Windowless Analysis of Speech for Automatic Recognition

by

Sungjae Lim

and

Mark A. Clements

Manuscript prepared for the *IEEE Trans. on Acoustics, Speech, and Signal Processing*

Abstract

Traditional hidden Markov model speech recognition is generally based upon a set of parameters which are extracted at discrete intevals. Such an analysis necessitates use of a discrete-transition hidden Markov model in which the underlying states can change only at intervals related to the frame rate of the analysis. The exact locations of the analysis windows can influence the front-end outputs. As a result, inconsistent performance can often be observed in discriminating words which differ only in short duration cues. In the current study, methods are explored which circumvent this framing effect by allowing state trasitions to occur at each sample. Efficient methods for implementing this strategy are derived, and testing of a variety of procedures using a set of highly confusable utterances is reported. Significantly superior performance was demonstrated both for quiet and noisy conditions.

1 Introduction

Over the past few years, the method of choice for many speech recognition applications has been on hidden Markov modelling. Steady improvement has been reported in such areas as speaker independence, noise handling, training and response times, as well as general performance. The first HMM based systems modeled speech as a discrete state discrete trial Markov process with discrete observations. More recently, models which allow a continuous distribution of observations have been presented. Throughout all these models, however, the assumption remains that sampling the parameterization of the speech (e.g., spectral or LPC based parameters) is only necessary every 10 to 30 milliseconds. When words differ only by a short duration interior consonant, however, the exact placement of the analysis windows can have an impact on performance.

The motivation for the current study came from our observations that although general performance of a recognizer may not depend highly on the exact placement of frames, the detailed error patterns often would. The methods explored are attempts at eliminating the apparent framing artifacts by, in essence, extracting a set of parameters for every sample of the digital speech. The recognition algorithm can then be considered a close approximation to a continuous transition hidden Markov model. This approach would not be feasible were it not for efficient algorithms we have been formulated for this specific problem.

In this paper, we will first discuss the aspects of hidden Markov models which are conducive to this strategy and discuss the issues involved in training, and recognition. Second we will describe three parameter extraction methods, one of which relies on a novel utilization of Kalman filtering, with others two involving more classical procedures. Third, we will examine experimental results and discuss the conclusion which can be drawn.

2 The Hidden Markov Model

A. Definitions :

Consider a discrete state discrete transition hidden Markov model for each pattern to be recognized. Assume the observations are drawn from a finite alphabet of size **M** , and a new observation is made for every sample of the digital speech. This would imply some form of vector quantizer continuously outputting a codeword sequence. Although the form and implementation of this process will be described in detail in section 3, for all systems considered, enough memory existed in the analysis to produce long sequences of the same codeword in a segment of an utterance. The importance of this result will become apparent below.

Denote the number of states in a model by n .

$$
\begin{aligned}
\pi_i &= \textit{probability the model starts in state } i, \\
\underline{\Pi^T} &= [\pi_1, \pi_2, \ldots, \pi_n] \\
\mathbf{A} &= \textit{transition probability matrix, where :} \\
a_{ij} &= \textit{probability of transition from state } i \textit{ to state } j \\
&\quad \textit{in one trial; } i, j = 1, 2, \ldots, n. \\
\mathbf{B} &= \textit{observation probability matrix where} \\
&\quad b_{jk} = \textit{probability of observing codeword } k \\
&\quad \textit{given state } j. \\
\mathbf{O}(t) &= \textit{codeword observed at time } t,\ 1 \leq t \leq F \\
\mathbf{R}(t) &= \textit{observation matrix, consisting of :} \\
&\quad \mathbf{R}(t) = diag[b_1(\mathbf{O}(t)), \ldots, b_n(\mathbf{O}(t))]
\end{aligned}
$$

For a given model **M**, and observations $\mathbf{O}(1), \mathbf{O}(2), \ldots, \mathbf{O}(F)$, we define

$$
\begin{aligned}
\underline{\alpha}^T(t) &= [\alpha_1(t), \ldots, \alpha_n(t)] \\
\alpha_i(t) &= prob[\mathbf{O}(1), \ldots, \mathbf{O}(t); \textit{state } i \textit{ at } t] \\
\underline{\beta}^T(t) &= [\beta_1(t), \ldots, \beta_n(t)] \\
\beta_i(t) &= prob[\mathbf{O}(t+1), \ldots, \mathbf{O}(F); \textit{ state } i \textit{ at } t]
\end{aligned}
$$

Then the probability that we observe the sequence from the model is

$$
Pr[\mathbf{O}(1), \ldots, \mathbf{O}(F)] = \sum_{i=1}^{n} \alpha_i(t)\beta_i(t) \qquad (1)
$$

We can rewrite $\underline{\alpha}(t)$, $\underline{\beta}(t)$, and Eq.(1) in matrix form such that

$$\begin{aligned} Pr[\mathbf{O}(1),\ldots,\mathbf{O}(F)] &= \underline{\Pi}^T\mathbf{R}(1)\mathbf{A}\mathbf{R}(2)\mathbf{A}\cdots\mathbf{A}\mathbf{R}(F)\underline{\beta}(F) \quad (2)\\ \underline{\alpha}^T(t) &= \underline{\Pi}^T\mathbf{R}(1)\mathbf{A}\mathbf{R}(2)\cdots\mathbf{A}\mathbf{R}(t) \quad (3)\\ \underline{\beta}^T(t) &= \mathbf{A}\mathbf{R}(t+1)\mathbf{A}\mathbf{R}(t+2)\cdots\mathbf{A}\mathbf{R}(F)\underline{\beta}(f) \quad (4) \end{aligned}$$

If the model is constrained to the left-to-right, **A** will be upper triangular. If the model demands the system to start in state 1 and end in state n, then

$$\begin{aligned} \underline{\Pi}^T &= [1,0,\ldots,0]\\ \underline{\beta}^T(F) &= [0,\ldots,0,1] \end{aligned} \quad (5)$$

B. Recognition :

For a given model, one needs to compute the probability of the observations. This can be accomplished, of course, through evaluation of Eq.(2). In our system, F is normally such a large number that dirrect evaluation of Eq.(2) would require tremendous amount of computation. In order to reduce this computational burden, we make use of the fact that usually a long run of the same codewords are observed, which makes Eq.(2) several long runs of the same matrix multiplications, and the constraint that the model be left-to-right which makes **A** upper-triangular. Let's assume that the codewords at time $t+1$ through $t+m$ are same. Then the partial product of Eq.(2) for the period of time,

$$[\mathbf{A}\mathbf{R}(t+1)\mathbf{A}\mathbf{R}(t+2)\cdots\mathbf{A}\mathbf{R}(t+m)],$$

is equal to

$$[\mathbf{A}\mathbf{R}(t+m)]^m$$

Since the matrix **A** is upper-triangular and $\mathbf{R}(t+m)$ is diagonal, the product, $[\mathbf{A}\mathbf{R}(t+m)]^m$ is an upper-triangular matrix. The upper-triangular matrix has a nice property that it can be diagonalized if the diagonal elements are distinct. In our case, if we assume that the diagonal elements of $\mathbf{A}\mathbf{R}(t+m)$ are distinct, it can be diagonalized in such a form that

$$\mathbf{A}\mathbf{R}(t+m) = \mathbf{P}\mathbf{D}\mathbf{P}^{-1} \quad (6)$$

where $\mathbf{D}$ is diagonal with its elements same as the diagonal elements of $\mathbf{AR}(t+m)$, $\mathbf{P}$ is a upper-triangular matrix with its diagonal elements equal to 1. Therefore,

$$[\mathbf{AR}(t+m)]^m = \mathbf{PD}^m\mathbf{P}^{-1} \tag{7}$$

And $\underline{\alpha}(t+m)$ can be computed directly from $\underline{\alpha}(t)$ without computing intermediate $\underline{\alpha}'s$ at $t+1, t+2, \ldots$, and $t+m-1$, that is,

$$\begin{aligned} \underline{\alpha}(t+m) &= \underline{\alpha}(t)[\mathbf{AR}(t+m)]^m \\ &= \underline{\alpha}(t)\mathbf{PD}^m\mathbf{P}^{-1} \end{aligned} \tag{8}$$

It seems that obtaining the matrices, $\mathbf{P}$ and $\mathbf{P}^{-1}$, require time-consuming computation, especially when the dimension of the matrix is large. This, however, is not so in our case. In fact, there exist very efficient ways using the property that $[\mathbf{AR}(t+m)]$ is upper-triangular. The efficient methods to compute $\mathbf{P}$ and $\mathbf{P}^{-1}$ are shown in Appendix A.

C. Training Algorithms :

In the previous section, we have shown an efficient way of computing $\underline{\alpha}'s$ without computing the intermediate ones when a long run of the same codewords are observed. $\underline{\beta}'s$ can also be computed in the same way. In this section, two different training methods are introduced in which we make use of the same method to efficiently carry out the restimation. The first one, denoted as " Algorithm 1 ", is strictly based on the Baum-Welch reestimation algorithm, while the second one,denoted as " Algorithm 2", is slightly varied version and yet performs better.

1). Algorithm 1 :

The Baum-Welch reestimation algorithm states that the estimates of a_{ij} and $b_j(v)$, denoted as $\hat{a}_{ij}$ and $\hat{b}_j(v)$ respectively, are updated at each iteration based on the previous estimates as follows :

$$\hat{a}_{ij} = \frac{\gamma_{ij}}{\gamma_i} \tag{9}$$

$$\hat{b}_j(v) = \frac{\sum_{t \in O(t)=k} \alpha_j(t)\beta_j(t)}{\sum_{t=1}^{F} \alpha_j(t)\beta_j(t)} \tag{10}$$

where

$$\gamma_{ij} = \frac{1}{p}\sum_{t=1}^{F-1} \alpha_j(t)\hat{a}_{ij}\hat{b}_j(O(t+1))\beta_j(t+1) \tag{11}$$

$$\gamma_i = \sum_{j=1}^{n} \gamma_{ij} \tag{12}$$

Let's consider the computation of γ_{ij}. If $O(k+1) = O(k+2) = \cdots = O(k+m)$, then $b_j(O(k+1)) = b_j(O(k+2)) = \cdots = b_j(O(k+m))$. Thus the partial summation of Eq.(11) for $k \leq t \leq k+m-1$, denoted as $\gamma_{ij}(k, k+m-1)$, can be written as

$$\gamma_{ij} = \frac{1}{p}[a_{ij}b_j(O(k+1))] \sum_{t=k}^{k+m-1} \alpha_i(t)\beta_j(t+1) \tag{13}$$

Computation of Eq.(13) in a straight forward way requires $\alpha_i(t)$ and $\beta_j(t+1)$ to be computed at $t = k, k+1, \ldots, k+m-1$, With a different manipulation, which will be shown in the following, this can be avoided and a lot of computation can also be saved,especially when m is large. First let's express $\underline{\alpha}(t)$ and $\underline{\beta}(t+1)$ for $k \leq t \leq k+m-1$ in terms of $\underline{\alpha}(k)$ and $\underline{\beta}(k+m)$ as follows ;

$$\underline{\alpha}^T(t) = \underline{\alpha}^T(k)[\mathbf{AR}(k+1)]^{t-k} \tag{14}$$

$$\underline{\beta}(t+1) = [\mathbf{AR}(k+1)]^{m-t+k-1}\underline{\beta}(k+m) \tag{15}$$

Then

$$\begin{aligned} \textstyle\sum_{t=k}^{k+m-1} \alpha_i(t)\beta_j(t+1) &= \textstyle\sum_{t=k}^{k+m-1}[\underline{\alpha}(t)\underline{\beta}^T(t+1)]_{ij} \\ &= \textstyle\sum_{t=k}^{k+m-1}[((\mathbf{AR})^{t-k})^T\underline{\alpha}(k)\underline{\beta}^T(k+m)((\mathbf{AR})^{m-t+k-1})^T]_{ij} \end{aligned} \tag{16}$$

where $[*]_{ij}$ denotes i-j component of matrix $[*]$ and $\mathbf{R} = \mathbf{R}(k+1)$ for simplicity. As shown in the previous section, $\mathbf{AR}$ can be decomposed such that $\mathbf{AR} = \mathbf{PDP}^{-1}$. Then Eq.(16) can be rewritten as follows;

$$\begin{aligned} &\textstyle\sum_{t=k}^{k+m-1} \alpha_i(t)\beta_j(t+1) \\ &= \textstyle\sum_{t=k}^{k+m-1}[\mathbf{P}^{-T}\mathbf{D}^{t-k}\mathbf{P}^T\underline{\alpha}(k)\underline{\beta}^T(k+m)\mathbf{P}^{-T}\mathbf{D}^{m-t+k-1}\mathbf{P}^T]_{ij} \\ &= [\mathbf{P}^{-T}(\textstyle\sum_{t=k}^{k+m-1}\mathbf{D}^{t-k}\mathbf{P}^T\underline{\alpha}(k)\underline{\beta}^T(k+m)\mathbf{P}^{-T}\mathbf{D}^{m-t+k-1})\mathbf{P}^T]_{ij} \end{aligned} \tag{17}$$

If we let

$$\hat{\underline{\alpha}}(k) = \mathbf{P}^T \underline{\alpha}(k) \tag{18}$$

$$\hat{\underline{\beta}}^T(k+m) = \underline{\beta}^T(k+m)\mathbf{P}^{-T}, \tag{19}$$

then Eq.(17) can be written more neatly such that

$$\sum_{t=k}^{k+m-1} \alpha_i(t)\beta_j(t+1) = [\mathbf{P}^{-T}\mathbf{M}\mathbf{P}^T]_{ij}, \tag{20}$$

where

$$\begin{aligned} \mathbf{M} &= \textstyle\sum_{t=k}^{k+m-1} \mathbf{D}^{t-k}\mathbf{P}^T\underline{\alpha}(k)\underline{\beta}^T(k+m)\mathbf{P}^{-T}\mathbf{D}^{m-t+k-1} \\ &= \textstyle\sum_{t=k}^{k+m-1} \mathbf{D}^{t-k}\hat{\underline{\alpha}}(k)\hat{\underline{\beta}}^T(k+m)\mathbf{D}^{m-t+k-1} \end{aligned} \tag{21}$$

Now let's consider the computation of $\mathbf{M}$. The $i-j^{th}$ component of $\mathbf{M}$, M_{ij}, can be expressed as

$$\begin{aligned} M_{ij} &= \textstyle\sum_{t=k}^{k+m-1} d_i^{t-k}\hat{\alpha}_i(k)\hat{\beta}_j(k+m)d_j^{m-t+k-1} \\ &= (\hat{\alpha}_i(k)\hat{\beta}_j(k+m))\textstyle\sum_{t=k}^{m-t+k-1} d_i^{t-k}d_j^{m-t+k-1} \end{aligned} \tag{22}$$

Since it was assumed that $d_i \neq d_j$ if $i \neq j$,the summation can be reduced such that

$$\sum_{t=k}^{k+m-1} d_i^{t-k}d_j^{m-t+k-1} = \begin{cases} \frac{d_j^m - d_i^m}{d_j - d_i} & \text{for } i \neq j \\ m(d_i)^{m-1} & \text{for } i = j \end{cases} \tag{23}$$

Thus

$$M_{ij} = \begin{cases} \frac{d_j^m - d_i^m}{d_j - d_i}\hat{\alpha}_i(k)\hat{\beta}_j(k+m) & \text{for } i \neq j \\ md_i^{m-1}\hat{\alpha}_i(k)\hat{\beta}_j(k+m) & \text{for } i = j \end{cases} \tag{24}$$

In summary,

$$\gamma_{ij}(k, k+m-1) = \frac{1}{p}[\mathbf{P}^{-T}\mathbf{M}\mathbf{P}^T]_{ij}(a_{ij}b_j(O(k+1))) \tag{25}$$

It is worth to be noted that only the upper triangular portions of $\mathbf{M}$ are necessary to be computed, since we only need $\gamma_{ij}(k, k+m-1)$, for $i \leq j$ and the matrices, $\mathbf{P}^{-T}$ and $\mathbf{P}^T$, are lower triangular.

Secondly, let's consider the numerator of Eq.(10) for the reestimation of $b_j(v)$. Under the same assumption that $O(k+1) = O(k+2) = \cdots = O(k+$

$m) = v$,the partial summation of the numerator, $\sum_{t\in O(t)=v} \alpha_j(t)\beta_j(t)$, for $k+1 \leq t \leq k+m$ can be expressed in terms of $\underline{\alpha}(k)$ and $\underline{\beta}(k+m)$,

$$\begin{aligned} &\textstyle\sum_{t=k+1}^{k+m} \alpha_j(t)\beta_j(t) = \sum_{t=k+1}^{k+m} [\underline{\alpha}(t)\underline{\beta}^T(t)]_{jj} \\ &= [\mathbf{P}^{-T} \textstyle\sum_{t=k+1}^{k+m} (\mathbf{D}^{t-k}\mathbf{P}^T\underline{\alpha}(k)\underline{\beta}^T(k+m)\mathbf{P}^{-T}\mathbf{D}^{k+m-t})\mathbf{P}^T]_{jj} \end{aligned} \tag{26}$$

Eq.(26) is very similar to Eq.(17), and can be evaluated similarly. In fact, if we denote the term in the summation of Eq.(26) as $\hat{\mathbf{M}}$, i.e.,

$$\hat{\mathbf{M}} = \sum_{k+1}^{k+m} \mathbf{D}^{t-k}\mathbf{P}^T\underline{\alpha}(k)\underline{\beta}^T(k+m)\mathbf{P}^{-T}\mathbf{D}^{k+m-t} \tag{27}$$

It can be observed that $\hat{\mathbf{M}}$ is the product of $\mathbf{D}$ and $\mathbf{M}$,i.e.,

$$\hat{\mathbf{M}} = \mathbf{DM} \tag{28}$$

Hence, once $\mathbf{M}$ is obtained to compute $\gamma_{ij}(k, k+m-1)$, Eq.(22) can be computed with only a few more computation as follows;

$$\sum_{t=k+1}^{k+m} \alpha_j(t)\beta j(t) = [\mathbf{P}^{-T}\mathbf{DMP}^T]_{jj} \tag{29}$$

As mentioned earlier, in the partial summations involved for the restimations of a_{ij} and $b_j(v)$, $\underline{\alpha}'s$ and $\underline{\beta}'s$ are not required to be computed at every time unit. For example, if we consider the assumption given above that $O(t+1) = O(t+2) = \cdots = O(t+m)$, only $\underline{\alpha}(k)$ and $\underline{\beta}(k+m)$ are required in the partial summations, that is, all the intermediate $\underline{\alpha}'s$ and $\underline{\beta}'s$ do not have to be computed, which contributes to the great saving of computation.

2). Algorithm 2 :

The algorithm presented here can be considered as the sampling version of Baum-Welch restimation algorithm. Unlike the Baum-Welch algorithm, which is formulated by Eq.(9) and Eq.(10), in the new algorithm only the samples of γ_{ij} are used. Eq.(11) can be rewritten as follows;

$$\gamma_{ij} = \sum_{t=1}^{F-1} \gamma_{ij}(t) \tag{30}$$

where

$$\gamma_{ij}(t) = \frac{1}{p}\alpha_i(t)\bar{a}_{ij}\bar{b}_j(O(t+1))\beta_j(t+1) \tag{31}$$

The restimation equations (9) and (10) can also be written in terms of $\gamma_{ij}(t)$.

$$\bar{a}_{ij} = \frac{\sum_{t=1}^{F-1} \gamma_{ij}(t)}{\sum_{j=1}^{n}(\sum_{t=1}^{F-1} \gamma_{ij}(t))} \tag{32}$$

$$\bar{b}_j(v) = \frac{\sum_{t \in O(t)=v}(\sum_{i=1}^{n} \gamma_{ij}(t-1))}{\sum_{t=1}^{F}(\sum_{i=1}^{n} \gamma_{ij}(t-1))} \tag{33}$$

In the new algorithm, we sample $\gamma_{ij}(t)$ at every k^{th} time unit, and assume that it stays same during the sampling interval. In other words, if $\gamma_{ij}(t)$ is sampled at $t = 1, k+1, 2k+1, \ldots$, then we assume that

$$\begin{array}{llll} \gamma_{ij}(1) & = \gamma_{ij}(2) & = \cdots & = \gamma_{ij}(k) \\ \gamma_{ij}(k+1) & = \gamma_{ij}(k+2) & = \cdots & = \gamma_{ij}(2k) \\ \gamma_{ij}(2k+1) & = \gamma_{ij}(2k+2) & = \cdots & = \gamma_{ij}(3k) \\ & \vdots & & \end{array} \tag{34}$$

Under this assumption and the assumption that $F = mk$ for some integer m, Eq.(32) becomes as follows,

$$\bar{a}_{ij} = \frac{\sum_{\tau=0}^{m-1} \gamma_{ij}(\tau k+1)}{\sum_{j=1}^{n} \sum_{\tau=0}^{m-1} \gamma_{ij}(\tau k+1)} \tag{35}$$

which can be seen as the sampled version. This algorithm is not proven mathematically to converge, but it has shown experimentally that it not only converges but also gives better results than the conventional Baum-Welch algorithm. It seems that this algorithm has a smoothing property which enables the algorithm to find a better local maximum point.

3 Front-End Analysis

The approach we are adopting is based on a linear model of speech which is time-invariant over short intervals. This is the traditional model often used in speech recognition and coding applications. However, we allow for natural smooth changes occuring in the system as well as additive uncorrelated noise. Our linear model may also have explicit modeling of time-varying system parameters. Since many phonemes are characterized by a particular evolution in time rather than by steady-state or target spectra, this model is more

powerful than more traditional ones. In particular our model is :

$$\begin{cases} \mathbf{X}(k) &= \Phi(k)\mathbf{X}(k-1) + \Gamma(k)w(k) \\ s(k) &= \mathbf{H}^T\mathbf{X}(k) + v(k) \end{cases} \tag{36}$$

where the vector $\mathbf{X}(k) = [x(k)x(k-1)\cdots x(k-p+1)]^T$, $x(k)$ is the speech without noise, $w(k)$ the noise input and $\Gamma(k)$ its gain, $\mathbf{H}^T = [1,0,0,\cdots,0]$, $v(k)$ the additive noise, and $\Phi(k)$ characterizes the time-varying vocal-tract filter.

Systems similar to this have been used to model many varied signals arising in sonar, heart monitoring, aircraft control,etc.. In the linear prediction synthesis model $\Phi(k)$ remains constant over 10 to 30 millisecond intervals, and $v(k)$ is zero. In the LPC analysis model, $v(k)$ is generally assumed to be zero so that $\Phi(k)$ can be estimated every 10 to 30 milliseconds. Recursive linear least square estimation based on our model falls within the general area of Kalman filtering, which allows one to efficiently compute the least squares estimate of $\mathbf{X}(k)$ from the least squares estimate of $\mathbf{X}(k-1)$ and $s(k)$. The property we wish to exploit is that if we have modeled the system correctly, the prediction error, $\epsilon(k)$, would be white, and it should have a predictable ratio of its power to the unfiltered signal's power. If there are L possible models from which the observed signals was generated, this idea can be used for computing the relative likelihood of each model given the observed signal. In the following our front-end process is explained in detail on the Kalman filtering process followed by decision making process.

3.1 Kalman Filtering

In the Kalman filtering process, we have L distinct competing models, each of which has the form,

$$\begin{cases} \mathbf{X}(k) &= \Phi\mathbf{X}(k-1) + \Gamma(k)\mathbf{w}(k) \\ s(k) &= \mathbf{H}\mathbf{X}(k) + v(k) \end{cases} \tag{37}$$

where

$$\Phi = \begin{bmatrix} a(1) & a(2) & \cdots & a(p) \\ 1 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 0 \end{bmatrix}$$

$$\mathbf{H} = [100\cdots0]$$

$$Ev(k) = 0 \; Ev(k)v(l) = \sigma_v^2(k)\delta_{kl}$$

$$E\mathbf{w}(k) = [0, 0, \cdots, 0]^T$$

$$E\mathbf{w}(k)\mathbf{w}(l) = \begin{bmatrix} 1 & 0 & \cdots & 0 \\ 0 & 0 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 0 \end{bmatrix} \delta_{kl}$$

$$\Gamma(k) = \begin{bmatrix} g(k) & 0 & \cdots & 0 \\ 0 & 0 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 0 \end{bmatrix}$$

and $a(1), a(2), \ldots, a(p)$ are linear prediction coefficients which characterize the model. This model results in the following time-recursive formula which gives the linear least squares estimate of $\mathbf{X}(k)$ given $s(k-1), s(k-2), \ldots, s(0)$.

$$\epsilon(k) = s(k) - \mathbf{H}\hat{\mathbf{X}}(k|k-1) \tag{38}$$

$$\sigma_\epsilon^2(k) = \mathbf{H}\mathbf{P}(k|k-1)\mathbf{H}^T + \sigma_v^2(k) \tag{39}$$

$$\mathbf{M}(k) = \frac{1}{\sigma_v^2}\mathbf{P}(k|k-1)\mathbf{H}^T \tag{40}$$

$$\hat{\mathbf{X}}(k|k) = \hat{\mathbf{X}}(k|k-1) + \mathbf{M}(k)\epsilon(k) \tag{41}$$

$$\hat{\mathbf{X}}(k+1|k) = \Phi\hat{\mathbf{X}}(k|k) \tag{42}$$

$$\mathbf{P}(k|k) = \mathbf{P}(k|k-1) - \mathbf{M}(k)\mathbf{M}^T(k)\sigma_v^2 \tag{43}$$

$$\mathbf{P}(k+1|k) = \Phi\mathbf{P}(k|k)\Phi^T + \Gamma(k)\Gamma^T(k) \tag{44}$$

where $\epsilon(k)$ is the innovations sequence, $\sigma_\epsilon^2(k)$ the variance of the innovations, $\mathbf{M}(k)$ the Kalman gain, and $\mathbf{P}(k|r)$ the covariance of the estimate error $\mathbf{X}(k) -$

$\hat{\mathbf{X}}(k|r)$. The initial condition is given as follows;

$$\begin{cases} \hat{\mathbf{X}}^T(0|0) = [s(0)s(-1)\cdots s(-p+1)] \\ \mathbf{P}(0|0) = \sigma_v^2(0)\mathbf{I} \end{cases} \tag{45}$$

With the innovations sequence obtained from each model, a likelihood test is performed in a recursive manner. If we denote $\epsilon_i(k)$ the innovation produced by model i at time k and $p_i(k)$ the probability that model i generate $s(k)$, then

$$p_i(k) = \frac{N(\epsilon_i(k), \sigma_{\epsilon_i}^2(k))p_i(k-1)}{\sum_{j=1}^{L} N(\epsilon_j(k), \sigma_{\epsilon_j}^2(k))p_j(k-1)} \tag{46}$$

where $\sigma_{\epsilon_j}^2(k)$ is the variance of $\epsilon_j(k)$ when model j is correct, and $N(a,b)$ represents the Gaussian density of zero mean with the variance b evaluated at a. We then choose the model with the largest p.

4 Experiments

Several recognition experiments were performed with clean speech, noisy speech of $SNR = 26dB$, and of $SNR = 20dB$. The isolated words used in the experiments are 'break', 'change', 'degree', 'eight', 'eighty', 'enter', 'fifty', 'fix', 'six', 'go'. Each word has 12 utterances, 6 of which were used for the training of HMM's. Each utterance was passsed through Kalman-filtering process with 3 different level of white Gaussian noises as stated above, which produced 3 different sets of codewords, one for clean speech, one for the noisy speech of $SNR = 26dB$, and one for the noisy speech of $SNR = 20dB$. In the Kalman-filtering process, the variances of the generating noise and the additive noise were updated at every 80 samples, and the initail conditions were reset accordingly at the same time. The filter order was 14 for each of the 64 different filters.

A. With Clean Speech : 2 errors out of 120 = 1.7 1 error from the set used for training : 'six' recognized as 'fix'
1 error from the set not used for training : 'eight' recoginzed as 'eighty'
B. With Noisy Speech of $SNR = 26dB$: 8 errors out of 120 = 6.7 0 error

from the set used for training

8 errors from the set not used for trianing : 'eight' recoginized as 'eighty' (4), 'fix' recognized as 'six' (1), 'six' recognized as 'fix' (3).

C. With Noisy Speech of $SNR = 26dB$ and Clean Speech :

i). the recognition of noisy speech : 6 errors out of 120 = 5 0 error from the set used for training

6 errors from the set not used for training : 'eight' recognized as 'eighty' (3), 'fix' recognized as 'six' (1), 'six' recognized as 'fix' (2)

ii). the recognition of clean speech : 7 errors out of 120 = 5.8 2 errors from the set used for training : 'eight' recognized as 'eighty', and 'six' recognized as 'fix'

5 errors from the set not used for training : 'eight' recognized as 'eighty' (5)

It is interesting to note that the models trained with both clean and noisy speech give higher recognition rate for noisy speech (compare the results of **B** and **C** i).) than the ones trained with only noisy speech, while giving lower recognition rate for clean speech (compare the results of **A** and **C** ii).) than the ones trained with only clean speech. It may be interpreted as clean speech giving positive information for the trtaining of noisy speech models, and noisy speech giving negative information for the training clean speech models. This behavior has been observed in several occasions. More comprehensive experiments are to be done with larger vocavulary and various SNR's.

Iterative Speech Enhancement With Spectral Constraints

John H. Hansen and Mark A. Clements

Georgia Institute of Technology
School of Electrical Engineering
Atlanta, Georgia 30332

Abstract

A new and improved iterative speech enhancement technique based on spectral constraints is presented in this paper. The iterative technique, originally formulated by Lim and Oppenheim, attempts to solve for the maximum likelihood estimate of a speech waveform in additive white noise. The new approach applies inter- and intra-frame spectral constraints to ensure convergence to reasonable values and hence improve speech quality. An extremely efficient technique for applying these constraints is in the use of line spectral pair (LSP) coefficients. The inter-frame constraints ensures more speech-like formant trajectories than those found in the unconstrained approach. Results from speech degraded by additive white Gaussian noise show noticeable quality improvement.

Introduction

The successfulness of an enhancement algorithm rests on the goals and assumptions used in deriving the approach. Depending on the application, a system may be directed at one or more objectives such as improving overall quality, increasing intelligibility, reducing listener fatigue, etc. Three assumptions normally made include: i) that the noise distortion be additive, ii) that only the degraded speech signal is available, and iii) that the noise and speech signals are uncorrelated. In general, constraints placed on the speech model improve the potential for separating speech from background noise. However, such systems are also more sensitive to "deviations" from these constraints. The degradation considered is additive white Gaussian noise. The basis of the technique is an iterative enhancement approach based on noncausal Wiener filtering originally formulated by Lim and Oppenheim [1]. This approach attempts to solve for the maximum likelihood estimate of a speech waveform in additive white noise using the constraint that the signal is an all-pole process. Crucial to the success of this approach is the accuracy of the estimates of the all-pole speech parameters at each iteration. One advantage of the Wiener filtering approach is that no "musical tone" artifacts are present after processing as can be observed in spectral subtraction techniques. In addition, under certain conditions, it can be shown that it is 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, it is an iterative scheme with sizable computational requirements as opposed to a direct form such as spectral subtraction. Second, although the original sequential MAP estimation technique was shown to increase the joint likelihood of the speech waveform and all-pole parameters, heuristic convergence criteria had to be employed. After an extensive investigation [2], this approach was found to produce significant levels of enhancement for white Gaussian noise in 3-4 iterations. The technique was generalized to allow for colored aircraft noise. Various spectral estimation techniques where employed for securing estimates of the colored background noise and although the noise was not stationary, estimates were performed prior to application of the algorithm. With these assumptions, good enhancement took place in 2-3 iterations. It is assumed that in a real-time environment however, noise spectral estimates could be gathered and updated during silent intervals. An important observation which could be made from this previous work was that as additional iterations were performed, individual formants of the speech decreased in bandwidth (see fig.1), resulting in unnatural sounding speech. Frame-to-frame pole jitter was also observed which contributed to unnatural sounding results. Also, the original technique employs no explicit frame-to-frame constraints. Since the original algorithm already constrains the speech to be the response from an all-pole system, applying further constraints on the pole movements may improve the algorithms performance. One set of constraints were applied directly to the LPC poles. These results were quite encouraging, yet computationally intensive. A new 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. This method of specification allowed constraints to be efficiently applied to the speech model pole movements across time (inter-frame) so that formants lay on smooth tracks. In addition, constraints could also be easily applied across iterations (intra-frame) on a frame-by-frame basis.

Iterative Speech Enhancement

Enhancement based on the estimation of all-pole speech parameters in additive white Gaussian noise was investigated by Lim and Oppenheim [1], and later for a colored noise degradation by Hansen and Clements [2]. It was shown that the estimation procedures which result in linear equations without background noise, become nonlinear when noise is introduced. However by allowing a suboptimal procedure, an iterative algorithm results which possesses the property that the estimation procedure is linear at each iteration.

Consider the statistical parameter estimation of speech in the presence of noise. Over a short-time basis, the speech signal can be represented as the following difference equation:

$$s(n) = \mathbf{a}^T s(n-1,n-p) + g\,w(n) \qquad (1)$$

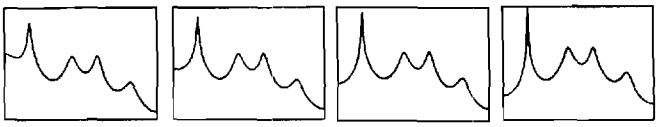
where $\mathbf{a}^T=[a_1,a_2,\ldots,a_p]$ represents the all-pole predictor coefficients. Substituting the degraded speech into the speech model gives the following equation for the observation vector:

$$\begin{aligned} Y_0 &= y(N-1,0) = s(N-1,0) + d(N-1,0) \\ Y_0 &= \mathbf{a}^T y(n-1,n-p) + g\,w(n) + d(n) - \mathbf{a}^T d(n-1,n-p) \end{aligned} \qquad (2)$$

where s(N-1,0) are N samples of original speech, and d(N-1,0) represents the additive background noise. The 2p + 1 unknowns include the predictor coefficients **a**, initial conditions for the predictor given by $S_i = s(-1,-p)$, and the gain factor g for the input excitation. Consider the case where all unknown parameters are random with a priori Gaussian probability density functions. The basic procedure used is a maximum a priori (MAP) estimator, which maximizes the probability density function of

CH2396-0/87/0000-0189 $1.00

the parameters given the observations. Therefore, $\mathbf{a}, g, \mathbf{S}_i$ are chosen to maximize the probability density function $p(\mathbf{a},g,\mathbf{S}_i|\mathbf{Y}_0)$. The procedure requires that $\mathbf{a}$ be chosen to maximize $p(\mathbf{a}|\mathbf{Y}_0)$, noting that the estimate is conditioned on the noisy observations $\mathbf{Y}_0$. Using Bayes' rule, $p(\mathbf{a}|\mathbf{Y}_0)$ can be written as a product of terms involving $p(\mathbf{Y}_0|\mathbf{a},g,\mathbf{S}_i)$. When the Gaussian density function $p(\mathbf{Y}_0|\mathbf{a},g,\mathbf{S}_i)$ is expanded, it can be shown that the mean and variance are functions of the predictor coefficients $\mathbf{a}$. Therefore the resulting equations for maximizing $p(\mathbf{a}|\mathbf{Y}_0)$ are nonlinear, involving partial derivatives with respect to $\mathbf{a}$. Lim and Oppenheim considered a suboptimal solution employing a two step approach based on MAP estimation of $\mathbf{S}_0$ given $\mathbf{Y}_0$, followed by MAP estimation of $\mathbf{a}$ given $\hat{\mathbf{S}}_0$, where $\hat{\mathbf{S}}_0$ is the result of the first estimation. Observations indicate that this algorithm converges to a local maximum of the joint density $p(\mathbf{a},\mathbf{S}_0|\mathbf{Y}_0,g,\mathbf{S}_i)$. In particular, if the probability density function is unimodal, and the initial estimate for $\mathbf{a}$ is such that the local maximum equals the global maximum, then the procedure is equivalent to the joint MAP estimate of $\mathbf{a}$ and $\mathbf{S}_0$. After some simplification, the MAP estimation of $\mathbf{S}_0$, based on maximizing the probability density function $p(\mathbf{S}_0|\mathbf{a}_i,\mathbf{Y}_0)$ which is jointly Gaussian in $\mathbf{Y}_0$, is equivalent to a minimum mean squared error (MMSE) estimate of $\mathbf{S}_0$. Therefore as the observation window increases in length, the procedure for obtaining a MMSE estimate of s(n) approaches a noncausal Wiener filter. With this, the implementation of the algorithm is presented in Figure 2. This approach can also be extended to the colored noise case as shown. As indicated, the background noise spectral density must be estimated during non-speech activity.



(a) Distorted Original (b) 2 Iterations (c) 4 Iterations (d) 8 Iterations

Figure 1: Variation in vocal tract response across iterations.

As indicated, the sequential MAP estimation technique increases the joint likelihood of the speech waveform and all-pole parameters, yet a heuristic convergence criterion had to be employed. Also, as additional iterations were performed, individual formants of the speech decrease in bandwidth as indicated in figure 1. Frame-to-frame pole jitter was also observed. Both effects contributed to unnatural sounding speech. The goal, therefore is to impose constraints on the pole movements across time (inter-frame) and iterations (intra-frame). An initial approach was to limit the poles from moving too close to the unit circle by performing an off-axis spectral evaluation where the z-transform is evaluated on a circle further away from the poles of the spectral model. Other approaches considered included applying constraints directly to the pole radii and/or angular displacements in the LPC model. Performance of such inter and intra-frame constraints lead to encouraging results, but at the expense of a pth order root-solve and a pole ordering step per frame for each iteration. 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. Previous success of the line spectral pair (LSP) transformation in speech coding by Crosmer [3], led to the use of LSP's for this purpose.

Line Spectral Pair Representation of Spectral Characteristics

The LSP transformation may be viewed as an alternative representation of the LPC spectrum. The LSP coefficients are obtained from the LPC prediction coefficients by combining the forward and backward predictor polynomials as follows:

$$P(z) = A(z) + B(z), \qquad Q(z) = A(z) - B(z). \qquad (3)$$

The vocal tract transfer function is given by $g/A(z)$, and M is the order of the LPC speech model. The resulting polynomials P(z) and Q(z), are symmetric and antisymmetric, respectively, with a root of P(z) at $z=+1$, and a root of Q(z) at $z=-1$. The remainder of the roots of P and Q all lie on the unit circle. Since the roots occur in conjugate pairs, the original polynomial can be represented by M real numbers. The angles of the roots, $\{\omega_i,\ i=1,2,\ldots,M\}$, are called the *line spectrum pairs*.

The LSP's possess several important properties which make them attractive for use in applying spectral constraints. One important characteristic is that if the vocal tract polynomial A(z) has all its roots inside the unit circle (i.e., a stable filter), then the roots of P and Q will be interleaved around the unit circle [3]. If two adjacent LSP frequencies are identical, it indicates that a root of A(z) lies on the unit circle.

In addition to their attractive representation of the LPC spectrum, the LSP coefficients offer the possibility of a more direct representation of perceptually important information. Specifically, their is a firm statistical relationship between the locations and bandwidths of the speech formants and the locations of the roots of P and Q respectively. Since roots of the P polynomial correspond approximately to locations of formant center frequencies (when a formant is present), the P polynomials' LSP coefficients are termed *position coefficients*. It can be shown that the closer two LSP coefficients are together, the narrower the bandwidth of the corresponding pole of the vocal tract filter. Therefore, formants are indicated when two LSP coefficients are close together. When LSP coefficients are far apart, they indicate poles which contribute only to the overall spectral shape. Because of their relationship to the presence or absence of a formant by their nearness to a position coefficient, the coefficients of Q are termed *difference coefficients*. Given the LSP coefficients, the position coefficients are simply the odd index LSP coefficients, $\{p_i=\omega_{2i-1},\ i=1,2,\ldots,M/2\}$. The difference coefficients are given as follows:

$$\{|\ d_i\ |\ =\ \underset{j=-1,1}{\text{MIN}}\ (\ |\ \omega_{2i+j} - \omega_{2i}\ |\ ),\ i = 1,2,\ldots,M/2\} \quad (4)$$

where the sign of d_i is positive if ω_{2i} is closer to ω_{2i+j}, and otherwise is negative. With this interpretation, a new enhancement technique based on Wiener filtering is now possible by imposing constraints on the LSP coefficients.

Step 1: Estimate $\mathbf{a}_i$ from $\mathbf{S}_{0,i}$.
Use either: i. first P values as the initial condition vector
or: ii. always assume $\mathbf{S}_i = \mathbf{0}^1$.

Step 2: i. Using $\hat{\mathbf{a}}_i$, estimate the speech spectrum:

$$P_S(\omega) = \frac{g^2}{\left|1 - \sum_{k=1}^{P} a_k e^{-jk\omega}\right|^2}$$

ii. Calculate gain term using Parseval's theorem.
iii. Estimate either the degrading
a.) white noise variance σ_d^2, or b.) colored noise spectrum $P_D(\omega)$
from a period of silence closest to the utterance.
iv. Construct the noncausal Wiener filter;

a.) $H(\omega) = \dfrac{P_S(\omega)}{P_S(\omega) + \sigma_d^2}$ b.) $H(\omega) = \dfrac{P_S(\omega)}{P_S(\omega) + P_D(\omega)}$

v. Filter the estimated speech $\hat{\mathbf{s}}_i$ to produce $\hat{\mathbf{s}}_{i+1}$.
vi. Repeat until some specified error criterion is satisfied, $\Delta\epsilon <$ THRESHOLD.

Figure 2: Enhancement Algorithm based on All-pole modeling/Wiener filtering. a) a AWGN distortion b) a non-white distortion

Enhancement with Spectral Constraints

Consider the statistical parameter estimation of speech in the presence of noise, where all unknown parameters are random with a priori Gaussian probability density functions. It can be shown that MAP estimation of $\mathbf{a}$, g, and $\mathbf{S}_i$ given the noisy observations $\mathbf{Y}_0$, results in a set of nonlinear equations. Therefore, instead of joint estimation of $\mathbf{a}$ and $\mathbf{S}_0$, a suboptimal solution is formulated employing a two step approach based on

MAP estimation of S_0 given Y_0, followed by MAP estimation of a given $\hat{S}_0$, where $\hat{S}_0$ is the result of the first estimation. Since speech can be considered short-time stationary, frame-to-frame spectral constraints may aid in enhancement. The new approach imposes such constraints on the vocal tract spectrum between MAP estimation steps. The procedure for obtaining the MAP estimate of a from MAX $p(a|\hat{S}_0;g,S_i)$ remains the same. The next step is to apply spectral constraints to $\hat{a}_i$ which will ensure that; i) the all-pole speech model is stable, ii) it possess speech-like characteristics (i.e., 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 this constrained approach, an improved estimate $\hat{a}_i'$ results. Given this new estimate, the second MAP estimation of S_0 given $\hat{a}_i'$ can be carried out by maximizing $p(S_0|\hat{a}_i',Y_0;g,S_i)$. Since $p(S_0|\hat{a}_i',Y_0;g,S_i)$ is still jointly Gaussian in Y_0, the resulting MAP estimate is equivalent to a MMSE estimate of S_0. Again, in the limiting case, the procedure for obtaining the MMSE estimate of s(n) approaches a noncausal Wiener filter. Once this new estimate of $\hat{S}_{0,i}$ is formed, the iterative procedure continues by re-estimating $\hat{a}_i$, applying constraints to $\hat{a}_i$, and then forming the noncausal filter using $\hat{a}_i'$ to re-estimate $\hat{S}_{0,i}$. This continues until some convergence criterion is satisfied. The procedure for implementing these constraints will now be addressed.

Two classes of spectral constraints are considered; inter-frame (across time), and intra-frame (across iterations). Two approaches are considered: a fixed frame rate, and a variable frame rate approach. In the first of these, the LPC predictor coefficients, a, are first converted to LSP position and difference coefficients. Next, each frame's energy is observed, and if it is above some threshold, it is classified as voiced speech; if it is below, then it is either noise or unvoiced speech. A local running count L_1, is kept for the number of consecutive frames which fall below the energy threshold. If L_1 reaches L_{MAX}, then all subsequent frames below the threshold are classified as noise. This allows for further smoothing for long periods of silence. The position coefficients 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 is performed. In addition, the lower formant frequencies are smoothed over a narrower triangle width than for those position coefficients at higher frequencies. This preserves perceptually important speech characteristics found in the lower formants. No smoothing is performed on the difference coefficients since they are more closely related to formant bandwidth than formant location. However, it is possible that a difference coefficient falls within a "forbidden zone," (i.e., the region within d_{MIN} of a position coefficient). When this occurs, the LPC analysis has most likely overestimated the Q of a particular pole. Since this causes unnatural sounding speech, (as in the unconstrained approach), the value of $|d_i|$ is set to d_{MIN}. Finally, the position and difference coefficients are combined to form the constrained LPC predictor coefficients $\hat{a}_i'$.

The second inter-frame constraint approach considered is a variable frame rate technique which takes advantage of the interpolation properties of the LSP coefficients. The speech signal is first divided into segments, where segments are chosen such that they are long when the speech spectrum is varying slowly and short when the speech spectrum is varying quickly. The LSP coefficients are reconstructed with linear interpolation used to compute the coefficients for intermediate frames.

The segmentation algorithm begins with a step to determine the onset/offset of speech. This is carried out by thresholding the LPC residual energy, which produces relatively long segments. Next, the long segments are subdivided based on the curvature of the position coefficients. 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 the distortion of position coefficients for successively longer segments until the distortion exceeds a threshold T_D. At that point, a subsegment boundary is set, with the intermediate position coefficients 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 that it incorporates more information from adjacent frames when the spectrum indicates similar characteristics. Yet, it also reduces the effects of adjacent frames when the spectrum is significantly different as in the case of a transition from unvoiced passages to noise. This in effect, distorts the position coefficients as little as possible when associated difference coefficients indicate the presence of formants. Difference coefficients for each frame, (or an average set across a segment) are used to compute the predictor coefficients $\hat{a}_i$. The difference coefficients are required to be at least d_{MIN} or greater in distance from adjacent position coefficients to ensure that poles from the LPC filter do not move too close to the unit circle.

Inter-frame constraints are applied to a single frame across iterations, and as such require the frames' previous estimates to be available. The motivation for such constraints is that under certain conditions, pole locations for the same frame vary significantly from their previous estimated values. Since the present estimate of $\hat{a}_i$ affects the next estimate of $\hat{S}_{0,i}$, sections of $\hat{S}_{0_i}$ will also vary significantly across iterations. In addition, previous results based on objective speech quality measures indicated that the unconstrained approach produced minimum objective measures at different iterations for different classes of speech. For example, maximum overall speech quality was observed for additive white Gaussian noise in three iterations. This was also true for vowels and fricatives. However, glides required two iterations, nasals, liquids, and affricates between five and six. It is therefore desirable to be able to affect the convergence rate so that the best objective measure of quality occurs at the same iteration across all classes of speech. Improved quality as measured by objective measures may also result in improved estimation of $\hat{a}_i$. By constraining the vocal tract filter to be a function of its previous estimates, it may be possible to accomplish this. Two approaches are considered, one applied to the autocorrelation lags, the other to the position coefficients. The first approach simply weights the present set of autocorrelation lags with the same frame from previous iterations. This technique is very easy to perform, since the autocorrelation lags must be computed in order to estimate the predictor coefficients a. The second approach weights position coefficients with those from the same frame but previous iteration. If the corresponding difference coefficient indicates the adjacent position coefficient to represent a formant, this approach has the effect of constraining the formants to lie along smooth tracks across iterations.

Results

Speech degraded by additive white Gaussian noise was processed using various configurations of the new constrained enhancement algorithm. Energy thresholds for inter-frame constraints were obtained from frame energy histograms at each signal-to-noise ratio. Excellent enhancement resulted for a wide range of threshold values. Intra-frame constraints were applied across two to three iterations. Informal listening tests indicated noticeable quality improvement, although no intelligibility testing has been performed. However, there has been extensive work carried out in the area of objective speech quality measures [4]. Good correlation has been shown to exist between subjective quality and objective measures. Therefore, objective measures including: the Itakura-Saito likelihood ratio, log area ratio, and weighted spectral slope measure where used for evaluation. Figure 3 illustrates a comparison of

typical results for the various constraint approaches. 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 inter-frame constraints (both fixed and variable rate), and intra-frame constraints (applied to position coefficients/autocorrelation lags). All configurations examined showed significant improvement in Itakura-Saito measures. Threshold settings for the variable frame rate inter-frame constraint were somewhat sensitive to varying noise levels. However, the fixed frame approach by itself, and with either autocorrelation or position intra-frame 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 **a** were used in the unconstrained algorithm. In essence, the two step MAP estimation approach is now reduced to a single MAP estimate of S_0, and therefore represents the theoretical limit for enhancement using Wiener filtering. Plot *f* indicates this limit. Although only Itakura-Saito measures are shown, similar improvement was also observed for log area ratios and weighted spectral slope measures. Figure 4 compares the new approach to existing techniques. Plot *b* shows results from spectral subtraction as formulated by Boll [5]. An evaluation was performed for both half and full-wave rectification, along with one to five frames of magnitude averaging; where these points represent the best results. Plot *c* is from the unconstrained Wiener filtering technique. Plots *d* and *e* are typical values for the inter-frame constraint (fixed frame rate), and inter plus intra-frame constraints (fixed frame and autocorrelation lags). Again *f* indicates the limit for the Wiener filtering approaches.

Sound Type	*Itakura-Saito Likelihood Measure*			
	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

SNR=+5dB

Table 1: Comparison of algorithms over sound types for white Gaussian noise.

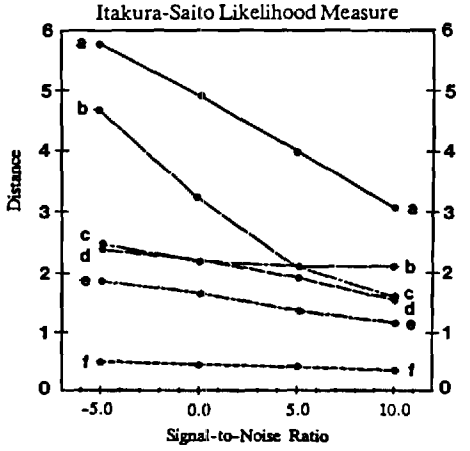


Figure 3: Comparison of constraint algorithms over SNR.
a.) Original Distorted Speech
b.) Inter-Frame Constraint: Variable Frame
c.) Inter-Frame Constraint: Fixed Frame
d.) Inter & Intra-Frame Constraints: Fixed Frame, Position
e.) Inter & Intra-Frame Constraints: Fixed Frame, Autocorrelation
f.) Theoretical limit: using undistorted LPC coefficients, **a**.

Performance evaluation over sound classes was accomplished by hand partitioning speech into segments. Entire sentences were processed, and objective measures from each class were computed. Table 1 summarizes this comparison between the unconstrained Lim-Oppenheim technique to that of the inter and intra-frame constraint approach. Measures for the theoretical limit using undistorted LPC predictor coefficients **a** are also indicated. Improvement is indicated for all types of speech. In addition, the constrained approach produced superior objective measures of quality across all speech classes at the same iteration. These results clearly indicate improvement over the unconstrained approach as well as spectral subtraction for additive white Gaussian noise.

Conclusions

The application of spectral constraints to noncausal Wiener filtering results in improved speech enhancement. Informal listening tests along with objective measures such as Itakura-Saito and log-area-ratio's show improvement over the unconstrained technique. By using the Line Spectral Pair transformation, a modest increase in computational requirements results in significant improvement in speech quality. This approach to pole movement constraints is quite robust over direct methods applied to pole radial/angular movements. Finally, this approach may be useful in enhancement for human listeners as well as a preprocessor for speech recognition.

References

[1] J.S. Lim, A.V. Oppenheim, " All-Pole Modeling of Degraded Speech," *IEEE Trans. Acoust., Speech, Sig. Proc.*, vol. ASSP-26, pp. 197-210, June 1978.

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

[3] J.R. Crosmer, " Very Low Bit Rate Speech Coding Using the Line Spectrum Pair Transformation of the LPC Coefficients, " Ph.D. dissertation, School of Electrical Engineering, Georgia Institute of Technology, Atlanta, June 1985.

[4] S.R. Quackenbush, " Objective Measures of Speech Quality, " Ph.D. dissertation, School of Electrical Engineering, Georgia Institute of Technology, Atlanta, May 1985.

[5] S.F. Boll, " Suppression of Acoustic Noise in Speech using Spectral Subtraction, " *IEEE Trans. Acoust., Speech, Sig. Proc.*, vol. ASSP-27, pp. 113-120, April 1978.

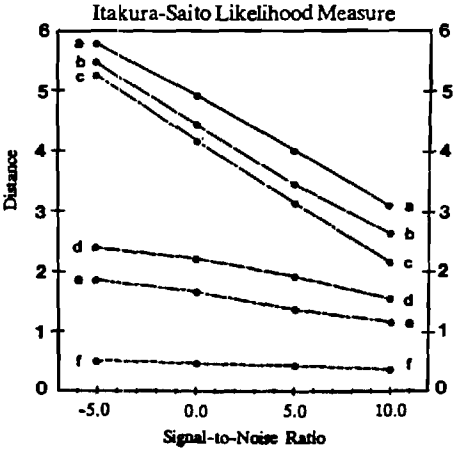


Figure 4: Comparison of enhancement algorithms over SNR.
a.) Original Distorted Speech
b.) Boll: Spectral Subtraction, using magnitude averaging
c.) Lim-Oppenheim: Unconstrained Wiener filtering
d.) Hansen-Clements: employing Inter-Frame constraints
e.) Hansen-Clements: employing Inter & Intra-Frame constraints
f.) Theoretical limit: using undistorted LPC coefficients, **a**

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 $\vec{S}_O$ 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.

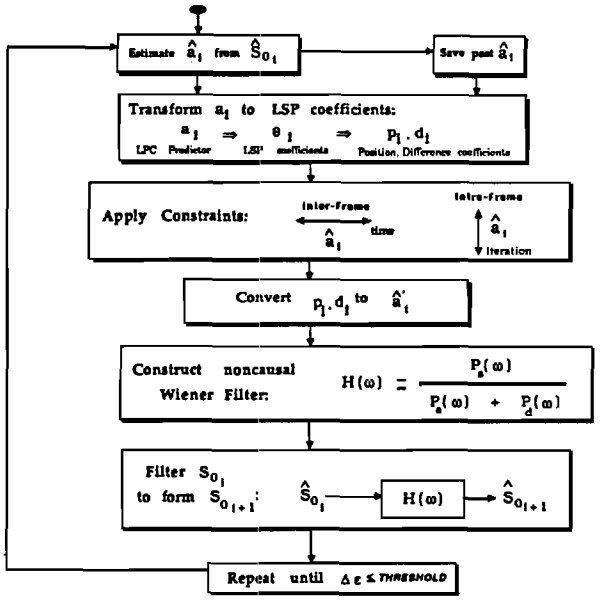


Figure 1: Framework for the constrained iterative enhancement algorithms.

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

rior to optimal, only minor differences in speech quality would esult. The results consistently suggested that the constrained enancement algorithms reach a maximum level of speech quality at he same iteration, independent of SNR and sound class concenrations.

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	49.80
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	81.87	—
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 or 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	8.88	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	1.33	1.21	1.19	1.16	♣1.15	1.32
Liquid	9.80	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 peech 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: -5 dB		-0 dB		+5 dB		+10 dB		OVERALL	
	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%
F-LSP:T,Auto:I	7	100%	7	100%	7	100%	7	88%	7	97%
							8	12%	8	3%
F-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-

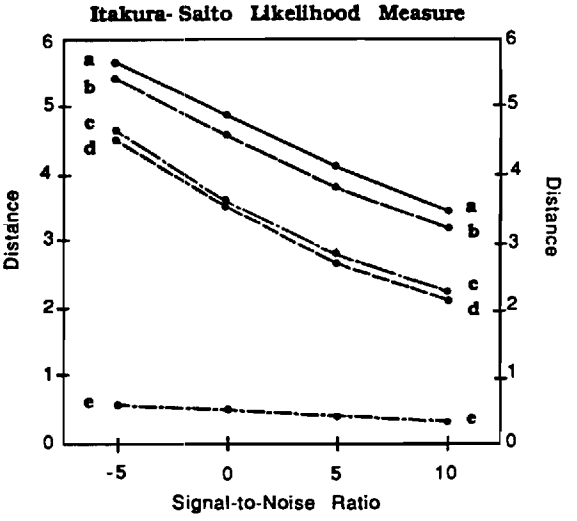


Figure 2: Comparison of inter & intra-frame constrained enhancement algorithms for colored aircraft noise over SNR.

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.

Constrained Iterative Speech Enhancement

by

John H. L. Hansen[4], Member, IEEE

and

Mark A. Clements[5], Member, IEEE

June 29, 1989

[1]Submitted March 20, 1988, Revised June 21, 1989

[2]This work sponsored in part by grants from U.S. Army Human Engineering Labs and Department of Defense.

[3]Please address all correspondence to Dr. Hansen.

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

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

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 intra- and inter-frame constraints 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, informed 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 non-white 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.

1 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 inter- and intra-frame 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 non-white distortion. Finally, a comparative evaluation is also performed to determine their usefulness as preprocessors for recognition in noisy environments.

2 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 $\vec{S}_O$ become nonlinear. Lim and Oppenheim considered a suboptimal solution employing sequential MAP estimation of $\vec{S}_O$ followed by MAP estimation of $\vec{a}$, g given $\hat{\vec{S}}_{O,i}$, where $\hat{\vec{S}}_{O,i}$ 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 $\vec{S}_O$ is equivalent to noncausal Wiener filtering of the noisy speech $\vec{Y}_O$. Lim and Oppenheim showed this technique, under certain conditions, increases the joint likelihood of $\vec{a}$ and $\vec{S}_O$ with 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 non-white, non-stationary 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 $\vec{S}_O$. 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 Figure 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. Lim and Oppenheim recognized these limitations and an improved method was formulated by Musicus and Lim [15] which addresses some of them. Even with their improvements, however, no explicit frame-to-frame constraints are employed. Since the original algorithm already constrains the speech to be the response from an all-pole system, applying further constraints on the pole movements imposes no new assumptions on the speech or noise, and may improve the algorithm's performance. The imposition of some relatively simple constraints turns out to improve speech quality results, even when directly attached to the original Lim-Oppenheim method.

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 parameters $\vec{a}$, gain g, and noise free speech $\vec{S}_O$) are random with a priori Gaussian probability density functions. It was shown that MAP estimation of $\vec{a}$, g, and

$\vec{S}_O$ given noisy observations $\vec{Y}_O$, results in a set of nonlinear equations. Therefore, instead of joint estimation of $\vec{a}$ and $\vec{S}_O$, a suboptimal solution was formulated employing a two-step approach based on MAP estimation of $\vec{S}_O$ given $\vec{Y}_O$, followed by MAP estimation of $\vec{a}$, g given $\hat{\vec{S}}_{O,i}$, where $\hat{\vec{S}}_{O,i}$ 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 $\vec{S}_O$ is performed by maximizing its conditional probability density function given $\hat{\vec{a}}_{i+1}$ and the observed noisy sequence $\vec{Y}_O$. Since this probability density function is jointly Gaussian, the resulting MAP estimate is equivalent to a MMSE estimate of $\vec{S}_O$. With further simplifying assumptions, it can be shown that MAP estimation of $\vec{S}_O$ 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 $\hat{\vec{S}}_{O,i}$ is formed, the iterative procedure continues by re-estimating $\hat{\vec{a}}_i$, applying constraints to $\hat{\vec{a}}_i$, and forming the noncausal filter using $\hat{\vec{a}}_{i+1}$ to re-estimate $\hat{\vec{S}}_{O,i}$. 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.

Figure 2 presents an overview of two classes of constraints which include inter-frame (across time) and/or intra-frame (across iterations). Each technique differs in the type of constraint and computational requirements. The present evaluation focuses on two representative inter-frame (FF-LSP:T) and combined inter-frame plus intra-frame (FF-LSP:T,Auto:I) based techniques. Further discussion of all techniques are found in [5, 6, 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 unconstrained 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 inter-frame versus intra-frame 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 two polynomials of order $p+1$ which have all roots on the unit circle.

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

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

The angles of the roots, $\{\ \omega_i, i = 1,2,\ldots,M\}$, 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 closest 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, $\{p_i = \omega_{2i-1}, i = 1, 2, \ldots, M/2\}$), and the separations mentioned above are the difference parameters, d_i.

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

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.

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 inter-frame 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 $E_{V/UV}$. 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 difference 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 $\hat{\tilde{a}}_{i+1}$.

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 of position 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 at 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 3). 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 parameters (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 $\vec{a}$. 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 refered 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 $\vec{a}$), improved convergence results. Results from inter-, intra-, and combined inter- plus intra-frame constraint approaches will be presented in the next section. Figure 3

illustrates the framework for the new set of constrained enhancement techniques.

3 Evaluation

We now evaluate the performance of the proposed algorithms for speech 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 inter-frame 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 inter- plus intra-frame 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 Intra-frame constraints were applied across two to three iterations.

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

[1]The signal-to-noise ratio is defined as $10\log\left(\frac{\sum_n s^2(n)}{\sum_n d^2(n)}\right)$, 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]

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 $.015 \leq d_{MIN} \leq .031$ radians, $.055 \leq d_{MAX} \leq .077$ radians, resulted in good quality improvement.

The value of 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 ($\geq$ +5dB). 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 [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 non-white 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.

3.1 Evaluation Using Additive White Gaussian Noise

Various configurations of the new constrained enhancement algorithms were evaluated in an additive white Gaussian 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. Figure 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 inter-frame constraints (both fixed and variable rate), and intra-frame constraints (applied to position parameters/autocorrelation lags). All configurations examined showed significant improvement in Itakura-Saito measures. Threshold settings for the variable frame rate inter-frame constraint were somewhat sensitive to varying noise levels. This indicates that although applying inter-frame 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 intra-frame 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 $\vec{S}_O$, and therefore represents the theoretical limit for enhancement using Wiener filtering. Plot *f* indicates this limit.

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 SNR 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 recti-

fication had greater improvement over the restricted SNR band of 5 to 10 dB. Magnitude averaging lead to improved 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. Figure 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 2. The average correlation between each objective quality measure and subjective quality as measured by the DAM (diagnostic acceptability test) is shown [17].

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 1 summarizes the comparison between the unconstrained technique, and an inter- plus intra-frame constrained 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.

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 3 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 4 presents results from an equivalent evaluation for one of the constrained enhancement algorithms (FF-LSP:T,Auto:I). A comparison between tables 3 and 4 show 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.

Termination Consistency Versus SNR

Further evaluations were performed to determine the consistency of the terminating iteration versus SNR. Table 5 summarizes optimum terminating points in terms of objective quality for some of the enhancement algorithms. Techniques employing only inter-frame constraints consistently resulted (94% 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, due to the relaxation of the iterative scheme as imposed by intra-frame constraints, adjacent iterations differ

only slightly in objective quality for the constrained techniques. Therefore, only minor 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.

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 Figure 1. Figure 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 overall spectral shape, especially in the region of the second formant. However, as additional iterations were performed, spectral distortions result, 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.

Computational Issues

Discussion of algorithm performance should also address computational issues as well as algorithm complexity. Naturally, there exists a trade-off between resulting speech

quality and each algorithm's computational complexity. It is clear that iterative techniques require greater computer resources than non-iterative 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 computer requirements. In Table 6, a comparison of 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 constrained approaches appears justifiable.

Time Versus Frequency Plots

Isometric plots of time versus frequency magnitude spectra were constructed. In Figure 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 inter- plus intra-frame constrained (FF-LSP:T,Auto:I) approach. These figures indicate that the considerable noise rejection achieved in the single frame noted in Figure 6, is generally true over time.

3.2 Evaluation Using Additive Non-White, Non-Stationary Noise

The enhancement techniques described for the white additive noise case were also tested using non-stationary, 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]. Energy thresholds for the inter-frame constraints were obtained from frame energy histograms at each signal-to-noise ratio. Intra-frame constraints were applied across two to three iterations. Figure 8 and Table 7 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.

3.3 Recognition Evaluation

One application for speech enhancement is a preprocessor for an automatic recognition system. For evaluation of the enhancement algorithms in this application, a 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, 5) and 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 twenty highly confusable words from a speech data base formulated for recognition evaluation in diverse environments [7]. These words are also used

[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 quality improvement for this particular distortion [3, 6].

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 twenty models employed by the HMM recognizer were trained using the forward-backward algorithm. Figure 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 +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). 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.

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

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 +30dB range.

4 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 inter- and intra-frame constraints on the input speech signal and were shown to be useful in enhancing speech for human listeners, and somewhat useful as preprocessing for recognition in noisy environments. Inter-frame constraints ensure more speech-like formant trajectories than those found in the unconstrained approach and thus reduce pole jitter on a frame-to-frame basis. Intra-frame 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 non-stationary colored noise. Third, the algorithms were shown to improve all segments of speech for both white and non-white 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 large enough to warrant enhancement preprocessing for recognition.

APPENDIX

All sentences were sampled at 8000 samples/sec.

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. on Acoustics, Speech, and Signal Processing,* Vol. ASSP-27, pp. 113-120, April 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 of Electrical Engineering, Georgia Institute of Technology, Atlanta, June 1985.

[3] 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.

[4] 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.

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

[6] J.H.L. Hansen, M.A. Clements, "Constrained Iterative Speech Enhancement with Application to Automatic Speech Recognition," *Proc. 1988 IEEE ICASSP,* pp. 44.S12.9.1-4, New York, NY, April 1988.

[7] J.H.L. Hansen, "Analysis and Compensation of Stressed and Noisy Speech With Application to Robust Automatic Recognition," Ph.D. Thesis, Georgia Institute of Technology, 396 pages, July 1988.

[8] J.H.L. Hansen, M.A. Clements, "Stress and Noise Compensation Algorithms for Robust Automatic Speech Recognition," *Proc. 1989 IEEE ICASSP,* pp. 266-269, Glasgow, Scotland, U.K., May 1989.

[9] F. Itakura, "Line Spectrum Representation of Linear Predictive Coefficients of Speech Signals, " *Journal of the Acoustical Society of America,* vol. 57, S35(A), 1975.

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

[11] 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.

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

[13] F.J. Malkin, K.A. Christ, "Human Factors Engineering Assessment of Voice Technology for the Light Helicopter Family," U.S. Army Human Engineering Laboratory Technical Report, pp. 1-20, June 1985.

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

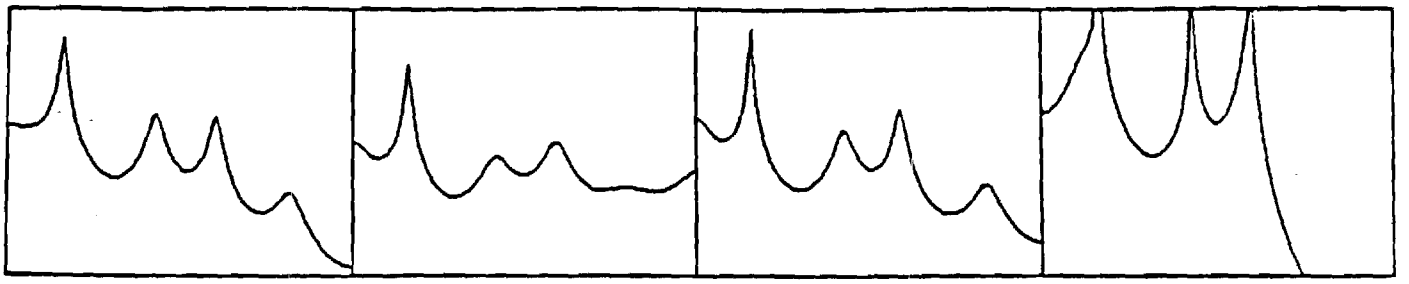
[15] B.R. Musicus, "An iterative technique for maximum likelihood parameter estimation on noisy data," S.M. Thesis, Massachusetts Institute of Technology, Cambridge, Mass., 1979.

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

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

[18] C.A. Simpson, "Speech Variability Effects on Recognition Accuracy Associated With Concurrent Task Performance by Pilots, " Psycho-Linguistic Research Associates, Technical Report, pp. 1-15, April 1985.

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



(a) Original (b) Distorted Original (c) 4 Iterations (d) 8 Iterations

Figure 1: Variation in vocal tract response across iterations.

Sound Type	*Itakura-Saito Likelihood Measure*			
	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 1: Comparison of unconstrained (Lim-Oppenheim) and inter- and intra-frame constrained (Hansen-Clements) algorithms over sound types for white Gaussian noise. SNR = +5 dB

	OBJECTIVE QUALITY MEASURE		
	Itakura-Saito	*log-area-ratio*	*Klatt*
$\lvert\hat{\rho}\rvert$	.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 2: 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, $\lvert\hat{\rho}\rvert$ is the average correlation coefficient between objective and subjective speech quality[17].

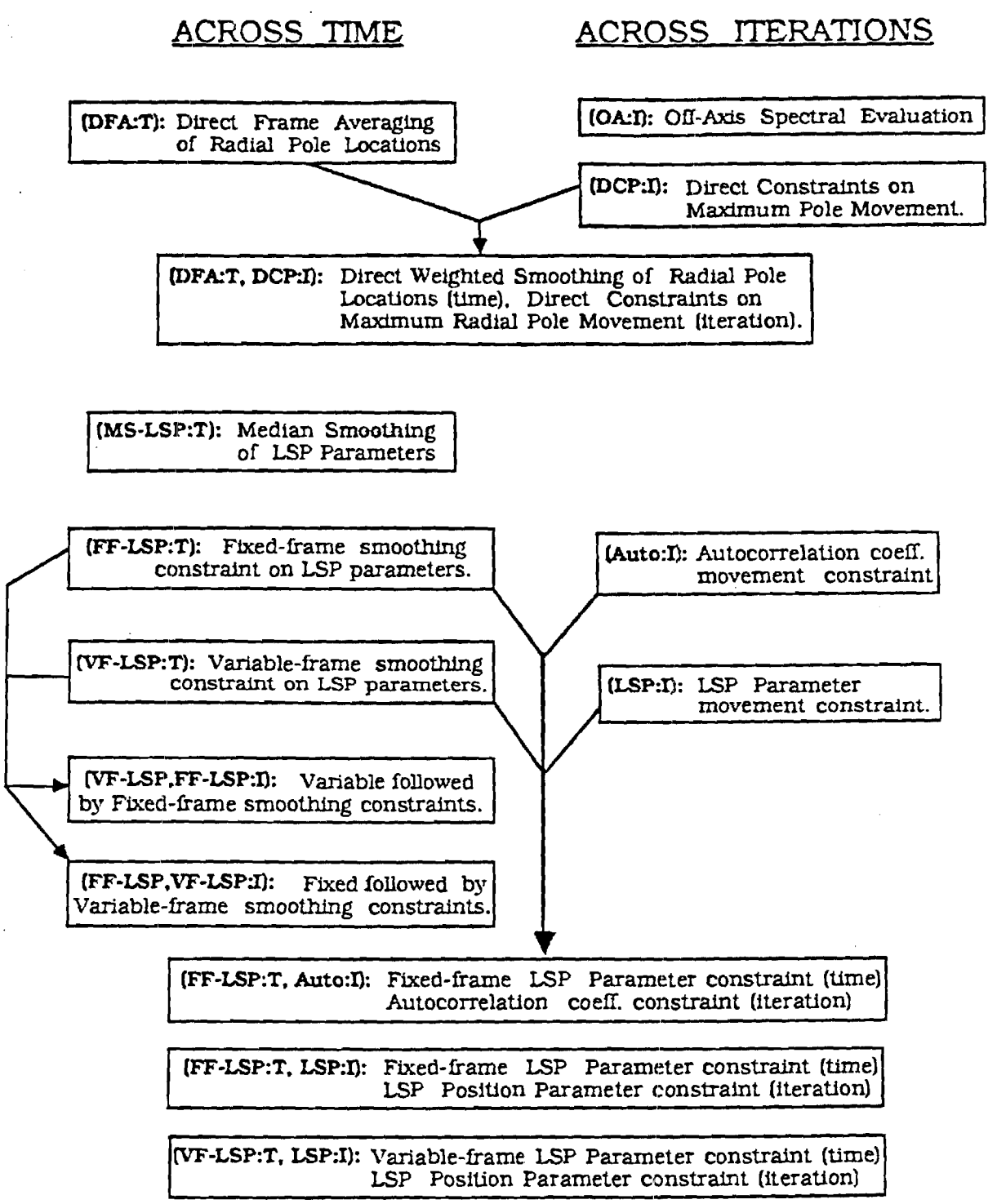


Figure 2: An overview of spectral constraints considered for the class of constrained speech enhancement algorithms.

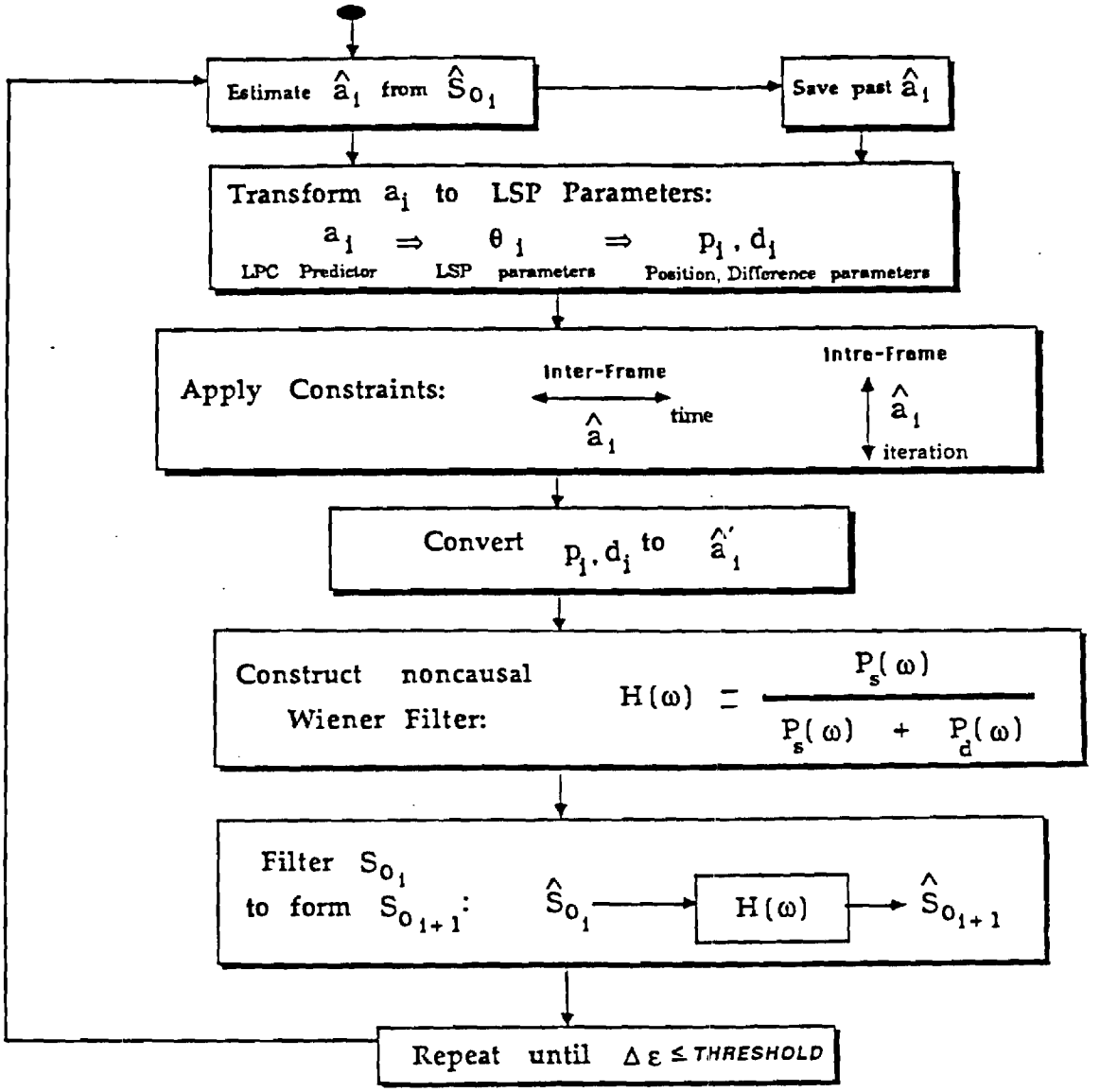


Figure 3: Framework for the new set of constrained enhancement algorithms.

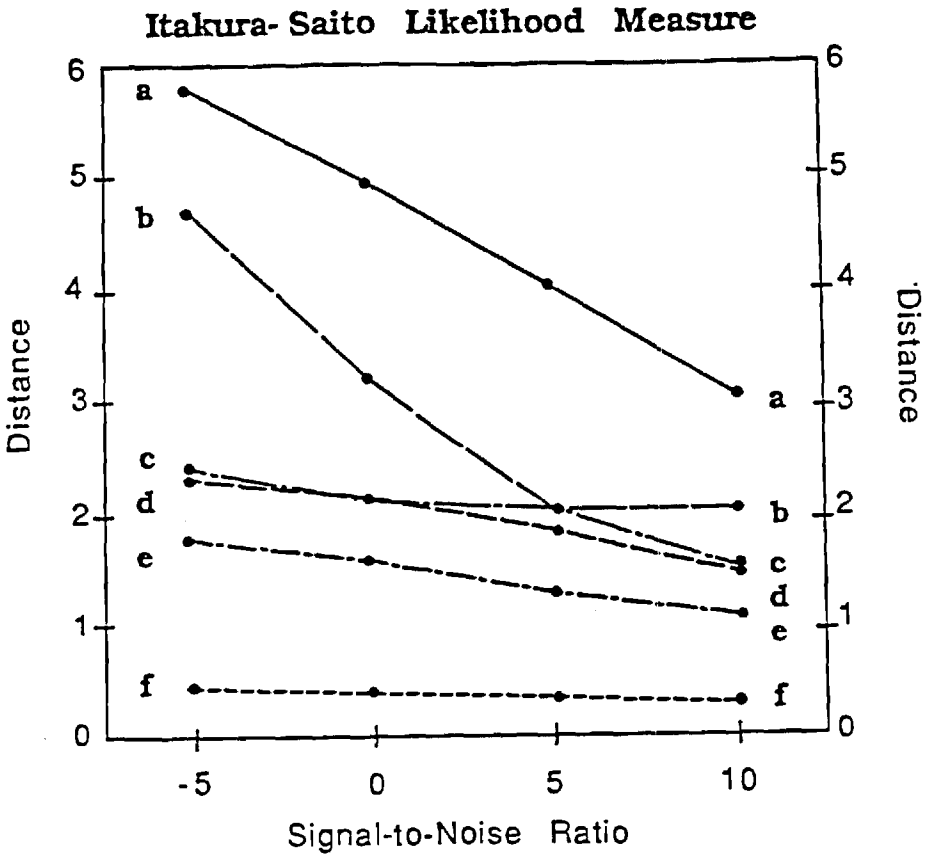


Figure 4: Comparison of constraint algorithms over SNR.

a.) **Original Distorted Speech**

b.) **Inter-Frame Constraint: Variable Frame (VF-LSP:T)**

c.) **Intra-Frame Constraint: Fixed Frame (FF-LSP:T)**

d.) **Inter & Intra-Frame Constraints: Fixed Frame, Position (FF-LSP:T,LSP:I)**

e.) **Inter & Intra-Frame Constraints: Fixed Frame, Autocorrelation (FF-LSP:T,Auto:I)**

f.) **Theoretical limit: using undistorted LPC coefficients $\bar{a}$.**

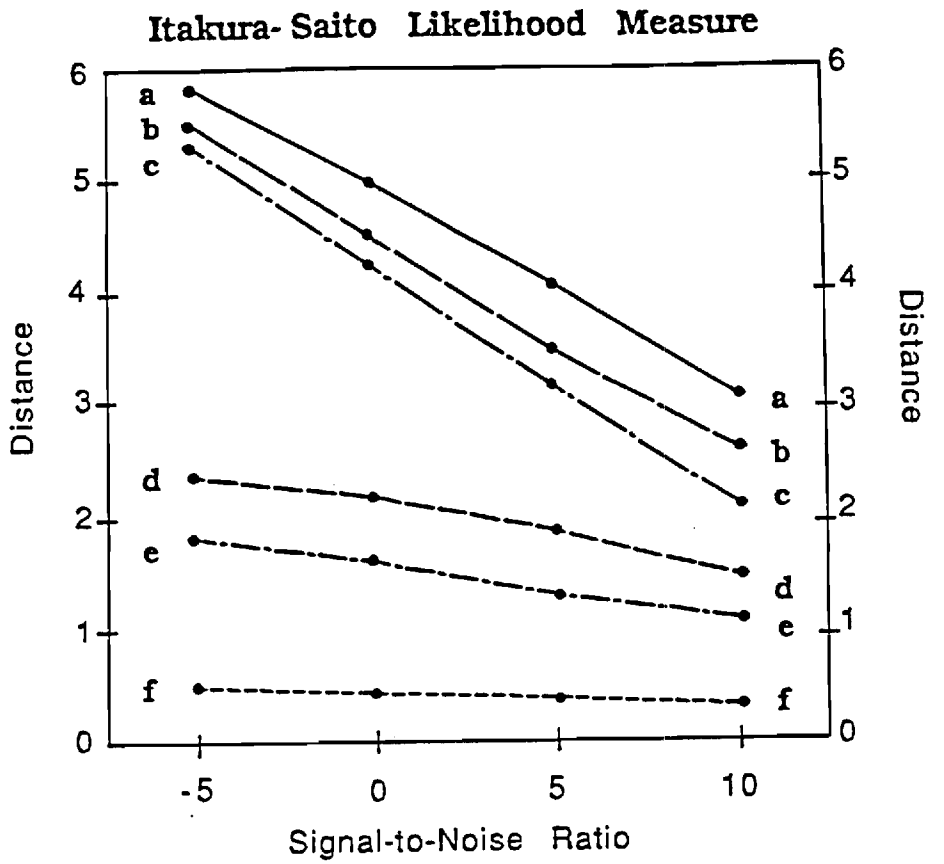


Figure 5: Comparison of enhancement algorithms over SNR.

a.) **Original Distorted Speech**

b.) **Boll: Spectral Subtraction, using magnitude averaging**

c.) **Lim-Oppenheim: Unconstrained Wiener filtering**

d.) **Hansen-Clements: employing Inter-Frame constraints (FF-LSP:T)**

e.) **Hansen-Clements: employing Inter & Intra-Frame constraints (FF-LSP:T,Auto:I)**

f.) **Theoretical limit: using undistorted LPC coefficients $\tilde{a}$.**

Sound Type	*Itakura-Saito Likelihood Measure (across iterations)*							
	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 3: 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=+5dB

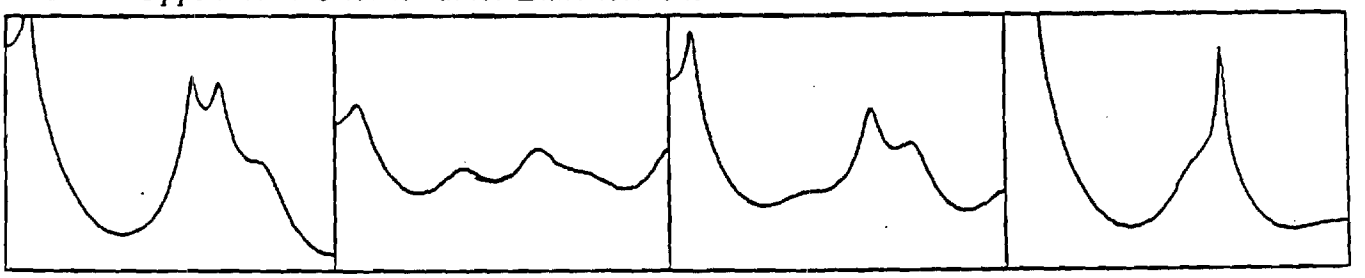
Sound Type	*Itakura-Saito Likelihood Measure (across iterations)*								
	Original	*#1*	*#2*	*#3*	*#4*	*#5*	*#6*	*#7*	*#8*
Silence	1.634	1.551	1.351	1.155	1.036	0.979	0.929	♣0.884	0.901
Vowel	4.020	3.319	2.865	2.394	1.863	1.677	1.571	♣1.565	1.828
Nasal	19.814	16.490	12.397	10.523	8.682	6.840	4.929	♣3.789	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	3.722	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 4: Hansen-Clements Inter & Intra-frame constrained speech enhancement for white Gaussian noise. Convergence for a particular speech class in terms of objective quality is indicated by a ♣. SNR=+5dB

Constrained Enhancement Algorithm	*Additive White Gaussian Noise SNR*									
	-5 dB		*-0 dB*		*+5 dB*		*+10 dB*			
	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:T,LSP:I	4	100%	4	100%	4	100%	4	100%	4	100%

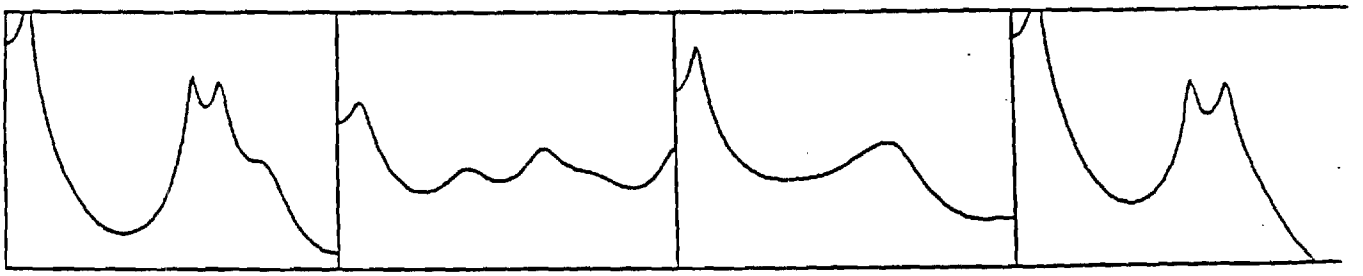
Table 5: Summary of optimal terminating iteration across SNR for AWGN.

Lim - Oppenheim: Unconstrained Enhancement



(1a) Original (1b) Distorted Original (1c) 4 Iterations (1d) 8 Iterations

Hansen - Clements: Constrained Enhancement (FF-LSP:T,Auto:I)



(2a) Original (2b) Distorted Original (2c) 4 Iterations (2d) 8 Iterations

Figure 6: Variation in vocal tract response across iterations for *1a-d)* unconstrained, and *2a-d)* constrained enhancement algorithms.

	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
(F-LPS:T,Auto:I)	1.36	9	10	7

Table 6: Comparison of enhancement algorithms in terms of quality, relative complexity, and relative computational resources. SNR = +5 dB, Additive white Gaussian noise distortion.

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

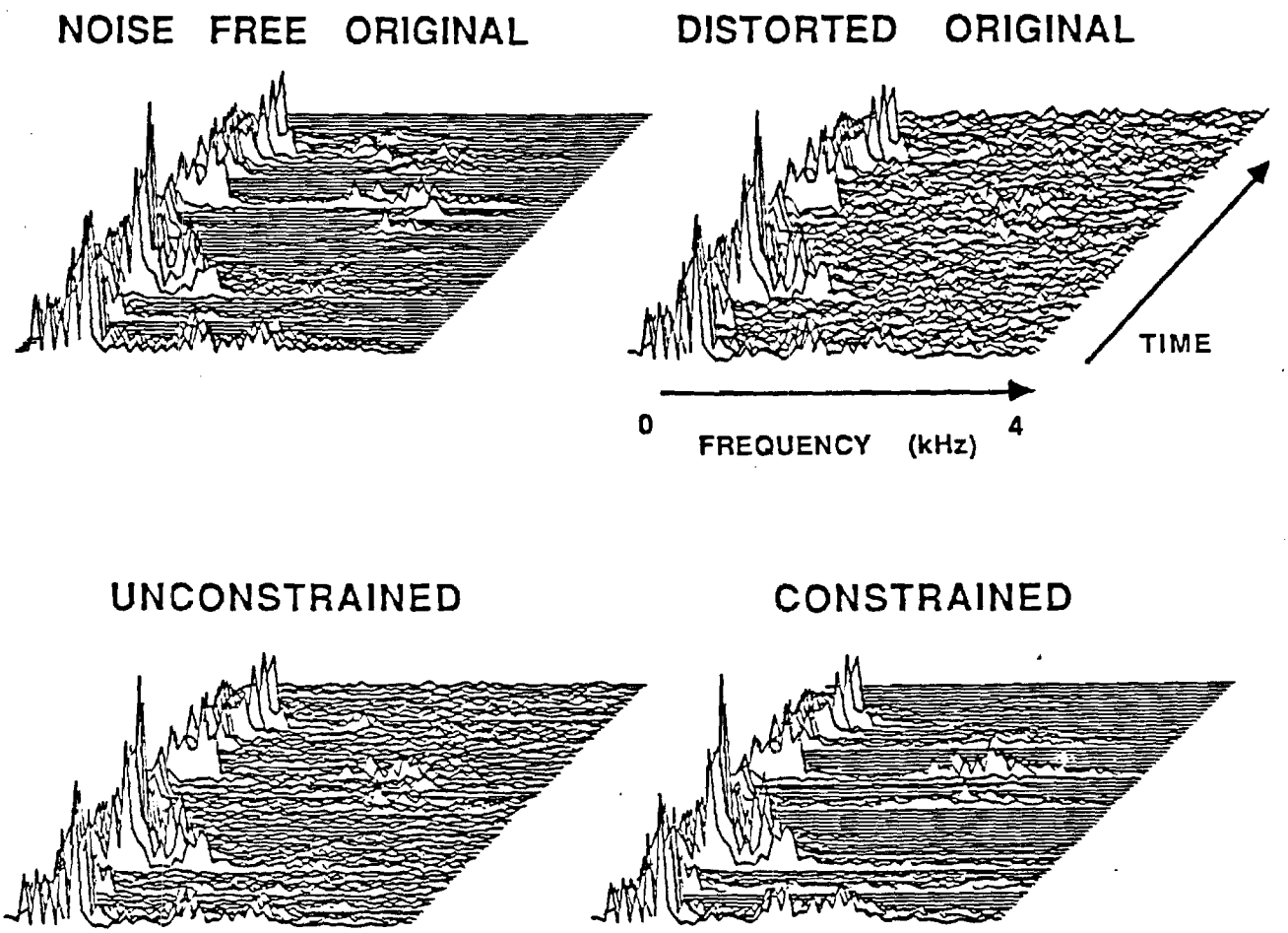


Figure 7: Time versus frequency plots of the sentence *Cats and dogs each hate the other.* The original and distorted original (additive white Gaussian noise, SNR = +5dB) are shown above. The lower left-hand plot is the response after three iterations of the unconstrained noncausal Wiener filtering approach. The lower right-hand plot is the frequency response after six iterations of an inter- plus intra-frame constrained (FF-LSP:T,Auto:I) approach.

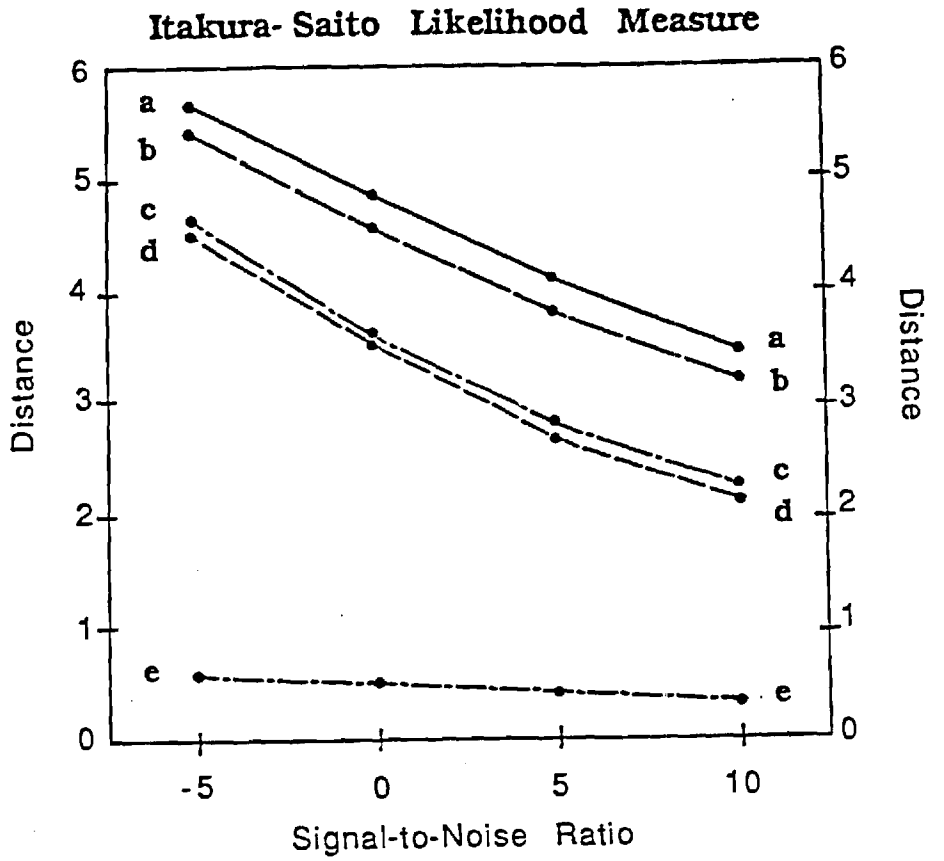


Figure 8: Comparison of inter & intra-frame constrained enhancement algorithms for colored aircraft noise over SNR.

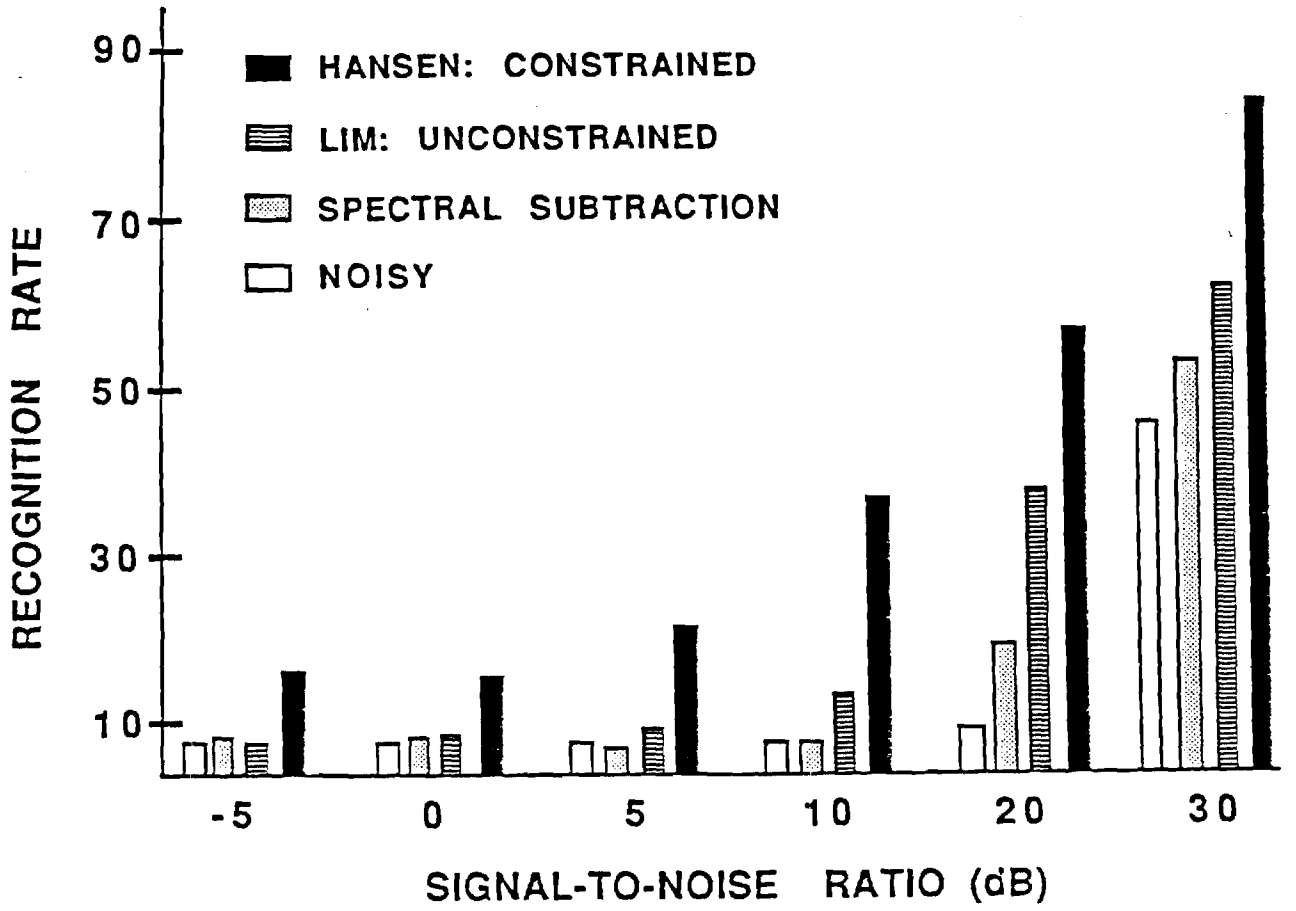
a.) **Original Distorted Speech**

b.) **Generalized unconstrained Wiener filtering**

c.) **Hansen-Clements: employing Inter-Frame constraints (FF-LSP:T)**

d.) **Hansen-Clements: employing Inter & Intra-Frame constraints (FF-LSP:T,Auto:I)**

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



Condition	*RECOGNITION RESULTS*						
		Signal-to-Noise Ratio					
(noise-free training)	*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%

Figure 9: Recognition performance using enhancement preprocessing in additive white Gaussian noise. †SNR = +10dB