Constrained Iterative Speech Enhancement with Application to Speech Recognition

John H. L. Hansen, *Member, IEEE,* and Mark A. Clements, *Senior Member, IEEE*

***Abstract*—In this paper, an improved form of iterative speech enhancement for single channel inputs is formulated. The basis of the procedure is sequential maximum *a posteriori* estimation of the speech waveform and its all-pole parameters as originally formulated by Lim and Oppenheim, followed by imposition of constraints upon the sequence of speech spectra. The new approaches impose intraframe and interframe constaints on the input speech signal to ensure more speech-like formant trajectories, reduce frame-to-frame pole jitter, and effectively introduce a relaxation parameter to the iterative scheme. Recently discovered properties of the line spectral pair representation of speech allow for an efficient and direct procedure for application of many of the constraint requirements. Substantial improvement over the unconstrained method has been observed in a variety of domains. First, informal listener quality evaluation tests and objective speech quality measures demonstrate the technique's effectiveness for additive white Gaussian noise. A consistent terminating point for the iterative technique is also shown. Second, the algorithms have been generalized and successfully tested for noise which is nonwhite and slowly varying in characteristics. The current systems result in substantially improved speech quality and LPC parameter estimation in this context with only a minor increase in computational requirements. Third, the algorithms were evaluated with respect to improving automatic recognition of speech in the presence of additive noise, and shown to outperform other enhancement methods in this application.**

I. Introduction

THE presence of background noise can seriously degrade the performance of many speech processing systems, since most digital voice communication and recognition systems have traditionally been formulated in noise-free tranquil environments. There are, however, many instances where such systems must perform reliably in noisy environments. As an example, consider the use of speech recognition in a noisy aircraft cockpit. It has been shown that recognition performance is severely reduced in such an environment due to background noise and pilot task requirements [8], [13], [18]. Since commonly used front ends do not usually take noise into account explicitly, recognition deteriorates rapidly. One alternative, which would benefit recognition as well as speech coding systems is to develop enhancement preprocessors that produce speech or recognition features less sensitive to background noise, so that existing recognition/communication systems may be employed. Such preprocessing systems would also benefit human listeners by improving speech characteristics in voice communications systems.

The problem of enhancing speech degraded by additive background noise covers a broad spectrum of applications and issues [12]. A system may be directed at one or more objectives such as improving overall quality, increasing intelligibility, or reducing listener fatigue. Assumptions made in this investigation include: i) the background noise distortion is additive, ii) only the degraded speech signal is available (i.e., single microphone environment), and iii) the noise and speech signals are uncorrelated.

This paper presents an improved method for iterative speech enhancement based on a set of vocal tract spectral constraints. The framework of this approach was adopted from all-pole modeling/noncausal Wiener filtering as formulated by Lim and Oppenheim [11]. The original iterative technique attempts to solve for the maximum *a posteriori* (MAP) estimate of a speech waveform in additive white noise. The improved techniques are formulated using interframe and intraframe constraints to ensure speech-like characteristics. An efficient technique for applying the spectral constraints is based on the line spectral pair (LSP) transformation of the LPC parameters. The paper is arranged as follows. First, the iterative unconstrained technique is discussed. Several anomalies are cited which motivate formulation of constrained enhancement techniques using the LSP transformation. Next, algorithm evaluation is performed for additive white Gaussian noise, and a slowly varying nonwhite distortion. Finally, a comparative evaluation is also performed to determine their usefulness as preprocessors for recognition in noisy environments.

II. Iterative Speech Enhancement

Enhancement based on the estimation of all-pole speech parameters in additive white Gaussian noise was investigated by Lim and Oppenheim [11], and later for a colored noise degradation by Hansen and Clements [3], [4], [6]. This approach attempts to solve for the maximum *a posteriori* 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. After some simplification, it can be shown that the resulting equations for the joint MAP estimate of the all-pole speech parameters $\vec{a}$, gain g, and noise free speech $\vec{S}_0$ become nonlinear. Lim and Oppenheim considered a suboptimal solution employing sequential MAP estimation of $\vec{S}_0$ followed by MAP estimation of $\vec{a}$, g given $\hat{\vec{S}}_{0,i}$, where $\hat{\vec{S}}_{0,i}$ is the result of the ith estimation. The sequential estimation procedure is linear at each iteration, and must continue until some criterion is satisfied. With further simplifying assumptions, it can be shown that MAP estimation of $\vec{S}_0$ is equivalent to noncausal Wiener filtering of the noisy speech $\vec{Y}_0$. Lim and Oppenheim showed this technique, under certain conditions, increases the joint likelihood of $\vec{a}$ and $\vec{S}$ with

Manuscript received March 20, 1988; revised December 20, 1989. This work was supported in part by grants from the U.S. Army Human Engineering Labs and Department of Defense.

J. H. L. Hansen was with the School of Electrical Engineering, Georgia Institute of Technology, Atlanta, GA. 30332. He is now with the Department of Electrical Engineering, Duke University, Durham, NC 27706.

M. A. Clements is with the School of Electrical Engineering, Georgia Institute of Technology, Atlanta, GA 30332.

IEEE Log Number 9042251.

each iteration. It can also be shown to be the optimal solution in the mean-squared sense for a white noise distortion.

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 represents a serious drawback if the approach is to be used in environments requiring automatic speech enhancement. Hansen and Clements performed an extensive investigation of this technique for additive white Gaussian (AWGN), and a generalized version for additive nonwhite, nonstationary aircraft interior noise [3], [4]. Objective speech quality measures, which have been shown to be correlated with subjective quality [17], were used in the evaluation. This approach was found to produce significant levels of enhancement for white Gaussian noise in 3–4 iterations. Improved all-pole parameter estimation was also observed in terms of reduced mean-squared error. Only if the probability density function is unimodal, and the initial estimate for $\vec{a}$ is such that the local maximum equals the global maximum, is the procedure equivalent to the joint MAP estimate of $\vec{a}$, g, and $\vec{S}_0$. Some interesting anomalies were noted which helped motivate development of the constrained approaches. First, as additional iterations were performed, individual formants of the speech consistently decreased in bandwidth and shifted in location as indicated in Fig. 1. Second, frame-to-frame pole jitter was observed across time. Both effects contributed to unnatural sounding speech. Third, although the 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 which was shown by Hansen and Clements to be dependent on speech class concentration [5].

Lim and Oppenheim later recognized that their method resulted in biased estimates of the all-pole speech parameters. This observation, could easily explain the variation in quality across different speech classes for the unconstrained technique. An improved maximum likelihood (ML) method for general AR and ARMA parameter estimation in noise was later formulated by Musicus and Lim [15] which addressed some of the limitations of the original technique. Although such a procedure might help in the estimation of speech parameters, it has never to our knowledge, been subjected to extensive testing on speech in noise. Even if this improved procedure were used, it would only serve to address the problem of better estimation of AR parameters in noise. Since speech is not truly all pole, there is no way of knowing whether or not the ML procedure is actually better for speech than the original Lim–Oppenheim procedure. On the other hand, performance of the unconstrained Lim–Oppenheim approach has been well documented for speech degraded by additive white Gaussian noise, thereby motivating its use as a basis of comparison.

The constrained iterative techniques presented here address the speech in additive noise problem by using speech-specific constraints via the line spectral pair parameters, which in themselves are different from the AR model alone. Hence a new set of constraints which add more knowledge of the type of signal being enhanced are added. Although it may be possible to improve the enhancement procedure further by employing our constraint approaches within the Musicus–Lim technique, we have shown that the imposition of some relatively simple constraints improves speech quality results, even when directly attached to the original Lim–Oppenheim method.

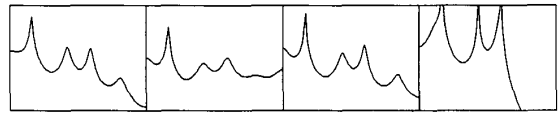


Fig. 1. Variation in vocal tract response across iterations. (a) Original. (b) Distorted original. (c) Four iterations. (d) Eight iterations.

A Enhancement with Spectral Constraints

Consider the statistical parameter estimation of speech in the presence of noise as formulated by Lim and Oppenheim where all unknown parameters over a short interval (all-pole speech paremeters $\vec{a}$, gain g, and noise free speech $\vec{S}_0$) are random with *a priori* Gaussian probability density functions. It was shown that MAP estimation of $\vec{a}$, g, and $\vec{S}_0$ given noisy observations $\vec{Y}_0$, results in a set of nonlinear equations. Therefore, instead of joint estimation of $\vec{a}$ and $\vec{S}_0$, a suboptimal solution was formulated employing a two-step approach based on MAP estimation of $\vec{S}_0$ given $\vec{Y}_0$, followed by MAP estimation of $\vec{a}$, g given $\hat{\vec{S}}_{0,i}$, where $\hat{\vec{S}}_{0,i}$ is the result of the ith estimation. In the currently reported work, constraints are imposed on the vocal tract spectrum between MAP estimation steps. The procedure for obtaining the MAP estimates of $\vec{a}$ and g remain the same, as that of Lim and Oppenheim. In the current system, constraints are applied to $\hat{\vec{a}}_i$ to ensure that i) the all-pole speech model is stable, ii) it possesses speech-like characteristics (e.g., poles are in reasonable places with respect to each other and the unit circle), and iii) the vocal tract characteristics do not vary by more than a prescribed amount from frame to frame when speech is present. Given the new estimate $\hat{\vec{a}}_{i+1}$, the second MAP estimation of $\vec{S}_0$ is performed by maximizing its conditional probability density function given $\hat{\vec{a}}_{i+1}$ and the observed noisy sequence $\vec{Y}_0$. Since this probability density function is jointly Gaussian, the resulting MAP estimate is equivalent to a MMSE estimate of $\vec{S}_0$. With further simplifying assumptions, it can be shown that MAP estimation of $\vec{S}_0$ reduces to a minimum mean-squared error (MMSE) estimate, and as the observation window increases, the procedure becomes a noncausal Wiener filter. Once the new estimate of $\hat{\vec{S}}_{0,i}$ is formed, the iterative procedure continues by reestimating $\hat{\vec{a}}_i$, applying constraints to $\hat{\vec{a}}_i$, and forming the noncausal filter using $\hat{\vec{a}}_{i+1}$ to reestimate $\hat{\vec{S}}_{0,i}$. The procedure continues until some convergence criterion is satisfied. Due to the flexibility of the enhancement framework, a variety of constraint options are possible between MAP estimation steps.

Fig. 2 presents an overview of two classes of constraints which include interframe (across time) and/or intraframe (across iterations). Each technique differs in the type of constraint and computational requirements. The present evaluation focuses on two representative interframe (FF-LSP : T) and combined interframe plus intraframe (FF-LSP : T, Auto : I) based techniques. Further discussion of all techniques are found in [5]–[7]. For historical purposes, several comments concerning the other approaches are summarized.

Since observations indicate that poles of the LPC filter often move unrealistically close to the unit circle when the uncon-

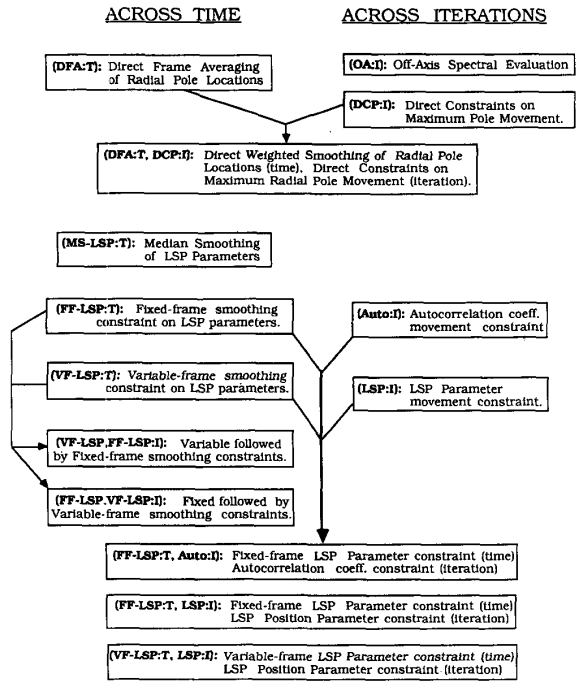


Fig. 2. An overview of spectral constraints considered for the class of constrained speech enhancement algorithms.

strained iterative technique is allowed to continue, initial techniques limited pole movement by applying constraints directly to radial and/or angular movements of the LPC poles across iterations and time. For these techniques, LPC predictor coefficients were obtained, a Pth-order root solve was performed and a pole ordering step applied. If pole movement fell within a movement constraint window, a constraint was applied, otherwise, no constraint was applied based on the assumption that either movement was allowable, or that the pole was mischaracterized due to the ordering step. Results showed substantial improvement in objective speech quality (as measured by Itakura–Saito, log-area-ratio, and weighted spectral slope (Klatt) measures [17]). Informal listening tests also revealed improvement, especially during vowels and vowel transitions toward nasals. Larger levels of quality improvement were observed using interframe versus intraframe constraints, thus suggesting that temporal variation in pole locations have a greater effect on overall quality.

Although successful in improving speech quality, constrained techniques based on direct pole location were computationally expensive. A Pth-order root-solve and a pole ordering step per frame for each iteration was required. Since root solving is not always numerically accurate and ordering can be inconsistent across frames, a more robust approach was sought to implement these constraints.

An alternative approach for implementing the spectral constraints was formed by employing the line spectral pair (LSP) transformation as a method for representing the vocal tract spectrum. Previous success of the LSP transformation in low-bit-rate speech coding by Crosmer [2] led to the use of LSP's for this purpose.

The line spectral pair (LSP) [9], [19] transformation comes from modifying the LPC polynomial, $A(z)$, in two ways: $P(z)$ and $Q(z)$ are obtained by augmenting $A(z)$'s PARCOR sequence with a $+1$ and -1, respectively. This results in the following two polynomials of order $p + 1$ which have all their roots on the unit circle:

$$P(z) = (1 - z^{-1}) \prod_{i=1,3,5,\cdots}^{M-1} (1 - 2\cos\omega_i z^{-1} + z^{-2}) \quad (1)$$

$$Q(z) = (1 + z^{-1}) \prod_{i=2,4,6,\cdots}^{M-1} (1 - 2\cos\omega_i z^{-1} + z^{-2}). \quad (2)$$

The angles of the roots, $\{\omega_i, i = 1, 2, \cdots, M\}$, are called the *line spectrum pairs*. In general, $A(z)$ will represent a stable LPC filter if and only if the roots of $P(z)$ and $Q(z)$ interleave. The angles of the roots of $P(z)$, correspond roughly to the angles of the roots of $A(z)$ (formant frequencies), and the separation of a particular root of $P(z)$ from the closet root of $Q(z)$ indicates in some sense the bandwidth of that resonance. The angle of the roots of $P(z)$ between 0 and π are termed the position parameters (i.e., the odd indexed LSP parameters, $\{p_i = \omega_{2i-1}, i = 1, 2, \cdots, M/2\}$), and the separations mentioned above the difference parameters d_i

$$\{|d_i| = \min_{j=-1,1} (|\omega_{2i+j} - \omega_{2i}|), i = 1, 2, \cdots, M/2\}. \quad (3)$$

The sign of d_i is positive if ω_{2i} is closer to ω_{2i+j}, and otherwise is negative. The useful properties of the LSP's include an easy check for stability, excellent interpolation properties, ease of computation (compared to roots of $A(z)$), some well-understood trajectories for speech, and the relative insensitivity of the auditory system under quantization of the difference parameters.

B. Enhancement Using the LSP Transformation

In these techniques, constraints are imposed on the LSP parameters directly. In the first technique (MS-LSP:T), a five frame median smoothing constraint was placed on the position parameters across time,with difference parameters restricted to be at least d_{MIN} in magnitude, ensuring the LPC poles of reasonable bandwidth. Good improvement resulted without the expense of root solving or pole ordering. Plots of LSP parameters versus time confirmed a reduction in frame-to-frame pole jitter with only a slight increase in computational requirements. Since vocal-tract characteristics and relative strength of background noise vary across time, the imposition of spectral constraints should be dependent on speech characteristics obtained during the enhancement procedure. Therefore, the remaining constraints are applied based on particular characteristics found in the speech waveform during enhancement.

Two interframe approaches are considered: a fixed frame rate (FF-LSP:T), and a variable frame rate approach (VF-LSP:T). In the first of these, the LPC predictor coefficients $\vec{a}$ are first converted to LSP parameters. Next, each frame's energy is observed, and classified as voiced or unvoiced speech according to some threshold $E_{V/\text{UV}}$. A local running count L_i is kept for the number of consecutive frames which fall below the energy threshold. If L_i reaches L_{MAX}, all subsequent frames below the threshold are classified as noise. This allows for a tighter pole movement constraint during long periods of silence. The position parameters for each frame are smoothed using a weighted triangular window with a variable base of support (1 to 5 frames). If a frame has been classified as noise, maximum smoothing (or tightest movement constraint) is performed. The

lower formant frequencies are smoothed over a narrower triangle width than for those position parameters at higher frequencies in order to preserve perceptually important speech characteristics found in the lower formants. No smoothing is performed on the differnce parameters since they are more closely related to formant bandwidth than formant location. However, it is possible that a difference parameter falls within a "forbidden zone." When this occurs, the LPC analysis has most likely underestimated a particular pole's bandwidth. Since this causes unnatural sounding speech, (as found in the unconstrained approach), the value of $|d_i|$ is set to d_{MIN}. Finally, the position and difference parameters are combined to form the constrained LPC predictor coefficients $\hat{\vec{a}}_{i+1}$.

The (FF-LSP : T) technique applies constraints across time on a frame-by-frame basis. Since phonetic transitions do not normally coincide with frame boundaries, an inter-frame approach (VF-LSP : T) based on constraints applied over speech segments was formulated. The technique is identical in theory to (FF-LSP : T), except for the front-end segmentation algorithm which divides the signal into speech segments. Segments are chosen to be long when the speech spectrum is slowly varying and short when the speech spectrum is varying quickly. The LSP parameters are reconstructed with linear interpolation used to compute the parameters for intermediate frames.

The segmentation algorithm begins by determining the onset/offset of speech by thresholding the LPC residual energy, which produces relatively long segments. Long segments are subdivided based on the curvature of the position parameters. This is performed by computing a gain-normalized Itakura–Saito measure of the spectral distance between the frequency response of two adjacent frames. The procedure continues by computing spectral distortion parameters for successively longer segments until the spectral distortion exceeds a threshold T_D. At that point, a subsegment boundary is set, with the intermediate position parameters reconstructed via linear interpolation. During this step, the length of a subsegment is also limited to L_{MAX} to prevent excessively long segments which might contribute to muffled or unnatural sounding speech. The advantage of this approach is to incorporate more information from adjacent frames when the spectrum indicates similar characteristics. This, in effect, distorts the position parameters as little as possible when associated difference parameters indicate the presence of formants. Difference parameters for each frame are used to compute the predictor coefficients $\hat{\vec{a}}_{i+1}$. The difference parameters are required to be least d_{MIN} or greater.

The convergence problems inherent in the unconstrained Wiener filtering approach which have been pointed out [5], [7], [15], are at least partially caused by bias in the MAP estimation. Although spectral constraints were originally constructed to be used across frames, it has been observed that if they are used across iterations, convergence to reasonable values occurs with much greater frequency and consistency. In particular, previous results based on objective speech quality measures show the unconstrained Wiener filtering approach to produce minimum objective measures at different iterations for different classes of speech [5], [7] (see Table III). By constraining the vocal tract filter to be a function of its values obtained from previous iterations, a much improved consistency in quality across speech classes and LPC parameter $\hat{\vec{a}}_i$ estimation resulted. Two approaches were considered, one applied to the autocorrelation lags (Auto : I), the other to the position parameter (LSP : I). The first approach simply weighted the present set of autocorrelation lags with the same frame from previous iterations. Such a technique is easy to perform, since the autocorrelation lags must be computed in order to estimate the predictor coefficients $\hat{\vec{a}}$. The second approach weighted position parameters with those from the same frame but previous iteration. If the corresponding difference parameter indicated the adjacent position parameter to represent a formant, this approach had the effect of constraining the formants to lie along smooth tracks across iterations. Such a procedure is generally referred to as introducing relaxation into the iterations [16]. If the iteration is producing results for which weighted averaging makes sense (e.g., LSP's but not $\hat{\vec{a}}$), improved convergence results. Results from interframe, intraframe, and combined interframe plus intraframe constraint approaches will be presented in the next section. Fig. 3 illustrates the framework for the new set of constrained enhancement techniques.

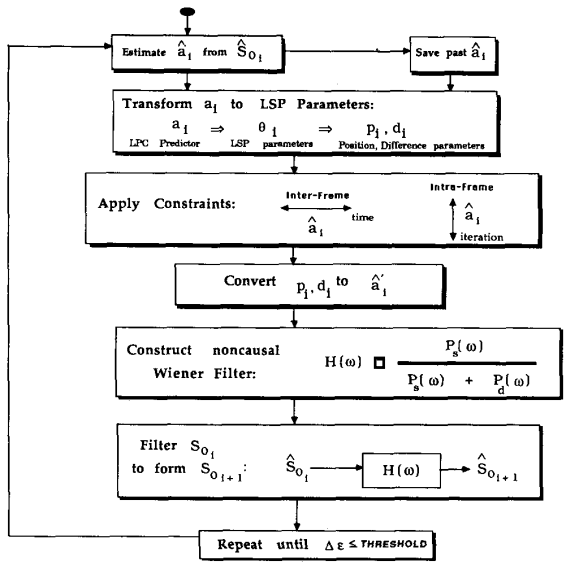


Fig. 3. Framework for the new set of constrained enhancement algorithms.

III. Evaluation

We now evaluate the performance of the proposed algorithms for spech enhancement alone, and as a preprocessor for word recognition in noisy environments. Speech was degraded by additive white or colored noise and processed. Enhancement algorithms evaluated include techniques incorporating interframe constraints applied on a fixed-frame (FF-LSP : T) or variable-frame (VF-LSP : T) basis to the LSP parameters, and algorithms incorporating combinations of interframe plus intraframe constraints (FF-LSP : T, Auto : I), (FF-LSP : T, LSP : I). Global estimates of SNR[1] were used in the evaluation, since the assumption of accurate local estimates is normally unrealistic in actual noisy environments. Further improvement is therefore possible if a continuous local SNR estimate is available. The intraframe constraints were applied across two to three iterations.

Several parameters must be addressed to ensure proper application of spectral constraints. These include the voiced/un-

[1]The signal-to-noise ratio is defined as 10 log $(\Sigma_n s^2(n)/\Sigma_n d^2(n))$, where the summation is over the entire length of the sentence. This definition was chosen in keeping with the format used in previous studies on noncausal Wiener filtering [11].

voiced energy threshold $E_{V/UV}$, silence frame count threshold L_{MAX}, LSP difference parameter thresholds d_{MIN}, d_{MAX}, and the accumulated frame-to-frame Itakura–Saito distance threshold T_D.

The energy threshold $E_{V/UV}$ is used to distinguish voiced from unvoiced or silent speech frames for use in applying interframe constraints. Values were obtained from frame energy histograms at each signal-to-noise ratio. Similar enhancement levels resulted for $E_{V/UV}$ in the range between average, and one standard deviation below average speech frame energy (e.g., average frame energy for sentence S6 was 7719. $E_{V/UV}$ set between 8000 and 5000 resulted in Itakura–Saito measures which ranged from 1.96 to 2.02).

The silence frame count threshold L_{MAX}, is used in conjunction with $E_{V/UV}$. If L_{MAX} consecutive frames fall below $E_{V/UV}$, that segment is classified as silence (or noise) so that tighter spectral constraints can be enforced. If $E_{V/UV}$ is set as above, similar speech quality measures resulted with L_{MAX} set between two and five frames. Reduced quality measures resulted with L_{MAX} in the eight to twelve frame range, thereby suggesting increased residual noise levels during silent portions.

The difference thresholds d_{MIN}, d_{MAX}, constrains the LSP difference parameters to ensure poles of reasonable bandwidths (e.g., the all-pole speech model is stable and that it possesses speech-like characteristics). Values in the range $0.015 \leq d_{MIN} \leq 0.031$ rad, $0.055 \leq d_{MAX} \leq 0.077$ rad, resulted in good quality improvement.

The value T_D (accumulated frame-to-frame Itakura–Saito distance threshold) greatly effects speech segment length. If set to high, small duration phonemes can be lost (e.g., an initial stop and final vowel joined to form one speech segment as in *be*). A value of 1.2 was found to produce segments of reasonable length and quality at higher SNR ($\geq +5$ dB). At lower SNR, frame-to-frame distance values were too large to reliably segment speech, resulting in decreased performance.

Generally speaking, substantial enhancement resulted for a wide range of $E_{V/UV}$, L_{MAX}, d_{MIN}, and d_{MAX} threshold settings, indicating the algorithms robust performance over estimated threshold values. Only T_D, the accumulated frame-to-frame Itakura–Saito distance threshold, proved to be sensitive, especially across varying SNR. Greater enhancement was observed when T_D was allowed to vary across iterations.

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 for speech coding [17], and the application of these measures to enhancement [3], [4]. Fair to good correlation has been shown to exist between subjective and objective quality measures, such as: the Itakura–Saito likelihood ratio, log area ratio, and weighted spectral slope measure. These measures have been shown to be a viable tool for use in evaluating speech enhancement algorithms for white and nonwhite additive noise [4]. In addition, the Itakura–Saito likelihood ratio is also a commonly used distance measure for speech recognition as well as for coding methods employing vector quantization. Therefore, improvement in Itakura–Saito distance might also suggest the possibility of improvement in automatic recognition. The speech data for enhancement evaluation is described in the Appendix.

A. Evaluation Using Additive White Gaussian Noise

Various configurations of the new constrained enhancement algorithms were evaluated in an additive white Guassian noise environment. Informal listening tests indicated noticeable quality improvement, although no intelligibility testing was performed. A variety of objective speech quality measures were used in the evaluation procedure. Fig. 4 illustrates a comparison of typical results for the various constraint approaches. The Itakura–Saito measure is plotted versus signal-to-noise ratio for a white noise distortion. Plot *a* represents the original distorted speech. Plots *b* through *e* represent combinations of interframe constraints (both fixed and variable rate), and intraframe constraints (applied to position parameters/autocorrelation lags). All configurations examined showed significant improvement in Itakura–Saito measures. Threshold settings for the variable frame rate interframe constraint were somewhat sensitive to varying noise levels. This indicates that although applying interframe constraints across speech segments is theoretically attractive and should aid in enhancement, in reality the speech segmentation step proves to be too sensitive to varying background noise levels. However, the fixed frame approach, by itself, and with either autocorrelation or position intraframe constraints gave impressive results with little sensitivity to varying levels of SNR. In order to determine a limit on the level of enhancement, the original undistorted predictor coefficients $\vec{a}$ were used in the unconstrained algorithm. In essence, the two step MAP estimation approach is now reduced to a single MAP estimate of $\vec{S}_0$, and therefore represents the theoretical limit for enhancement using Wiener filtering. Plot *f* indicates this limit.

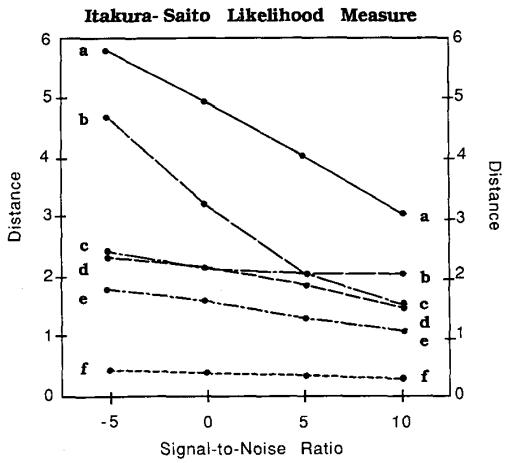


Fig. 4. Comparison of constraint algorithms over SNR. (*a*) Original distorted speech. (*b*) Interframe constraint: Variable frame (VF-LSP:T). (*c*) Intraframe constraint: Fixed frame (FF-LSP:T). (*d*) Interframe and intraframe constraints: Fixed frame, position (FF-LSP:T, LSP:I). (*e*) Interframe and intraframe constraints: Fixed frame, autocorrelation (FF-LSP:T, LSP:I). (*f*) Theoretical limit: Using undistorted LPC coefficients $\vec{a}$.

One advantage of the general class of Wiener filtering approaches is that no "musical tone" artifacts are present after processing as observed in spectral subtraction techniques [1], [3], [12]. To determine performance versus spectral subtraction, a series of enhancement evaluations under identical conditions (same distorted utterances, same global estimates) were performed. Evaluation was performed for both half-and full-wave rectification over a SNR range of -20 to $+20$ dB, and employed one to five frames of magnitude averaging (as defined by Boll [1]). See Hansen [7] for details. Full-wave rectification resulted in improvement over a wider range of SNR, however half-wave rectification had greater improvement over the re-

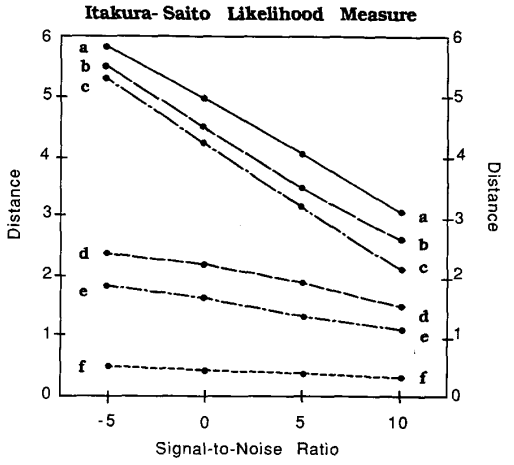


Fig. 5. Comparison of enhancement algorithms over SNR. (*a*) Original distorted speech. (*b*) Boll: Spectral subtraction, using magnitude averaging. (*c*) Lim–Oppenheim: Unconstrained Wiener filtering. (*d*) Hansen–Clements: Employing interframe constraints (FF-LSP:T). (*e*) Hansen–Clements: Employing interframe and intraframe constraints (FF-LSP:T, Auto:I). (*f*) Theoretical limit:Using undistorted LPC coefficients $\vec{a}$.

stricted SNR band of 5 to 10 dB. Magnitude averaging lead to improve enhancement for both rectification approaches.

Next, the constraint approaches were compared to spectral subtraction and unconstrained noncausal Wiener filtering. All systems performed enhancement on the same speech, with the same global estimates of SNR. Fig. 5 compares quality improvement for each technique. Although only Itakura–Saito measures are shown, similar improvement was observed for log area ratios and weighted spectral slope measures (Klatt). Itakura–Saito measures are presented since they are widely accepted as a spectral distance measure and have been used extensively for speech recognition applications. A comparison of the three speech quality measures is shown in Table I. The average correlation between each objective quality measure and subjective quality as measured by the diagnostic acceptability test (DAM) is shown [17].

1) Quality Improvement Over Speech Classes: To determine individual quality improvement, an evaluation over sound classes was performed by hand partitioning speech into segments, processing entire sentences, and computing objective measures from each class. Table II summarizes the comparison between the unconstrained technique, and an interframe plus intraframe constained approach (FF-LSP:T, Auto:I). Measures for the theoretical limit using undistorted LPC predictor coefficients $\vec{a}$ are also indicated. Improvement is indicated for all classes of speech. These results show that the constraint techniques are enhancing all aspects of the speech signal.

2) Termination Criterion: 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 III illustrates this behavior over the indicated sound classes. As shown, maximum overall speech quality is obtained at the third iteration, with considerable variation across sound types. Glides required two iterations for maximum quality, 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. Observations from a previous investigation indicate that the optimal iteration varies between the second and sixth and that it is also somewhat dependent on SNR [3].

The new constrained enhancement algorithms have less sensitivity to sound class. Table IV presents results from an equivalent evaluation for one of the constrained enhancement algorithms (FF-LSP:T, Auto:I). Comparing Tables III and IV shows that the constrained approach produces superior quality measures across all speech classes at the same iteration. The 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 I
A COMPARISON OF OBJECTIVE SPEECH QUALITY MEASURES FOR NOISY AND ENHANCED SPEECH EMPLOYING THE UNCONSTRAINED (LIM–OPPENHEIM) AND CONSTRAINED FF-LSP:T, AUTO:I (HANSEN–CLEMENTS) ALGORITHMS FOR WHITE GAUSSIAN NOISE. SNR = +5 dB, $|\hat{\rho}|$ IS THE AVERAGE CORRELATION COEFFICIENT BETWEEN OBJECTIVE AND SUBJECTIVE SPEECH QUALITY [17]

	Objective Quality Measure		
	Itakura–Saito	Log-Area Ratio	Klatt
$\|\hat{\rho}\|$	.59	.62	.74
Noisy Original	4.02	15.27	2.39
(Lim-Oppenheim)	3.15	8.78	2.19
(Hansen-Clements)	1.38	5.56	1.62

TABLE II
COMPARISON OF UNCONSTRAINED (LIM–OPPENHEIM) AND INTERFRAME AND INTRAFRAME CONSTRAINED (HANSEN–CLEMENTS) ALGORITHMS OVER SOUND TYPES FOR WHITE GAUSSIAN NOISE. SNR = +5 dB

	Itakura-Saito Likelihood Measure			
Sound Type	Original	Lim–Oppenheim	Hansen–Clements	True LPC
Silence	1.634	1.649	0.842	0.319
Vowel	4.020	3.299	1.651	0.582
Nasal	19.814	17.656	3.968	0.324
Stop	7.261	3.979	1.099	0.435
Fricative	3.739	3.509	1.766	0.649
Glide	1.525	1.442	1.131	0.705
Liquid	9.597	4.545	0.998	0.303
Affricate	3.924	2.702	2.229	0.323
Voiced + Unvoiced	5.838	4.293	1.761	0.519
Total	4.022	3.151	1.364	0.433

TABLE III
LIM-OPPENHEIM UNCONSTRAINED SPEECH ENHANCEMENT FOR WHITE GAUSSIAN NOISE. OPTIMUM PERCEIVED QUALITY FOR A PARTICULAR SPEECH CLASS IN TERMS OF OBJECTIVE MEASURES IS INDICATED BY A ♣. SNR = +5 dB

	Itakura-Saito Likelihood Measure (across iterations)							
Sound Type	Original	#1	#2	#3	#4	#5	#6	#7
Silence	1.634	1.615	♣1.608	1.649	1.933	3.756	20.360	49.884
Vowel	4.020	3.721	3.445	♣3.299	3.720	8.319	121.82	
Nasal	19.814	19.154	18.416	17.656	17.009	16.593	♣15.192	15.697
Stop	7.261	6.114	4.926	3.979	♣3.822	6.889	25.515	29.694
Fricative	3.739	3.637	3.532	♣3.509	3.902	7.658	47.829	94.106
Glide	1.525	1.414	♣1.333	1.442	2.231	4.300	8.391	15.561
Liquid	9.597	8.241	6.546	4.545	2.606	♣1.676	6.381	30.001
Affricate	3.924	3.609	3.213	2.702	2.091	♣1.552	2.911	2.975
Voiced + Unvoiced	5.838	5.321	4.767	4.293	♣4.289	7.346	61.865	
Total	4.022	3.720	3.402	♣3.151	3.271	5.795	43.457	

TABLE IV
HANSEN-CLEMENTS INTERFRAME AND INTRAFRAME CONSTRAINED SPEECH ENHANCEMENT FOR WHITE GAUSSIAN NOISE. CONVERGENCE FOR A PARTICULAR SPEECH CLASS IN TERMS OF OBJECTIVE QUALITY IS INDICATED BY A ♣. SNR = +5 dB

	Itakura-Saito Likelihood Measure (across iterations)								
Sound Type	Original	#1	#2	#3	#4	#5	#6	#7	#8
Silence	1.634	1.551	1.351	1.155	1.036	0.979	0.929	♣0.884	0.901
Vowel	4.020	3.319	2.865	2.394	1.836	1.677	1.571	♣1.565	1.828
Nasal	19.814	16.490	12.397	10.523	8.682	6.840	4.929	♣3.789	5.548
Stop	7.261	6.246	4.840	3.492	2.668	1.812	1.383	♣1.129	1.435
Fricative	3.739	3.432	3.027	2.612	2.245	1.948	1.729	♣1.615	1.844
Glide	1.525	1.389	1.275	1.232	1.219	1.189	1.161	♣1.153	1.217
Liquid	9.597	6.481	3.382	2.243	1.612	1.209	0.943	♣0.926	1.211
Affricate	3.924	3.772	3.447	3.117	2.806	2.598	2.472	♣2.368	3.966
Voiced + Unvoiced	5.838	4.642	3.658	3.006	2.501	2.131	1.865	♣1.740	1.953
Total	4.022	3.026	2.441	2.069	1.801	1.611	1.457	♣1.381	1.498

TABLE V
SUMMARY OF OPTIMAL TERMINATING ITERATION ACROSS SNR FOR ADDITIVE WHITE GAUSSIAN NOISE

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

3) Termination Consistency Versus SNR: Further evaluations were performed to determine the consistency of the terminating iteration versus SNR. Table V summarizes optimum terminating points in terms of objective quality for some of the enhancement algorithms. Techniques employing only interframe constraints consistently resulted (94% occurrence) in maximum quality at the third iteration. Techniques employing interframe and intraframe constraints had a 97% occurrence of maximum quality at the seventh iteration. In addition, due to the relaxation of the iterative scheme as imposed by intraframe constraints, adjacent iterations differ only slightly in objective quality for the constrained techniques. Therefore, only minor

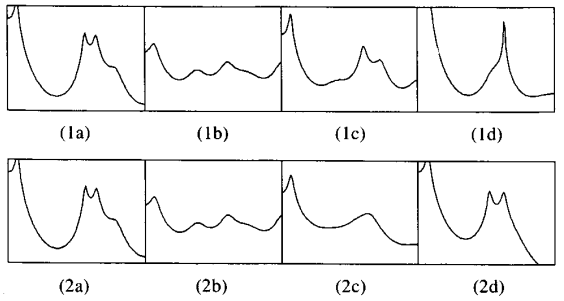


Fig. 6. Variation in vocal tract response across iterations for (1a)-(1d) unconstrained Lim-Oppenheim, and (2a)-(2d) Hansen-Clements constrained enhancement (FF-LSP:T, Auto:I) algorithms. (1a), (2a) Original. (1b), (2b) Distorted original. (1c), (2c) Four iterations. (1d), (2d) Eight iterations.

TABLE VI
COMPARISON OF ENHANCEMENT ALGORITHMS IN TERMS OF QUALITY, RELATIVE COMPLEXITY, AND RELATIVE COMPUTATIONAL RESOURCES. SNR = +5 dB, ADDITIVE WHITE GAUSSIAN NOISE DISTORTION

	Itakura-Saito Measure	Relative Complexity (1-10)	Relative Computation (1-10)	Terminating Iteration
Noisy Original	4.02			
Spectral Subtraction	3.36	2	1.5	
Lim-Oppenheim	3.15	5	3	3
(MS-LPS:T)	2.68	6	4	4
(FF-LPS:T)	1.96	7	6	3
(FF-LPS:T, Auto:I)	1.36	9	10	7

differences in speech quality would result if the iterative scheme were halted one iteration prior to optimum. The results consistently suggest that the constrained enhancement algorithms reach a maximum level of speech quality at the same iteration, independent of SNR and sound class concentrations. Thus, a convenient terminating criterion may be determined under simulated conditions and employed in actual noisy environments.

4) Vocal Tract Estimation: In addition to the problem of a terminating point dependent on speech class concentration and SNR, the unconstrained approach also suffered from undesirable movements of the LPC poles. Specifically, it was observed that as additional iterations were performed, individual formants of the speech consistently decreased in bandwidth and shifted in location as shown in Fig. 1. Fig. 6 illustrates results from a single frame of speech for the unconstrained and constrained approaches. The original and distorted original spectra are the same for both approaches. Results from 4 iterations and 8 iterations are presented for both approaches. For the unconstrained approach, the terminating point is the fourth iteration. For this example the unconstrained approach was somewhat successful in improving spectral shape, especially in the region of the second formant. However, as additional iterations were performed, spectral distortions resulted, especially with respect to bandwidth information. The constraint approach (FF-LSP:T, Auto:I) is able to eliminate these undesirable effects. The terminating point for this approach was the seventh iteration. The change in spectral shape between the seventh and eighth iterations were minor, based on visual observation and objective speech quality measures. As this figure indicates, fine characteristics of the speech spectrum result only in the later iterations.

5) Computational Issues: Discussion of algorithm performance should also address computational issues as well as algorithm complexity. Naturally, there exists a tradeoff between resulting speech quality and each algorithm's computational complexity. It is clear that iterative techniques require greater computer resources than noniterative approaches such as spectral subtraction and correlation subtraction. However, improvement in speech quality for the constraint approaches may be substantial enough to justify the additional computational requirements. In Table VI, a comparison of the enhancement algorithms are made with respect to speech quality, relative computer resources and memory requirements, and algorithm complexity. By applying constraints to the LSP parameters, a modest increase in computer resources results in a marked increase in speech quality. For example, median smoothing of the LSP parameters (MS-LSP:T) increases speech quality with only slight increases in computation and complexity. If greater resources are available, more sophisticated constraint approaches may be chosen. If memory and computational resources are available, use of the contrained approaches appears justifiable.

6) Time Versus Frequency Plots: Isometric plots of time versus frequency magnitude spectra were constructed. In Fig. 7, each line represents a 128-point frequency analysis. The top two graphs are the original and distorted cases. The lower left graph is the time versus frequency response for the unconstrained approach, terminated at the third iteration. The lower right graph is the frequency response after six iterations of an interframe plus intraframe constrained (FF-LSP:T, Auto:I) approach. These figures indicate that the considerable noise rejection achieved in the single frame noted in Fig. 6, is generally true over time.

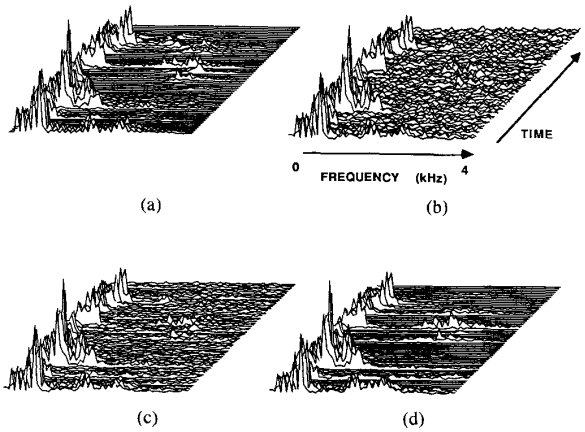


Fig. 7. Time versus frequency plots of the sentence, "Cats and dogs each hate the other." The original and distorted original (additive white Gaussian noise, SNR = +5 dB) are shown above. The lower left-hand plot is the response after three iterations of the unconstrained noncausal Wiener filtering approach. The lower right-hand plot is the frequency response after six iterations of an interframe plus intraframe constrained (FF-LSP:T, Auto:I) approach. (a) Noise free original. (b) Distorted original. (c) Unconstrained. (d) Constrained.

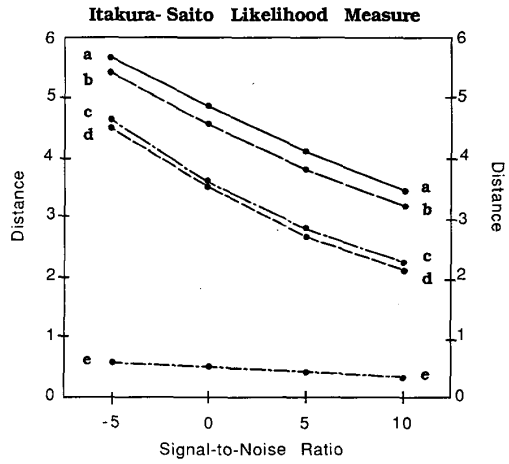


Fig. 8. Comparison of interframe and intraframe constrained enhancement algorithms for colored aircraft noise over SNR. (a) Original distorted speech. (b) Generalized unconstrained Wiener filtering. (c) Hansen-Clements: Employing interframe constraints (FF-LSP:T). (d) Hansen-Clements: Employing interframe and intraframe constraints (FF-LSP:T, Auto:I). (e) Theoretical limit: Using undistorted LPC coefficients $\vec{a}$.

B. Evaluation Using Additive Nonwhite, Nonstationary Noise

The enhancement techniques described for the white additive noise case were also tested using nonstationary, colored noise recorded from the interior of a Lockheed C130 aircraft. Estimates for the noise spectrum were made using Bartlett's method [10], [14] over long intervals.[2] Only two spectral estimates were used across each processed sentence. Further improvement is possible if noise characteristics are updated more frequently. Energy thresholds for the interframe constraints were obtained from frame energy histograms at each signal-to-noise ratio. Intraframe constraints were applied across two to three iterations. Fig. 8 and Table VII list the results of the analysis, presented in a manner consistent with the white noise descriptions. Although only Itakura-Saito measures are shown, similar improvement was observed for log-area-ratio and weighted spectral slope distance measures [7]. As seen, consistent improvement over all SNR's and speech sounds resulted, although the improvement was not as much as the white noise case.

TABLE VII
COMPARISON OF GENERALIZED UNCONSTRAINED (LIM-OPPENHEIM) AND INTERFRAME AND INTRAFRAME CONSTRAINED (HANSEN-CLEMENTS) ALGORITHMS OVER SOUND TYPES FOR SLOWLY VARYING COLORED NOISE. SNR = +5 dB

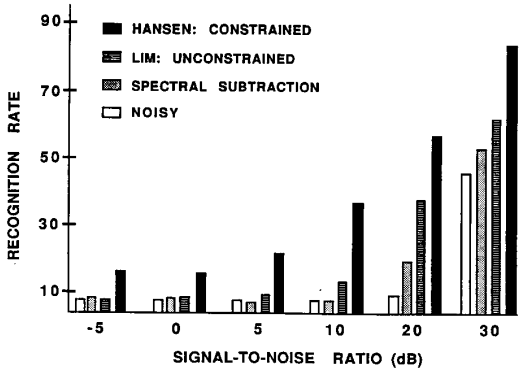
	Itakura-Saito Likelihood Measure			
Sound Type	Original	Lim-Oppenheim	Hansen-Clements	True LPC
Silence	6.63	6.33	4.32	2.03
Vowel	3.23	2.54	1.44	0.53
Nasal	4.03	3.26	2.13	0.45
Stop	1.58	1.29	0.66	0.61
Fricative	1.37	1.09	0.85	0.65
Glide	1.14	1.04	0.52	0.51
Liquid	1.22	0.55	0.22	0.18
Affricate	0.90	0.51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

C. Recognition Evaluation

One application for speech enhancement is a preprocessor for an automatic recognition system. For evaluation of enhancement algorithms in this application, as set of recognition experiments were performed, including: 1) the no noise condition (in order to set an upper limit of recognition performance), 2) distorted condition with no preprocessing (in order to set an assumed lower limit of recognition), 3) the best performing spectral subtraction preprocessing (i.e., the configuration employing either half or full-wave rectification and 1 to 5 frames of magnitude averaging, which gave the highest quality improvement for the given vocabulary), 4) unconstrained Lim-Oppenheim preprocessing, and 5) constrained preprocessing. The evaluation was performed at six levels of SNR (−5, 0, +5, +10, +20, +30 dB) for the additive white Gaussian noise degradation.

A fairly standard, isolated-word, discrete-observation hidden Markov model recognition system was used for evaluation. This system was LPC based with no embellishments. In all experiments, a five state, left-to-right model was used. The system dictionary consisted of 20 highly confusable words from a speech data base formulated for recognition evaluation in diverse environments [7]. These words are also used by Texas Instruments and Lincoln Labs to evaluate recognition systems. Subsets include /go-oh-no-hello/, /six-fix/, /wide-white/, and /degree-freeze-three/. 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 20 models employed by the HMM recognizer were trained using the for-

[2]Previous enhancement investigations employing colored aircraft background noise indicated that of the spectral estimation techniques considered (maximum entropy method, maximum likelihood method, Burg's method, Bartlett's method, Pisarenko harmonic decomposition, and the Periodogram method [10], [14]), Bartlett's method produced estimates resulting in highest improvement for this particular distortion [3], [6].



(a)

Condition (noise-free training)	RECOGNITION RESULTS Signal-to-Noise Ratio						
	Original	-5dB	0dB	+5dB	+10dB	+20dB	+30dB
Noise-free	88%						
Noisy		5%	5%	6.7%	5%	8%	49%
Spectral Subtraction		5.8%	7.1%	5%	5.4%	20%	55%
Lim-Oppenheim		5.4%	5.8%	7.5%	12.5%	41%	64%
Hansen-Clements		15%	14%	19.5%	34.5%	59%	83%

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

(b)

Fig. 9. Recognition of speech in noise performance using enhancement preprocessing in additive white Gaussian noise. †SNR = +10 dB. (a) Bar graph. (b) Table.

ward–backward algorithm. Fig. 9 presents results from five scenarios using a noise-free codebook and noise-free trained system. The 88% recognition rate clearly indicates the difficulty (confusability) of the chosen vocabulary.[3] Spectral subtraction preprocessing employed three frames of magnitude averaging. The unconstrained Lim–Oppenheim approach was terminated at the third iteration. The constrained (FF-LSP:T, Auto:I) approach was terminated at the seventh iteration. Results show that recognition was reduced to chance for noisy, spectral subtraction, and Lim–Oppenheim preprocessed speech in the SNR range of (−5, 0, 5 dB). The constrained approach resulted in improved recognition across all SNR's considered, which is quite encouraging in light of the severe levels of noise, and difficulty of dictionary employed. An increased number of training tokens as well as a less confusable vocabulary would at the very least be required if recognition in such hostile environments is to be feasible with enhancement preprocessing. In this first set of tests, all recognition training was performed on undegraded speech. This serves to model the case of training a recognizer in advance in quiet surroundings (off line) and using it in a noisy environment. As a final comparison, recognizer training was carried out using enhanced speech, which models training in the field. Three tests were performed using noisy and enhanced speech at a SNR of +10 dB. 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). The results indicate that the new constrained enhancement algorithms improve recognition performance over the unconstrained Lim–Oppenheim approach. Although the scenario of training in noise, and recognizing in noise shows improvement, the recognition system is now dedicated to a specific SNR. If noise characteristics or SNR should change over time, recognition performance would seriously degrade. The constraint approaches have been shown to be robust over varying SNR, and therefore should result in higher recognition rates with changing levels of SNR.

It is worth noting that, although performance is poor for apparently high SNR's, the SNR computation was performed over entire words. For low energy consonantal portions, the SNR's may well be 20 dB lower; and for highly confusable word pairs (e.g., /six-fix/, /go-oh-no/), errors are understandable. A detailed analysis of the error patterns bears out this hypothesis since almost all confusions were between such pairs. For example, in one noisy speech recognition test, 43 of 61 recognition errors (70%) were caused by misclassification of distinguishing consonants, many of which were leading consonants (especially fricatives). Constrained enhancement significantly reduces these errors (e.g., one test using (FF-LSP:T, Auto:I) resulted in 16 of 21 recognition errors (with 120 test tokens) caused by misclassification of distinguishing consonants). The noise-free case itself, gave 12% errors due to the difficulty of the test set, and the small number of tokens (6) per word used for training. These results show that the new constrained techniques are valuable for recognition, especially at SNR's in the +10 to +30 dB range.

IV. Conclusions

The problem of enhancing speech degraded by additive white and slowly varying colored background noise was addressed. In addition, algorithm performance as a preprocessor for speech recognition was also considered. The set of enhancement algorithms presented impose interframe and intraframe constraints on the input speech signal and were shown to be useful in enhancing speech for human listeners, and as a preprocessor for recognition in noisy environments. Interframe constaints ensure more speech-like formant trajectories than those found in the unconstrained approach and thus reduce pole jitter on a frame-to-frame basis. Intraframe constraints ensure relaxation of the iterative scheme so that overall maximum speech quality is obtained across all classes of speech. In order to increase numerical accuracy, reduce computational requirements, and eliminate inconsistencies in pole ordering across frames, the line spectral pair (LSP) transformation of the LPC coefficients was used to implement many of the constraint requirements. The new set of constrained algorithms were shown to be effective in several domains. First, improvement in objective speech quality measures was shown. Improved LPC parameter estimation was also observed. Second, the algorithms were extended and shown to be effective on nonstationary colored noise. Third, the algorithms were shown to improve all segments of speech for both white and nonwhite noise. Fourth, the current algorithms have been shown to possess a consistent terminating criterion. Specifically, the optimum terminating iteration was shown to be consistent over all speech sound classes, and virtually all tested SNR's. Finally, the constrained algorithms have shown improvement as a preprocessor for speech recognition. Their ability to bring performance up to an acceptable level in SNR's between −5 and +5 dB is questionable. This may be due in part to the difficulty of the highly confusable test set, the small number of tokens per word used for training, and the observation that SNR's in low energy consonantal portions which discriminate confusable pairs may well be 20 dB lower. Recognition improvement in SNR's between +10 and +30 dB may be

[3]On isolated digit tasks in quiet, the recognizer consistently scored 100% [7].

large enough to warrant enhancement preprocessing for recognition.

APPENDIX
SPEECH DATA USED IN THE EVALUATION

All sentences were sampled at 8000 samples/s.

Speech Data

S1: *The pipe began to rust while new.*	Female Speaker
S2: *Thieves who rob friends deserve jail.*	Male Speaker
S3: *Add the sum to the product of these three.*	Female Speaker
S4: *Open the crate but don't break the glass.*	Male Speaker
S5: *Oak is strong and also gives shade.*	Male Speaker
S6: *Cats and dogs each hate the other.*	Male Speaker

REFERENCES

[1] S. F. Boll, "Suppression of acoustic noise in speech using spectral subtraction," *Trans. Acoust., Speech, Signal Processing*, vol. ASSP-27, pp. 113-120, Apr. 1979.

[2] J. R. Crosmer, "Very low bit rate speech coding using the line spectrum pair transformation of the LPC coefficients," Ph.D. dissertation, School Elec. Eng., Georgia Inst. Technology, Atlanta, June 1985.

[3] J. H. L. Hansen and M. A. Clements, "Enhancement of speech degraded by nonwhite additive noise," Final Tech. Rep. DSPL-85-6, Georgia Inst. Technology, Atlanta, Aug. 1985.

[4] J. H. L. Hansen and M. A. Clements, "Objective quality measures applied to enhanced speech," in *Proc. Acoust. Soc. Amer.*, 110th Meeting, C11 (Nashville, TN), Nov. 1985.

[5] J. H. L. Hansen and M. A. Clements, "Iterative speech enhancement with spectral constraints," in *Proc. 1987 IEEE ICASSP* (Dallas, TX), Apr. 1987, pp. 189-192.

[6] J. H. L. Hansen and M. A. Clements, "Constrained iterative speech enhancement with application to automatic speech recognition," in *Proc. 1988 IEEE ICASSP* (New York, NY), Apr. 1988, pp. 44.S12.9.1-4.

[7] J. H. L. Hansen, "Analysis and compensation of stressed and noisy speech with application to robust automatic recognition," Ph.D. dissertation, Georgia Inst. Technology, July 1988.

[8] J. H. L. Hansen and M. A. Clements, "Stress and noise compensation algorithms for robust automatic speech recognition," in *Proc. 1989 IEEE ICASSP* (Glasgow, Scotland), May 1989, pp. 266-269.

[9] F. Itakura, "Line spectrum representation of linear predictive coefficients of speech signals," *J. Acoust. Soc. Amer.*, vol. 57, no. S35(A), 1975.

[10] S. Kay, *Modern Spectral Estimation: Theory and Application*. Englewood Cliffs, NJ: Prentice-Hall, 1988.

[11] J. S. Lim and A. V. Oppenheim, "All-pole modeling of degraded speech," *Trans. Acoust., Speech, Signal Processing*, vol. ASSP-26, pp. 197-210, June 1978.

[12] J. S. Lim, Ed., *Speech Enhancement*. Englewood Cliffs, NJ: Prentice-Hall, 1983.

[13] F. J. Malkin and K. A. Christ, "Human factors engineering assessment of voice technology for the light helicopter family," U.S. Army Human Eng. Lab. Tech. Rep., June 1985, pp. 1-20.

[14] S. L. Marple, *Digital Spectral Analysis with Applications*. Englewood Cliffs, NJ: Prentice-Hall, 1987.

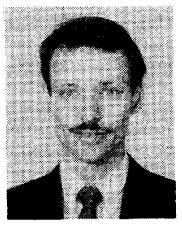
[15] B. R. Musicus, "An iterative technique for maximum likelihood parameter estimation on noisy data," S. M. thesis, Massachusetts Inst. Technology, Cambridge, MA, 1979.

[16] J. M. Ortega and W. C. Rheinbolt, *Iterative Solutions of Nonlinear Equations in Several Variables*. New York: Academic, 1970.

[17] S. R. Quackenbush, T. P. Barnwell, and M. A. Clements, *Objective Measures of Speech Quality*. Englewood Cliffs, NJ: Prentice-Hall, 1988.

[18] C. A. Simpson, "Speech variability effects on recognition accuracy associated with concurrent task performance by pilots," Psycho-Linguistic Research Associates, Tech. Rep., Apr. 1985, pp. 1-15.

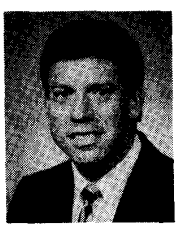
[19] F. K. Soong and B. H. Juang, "Line spectrum pair (LSP) and speech compression," in *Proc. 1984 IEEE ICASSP* (San Diego, CA), Mar. 1984, pp. 705-708.



John H. L. Hansen (S'81-M'82) was born in Plainfield, NJ, on November 17, 1959. He received the M.S. and Ph.D. degrees in electrical engineering from the Georgia Institute of Technology, Atlanta, in 1983 and 1988, respectively, and the B.S.E.E. degree from Rutgers University in 1982.

He was employed by the RCA Solid State Division from 1981 to 1982. In 1988, he joined the faculty of Duke University as an Assistant Professor in the Department of Electrical Engineering. His research interests include digital signal processing, speech analysis, speech enhancement, robust speech recognition, and statistical pattern recognition.

Dr. Hansen is presently serving as Secretary and member of the ADCOM committee for the IEEE Communication Society of North Carolina. He is coauthor of *Digital Processing of Speech Signals* (Macmillan), and received the National Science Foundation's Research Initiation Award in 1990. He is a member of the Acoustical Society of America, Sigma Xi, Tau Beta Pi, Eta Kappa Nu, and Pi Mu Epsilon.



Mark A. Clements (M'82-SM'89) received the S.B., S.M., and Sc.D. degrees, all in electrical engineering and computer science, from the Massachusetts Institute of Technology in 1976, 1978, and 1982, respectively.

In 1982, he joined the Faculty of the School of Electrical Engineering at the Georgia Institute of Technology where he is now an Associate Professor. His interests include speech analysis, robust automatic speech recognition, enhancement of speech for the hearing impaired, pattern recognition, and spectral estimation. He is the author of numerous papers and technical reports and is coauthor of *Objective Measures of Speech Quality* (Prentice-Hall).

Dr. Clements currently is serving on the IEEE SP Speech Technical Committee.