UUCS-79-114

SUPPRESSION OF ACOUSTIC NOISE IN SPEECH USING TWO MICROPHONE ADAPTIVE NOISE CANCELLATION

Steven F. Boll

Dennis Pulsipher

TABLE OF CONTENTS

Page

I. DD Form 1473

II. List of Figures iii

III. REPORT SUMMARY

Section I Summary of Program for Reporting Period

IV. RESEARCH ACTIVITIES

Suppression of Acoustic Noise in Speech Using two Microphone Adaptive Noise Cancellation 3

(Steven F. Boll, Dennis Pulsipher)

I. Abstract 3

II. Two Microphone Signal Generation Model . . . 5

III. The Wiener Solution for Acoustic Noise Suppression Using two Inputs . . . 6

IV. Matrix Form of Wiener-Hoff Equation and the Gradient Vection 9

V. Interative Methods for Estimating the Optimal Weight Vector 12

A. Method of Steepest Decent 12

B. Newton's Method 14

VI. Interative Solutions Based on Approximations to the Ensemble Averages 15

A. The LMS Algorithm 15

B. Approximate Newton's Method 17

C. Orthoganalization Using The Lattice Structure 19

D. Adaptive Lattice Algorithm 21

E. Adaptive Noise Filter In Orthoganal Coordinates 23

TABLE OF CONTENTS

		Page
VII.	Experiments and Results	24
	A. Introduction	24
	B. Data Base Generation	25
	C. Digital Simulation	26
	D. Results Using the LMS Algorithm On the Acoustic Data	29
	E. Results Using teh Adaptive Lattice Algorithm	31
VIII.	Conclusions	33
	A. Comparison of Methods	33
	B. Summary	35
	References	36
	Figures	38
	Approaches Towards a Linear Narrow Band Digital Voice Analysis/Synthesis Algorithm	51
	(H. Ravindra)	
	Abstract	52
I.	Introduction	53
II.	Articulation Rate Change Techniques	55
	Theory	55
	Implementation and Results	58
III.	Analytic Continuation of Band Limited Signals or Signal Extrapolation	60
IV.	2-D AGC Techniques	62
V.	Future Work	63
VI.	References	64

LIST OF FIGURES

Figure	Page
Two Microphone Signal Generation Model	38
General Noise Filtering Model	39
Adaptive Noise Cancelling Lattice Filter Algorithm	40
Noise Power Reduction versus Processing Time for the LMS Algorithm for Misadjustments of 10%, 5%, and 1%	41
Short Time Spectrum of the Unprocessed Speech "The pip began to"	42
Short Time Spectrum of the LMS Algorithm Output	43
Noise Shaping Filter, W Estimated by the LMS Algorithm	44
Noise Power Reduction Versus Processing Time for the Adaptive Lattice Algorithm for Misadjustments of 10%, 3.3%, and 1%	45
Short Time Spectrum of the Lattice Gradient Output	46
Noise Shaping Filter H Estimated by the Adaptive Lattice Algorithm in the Orthoganal Basis	47
K-Parameter Estimates for the Lattice Filter	48
Average Backward Prediction Energies for the Lattice Filter	49
Noise Shaping Filter, W Estimated by the Lattice Gradient Algorithm	50

Section I

Summary of Program for Reporting Period

Program Objectives

To develop practical, low cost, real-time methods for suppressing noise which has been acoustically added to speech.

To demonstrate that through the incorporation of the noise suppression methods,speech can be effectively analysed for narrow band digital transmission in practical operating environments.

Summary of Tasks and Results

Introduction

This semi-annual technical report describes the current status in the research areas for the period 1 April 1979 through 30 September 1979.

SUPPRESSION OF ACOUSTIC NOISE IN SPEECH USING TWO MICROPHONE ADAPTIVE NOISE CANCELLATION

Steven F. Boll
Dennis Pulsipher

ABSTRACT

Acoustic noise with energy greater or equal to the speech is suppressed by filtering a separately recorded correlated noise signal and subtracting it from the speech waveform. This approach was investigated to determine the degree of noise suppression possible using an external correlated input. The second reference noise signal is adaptively filtered using the least mean squares, LMS and the lattice gradient algorithms. These two approaches are developed and compared in terms of degree of noise power reduction, algorithm convergence time, and degree of speech enhancement. Both methods were shown to reduce ambient noise power by at least 20dB with minimal speech distortion and thus to be potentially powerful as noise suppression preprocessors for voice communication in severe noise environments.

I. INTRODUCTION

It has been shown that there is a significant reduction in measured speech intelligibility and quality due to the ambient background noise generated in many operating environments [1], [2]. A number of single microphone approaches for reducing the background noise added to speech have been developed [3], [4]. These methods become ineffective when the noise power is equal to or greater than the signal power or when the noise characteristics, e.g. mean, variance etc., change rapidly in time. This paper studies the performance of an approach to noise suppression in which a second correlated noise source is recorded and used to reduce the noise added to the speech. This second noise source is adaptively filtered to minimize the output power between the two microphone signals. This approach generates an output which is the least squares estimate of the speech waveform. Two adaptive algorithms used to filter the correlated noise are investigated; the LMS approach, [5], [6] and the lattice gradient approach [7], [8], [9]. Both methods approximate the least squares, Wiener solution. The LMS algorithm uses the method of steepest descent and approximates the ensemble gradient with the instantaneous gradient. The lattice gradient algorithm uses Newton's

method in an orthogonal basis generated by the lattice filter.

A severely degraded noisy speech signal was recorded for testing the performance of each method. The ambient energy noise was amplified to mask the recorded speech signal. Both approaches are compared in terms of degree of noise power reduction, algorithm settling time, degree of speech enhancement, and computational complexity.

The paper is divided into sections which develop the two adaptive least squares estimators, describe the experiments conducted, and demonstrate the algorithm performance.

II. TWO MICROPHONE SIGNAL GENERATION MODEL

The noise suppression experiments were based on the model shown in Figure 1. The primary signal x(j) consists of the common noise signal n(j) filtered through a transmission channel $G_1(z)$ and added to the speech signal s(j) plus another independent noise signal $m_1(j)$. The reference signal v(j) consists of the common noise signal n(j) filtered through a transmission channel $G_2(z)$, added to

a second independent noise signal $m_2(j)$. The signals $s(j)$, $n(j)$, $m_1(j)$ and $m_2(j)$ are assumed independent of each other and $G_1(z)$ and $G_2(z)$ are assumed constant in time.

III. THE WIENER SOLUTION FOR ACOUSTIC NOISE SUPPRESSION USING TWO INPUTS

A general filtering model used to suppress the noise component in $x(j)$ which is correlated with $v(j)$ is shown in Figure 2. As is discussed in [5], minimizing the output power of $e(j)$, minimizes the output noise power, and results in $e(j)$ being the minimum mean square estimate of $s(j)$.

The tap weights of the all zero filter $W(z)$ are computed to minimize the total expected output power. Using the orthogonal projection theorem [10], $E[e^2(j)]$ will be minimized when

$$E[e(j+k)v(j)] = 0 \quad \text{for all } k.$$

where

$$e(j) = x(j) - \sum_{i=-\infty}^{\infty} w(i)v(j-i)$$

This orthoganality relation results in the Wiener-Hopf equation:

$$i \sum_{-\infty}^{\infty} w(i) R_{vv}(k-i) = R_{xv}(k) \qquad \text{for all } k$$

where

$$R_{vv}(k) = E[v(j+k)\,v(j)]$$

$$R_{xv}(k) = E[v(j+k)\,x(j)]$$

The z transform, W(z) of the Wiener filter is given by:

$$W(z) = \frac{P_{xv}(z)}{P_{vv}(z)}$$

where

$$P_{xv}(z) = \sum_{k=-\infty}^{\infty} R_{xv}(k)\,z^{-k}$$

$$P_{vv}(z) = \sum_{k=-\infty}^{\infty} R_{vv}(k)\,z^{-k}$$

For the signal model shown in Figure 1, the Wiener filter reduces to

$$W(z) = \frac{P_{nn}(z)G_1(z)\ G_2^*(z)}{P_{m_2m_2}(z) + P_{nn}(z)|G_2(z)|^2}$$

where $P_{nn}(z)$ and $P_{m_2m_2}(z)$ are the power spectra of $n(j)$ and $m_2(j)$ respectively.

The output power spectrum $P_{ee}(z)$ using the Wiener filter is given by:

$$P_{ee}(z) = P_{ss}(z) + P_{m_1m_1}(z) + P_{nn}(z)|G_1(z)|^2\left[1- \frac{P_{nn}(z)|G_2(z)|^2}{P_{m_2m_2}+|G_2(z)|^2P_{nn}(z)}\right]$$

From the Wiener filter equation, note that if $P_{m_2m_2}(z)$ is small compared to $P_{nn}(z)|G_2(z)|^2$ then

$$W(z) = G_1(z)G_2^{-1}(z)$$

This is exactly the linear system required to transform v(j) into the correlated noise component which was added to s(j). Also under this condition:

$$P_{ee}(z) = P_{ss}(z) + P_{m_1 m_1}(z)$$

Thus if the independent noise sources $m_1(j)$ and $m_2(j)$ are negligible with respect to the signal s(j), and the common noise signal n(j), the output signal e(j) will match to s(j) in the mean squared sense.

IV. MATRIX FORM OF WIENER-HOPF EQUATION AND THE GRADIENT VECTOR

Define the reference signal vector V(j) as

$$V(j) = [v(j-1)\ v(j-2) \ldots v(j-N)]^T$$

and the filter weight vector as

$$W = [w_1\ w_2\ \ldots w_N]^T$$

The noise cancelled output e(j) is given in vector form as:

$$e(j) = x(j-1) - W^T V(j) = x(j-1) - V^T(j) W$$

The mean square error is given by:

$$\xi = E[e^2(j)] = E[x^2(j-1)] - 2P^T W + W^T R W$$

where

$$P = E\{x(j-1)V(j)\} = E\begin{bmatrix} x(j-1)v(j-1) \\ x(j-1)v(j-2) \\ \vdots \\ x(j-1)v(j-N) \end{bmatrix}$$

$$R = E\{V(j)V^T(j)\} = \begin{bmatrix} v^2(j-1) & v(j-1)v(j-2) & \cdots & v(j-1)v(j-N) \\ v(j-2)v(j-1) & v^2(j-2) & \cdots & v(j-2)v(j-N) \\ \vdots & \vdots & \cdots & \vdots \\ v(j-N)v(j-1) & v(j-N)v(j-2) & \cdots & v^2(j-N) \end{bmatrix}$$

The optimal weight vector, W^* orthogonalizes the error, e(j) with respect to the reference signal vector V(j), thus the optimate estimate of W satisfies:

$$E[V(j)e(j)] = 0$$

then

$$W^* E[V(j)V^T(j)] = E[x(j-1)V(j)]$$

or

$$W^* = R^{-1}P$$

The optimal weight vector can also be derived by first calculating the gradient of the mean square error surface and using the value which forces it to zero. Thus define:

$$\nabla = \frac{\partial \xi}{\partial W} = -2P + 2RW$$

then $\nabla = 0$ when $W = R^{-1}P$

The minimum mean square error is given by:

$$\xi_{min} = E[x^2(j-1)] - P^T W$$

The mean squared error ξ can also be expressed as:

$$\xi = \xi_{min} + \Delta W^T R \Delta W$$

Where $\Delta W = W - W^*$

Taking the gradient ∇ of ξ with respect to W gives an alternative expression for ∇ as

$$\nabla = 2R \Delta W$$

V. ITERATIVE METHODS FOR ESTIMATING THE OPTIMAL WEIGHT VECTOR

A. Method of Steepest Decent [5], [6]

Let W(j) denote an estimate of the optimal weight vector at time index j. The gradient ∇ (j) evaluated at W(j) points in the direction of greatest rate of increase in ξ. In the method of steepest descent a new estimate of W(j) equals the old estimate plus a term proportional to the negative of the gradient.

Thus

$$W(j+1) = W(j) + \mu\,(-\nabla(j))$$

where μ is a positive constant called the step size.

Subtracting W^* from both sides gives

$$\Delta W(j+1) = (I-\mu R)\Delta W(j)$$

or

$$\Delta W(j) = (I-\mu R)^j \Delta W(0)$$

By diagonalizing R using the eigenvector decomposition [6] it is shown that for convergence it is necessary that:

$$\frac{1}{\lambda_{max}} > \mu > 0$$

and that the convergence rate time constant for the pth tap weight is given by

$$\tau_p \simeq \frac{1}{2\mu\lambda_p}$$

where λ_p is the pth eigenvalue of R.

B. Newton's Method

In the scalar case the root of a function f(x) is iteratively estimated according to the rule:

$$x(j) = x(j-1) - \frac{f(x(j-1))}{f'(x(j-1))}$$

In noise cancellation, the weight vector is updated to force the gradient to zero. Thus the gradient has the role of the function f. The derivative of gradient is given by:

$$\frac{\partial \nabla}{\partial w} = 2R$$

The weight vector is updated as:

$$W(j+1) = W(j) - \frac{R^{-1}\nabla(j)}{2}$$

Substituting for the gradient gives:

$$W(j+1) = W(j) - \frac{R^{-1}}{2}(-2P+2RW(j))$$

or

$$W(j+1) = R^{-1}P = W^{*}$$

Thus Newton's method converges in one iteration. This approach is also referred to as the fast start-up equalizer [8], [11].

VI. ITERATIVE SOLUTIONS BASED ON APPROXIMATIONS TO THE ENSEMBLE AVERAGES

A. The LMS Algorithm [5], [6]

In the LMS Algorithm the method of steepest descent is used with the ensemble gradient approximated by the instantaneous gradient given by:

$$\hat{\nabla}(j) = \frac{\partial e^{2}(j)}{\partial W} = -2e(j)V(j)$$

The LMS algorithm is given as:

$$W(j+1) = W(j) + 2\mu\, e(j)V(j)$$

It can shown [5], [6] that the expected value of the LMS weight vector converges to the Wiener solution. Using the instantaneous gradient introduces an error called gradient noise which results in an excess mean squared error over that obtainable with the Wiener solution. A figure of merit for the estimation process is defined as misadjustment. It is equal to the average excess mean squared error divided by the minimum mean squared error. It can be shown [6] that misadjustment is equal to:

$$M = \frac{1}{2} \sum_{i=1}^{N} \frac{1}{\tau_i} = \mu \mathrm{tr} R, \text{ tr equals trace}$$

As is discussed in the section of results, misadjustment is an important design factor in noise suppression since large misadjustment manifest itself as a pronounced echo in the speech waveform. The echo is removed by reducing the step size, and thus the misadjustment. This reduction of course conflicts with the requirement of quick

settling time for the algorithm which can be shortened by having a large step size. The trade-off between misadjustment and settling time are discussed in the results section.

B. Approximate Newton's Method

Newton's method can be approximated just as the steepest descent method by replacing the ensemble gradient with the instantaneous gradient. The Newton's method approximation would then be:

$$W(j+1) = W(j) + \sigma R^{-1} e(j) V(j)$$

where σ is the normalized step size.

Ignoring for the moment how the autocorrelation matrix, R would be calculated and inverted, this approach offers certain advantages over the method of steepest descent.

The convergence properties of this approach can be estimated using the same approach as with the method of steepest descent. Specifically replacing the noisy gradient with the true gradient and subtracting the optimal weight vector from both sides of the above expression gives:

$$\Delta W(j+1) = (1-\sigma)\ \Delta W(j)$$

Thus for convergence it is necessary that:

$$2 > \sigma > 0$$

The adaptation time constant would be the same for each tap weight and be appropriately equal to:

$$\tau \simeq \frac{1}{\sigma}$$

Using a diagonalization analysis similar to that used for the LMS algorithm, the misadjustment due to gradient noise can be shown to be approximately equal to:

$$M \simeq \frac{N\sigma}{2-\sigma}$$

This approach to tap weight estimation has the advantage over LMS that all tap weights have essentially the same adaptation time constant, but the disadvantage that the gradient estimate must be multiplied by the inverse of R at

each iteration.

The next sections discusses how the tap weights can be estimated using an orthogonal basis which has an auto-correlation matrix that is diagonal. With this diagonalization the number of operations per update is linear with respect to the number of tap weights.

C. Orthoganalization Using the Lattice Structure [7], [8], [9]

To generate an orthogonal basis for estimating the tap weights requires a transformation to map the reference signal $\{v(j-m)\}$ into an orthogonal signal $\{g_m(j)\}$ where:

$$E\{g_m(j)g_k(j)\} = \beta_k \delta_{mk}$$

The lattice structure provides this transformation.

The mathematics describing this orthogonalization for stationary reference signals is well developed in the theory of linear prediction of speech [12], [13]. The $\{g_m(j)\}$ basis called the backward prediction error can be generated recursively using the lattice filter structure:

$$g_m(j+1) = k_m f_{m-1}(j) + g_{m-1}(j)$$

$$f_m(j) = f_{m-1}(j) + k_m g_{m-1}(j)$$

$$m = 1,2, \ldots,N$$

where:

$$f_o(j) = v(j)$$

$$g_o(j) = v(j-1)$$

The sequence $\{f_m(j)\}$ is called the forward prediction error and the sequence $\{k_m\}$ are known as either k-parameters, reflection coefficients or PARCOR parameters. It can be shown that if k_m is estimated to minimize the forward prediction energy $\alpha(m)$ at stage m, where:

$$\alpha(m) = E\left[f_m^2(j)\right]$$

then the backward prediction sequence will be orthogonal:

$$E\{g_m(j)g_k(j)\} = \beta_k \delta_{mk}$$

where:

$$\beta_m = E\left[g_m^2(j)\right]$$

D. Adaptive Lattice Algorithm [7], [8], [9]

To generate the orthogonal basis needed to estimate the Wiener noise cancelling filter with Newton's Method, the k-parameters must first be estimated to minimize the forward prediction error energy. These k-parameters can then be used to generate the backward prediction sequence using the lattice structure. Since estimating the k-parameters is just another least squares problem, Newton's method is used here too.

The derivative of the forward prediction energy is given by, (i.e.the instantaneous gradient in orthogonal basis):

$$\frac{\partial f_m^2(j)}{\partial k_m} = 2 f_m(j) g_{m-1}(j)$$

The adaptive lattice algorithm is then defined as:

$$k_m(j+1) = k_m(j) - \frac{\sigma f_m(j) g_{m-1}(j)}{\beta_{m-1}(j)}$$

where:

$$\beta_m(j+1) = (1-\sigma)\beta_m(j) + \sigma g_m^2(j)$$

σ is the normalized step size. It is determined by the degree of relative misadjustment desired. The signals $f_m(j)$ and $g_m(j)$ are generated from the lattice filter. The vector $\beta_m(j)$ represents a single pole filtered estimate of the average backward prediction error energy:

$$\beta_m(j) \simeq E\{g_m^2(j)\}$$

Dividing by $\beta_m(j)$ in the orthogonal basis $\{g_m(j)\}$ is equivalent to multiplying by R^{-1} in the $\{v(m-j)\}$ basis. The vectors $\{g_m(j)\}$ will approach orthogonality only in steady state.

E. Adaptive Noise Filter in Orthoganal Coordinates

Define H as the N x 1 noise filter vector to be estimated in the $\{g_m(j)\}$ basis. The output error at the mth stage is given by:

$$s_m(j) = s_{m-1}(j) - h_m(j) g_{m-1}(j)$$

where: $s_o(j) = x(j-1)$

Taking the derivative of the prediction error gives the expression for the instantaneous gradient at the mth stage:

$$\frac{\partial s_m^2(j)}{\partial h_m} = -2 s_m(j) g_{m-1}(j)$$

The adaptive algorithm is then defined as:

$$h_m(j+1) = h_m(j) + \frac{\sigma s_m(j) g_{m-1}(j)}{\beta_{m-1}(j)} \quad m = 1,2,\ldots N$$

The adaptive filter estimate of the speech waveform is equal to $s_N(j)$.

Figure 3 shows the composite adaptive noise cancelling lattice filter algorithm.

VII. EXPERIMENTS AND RESULTS

A. Introduction

A controlled data base was generated and a series of experiments were conducted to determine the performance of the two adaptive estimation methods for removing noise from speech. A two input signal data base was recorded with a high degree of control over critical environment factors. The expected performance of the algorithm was then predicted using a digital simulation of the acoustic environment. The performance of the LMS and adaptive lattice methods were measured in terms of degree of noise power reduction, algorithm settling time and amount of echo present. These results are summarized as well as the advantages and limitations of each approach.

B. Data Base Generation

When the noise added to the speech at the primary input differs from the noise at the reference input by a single linear stationary system, $G_1(z)$ the adaptive filter will converge to this linear system and complete noise cancellation results [14]. Referring back to the Wiener solution development given in Section III, this type of experiment would correspond to a situation where the added independent noise sources, $m_1(j)$ and $m_2(j)$ are absent, and $G_2(z) = 1$. Since the intent of this paper is to investigate the degree of noise suppression possible using an external correlated input, it was decided to construct a recording environment as close as possible to above ideal situation.

An acoustically shielded hard-walled room having an ambient noise level of approximately 26dB SPL was used for recording the signals. The room contains audio recording and playback equipment, a computer terminal, and connections to the stereo analog to digital and digital to analog converters. The acoustic shielding prevented independent noise (modeled as $m_1(j)$ and $m_2(j)$) from interfering in the estimation process.

A stationary white noise source was recorded from an analog noise generator onto audio tape. The acoustic noise was generated by playing the audio tape out through a loud speaker into the room. The reference signal microphone was placed next to the speaker, while the primary microphone was placed twelve feet away next to the control terminal. The speaker spoke into the primary microphone while controlling the stereo recording program. The noise power was adjusted to such a level that the recorded speech was completely masked. The signals were filtered at 3.2kHz, sampled at 6.67kHz, and quantized to fifteen bits. Recordings were made with and without speech present, each lasting 23.4 sec.

C. Digital Simulation

Before processing the acoustically recorded data described above, a digital, nonacoustic, simulation was conducted. Two estimates of the rooms impulse response were available from a previous experiment [15]. In this experiment each impulse response was estimated empirically by playing an electrical pulse through the loud speaker. The acoustic response of this pulse was then recorded by two microphones placed eight feet from the speaker. The two

microphone signals were digitized and stored on disk. Each of the measured room impulse responses was digitally convolved with the digitized white noise source to form the primary and reference inputs. When these signals were processed through the LMS algorithm having a step size corresponding to a misadjustment setting of 1%, using 3000 tap weights, the noise power at the output was reduced by 12dB after 23.4 seconds.

The experiment points out some of the problems to contend with in using the two microphone approach for noise suppression. First since $G_2(z)$ is not an identity, the optimal filter must approximate $G_1(z)/G_2(z)$. A long all-zero filter is required to approximate the poles induced by $G_2^{-2}(z)$. A series of experiments [16] measuring noise power reduction verses filter length showed that 3000 tap weights with a 1% misadjustment setting resulted in 12 dB noise reduction after 23.4 seconds. When only 1000 tap weights were used the noise power was reduced by 6 dB and when 500 tap weights were used the noise power was reduced by only 4 dB. Long filter lengths, in-turn, induce more excess mean squared error and increase misadjustment. The increased misadjustment can be minimized by decreasing the step size, but at the expense of increasing the algorithm's

settling time.

The second problem concerns the non-causality of the estimated filter. There is no guarantee that $G_2(z)$ will be minimum phase and thus a stable estimate of $G_1(z)/G_2(z)$ may be non-causal. Non-causal adaptive filter estimates are easily generated by placing a delay into the primary channel [5]. However more tap weights are then required with the accompanying misadjustment problems described above. Also, the amount of delay depends on the microphone placement with respect to the noise source. In the digital simulation experiment both microphones were placed approximately eight feet from the loudspeaker. The estimated adaptive filter impulse response then required a delay of 1500 points. To minimize this non-causal delay requirement for the acoustic experiment, the reference microphone was moved next to the loudspeaker. As is seen in the section on results, placing the reference microphone close to the noise source removed the non-casual filter effects.

This simulation predicted the potential performance achievable. In fact considerably better performance was measured in the actual acoustic experiments described below.

D. Results Using the LMS Algorithm on the Acoustic Data

The algorithm's performance is measured in terms of the degree of steady-state noise power reduction during non-speech activity, the time it takes to reach this steady state value, (algorithm settling time), and the amount of echo induced when speech is present. The first two factors were measured quantitatively while the third factor was determined from listening tests.

Algorithm settling time can be minimized by choosing a large step size value. This however will increase the echo present in the speech output due to the fact that the output is fed back when estimating the tap weights. A large echo is unacceptable in the noise suppression algorithm. Three experiments were conducted to measure algorithm settling time. The experiments differed by the amount of misadjustment specified.

Step sizes were used corresponding to misadjustments of 1%, 5%, and 10%. Based on the simulation experiment, fifteen hundred tap weights were used for estimating the noise filter. The results of steady-state noise reduction for the LMS algorithm are shown in Figure 4. The results show that the algorithm converges to a steady-state noise

power reduction of -20db in approximately 15 seconds for 10% misadjustment and 21 seconds for 5% misadjustment. At 1% misadjustment the step size was so small that the noise power was reduced by only -10dB before the data ran out.

In listening to the output during speech activity it was judged that at a 10% misadjustment setting an unacceptable amount of echo was present and that a 5% setting the echo was just noticeable. For each misadjustment setting there was significant noise suppression and corresponding speech enhancement. At the 1% misadjustment setting the output had a noise floor which was 10dB higher than the 5% and 10% misadjustment outputs due to slow settling time. To illustrate this noise suppression capability, isometric plots of time versus frequency magnitude spectra of speech with and without noise suppression are shown in figures 5 and 6. The plots were constructed by computing magnitude spectra from 64 half overlapped hanning windowed data sets. Each line represents a 128-point frequency analysis. Time increases from bottom to top and frequency from left to right. Figure 5 corresponds to the unprocessed speech signal "The pipe began to". Figure 6 corresponds to the processed speech signal using a 5% misadjustment step size. This phrase occurs 17.5

seconds after startup. Finally Figure 7 shows the noise shaping filter, W, estimated by the LMS algorithm after processing 23.4 seconds of noise only signal.

E. Results Using the Adaptive Lattice Algorithm

A similar set of experiments were made to measure algorithm settling time and amount of echo present for three representative misadjustment step sizes. In section IV. an approximate expression for misadjustment was given as:

$$M \simeq \frac{N\sigma}{2-\sigma}$$

Step sizes, σ , corresponding to misadjustments of 10%, 3.3% and 1% were used. The convergence characteristics for the algorithm are shown in figure 8.

In listening to the output during speech activity it was judged that at the 10% misadjustment setting the amount of echo present was unacceptable (actually worse than the 10% case for LMS), that at 3.3% the echo was just noticeable (judged equal to the 5% case for LMS) and that at 1% there was so little noise reduction that echo present, if any, was

irrelevant.

For the 10% and 3.3% misadjustment settings there was significant noise suppression during speech activity. Figure 9 shows the time verses frequency magnitude spectrum of the output of adaptive lattice algorithm at the 3.3% misadjustment settings for the same speech phrase.

There are four sets of filter parameters generated by the adaptive lattice algorithm. Figure 10 shows the tap weights H in the orthogonal basis $\{g_m(j)\}$. Figure 11 shows the k-parameters for the lattice filters and figure 12 shows the average backward prediction energies, $\{\beta_m\}$. The β_m's are not strictly manotonically decreasing due to the one-pole digital smoothing. The corresponding tap weights in the reference signal basis can be obtained by multiplying the H vector by the matrix which transforms k-parameters into the linear prediction coefficients. This matrix is defined in [12] and can be generated by the STEPUP procedure given in [12]. Figure 13 shows the tap weights obtained from this transformation. Each of these parameter sets were recorded at the end of the 23.4 second noise only data segment using the 3.3% misadjustment step size. Note that the filters shown in Figures 7 and 13 are quite similar

(differing primarily due to tap weight noise). This is to be expected since they both represent estimates of the Wiener filter in the same basis.

VIII. CONCLUSIONS

A. Comparison of Methods

In terms of noise power reduction and amount of echo present, both approaches can be adjusted to give equivalent results. Using step sizes corresponding to approximately 5% and 3.3% misadjustments, each algorithm converges (noise power down 20dB) after 20 seconds of processing, with a just noticeable amount of echo. The adaptation rates are not significantly different. These equivalent results between the two methods is to be expected since the reference signal is just white noise, colored by the room's acoustics. The averaged backward prediction error energies and the k-parameters are nearly constant after the first one hundred values. Thus for this environment, the normalization offered by the gradient lattice offers little advantage over LMS. For environments with a large ratio between the smallest to largest eigenvalue, the gradient lattice method has been shown to converge faster [9].

The computational price payed for the orthogonalization and normalization is high compared to the LMS approach. The LMS requires 2N multiply-adds per sample while the gradient lattice requires 10N multiplies, 6N adds, and 2N divides per point. For fifteen hundred tap weights at a sampling rate of 6.67 kHz, LMS requires 20 million multiply-adds per second and gradient lattice requires at least 120 million multiply-adds per second to process this data in real time. The enormous computation requirement necessitated implementing both algorithms on an FPS 120-B array processer. These micro-coded implementations resulted in a 30 to 1 speed-up over that achievable on a conventional general purpose DEC-10 processor. Both algorithms of course still did not run in real-time but were processed in a non-real time disk to disk configuration.

In addition the gradient lattice method has the disadvantage that it requires an estimate of the average backward prediction error energy. For this implementation these estimates were obtained by smoothing the squared backward prediction values through a single pole filter. For nonstationary reference signals with a large dynamic range, this smoothing approach may be unable to track the gain variations thus resulting in an unstable adaptive

filter.

B. Summary

This paper addressed the problem of reducing the acoustic noise added to speech by subtracting off an adaptively filtered second correlated noise source. Two adaptive algorithms were developed and their performance characteristics measured using an acoustic signal in which the noise power was equal or greater than the speech power. In both approaches it was shown that significant noise reduction is possible with minimal distortion to the speech waveform.

In summary, though this two-microphone approach to noise suppression requires a second signal and considerable computation, it offers a potentially powerful alternative approach for speech enhancement in severe noise environments.

REFERENCES

[1] C.F. Teacher and D. Coulter, "Performance of LPC Vocoders in a Noisy Environment," in Proc. IEEE Conf. Acoust., Speech, and Signal Processing, Wash. D.C., April 1979, pp. 216-219.

[2] C.F. Teacher, et.al., ANDVT Microphone and Audio System Study, Final Report, NAVELEX N00039-77-C-0365, 1978.

[3] S.F. Boll, "Suppression of Acoustic Noise in Speech Using Spectral Subtraction," IEEE Trans. Acoust., Speech, and Signal Processing, Vol. ASSP-27, No. 2, pp. 113-120, April 1979.

[4] J.S. Lim and A.V. Oppenheim, "Enhancement and Bandwidth Compression of Noisy Speech," accepted for publication Proc. IEEE, July 1979.

[5] B. Widrow, et.al., "Adaptive Noise Cancelling: Principles and Applications," Proc. IEEE, Vol. 63, pp. 1692-1716, Dec. 1975.

[6] B. Widrow, J. McCool, M. Larimore, and C.R. Johnson, Jr., "Stationary and Nonstationary Learning Characteristics of LMS Adaptive Filter, Proc. IEEE, Vol 64, pp. 1151-1162, Aug. 1976.

[7] L. Griffths, "An Adaptive Lattice Structure Used for Noise-Cancelling Applications," in Proc. IEEE Conf. Acoust., Speech and Signal Processing, Tulsa, Ok., April 1978, pp. 87-90.

[8] J. Makhoul, "A Class of All-Zero Lattice Digital Filters: Properties and ApplicationS," IEEE Trans. Acoustics, Speech and Signal Processing, Vol.

ASSP-26, No. 4, pp. 304-314, Aug. 1978.

[9] E.H. Satorius and S.T. Alexander, "Channel Equalization Using Adaptive Lattice Algorithms," IEEE Trans. Communications, Vol. COM-27, No. 6, June 1979, pp. 899-905.

[10] A Papoulis, Probability, Random Variables, and Stochastic Processes, New York, New York, McGraw-Hill, 1965.

[11] R.W. Chang, "A New Equalizer Structure for Fast Start-Up Digital Communication," Bell Syst. Tech. J. Vol. 50, pp. 1969-2014, July-Aug. 1971.

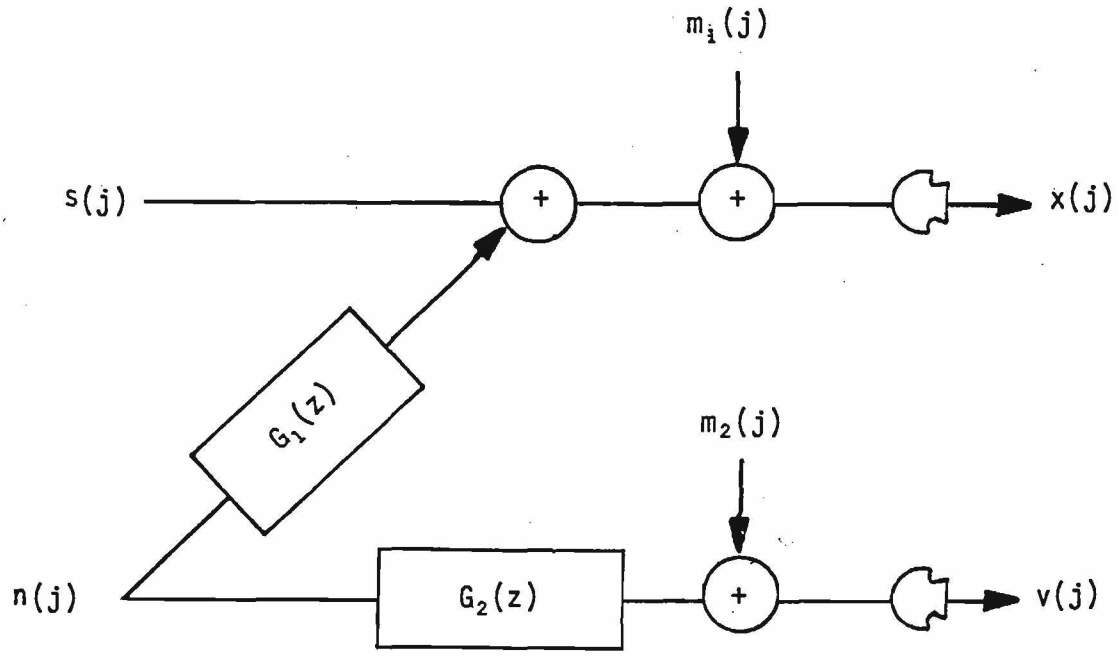
[12] J.D. Markel and A. H. Gray, Linear Prediction of Speech, New York, New York, Springer-Verlag, 1976.

[13] J. Makhoul, "Linear Prediction: A Tutorial Review," Proc. IEEE, Vol. 63, pp. 561-580, April 1975.

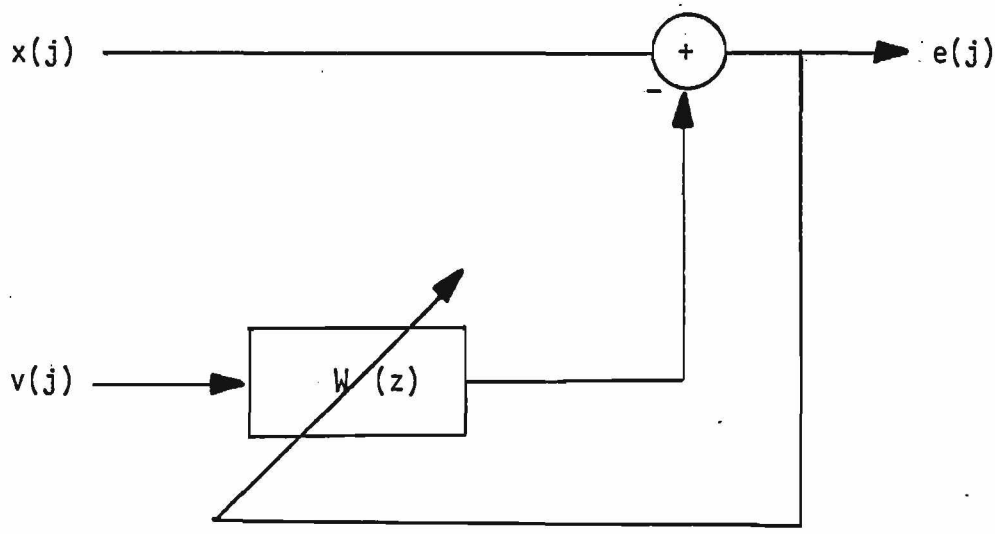
[14] D. Pulsipher, S.F. Boll, C.K. Rushforth, and L.K. Timothy, "Reduction of Nonstationary Acoustic Noise in Speech Using LMS Adaptive Noise Cancelling" in Proc. IEEE Conf. Acoust. Speech and Signal Processing, Wash. D.C. April 1979, pp. 204-208.

[15] S.F. Boll, E. Ferretti, and T. Petersen, "Improving Synthetic Speech Quality Using Binaural Reverberation" in Proc. IEEE Conf. Acoust. Speech and Signal Processing, Phila. PA., April 1976, pp. 705-708

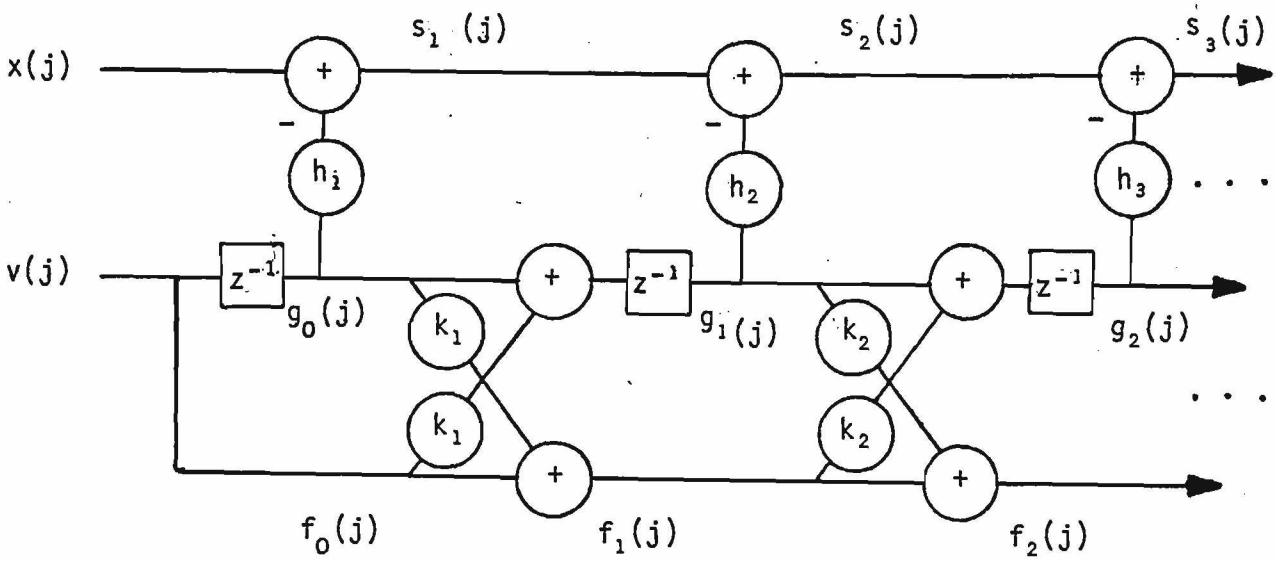
[16] D.C. Pulsipher, Application of Adaptive Noise Cancellation to Noise Reduction in Audio Signals, Ph.D Dissertation, University of Utah, 1979, also Computer Science Dept. Tech Report No. UTEC-CSc-79-022, March 1979.



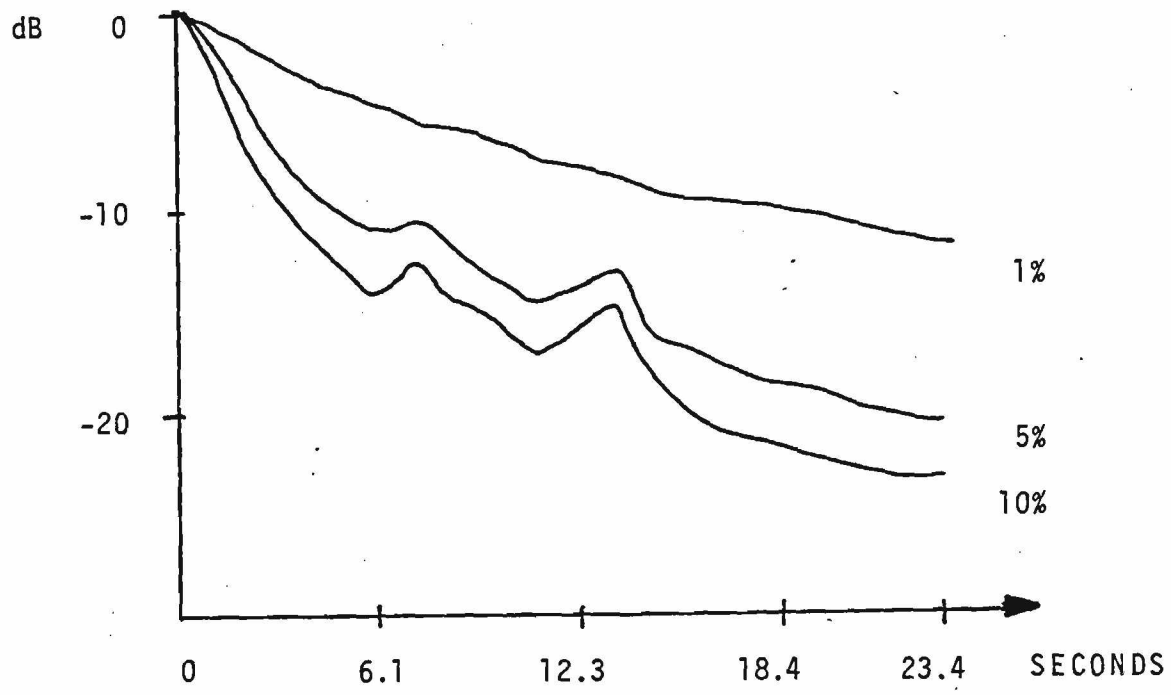
1. Two Microphone Signal Generation Model



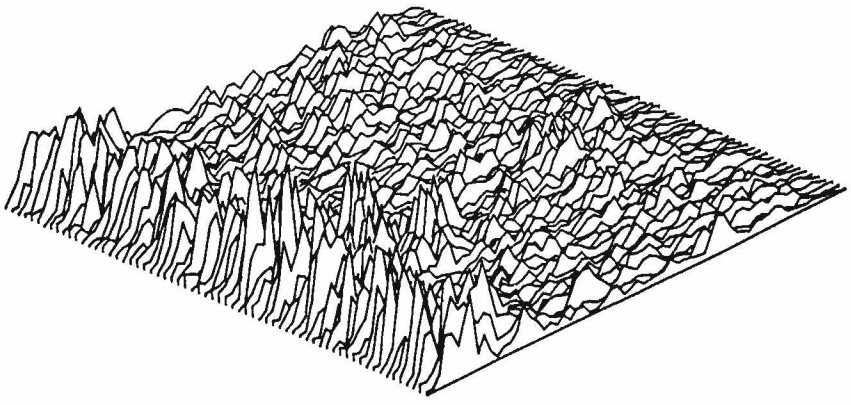
2. General Noise Filtering Model.



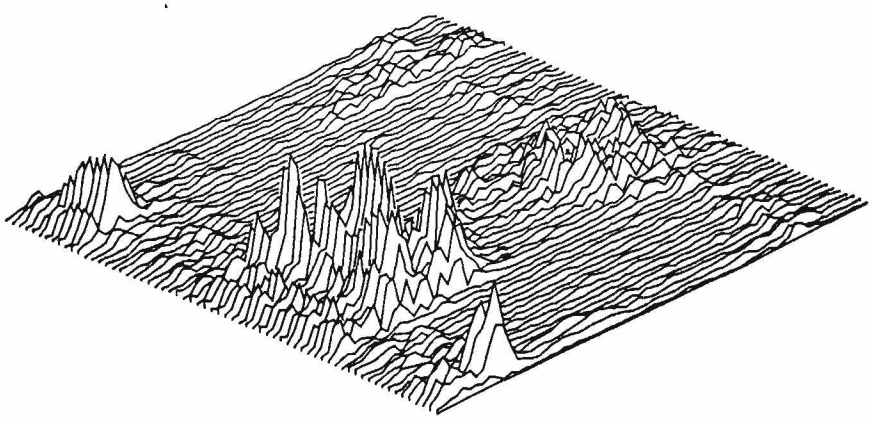
3. Adaptive Noise Cancelling Lattice Filter Algorithm.



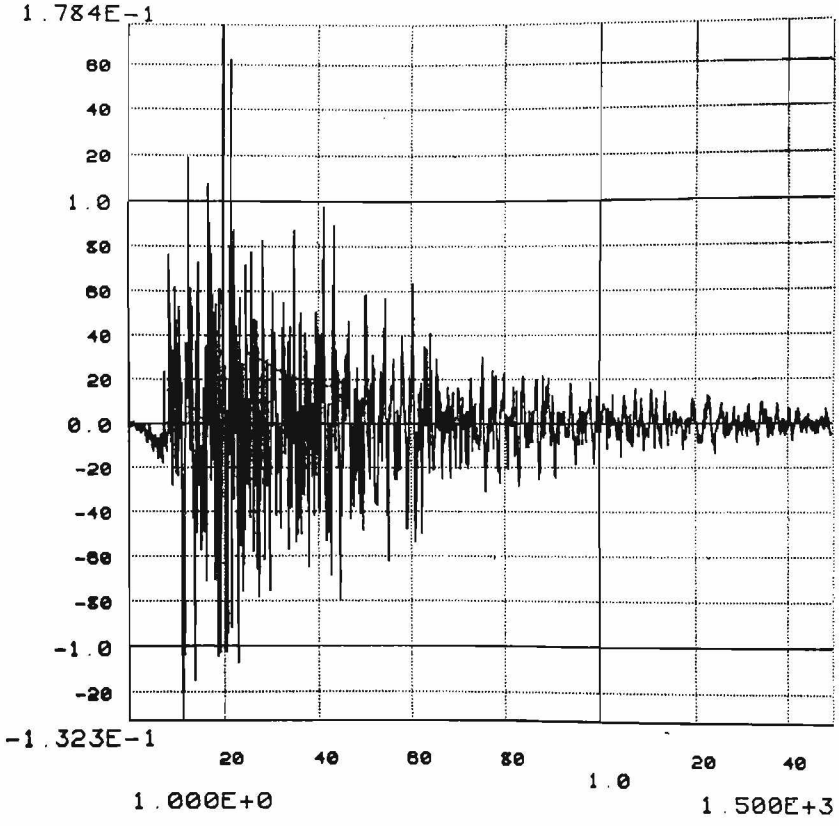
4. Noise Power Reduction versus Processing Time for the LMS Algorithm for Misadjustments of 10%, 5%, and 1%.



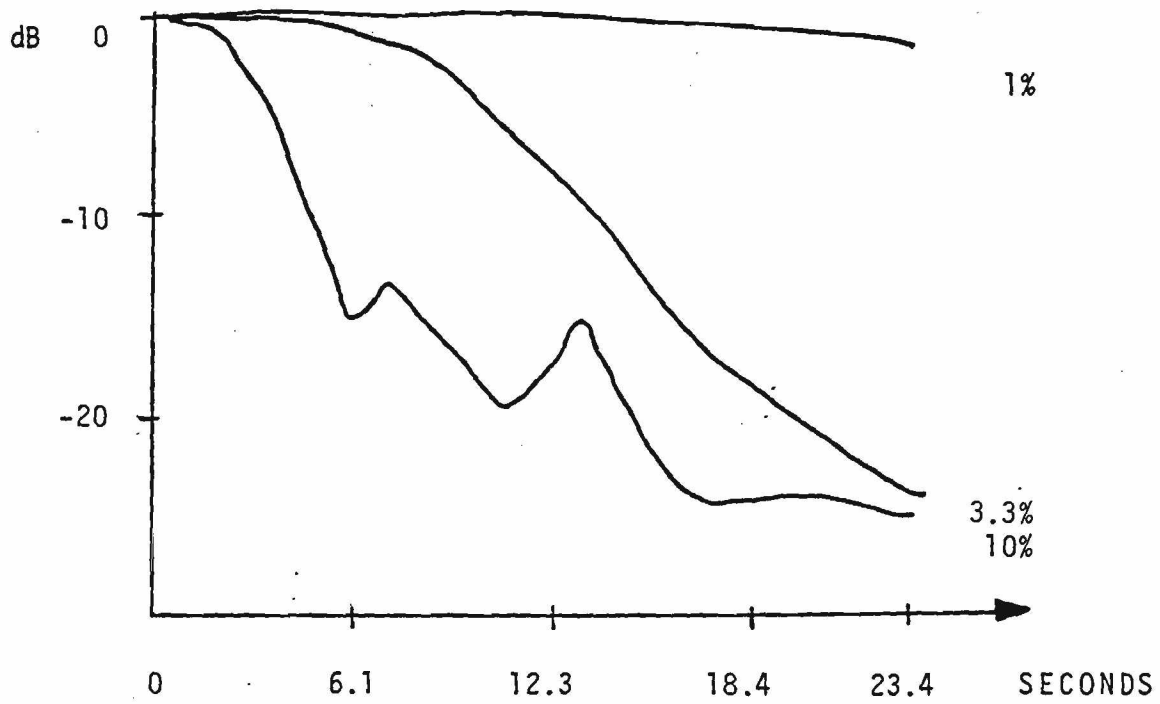
5. Short Time Spectrum of the Unprocessed Speech "The pipe began to"



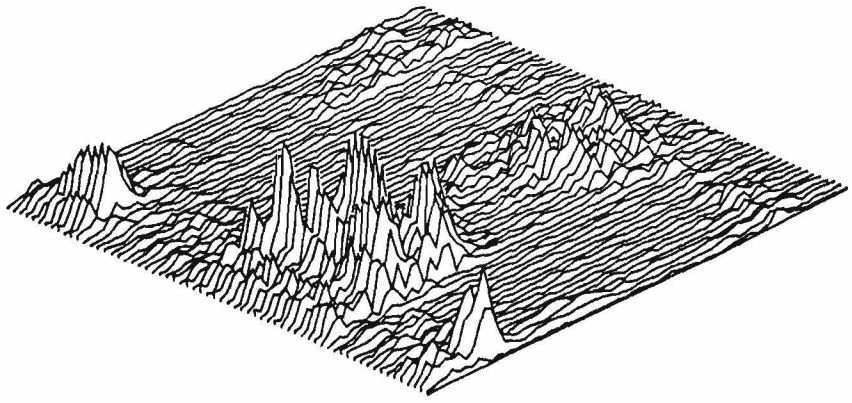
6. Short Time Spectrum of the LMS Algorithm Output.



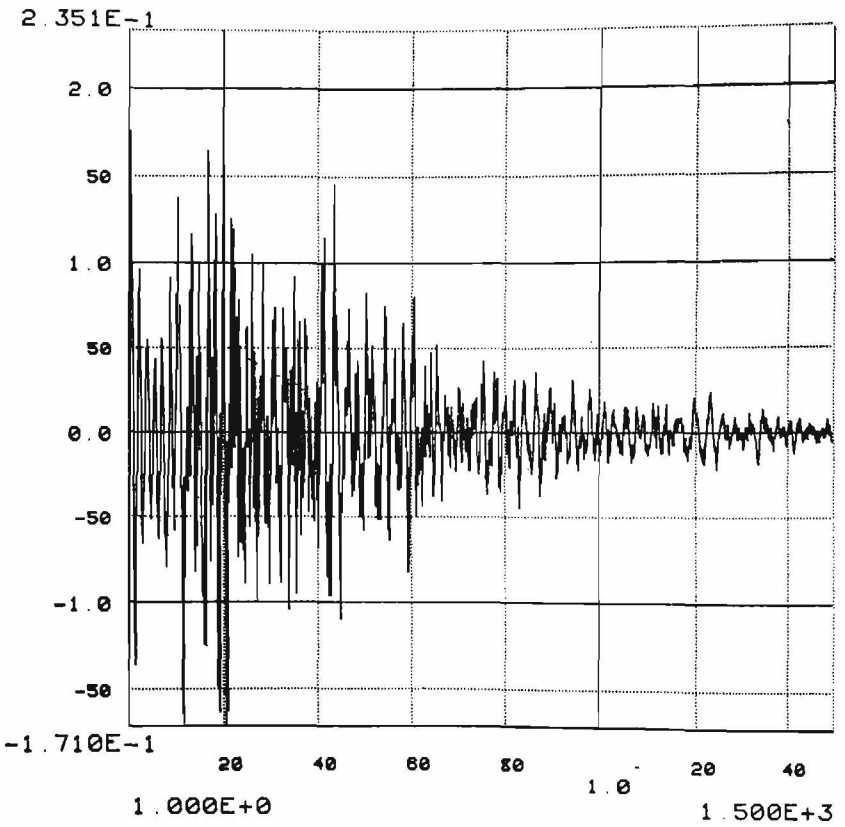
7. Noise Shaping Filter, W Estimated by the LMS Algorithm.



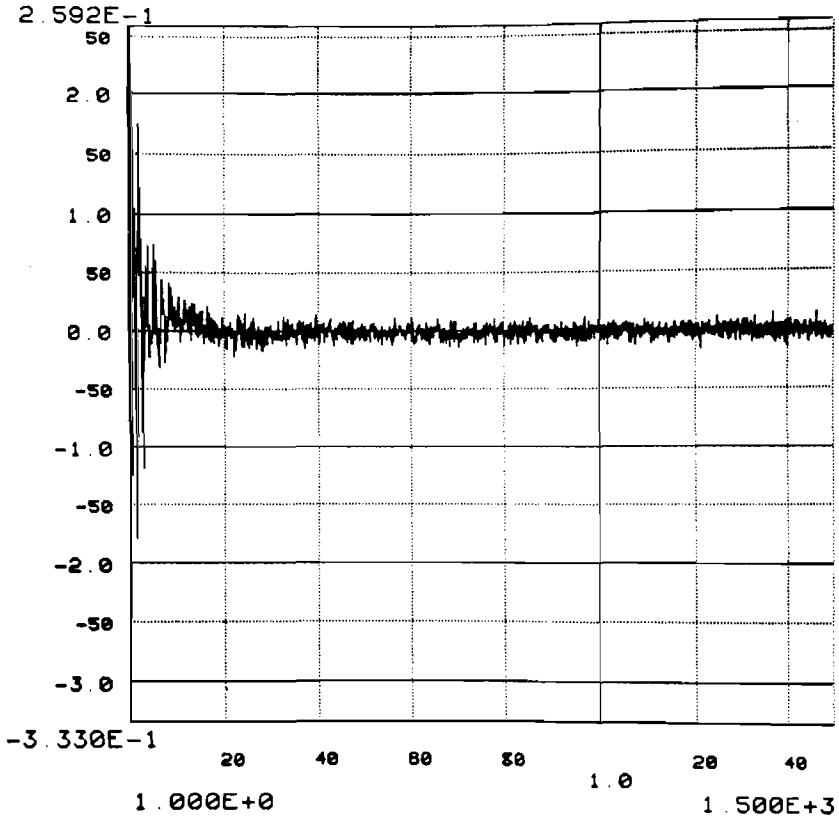
8. Noise Power Reduction Versus Processing time for the Adaptive Lattice Algorithm for Misadjustments of 10%, 3.3%, and 1%.



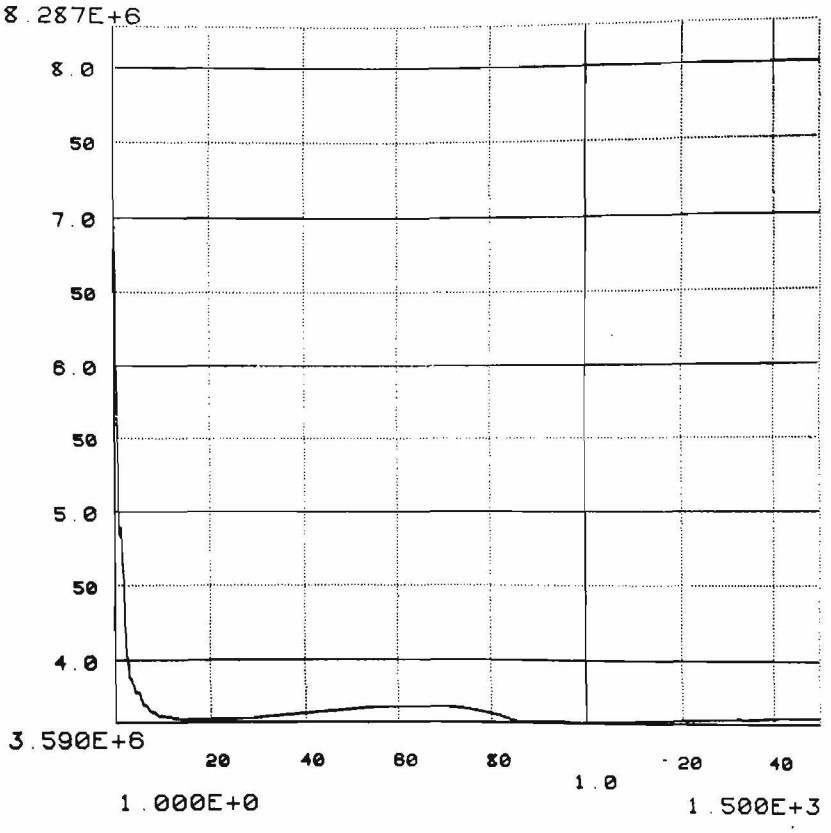
9. Short Time Spectrum of the Lattice Gradient Output.



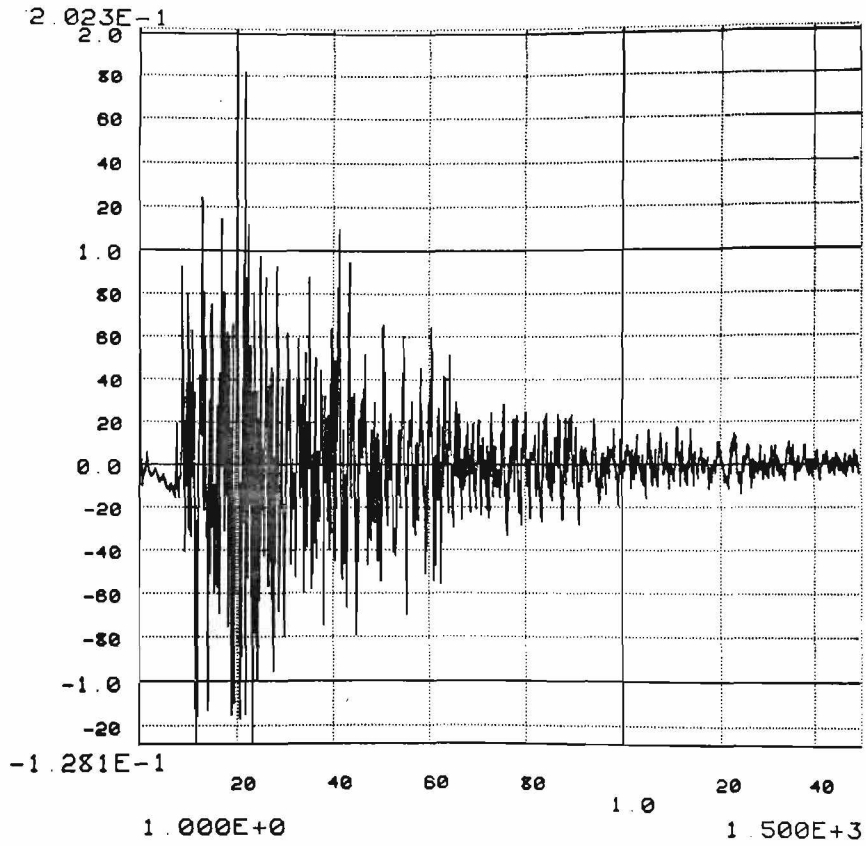
10. Noise Shaping Filter H Estimated by the Adaptive Lattice algorithm in the Orthoganal Basis.



11. K-Parameter Estimates for the Lattice Filter.



12. Average Backward Prediction Energies for the Lattice Filter.



13. Noise Shaping Filter, W Estimated by the Lattice Gradient Algorithm.

APPROACHES TOWARDS A LINEAR NARROW BAND DIGITAL VOICE ANALYSIS/SYNTHESIS ALGORITHM

H. Ravindra

APPROACHES TOWARDS A LINEAR NARROW BAND DIGITAL VOICE ANALYSIS/SYNTHESIS ALGORITHM

H.Ravindra

Abstract

Certain key ideas towards the development of a linear narrow band digital voice analysis/synthesis algorithm which can be used in multiple talker and conferencing environments, are presented. The use of articulation rate change, signal extrapolation (analytic continuation) and 2-D AGC techniques in combination is discussed, problems highlighted and current results presented in some of the areas. This approach does not parameterize speech as most narrow band vocoder algorithms do, but uses data compression ideas on the speech waveform which lends it the property of linearity which makes it suitable for use in conferencing and multiple-talker environments. Also, such a system is expected to degrade gracefully with noise.

I. Introduction

Vocoder algorithms which operate at rates of 2.4 to 4.8 kilobits/sec. are considered narrowband. These vocoders such as LPC vocoder, channel vocoder, etc. parameterize the speech signal, attempting to extract the parameters in such a way that a good fit of the output of the model to the actual signal is obtained. The presence of noise in the speech signal leads to great difficulty in extracting exactly the parameters of the model. Thus, these vocoders degrade drastically with increased noise level. Also, they are not linear because of the parameterization and hence cannot be used in conferencing environments.

This note presents several ideas which in combination point to the possibility of developing of a linear, narrow band voice analysis/synthesis algorithm which possesses a graceful degradation characteristic with noise. Because of linearity, the algorithm satisfies the superposition principle and hence can be used in conferencing environments.

The approach considered consists of the following steps. The speech signal is band limited and transformed into the short time Fourier domain. Two-dimensional automatic gain control (2-D AGC) is then applied which results in a modified speech signal in the time domain. The number of bits required to quantize each sample of this

signal has been shown by Mike Callahan [1] to be less than half of what is required by log PCM techniques, for the same quantization noise levels. In addition, the instantaneous phase and center frequency of each channel in the short time Fourier domain are scaled by a factor less than unity which leads to a reduction in the bandwidth occupied by the resulting signal. Hence, this signal requires a lower sampling rate. The final signal in the time domain is divided into short segments and only one in every few (2 or 3) segments is transmitted. At the receiver, signal extrapolation techniques are applied to recover the segments which were not transmitted, using the segments transmitted. Then, the first two processes are inverted to realize a signal which is a close approximation to the original. Since no parameter extraction is involved and since the coder is of the waveform type, the algorithm will be linear and will exhibit graceful degradation with noise.

The technique for articulation rate change and the problems involved are discussed in Section II. In section III, signal extrapolation techniques are discussed. Section IV briefly summarizes the results of Callahan with the 2-D AGC experiments. Problems which require further research are highlighted in Section V.

II. Articulation Rate Change Techniques

Theory:

The speech signal is analysed into several bands in the frequency domain using a Short-Time Fourier analyzer or a Constant-Q analyzer. The instantaneous phase and center frequency of each channel are both scaled by the same factor and a time domain signal is synthesized from the resulting channel signals. The process defined by these steps leads to the scaling of the bandwidth of the synthesized signal by a factor equal to that used to scale the phase and center frequencies. To be used as a bandwidth compression technique, a scale factor less than unity should be used at the sender and the reciprocal of that factor at the receiver. Some of the fundamental limitations and other problems associated with this approach to bandwidth scaling are discussed below.

The procedure involves dividing a speech signal into several bandpass signals using any of the analysis techniques. The signal in the nth band, $f_n(t)$, can be modeled as a simultaneously amplitude and angle modulated wave (AAM) as below.

$$f_n(t) = q_n(t) \text{ Cos } (\omega_n t + \phi_n(t))$$

where:

$q_n(t)$ is the magnitude signal in the nth channel.

$\phi_n(t)$ is the phase signal in the nth channel and ω_n is the center frequency of the nth channel. The complete signal, f(t) is given by

$f(t) = \sum_{n=o}^{N-1} f_n(t)$ where N is the number of channels.

Kahn and Thomas [2] have studied the bandwidth properties of such signals and they have derived the following equation for the instantaneous bandwidth of the AAM wave:

$$\Omega_{f_n} = \left[\Omega_{q_n}^2 + \frac{||q_n(t)\dot{\phi}_n(t)||^2}{||q_n(t)||^2}\right]$$

where Ω_{f_n} is the second moment bandwidth of the signal in the nth channel.

Ω_{q_n} is the second moment bandwidth of the magnitude signal in the nth channel.

$q_n(t)$ is the magnitude signal in the nth channel.

$\dot{\phi}_n(t)$ is the derivative of the phase signal in the nth channel.

$||.||$ is the norm of the vector in the function space.

Second moment bandwidth of a signal f(t) is defined as $\Omega_f = \left|\left|\frac{\dot{f}}{f}\right|\right|$. Ensemble averages are used for non-deterministic signals. It is clear from the above

expression that scaling $\phi_n(t)$ leads to scaling of the quantity Ω_{f_n}. But the amount by which Ω_{f_n} is scaled depends on the relative energies in the magnitude and phase signals. Linearity results only when $\Omega_{q_n} = 0$, which in general is not true. So, this non-linearity results in incorrect scaling of the bandwidths of the channel signals which can lead to frequency aliasing on bandwidth compression and reverberation on expansion. This we will call the "Kahn-Thomas effect". Thus, it is not only necessary to scale the phase signal but also scale the bandwidth of the magnitude signal in each channel. The approach considered in this research is to apply the bandwidth compression process defined above recursively to each channel magnitude signal. That is, each channel magnitude signal is further analysed into subchannels, each subchannel consisting of a magnitude and phase signal. Scaling the phase of the subchannels leads to the scaling of the magnitude signals at the next higher level. This idea can be carried further down by analyzing the subchannel magnitude signals further into narrower channels and scaling the phase at this level. Of course, the depth of recursion is limited by the difficulty in the implementation of the analysis filters.

A fundamental limitation arises when this type of bandwidth compression is attempted. The technique described attempts to discard redundant information in speech, like extra pitch periods. Since speech is only a quasi-stationary signal, using a large scale factor it

causes excessive loss of information in each assumed stationary section of the signal. On subsequent expansion, the lost information is not recovered. This type of distortion is perceived as loss of voicing in voiced sections of speech.

Implementation and Results:

Three different analysis techniques were used to implement the rate change. In each case, the bandwidth of a speech signal was compressed and re-expanded and the resulting speech compared with the original. The three schemes are described below.

(a) Using a Constant-Q Analyzer:

A Constant-Q filter bank was used in this scheme. A Constant-Q analyzer has a frequency resolution which decreases with increasing frequencies somewhat similar to the resolution properties of the ear, whereas a short-time Fourier analyzer has constant frequency resolution. The distortions were severe for compression factors greater than 2. Also a signal dependent background noise was observed in the processed signal which can be attributed to the Kahn-Thomas effect. Using the recursive approach described on the channel magnitude signals, with a recursion depth of two, it was found that the signal dependent noise was reduced but the distortions due to the fundamental limitation noted above still prevailed.

(b) Using a Filter Bank Made Up of Constant Bandwidth, Sharp Cutoff Filters.

The overlap between adjacent channels was reduced by using this type of filter bank. The basic quality of processed speech remained as in case (a) except for reduced background noise.

(c) Using a Short Time Fourier Analyzer

In this case a narrow band analysis system was used, and essentially similar quality results were obtained. Application of the recursive procedure described earlier to compensate for the Kahn-Thomas effect is under study.

The experiments suggest that this technique can be used, without introducing serious degradation of the signal, with compression factors less than 2.

III. Analytic Continuation of Band Limited Signals or Signal Extrapolation

An analytic signal can be recovered completely given only a section of the signal. This problem can be characterized as follows. Let P_a and P_b be two subspaces of a parent Hilbert Space H. If the projection of a signal $f \epsilon P_b$, on the subspace P_a is given, then under certain conditions, it is possible to realize an inverse operator by recursive techniques. With this operator, the signal f can be recovered from its projection. Papoulis and Gerchberg [3] have independently proposed similar algorithms based on the above formulation. They attempt to obtain the signal f when P_a is the subspace of all signals band limited to a particular frequency and P_a is the subspace of all signals time limited to a particular time interval. Youla [4] has shown that these are special cases of a more general problem of solving operator equations in Hilbert spaces and has derived certain important conclusions. He shows that the problems of Papoulis and Gerchberg are not well posed. He shows and we have found that applying these algorithms to noisy data leads to serious noise amplification problems. Richard Frost [5] has modified the above to come up with a new algorithm which performs the extrapolation by small amounts with each iteration and does not add back the distortion energy at each step (Gerchberg's algorithm does add back the distortion energy at each step). This makes the algorithm stable in the presence of noise. He has

applied it to the restoration of astronomical image data and proved that it has better convergence properties in the presence of noise. Signal extrapolation is expected to be harder in the case of speech signals as no assumptions can be made about the sign of the signal being extrapolated. (This is always non-negative in the case of images.)

In speech application, the algorithm is used to recover the missing signal segment between two successive segments. So, the problem can be characterized as follows: The signal f belongs to the subspace of band limited functions. Its projections onto two mutually orthogonal subspaces are given. These subspaces consist of functions limited over two different intervals of time. The problem is then to find the projection of f onto a third subspace of functions which is orthogonal to both the given subspaces. The two key issues to be addressed are the stability of reconstruction in the presence of noise, and convergence rate. Preliminary experiments with Gerchberg's algorithm seem to indicate that the segments of the speech signal must be very short (much less than a pitch period). Currently, other algorithms based on Richard Frost's step by step extrapolation and the three orthogonal subspaces formulation derived above are under study.

IV. 2-D AGC Techniques

Mike Callahan [1] has developed a AGC technique to be applied in the short-time Fourier domain. Essentially, he models the short-time Fourier Transform $F(\omega,t)$ of a speech signal $f(t)$ as the product of an envelope function $E(\omega,t)$ and a vibratory function $V(\omega,t)$ and notes that E is slowly varying and positive, and V is fast varying and complex.

Then,

$$\log[F(\omega,t)] \approx \log|E(\omega,t)| + \log|V(\omega,t)| + j \arg[V(\omega,t)]$$

So, passing $|\log F(\omega,t)|$ through a high pass filter with a low pass gain of $p<1$ and then undoing the effect of the logarithm leads to a Short-time Fourier transform given by E^pV. The time signal synthesized from E^pV is the original signal with its dynamic range compressed. Callahan has shown that this signal can be quantized with 2 to 3 bits/sample to achieve the same signal/quantization noise ratio as with ordinary 8-bit PCM techniques.

This technique can be applied to achieve reduced bit rate requirements per sample of speech signal independent of the techniques described in the previous sections which attempt to reduce the effective sampling rate. Hence it may be possible to use the compression ideas described in tandem to achieve low bit rates.

V. Future Work

In the area of articulation rate change, the effect of recursive correction for the Kahn-Thomas effect, when using Short-Time Fourier analysis, is to be studied.

Work needs to be done in the area of signal extrapolation to study the performance of various existing algorithms when applied to speech and develop modifications. More research needs to be done to develop new algorithms to suit the speech application. The application of existing one step extrapolation procedures to speech reconstruction is to be studied.

In all the cases, work is required to better condition the problem in the presence of noise even at the cost of imperfect, but acceptable, reconstruction of noise free signals.

VI. References

[1] M.W. Callahan, "Acoustic Signal Processing Based on the Short-Time Spectrum," Ph.D. Thesis, Computer Sci. Dept., Univ. of Utah, Tech. Rept. No. UTEC-CSc-76-209, March 1976.

[2] R.E. Kahn and J.B. Thomas, "Some Bandwidth Properties of Simultaneous Amplitude and Angle Modulation," IEEE Trans. on Inf. Theory, Vol. IT-11, No. 4, pp. 516-520.

[3] R.W. Gerchberg, "Super-resolution through error energy Reduction," Optica Acta, 1974, Vol. 21, No. 9, pp. 709-720.

[4] Richard L. Frost, "High Resolution Astronomical Imaging Through the Turbulent Atmosphere," Ph.D. Thesis, Dept. Elec. Eng., Univ. of Utah, July 1979.

[5] D.C. Youla, "Generalized Image Restoration by the Method of Attenuating Orthogonal Projections," pp. 176-185, Image Science Mathematics.