# Constrained Iterative Speech Enhancement with Application to Automatic Speech Recognition

**S12.9**

John H. L. Hansen and Mark A. Clements
School of Electrical Engineering
Georgia Institute of Technology
Atlanta, GA 30332

## 1 Abstract

A set of iterative speech enhancement techniques employing spectral constraints is extended and evaluated in this paper. The original unconstrained technique attempts to solve for the maximum likelihood estimate of a speech waveform in additive noise. The new approaches (presented in ICASSP-87 [3]), apply inter- and intra-frame spectral constraints to ensure optimum speech quality across all classes of speech. Constraints are applied based on the presence of perceptually important speech characteristics found during the enhancement procedure. Previous results show improvement over past techniques for additive white noise distortions. Three points are addressed in the present study. First, a convenient and consistent terminating point for the iterative technique is presented which was previously unavailable. Second, the techniques have been generalized to allow for slowly varying, colored noise. And finally, a comparative evaluation was performed to determine their usefulness as preprocessors for recognition in extremely noisy environments in the vicinity of 0 dB SNR.

## 2 Introduction

The general problem of automatic speech recognition is one which requires several alternatives to be specified prior to formulation of a solution. The type of speech, restrictions on speakers, vocabulary size, and environment all ultimately affect recognition performance. The specific problem of limited vocabulary, speaker dependent, isolated word recognition has to varying degrees been solved. In the past, approaches such as dynamic time warping or hidden Markov modeling have largely been applied in tranquil environments. Studies have shown that recognition accuracy is severely reduced when speech is uttered in noisy, stressful environments. One alternative is to reformulate previous approaches to the recognition problem assuming a noisy environment. Unfortunately, many systems are LPC based which, from research in speech enhancement and coding are known to deteriorate rapidly in noise. Another alternative, which would be beneficial for recognition as well as speech transmission systems is to develop robust enhancement preprocessors. Such preprocessors would produce speech or recognition features which are less sensitive to background noise so that existing recognition systems may be employed.

The set of speech enhancement algorithms under consideration were previously developed for improving both speech quality and all-pole speech parameter estimation [3,4]. The basis of these algorithms is to form a maximum likelihood estimate of the speech waveform in additive noise with the constraint that the signal be an all-pole process. In section 3, a review of the constrained techniques is presented. A comparative evaluation is presented in section 4 which include; additive white Gaussian noise, and slowly varying colored aircraft interior noise. Finally, the enhancement algorithms are evaluated to determine their ability as preprocessors for automatic recognition in extremely noisy environments.

## 3 Iterative Speech Enhancement

The success of a speech enhancement algorithm is dependent on the objectives made in deriving an approach. Assumptions made in this environment include: i) the noise distortion is additive, ii) only the degraded speech signal is available, and iii) the noise and speech signals are uncorrelated. The basis of the original unconstrained iterative enhancement approach is noncausal Wiener filtering [5]. This approach attempts to solve for the maximum likelihood estimate of a speech waveform in additive white Gaussian noise with the requirement that the signal be the response from an all-pole process. Crucial to the success of this approach is the accuracy of the estimates of the all-pole parameters at each iteration. The algorithm is formulated by considering the case where all unknowns (all-pole speech parameters $\vec{a}$, noise free speech $\vec{S}_O$) are random with a priori Gaussian probability density functions. The basic procedure used is a maximum a posteriori (MAP) estimator, which maximizes the probability density function of the unknown parameters given the noisy observations. After some simplification, it can be shown that the resulting equations for the joint MAP estimate of $\vec{a}$ and $\vec{S}_O$ become nonlinear, involving partial derivatives with respect to $\vec{a}$. Lim and Oppenheim considered a suboptimal solution employing a sequential two step approach based on MAP estimation of $\vec{S}_O$ followed by MAP estimation of $\vec{a}$ given $\hat{\vec{S}}_{O,i}$, where $\hat{\vec{S}}_{O,1}$ is the result of the first estimation. This sequential estimation procedure is linear at each iteration, and continues until some convergence criterion is satisfied. After further simplifying assumptions, it can be shown that the MAP estimation of $\vec{S}_O$ is equivalent to a minimum mean squared error (MMSE) estimate. In addition, as the observation window increases, the procedure for obtaining a MMSE estimate approaches a noncausal Wiener filter.

Although successful in a mathematical sense, this technique has received little application due to several factors. First, the scheme is iterative with sizable computational requirements. Second and most important, is that although the original sequential MAP estimation technique was shown to increase the joint likelihood of the speech waveform and all-pole parameters, a heuristic convergence criterion had to be employed. This is a serious drawback if the approach is to be used in environments requiring automatic speech enhancement. After an extensive investigation [1], this approach was found to produce significant levels of enhancement for white Gaussian noise in 3-4 iterations. Some interesting anomalies were noted



CH2561-9/88/0000-0561 $1.00 © 1988 IEEE

which helped motivate development of the constrained approaches. First, as additional iterations were performed, individual formants of the speech decreased in bandwidth and shifted in location. Second, frame to frame pole jitter was observed across time. Both effects contributed to unnatural sounding speech. The goal therefore was to formulate a new set of enhancement algorithms which impose constraints on pole locations across time (inter-frame) and iterations (intra-frame). Spectral constraints are applied to the all-pole parameters $\hat{\vec{a}}_i$ which ensure that; i) the all-pole speech model is stable, ii) it possess speech-like characteristics (e.g., poles are not too close to the unit circle causing narrow bandwidths), and iii) the vocal tract characteristics do not vary wildly from frame to frame when speech is present. Due to the constraints imposed, improved estimates of $\hat{\vec{a}}_{i+1}$ result. Given this new estimate, the second MAP estimation of $\hat{\vec{S}}_O$ can be carried out. In order to increase numerical accuracy, reduce computational requirements, and eliminate inconsistencies in pole ordering across frames, the line spectral pair (LSP) transformation was used to implement most of the constraint requirements. Figure 1 illustrates the framework for the constrained enhancement algorithms.

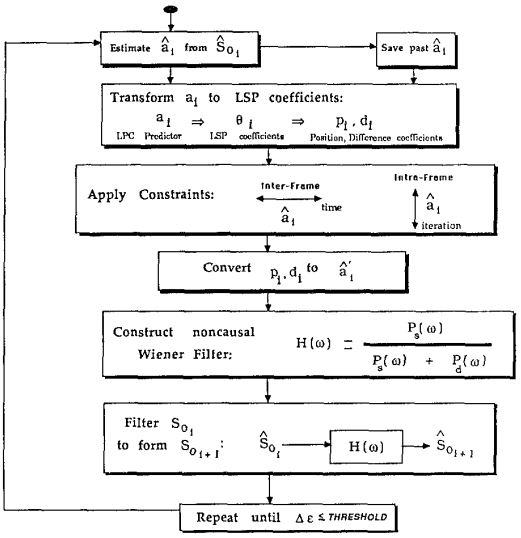


Figure 1: Framework for the constrained iterative enhancement algorithms.

## 4 Evaluation

Speech degraded by additive noise was processed using various configurations of the constrained algorithms. Enhancement algorithms evaluated include: algorithms incorporating inter-frame constraints applied on a fixed-frame (FF-LSP:T) or variable-frame (VF-LSP:T) basis to the LSP coefficients, algorithms incorporating intra-frame constraints applied to autocorrelation coefficients (Auto:I) or LSP coefficients (LSP:I), along with combinations (FF-LSP:T,Auto:I), (FF-LSP:T,LSP:I), (VF-LSP:T,LSP:I). In the evaluation, global estimates of SNR were employed since the assumption of accurate local estimates is normally unrealistic in actual enhancement environments. Also, energy thresholds for inter-frame constraints were obtained from frame energy histograms at each SNR. In this study, the primary tool for quantitative enhancement evaluation has been objective quality measures. This is based on extensive work carried out in the formulation of objective speech quality measures [6], and the application of these measures to enhancement [2]. Fair to good correlation has been shown to exist between subjective and objective quality measures.

### Evaluation Using Additive White Gaussian Noise

As previously reported, the constrained enhancement algorithms have been shown to significantly improve speech quality over such past techniques as the unconstrained Lim-Oppenheim technique as well as spectral subtraction with magnitude averaging [3]). Although significant improvement was noted, it was possible the algorithms were improving one or two particular speech classes which had high concentrations over the speech considered. Therefore, a comparative evaluation over speech sound classes was performed. Improvement over all classes of speech was reported.

As mentioned, the iterative enhancement algorithms must be suspended at some iteration. In order to determine a terminating iteration, a criterion must be selected to evaluate levels of improvement as the iterative scheme progresses. The criterion chosen is based on objective speech quality measures. Such measures are formed by a weighted comparison of actual and resulting estimated LPC predictor coefficients found during enhancement. The obvious problem with such a criterion is that, outside of simulation, the actual speech is unknown during the procedure. If, however, simulations were to show a consistent value for the best iteration in terms of this criterion, a convenient stopping condition would exist. Previous results based on objective quality measures indicate the unconstrained approach to produce maximum objective quality at different iterations for different classes of speech. Table 1 illustrates this behavior over the indicated sound classes. As this table shows, maximum overall speech quality is obtained at the third iteration, with considerable variation across sound types. For example, glides required two iterations, with nasals, liquids, and affricates requiring between five and six. Therefore, depending on sound class concentration, the optimal iteration (in terms of minimum distance) would vary considerably. This result indicates the inability to determine in advance a terminating iteration for the unconstrained approach since it is highly dependent on sound class and to a lesser degree on SNR.

The new constrained enhancement algorithms appear to solve this problem of sound class dependency. Table 2 presents results from an equivalent evaluation for one of the constrained enhancement algorithms (FF-LSP:T,Auto:I). A comparison between tables 1 and 2 show that the constrained approach produces superior quality measures across all speech classes at the same iteration. This improvement surpasses even combined individual maximum quality measures found across the unconstrained approach. Thus, the constrained enhancement algorithm does more than simply impose a constraint to adjust the rate of improvement: the constrained approaches consistently result in superior objective speech quality at the same iteration over all sound classes, independent of SNR. Table 3 summarizes optimum terminating points in terms of objective quality for the enhancement algorithms. Techniques employing only inter-frame constraints consistently resulted (93% occurrence) in maximum quality at the third iteration. Techniques employing inter- and intra-frame constraints had a 97% occurrence of maximum quality at the seventh iteration. In addition, adjacent iterations differ only slightly in objective quality for the constrained techniques. This is in sharp contrast to the large variations in adjacent iterations for the unconstrained technique. Therefore, if the iterative scheme were allowed to continue or halted one iteration

prior to optimal, only minor differences in speech quality would result. The results consistently suggested that the constrained enhancement algorithms reach a maximum level of speech quality at the same iteration, independent of SNR and sound class concentrations.

| *Sound Type* | *Itakura-Saito Likelihood Measure (across iterations)* | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | *Original* | *#1* | *#2* | *#3* | *#4* | *#5* | *#6* | *#7* |
| Silence | 1.63 | 1.62 | ♣1.61 | 1.65 | 1.93 | 3.76 | 20.36 | 49.88 |
| Vowel | 4.02 | 3.72 | 3.45 | ♣3.30 | 3.72 | 8.32 | 121.8 | — |
| Nasal | 19.81 | 19.15 | 18.42 | 17.66 | 17.01 | 16.59 | ♣15.19 | 15.70 |
| Stop | 7.26 | 6.11 | 4.93 | 3.98 | ♣3.82 | 6.89 | 25.52 | 29.69 |
| Fricative | 3.74 | 3.64 | 3.53 | ♣3.51 | 3.90 | 7.66 | 47.83 | 94.11 |
| Glide | 1.53 | 1.41 | ♣1.33 | 1.44 | 2.23 | 4.30 | 8.39 | 15.56 |
| Liquid | 9.60 | 8.24 | 6.55 | 4.55 | 2.61 | ♣1.68 | 6.38 | 30.00 |
| Affricate | 3.92 | 3.61 | 3.21 | 2.70 | 2.09 | ♣1.55 | 2.91 | 2.98 |
| Voiced + Unvoiced | 5.84 | 5.32 | 4.77 | 4.29 | ♣4.29 | 7.35 | 61.87 | — |
| Total | 4.02 | 3.72 | 3.40 | ♣3.15 | 3.27 | 5.80 | 43.46 | — |

Table 1: Lim-Oppenheim unconstrained speech enhancement for AWGN, SNR=+5dB. Optimum perceived quality for a particular speech class is indicated by a ♣.

| *Sound Type* | *Itakura-Saito Likelihood Measure (across iterations)* | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | *Original* | *#1* | *#2* | *#3* | *#4* | *#5* | *#6* | *#7* | *#8* |
| Silence | 1.63 | 1.55 | 1.35 | 1.16 | 1.03 | 0.98 | 0.93 | ♣0.88 | 0.90 |
| Vowel | 4.02 | 3.32 | 2.87 | 2.39 | 1.86 | 1.68 | 1.57 | ♣1.56 | 1.83 |
| Nasal | 19.81 | 16.49 | 12.40 | 10.52 | 8.68 | 6.84 | 4.93 | ♣3.79 | 5.55 |
| Stop | 7.26 | 6.25 | 4.84 | 3.49 | 2.67 | 1.81 | 1.38 | ♣1.13 | 1.43 |
| Fricative | 3.74 | 3.43 | 3.03 | 2.61 | 2.24 | 1.95 | 1.73 | ♣1.61 | 1.84 |
| Glide | 1.53 | 1.39 | 1.28 | 1.23 | 1.21 | 1.19 | 1.16 | ♣1.15 | 1.22 |
| Liquid | 9.60 | 6.48 | 3.38 | 2.24 | 1.61 | 1.21 | 0.94 | ♣0.92 | 1.21 |
| Affricate | 3.92 | 3.72 | 3.45 | 3.12 | 2.80 | 2.60 | 2.47 | ♣2.37 | 3.96 |
| Voiced + Unvoiced | 5.84 | 4.64 | 3.66 | 3.01 | 2.50 | 2.13 | 1.86 | ♣1.74 | 1.95 |
| Total | 4.02 | 3.03 | 2.44 | 2.07 | 1.80 | 1.61 | 1.46 | ♣1.38 | 1.49 |

Table 2: Hansen-Clements Inter & Intra-frame constrained speech enhancement for AWGN, SNR=+5dB. Optimum perceived quality for a particular speech class is indicated by a ♣.

| *Constrained Enhancement Algorithm* | *Additive White Gaussian Noise SNR* | | | | | | | | *OVERALL* | |
|---|---|---|---|---|---|---|---|---|---|---|
| | *-5 dB* | | *-0 dB* | | *+5 dB* | | *+10 dB* | | | |
| | *Optimal Iteration using Itakura-Saito Likelihood Measure* | | | | | | | | | |
| | *Iter.* | *Freq.* | *Iter.* | *Freq.* | *Iter.* | *Freq.* | *Iter.* | *Freq.* | *Iter.* | *Freq.* |
| *FF-LSP:T* | 3 | 100% | 3 | 87% | 3 | 87% | 3 | 100% | 3 | 91% |
| | | | 4 | 13% | 4 | 13% | | | 4 | 9% |
| *VF-LSP:T* | 3 | 90% | 3 | 85% | 3 | 94% | 3 | 100% | 3 | 93% |
| | 4 | 10% | 4 | 15% | 4 | 6% | | | 4 | 7% |
| *FF-LSP:T,Auto:I* | 7 | 100% | 7 | 100% | 7 | 100% | 7 | 88% | 7 | 97% |
| | | | | | | | 6 | 12% | 6 | 3% |
| *FF-LSP:T,LSP:I* | 4 | 100% | 4 | 100% | 4 | 100% | 4 | 100% | 4 | 100% |
| *VF-LSP:T,LSP:I* | 4 | 100% | 4 | 100% | 4 | 100% | 4 | 100% | 4 | 100% |

Table 3: Summary of optimal terminating iteration across SNR for AWGN.

### Additive Non-White, Non-Stationary Noise

The unconstrained Wiener filtering/all-pole modeling approach was previously generalized for colored aircraft noise [1]. In that study, an extensive investigation was performed using various spectral estimation techniques (MEM, MLM, Burg, Bartlett, Pisarenko, Periodogram) for securing estimates of colored background noise, along with varying SNR (-20dB to +20dB). Results indicated that Bartlett's method produced spectral estimates which resulted in highest quality improvement for this particular distortion.

Noise recorded from a Lockheed C130 aircraft interior was used to degrade noise free utterances. For these simulations, two Bartlett spectral estimates found from the original noise waveform (to avoid complications in silence detection) were used across each sentence. The noise was both colored and non-stationary, so increasing the number of spectral estimates across the utterance should improve enhancement performance. An analysis was performed for an inter-frame (FF-LSP:T), and a combined inter and intra-frame (FF-LSP:T,Auto:I) approach. Informal listening tests indicated noticeable quality improvement. Figure 2 illustrates results from this study. All configurations examined showed significant improvement in Itakura-Saito measures. Plot *a* shows Itakura-Saito measures for the original distorted speech. Plot *b* is from the unconstrained Wiener filtering technique. Plots *c* and *d* are typical values for the inter-frame constraint (FF-LSP:T), and inter- plus intra-frame constraint (FF-LSP:T, Auto:I) approaches. In order to determine limits on the level of enhancement, the original undistorted predictor coefficients were used in the unconstrained algorithm. In essence, the two step MAP estimation approach is now reduced to a single MAP estimate of $\vec{s}_O$, and therefore represents the theoretical limit for enhancement using Wiener filtering. Plot *e* indicates this limit. Although only Itakura-Saito measures are shown, similar improvement was observed for log area ratio and weighted spectral slope distance measures. As this figure indicates, significant levels of enhancement result for the constrained enhancement algorithms.

These results show that the constraint algorithms outperform the unconstrained approach for a colored distortion. However, it is possible that the constrained techniques are improving only particular speech classes which may have high concentrations in the test utterances. Therefore, a performance evaluation over sound classes was performed by hand partitioning speech into segments, pro-

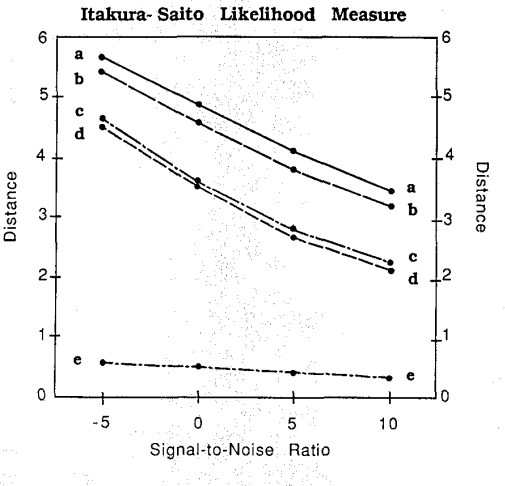


Figure 2: Comparison of inter & intra-frame constrained enhancement algorithms for colored aircraft noise over SNR.

a.) Original Distorted Speech

b.) Lim-Oppenheim: Unconstrained Wiener filtering

c.) Hansen-Clements: employing Inter-Frame constraints

d.) Hansen-Clements: employing Inter & Intra-Frame constraints

e.) Theoretical limit: using undistorted LPC coefficients $\vec{a}$.

ing entire sentences, and computing objective measures from each class. Table 4 summarizes this comparison between the unconstrained technique to that of the inter- and intra-frame constraint approach (FF-LSP:T,Auto:I). Measures for the theoretical limit using undistorted LPC coefficients are also indicated. It should be noted that voiced plus unvoiced measures give a better indication of quality improvement due to the time varying nature of the interfering background noise. Improvement is indicated for all types of speech. This shows that the constrained techniques are enhancing all aspects of the speech signal.

| *Sound Type* | *Itakura-Saito Likelihood Measure* | | | |
|---|---|---|---|---|
| | *Original* | *Lim-Oppenheim* | *Hansen-Clements* | *True LPC* |
| Silence | 6.63 | 6.33 | 4.32 | 2.03 |
| Vowel | 3.23 | 2.54 | 1.44 | 0.53 |
| Nasal | 4.03 | 3.26 | 2.13 | 0.45 |
| Stop | 1.58 | 1.29 | 0.66 | 0.61 |
| Fricative | 1.37 | 1.09 | 0.85 | 0.65 |
| Glide | 1.14 | 1.04 | 0.52 | 0.51 |
| Liquid | 1.22 | 0.55 | 0.22 | 0.18 |
| Affricate | 0.90 | 0.51 | 0.33 | 0.16 |
| Voiced + Unvoiced | 2.27 | 1.76 | 1.08 | 0.52 |
| Total | 4.15 | 3.86 | 2.74 | 1.17 |

Table 4: Comparison of unconstrained (Lim-Oppenheim) and inter- and intra-frame constrained (Hansen-Clements) algorithms over sound types for slowly varying colored noise. SNR = +5 dB

### Recognition Evaluation

A fairly standard, isolated-word, discrete-observation hidden Markov model recognition system was used for evaluation. This system was LPC based and had no embellishments. In all experiments, a five state, left-to-right model was used. System dictionary consisted of twenty highly confusable words used by Texas Instruments and Lincoln Labs to evaluate recognition systems. Subsets include {go,oh,no,hello} and {six,fix}. Twelve examples of each word were used, six for training, six for recognition (i.e., all tests fully open). A vector quantizer was used to generate a 64 state codebook using two minutes of noise free training data. The twenty models employed by the HMM recognizer were trained using the forward-backward algorithm. Table 5 presents results from five scenarios using a noise free codebook and noise free trained system. Spectral subtraction preprocessing employed three frames of magnitude averaging. The unconstrained Lim-Oppenheim approach was terminated at the third iteration. The constrained Hansen-Clements (FF-LSP:T,Auto:I) was terminated at the seventh. As these results indicate, recognition was reduced to chance for noisy, spectral subtraction, and Lim-Oppenheim (-5,0,5 dB) speech. The constrained approach resulted in improved recognition across all SNR considered, which is quite remarkably in light of the severe levels of noise, and difficulty of dictionary employed. However, reliable recognition in such a hostile environment may require more than merely extending existing techniques. As a final comparison, three tests were performed using noisy and enhanced speech (SNR=+10dB). For the noisy case, speech was coded using a noisy codebook, and recognition performed using a noisy trained HMM recognizer. Similar tests were performed for two enhancement techniques, (i.e., enhanced words coded using enhanced codebook, and tested using enhanced speech trained HMM recognizer). 40% of the errors in recognition were caused by misclassification of leading consonants (especially fricatives).

| *Condition* | *RECOGNITION RESULTS* | | | | |
|---|---|---|---|---|---|
| | | *Signal-to-Noise Ratio* | | | |
| *(noise free training)* | *Original* | *-5dB* | *0dB* | *+5dB* | *+10dB* |
| *Noise free* | 88% | | | | |
| *Noisy* | | 5% | 5% | 6.7% | 5% |
| *Spectral Subtraction* | | 5.8% | 7.1% | 5% | 5.4% |
| *Lim-Oppenheim* | | 5.4% | 5.8% | 7.5% | 12.5% |
| *Hansen-Clements* | | 15% | 14% | 19.5% | 34.5% |

| *Train & Recognize In Same Environment* | | | |
|---|---|---|---|
| *Noise free* | *Noisy †* | *Hansen-Clements †* | *Lim-Oppenheim †* |
| 88% | 90% | 77% | 23% |

Table 5: Recognition performance using enhancement preprocessing in AWGN. † SNR = +10dB

## 5 Conclusions

The constrained speech enhancement algorithms have been shown to improve speech quality across all classes of speech for both additive white Gaussian and slowly varying, non-white degradations. In addition, a consistent terminating procedure has been identified which is independent of sound class concentration and relatively insensitive to varying SNR. Finally, the constrained algorithms have shown improvement as a preprocessor for speech recognition, although their ability to bring performance up to an acceptable level in SNR's low as those considered is questionable. Though the enhancement procedures improved LPC parameter estimation substantially, LPC-based strategies may simply be inappropriate for SNR's of roughly 0dB. Further work in this SNR range will require as a minimum, different front end processing.

This work sponsored in part by U.S. Army Human Engineering Labs.

## References

[1] J.H.L. Hansen, M.A. Clements, " Enhancement of Speech Degraded by Non-White Additive Noise," Final Technical Report DSPL-85-6, Georgia Institute of Technology, Atlanta, August 1985.

[2] J.H.L. Hansen, M.A. Clements, " Objective Quality Measures Applied to Enhanced Speech," *Proc. of the Acoustical Society of America,* 110th Meeting, C11, Nashville, Tenn., Nov. 1985.

[3] J.H.L. Hansen, M.A. Clements, " Iterative Speech Enhancement With Spectral Constraints," *Proc. 1987 IEEE ICASSP,* pp. 189-192, Dallas, TX, April 1987.

[4] J.H.L. Hansen, M.A. Clements, " Constrained Iterative Speech Enhancement," *IEEE Trans. on Acoust., Speech, Signal Processing,* submitted, Dec. 1987.

[5] J.S. Lim, A.V. Oppenheim, " All-Pole Modeling of Degraded Speech," *Trans. on Acoustics, Speech, and Signal Processing,* Vol. ASSP-26, pp. 197-210, June 1978.

[6] S.R. Quackenbush, " Objective Measures of Speech Quality," Ph.D. Thesis, Georgia Institute of Technology, May 1985.