LMS. Over the short-term one may consider the speech as quasi-stationary and the LMS will begin to converge toward some "steady state" solution. However, as the power level, formant frequencies, pitch periods, etc., begin to change, the LMS filter will begin to converge toward a "new" steady-state solution. Thus, as long as the main speaker is present the LMS weight vector will be largely adapting to this dominant speaker and in the desired mode of operation.

The next section examines (2-4) and (2-13) in detail and gives a quantitative illustration of the above property for the simple, but important, case of $N=2$ coefficients.

3.0 Separation of Speaker and Interferer

The separation of speaker and interferer components using the LMS algorithm is best seen by considering the simple, although important, example of a two-coefficient predictor.

Consider the case of a segment of speech, $x(k)$, consisting of a main speaker, $s(k)$, plus an interfering speaker, $z(k)$. Thus,

$$x(k) = s(k) + z(k) \quad (3-1)$$

Assuming the speaker and interferer are uncorrelated (i.e., their speech has different unrelated information contents), the autocorrelation function becomes

$$r_x(n) = r_s(n) + r_z(n) \quad (3-2)$$

where $r_s(n)$ and $r_z(n)$ are the autocorrelation functions at the n th lag of the speaker and interferer sequences, respectively. Before we can proceed with the analysis of the LMS filter, the delay values from (2-1) must be chosen. Although there are many values which could be chosen, it was decided to estimate the current sample using the $(k-1)$ st and $(k-p)$ th samples, where p corresponds to the value of pitch period.* For the quasi-stationary case under analysis, p is assumed to be constant. Thus, the data vector $\underline{X}(k)$ becomes

$$\underline{X}(k) = \begin{bmatrix} x(k-1) \\ x(k-p) \end{bmatrix} \quad (3-3)$$

*It will be assumed that another signal processing structure provides us with this value of p . It should be noted that an accurate estimate of p is, in itself, a difficult problem [9].

For this $N=2$, it is a simple matter to compute the quantities in (2-13). For simplicity in (2-18), we will assume that $\underline{v}(0) = \underline{0}$. The R_{xx} matrix is easily found to be

$$R_{xx} = \begin{bmatrix} r_x(0) & r_x(p-1) \\ r_x(p-1) & r_x(0) \end{bmatrix} \quad (3-4)$$

The eigenvalues and eigenvectors of R_{xx} are also easily computed as

$$\lambda_1 = r_x(0) + r_x(p-1) \quad (3-5)$$

$$\lambda_2 = r_x(0) - r_x(p-1) \quad (3-6)$$

$$A = \frac{1}{\sqrt{2}} \begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix} \quad (3-7)$$

Thus, $v_1(k)$ and $v_2(k)$ from (2-14) become

$$v_1(k) = \frac{r_x(1) + r_x(p)}{\sqrt{2} [r_x(0) + r_x(p-1)]} (1 - \gamma_1^k) \quad (3-8)$$

$$v_2(k) = \frac{r_x(1) - r_x(p)}{\sqrt{2} [r_x(0) - r_x(p-1)]} (1 - \lambda_2^k) \quad (3-9)$$

The convergence of this set of $v_1(k)$ to the SPEAKER-ONLY $v_1(k)$ may be illustrated by considering only $v_1(k)$. A similar development holds for $v_2(k)$.

Using the binomial expansion, we may approximate the exponential term in (3-7) as follows:

$$\lambda_1^k = (1 - \alpha\lambda_1)^k \cong 1 - k\alpha\lambda_1 \quad (3-10)$$

for $\alpha\lambda \ll 1$, which is a valid assumption for this application.

Also, the approximation

$$e^{-k\alpha\lambda_1} \cong 1 - k\alpha\lambda_1 \quad (3-11)$$

is valid in this region. Equating the results from (3-10) and (3-11) we have

$$\gamma_1^k \cong e^{-k\alpha\lambda_1} \quad (3-12)$$

Substituting this approximation and the definition of $r_{xx}(n)$ from (2-18) into (3-8) we have the following expression for $v_1(k)$:

$$v_1(k) = \left\{ \frac{r_s(1) + r_z(1) + r_s(p) + r_z(p)}{\sqrt{2} [r_s(0) + r_z(0) + r_s(p-1) + r_z(p-1)]} \right\} \\ \times \left\{ 1 - e^{-k\alpha[r_s(0) + r_z(0) + r_s(p-1) + r_z(p-1)]} \right\} \quad (3-13)$$

It is useful to normalize the factors in (3-13) by $r_s(0) = \sigma_s^2$, the mean square power of the main speaker. Redefining the correlations in (3-13) in terms of a set of normalized correlation coefficients,

$$r_s(n) = \sigma_s^2 \tau_s(n) \quad (3-14)$$

$$r_z(n) = \sigma_z^2 \tau_z(n) , \quad (3-15)$$

Substituting in (3-13) and rearranging somewhat produces

$$v_1(k) = \left\{ \frac{\tau_s(1) + \tau_s(p) + \sigma^2[\tau_z(1) + \tau_z(p)]}{\sqrt{2} [1 + \tau_s(p-1) + \sigma^2[1 + \tau_z(p-1)]]} \right\}$$

$$x \left[1 - e^{-k\alpha\delta_s^2 \{ 1 + \tau_s(p-1) + \sigma^2 [1 + \tau_z(p-1)] \}} \right] \quad (3-15)$$

where

$$\sigma^2 = \sigma_z^2 / \sigma_s^2 \quad (3-16)$$

The parameter σ^2 may be thought of as the inverse of the speaker-to-interferer ratio (SIR), and for many environments will be

$$\sigma^2 \ll 1 \quad (3-17)$$

This is in keeping with the original assumption that the main speaker signal was the dominant component in the composite waveform.

Note the consequence of (3-17) in the solution for $v_1(k)$ from (3-15). The normalized correlation coefficients, $\tau_s(n)$ and $\tau_p(n)$, are less than unity, and thus all terms multiplied by $\sigma_s^2 \ll 1$ may be neglected without substantial error for a wide range of k . From the exponential argument in (3-15), it is seen that this condition is satisfied as long as the convergence contributions due to the $\tau_z(n)$ terms are negligible. This holds if

$$k\alpha\sigma_z^2 [1 + \tau_z(p-1)] \ll 1 \quad (3-18)$$

or, solving for k ,

$$k \ll \frac{1}{\alpha\sigma_z^2 [1 + \tau_z(p-1)]} \quad (3-19)$$

Thus, as long as (3-19) holds, the terms in (3-15) which are multiplied by σ^2 may be neglected and the solution for $v_1(k)$ becomes

$$v_1(k) \cong \frac{\tau_s(1) + \tau_s(p)}{\sqrt{2} [1 + \tau_s(p-1)]} \{ 1 - e^{-k\alpha\sigma_s^2 [1 + \tau_s(p-1)]} \} \quad (3-20)$$

A similar reasoning and analysis holds for $v_2(k)$ and gives

$$v_2(k) \cong \frac{\tau_s(1) - \tau_s(p)}{\sqrt{2}[1 - \tau_s(p-1)]} \{1 - e^{-k\alpha\sigma_s^2[1-\tau_s(p-1)]}\} \quad (3-21)$$

These equations (3-20) and (3-21) are identical to the $v_1(k)$ solutions which would be obtained for an input sequence consistence of SPEAKER-ONLY; that is,

$$x(k) = s(k)$$

Thus, we have shown that for the dominant speaker plus interfer, the adaptive LMS weights initially converge to the same weights as would be produced by the speaker-only case.

4.0 Simulation Results

This section presents results of applying the LMS filter as a preprocessor to reduce the interfering signal power for simulated speech signals. The speech model chosen is the common cascade of second-order sections excited by an impulse at the pitch frequency [9]. This is a standard model for voiced-speech and utilizes second-order transfer functions of the form

$$V_p(z) = \frac{1 - 2r_p \cos \sigma_p + r_p^2}{1 - (2r_p \cos \sigma_p)z^{-1} + r_p^2 z^{-2}} \quad (4-1)$$

where, for the p th section, the location of the p th pole conjugate-pair is given by

$$z_p = r_p e^{j\sigma_p} \quad (4-2)$$

$$z_p^* = r_p e^{-j\sigma_p} \quad (4-2b)$$

Thus, the p th formant frequency is selectable via σ_p and the formant bandwidth is controlled by r_p . For a P -stage model of the voiced speech process, the total transfer function becomes

$$V(z) = \prod_{p=1}^P V_p(z) \quad (4-3a)$$

The output simulated speech is then given by

$$Y(z) = V(z)X(z) \quad (4-3b)$$

where $X(z)$ is the z -transform of $x(n)$, the input signal. For voiced speech,

$$x(k) = v \sum_{n=0}^{\infty} \delta(k-nT_p); \quad n = 0, 1, 2, \dots \quad (4-4)$$

where v is a power scaling factor and T_p is the pitch period. In the simplistic model of (4-4), the pitch period is assumed to be an integer number of samples. To produce simulated speech via the preceding model, one simply

defines a filter order P , gives numerical values to the (σ_p, r_p) parameters in (4-1) for each second order section, and then takes the inverse z -transform of (4-3a) to produce a difference equation. Then, when excited by the pitch period model of (4-4), the simulated speech will be the output sequence $y(k)$. This procedure is equivalent to exciting the linear discrete filter with the impulse train, producing the output simulated speech as shown in figure 2.

For the work of this paper, the parameters shown in Table 1. were chosen for the speaker and interferer. A simple second order resonator ($P = 1$) was chosen for simplicity to be the transfer function. Thus, $V(z) = V_1(z)$ in (4-3a).

	σ_1	r_1	PITCH FREQ. (Hz)
Speaker	900	300	variable
Interferer	1800	1000	300

Table 1. Parameters of Second-Order Resonator.

The sampling frequency was 8 KHz. The σ_1, r_1 parameters in Table 1 correspond to the formant frequency location and the formant bandwidth, respectively. The speaker pitch was allowed to vary to show the tracking adaptability of the LMS algorithm. The trajectory of the speaker pitch is shown in Figure 3.

The preceding model is used since we may create known discrete data sequences which have the same statistical and spectral parameters as actual speech. Thus, known sequences corresponding to the main speaker and interferer may be created independently with selected formants, bandwidths and pitch excitations. One may then crosscorrelate the output sequence resulting

Figure 2. Synthetic Voiced Speech Generation

Figure 3. Pitch Trajectory for Simulations.

from LMS processing with the above known speaker and interferer sequences, producing a speaker correlaton coefficient, $\rho_s(k)$ and an interferer correlation coefficient, $\rho_I(k)$. An output Signal-to-Interferer Ratio, $SIR_0(k)$, may then be calculated, according to

$$SIR_0(k) = 10 \log \left| \frac{\overline{\rho_s(k)}}{\overline{\rho_I(k)}} \right|_{OUT} \quad (4-5)$$

In (4-5), the "k" argument signifies that this evaluation is made at time k. A similar cross-correlation of known speaker and interferer sequences with the composite input sequence produces

$$SIR_i(k) = 10 \log \left| \frac{\overline{\rho_s(k)}}{\overline{\rho_I(k)}} \right|_{IN} \quad (4-6)$$

Thus, we may use the "before and after" SIR values to compute the speaker enhancement due to using LMS.

Concerning the specific computation of the $\rho(k)$, recall from the previous sections that $\hat{x}(k) = s(k) + z(k)$ and the output of the LMS filter is $\hat{x}(k)$. The output speaker correlation is therefore calculated by products of the form $\hat{x}(k)s(k)$ and the interferer correlation is calculated by products of the form $\hat{x}(k)z(k)$. In this work, frames of 30 samples were used to form the output correlation coefficients as follows:

$$\rho_s(k) = \frac{1}{P_x(k)} \sum_{i=1}^{30} \hat{x}(k-i) s(k-i) \quad (4-7a)$$

$$\rho_I(k) = \frac{1}{P_x(k)} \sum_{i=1}^{30} \hat{x}(k-i) z(k-i)$$

where

$$P_x(k) = \sum_{i=1}^{30} x^2(k-i)$$

is the total power in the input signal frame at time k . The coefficient values above were computed at $k=30, 60, 90, \dots$

The results of processing simulated speech with the LMS preprocessor are shown in Figure 4. The input and LMS output SIR values in dB are plotted for approximately 2400 samples of simulated speech. Power values σ_s^2 and σ_z^2 were scaled such that the input SIR was approximately 6 dB. From Figure 4, we can see that over the entire block, the LMS processed sequence has a higher SIR (i.e., higher correlation with actual speaker) than does the raw input sequence. The difference between the two curves is the gain provided by LMS preprocessing. Over the entire block, the LMS gain is approximately 7 dB.

It should be noted that these results are presented to verify the preceding analytical work and provide a proof of concept of using the LMS preprocessor. It is not intended to give an indication of performance on actual speech. However, if actual speech indeed adheres to the all-pole resonator model of (4-1) then these results should be representative.

Figure 4. Signal - to - Interference Ratio (SIR) for LMS Filter Output (Top Curve) and Original Speech Input (Bottom Curve)

5.0 References

1. B. Widrow, et. al., "Stationary and Nonstationary Learning Characteristics of the LMS Adaptive Filter," Proc. IEEE, Vol. 64, No. 8, Aug. 1976.
2. L. J. Griffiths, "Rapid Measurement of Digital Instantaneous Frequency," IEEE Trans. Acous., Speech, and Sig. Proc., vol. ASSP-23, pp. 207-222, April 1975.
3. B. Widrow, et al., "Adaptive Antenna Systems," Proc. IEEE, vol. 55, pp. 2143-2159, Dec. 1967.
4. S. T. Alexander and S. A. Rajala, "An Analysis and Simulation of an Adaptive Image Coding System Using the LMS Algorithm," IEEE Int. Conf. on Acous., Speech, and Sig. Proc., Paris, France, May 1982.
5. M. R. Sambur, "Adaptive Noise Canceling for Speech Signals," IEEE Trans. on Acous., Speech and Signal Processing, vol. ASSP-26, No. 5, pp. 419-423, October 1978.
6. J. B. O'Neal and R. W. Stroh, "Differential PCM for Speech and Data Signals," IEEE Trans. Commun., vol. COM-25, pp. 900-912, Oct. 1972.
7. J. D. Gibson, "Adaptive Prediction in Speech Differential Encoding Systems," Proc. IEEE, vol. 68, No. 4, pp. 486-525, April 1980.
8. J. R. Treichler, "Response of the Adaptive Line Enhancer to Chirped and Doppler-Shifted Sinusoids," IEEE Trans. Acous., Speech and Sig. Proc., vol. ASSP-28, no. 3, pp. 343-349, June 1980.
9. L. R. Rabiner and R. W. Schafer, Digital Processing of Speech Signals, Prentice-Hall, Englewood Cliffs, NJ, 1978.
10. N. J. Bershad, et al., "Tracking Characteristics of the LMS Adaptive Line Enhancer Response to a Linear Chirp Signal in Noise," IEEE Trans. Acous., Speech and Sig. Proc., vol. ASSP-28, No. 5, pp. 504-517 October 1980.