Codebook Constrained Wiener Filtering for Speech Enhancement

T. V. Sreenivas and Pradeep Kirnapure

Abstract— **Speech enhancement using iterative Wiener filtering has been shown to require interframe and intraframe constraints in all-pole parameter estimation. In this correspondence, we show that a clean speech VQ codebook is more effective in providing intraframe constraints and, hence, better convergence of the iterative filtering scheme. Satisfactory speech enhancement results are obtained with a small codebook of 128, and the algorithm is effective for both white noise and pink noise up to 0 dB SNR.**

I. Introduction

Speech processing systems, viz., speech coders, speech recognizers, etc., have traditionally been developed for a noise-free environment. The presence of background noise can severely degrade the performance of such systems. One way to improve the performance of such systems is to develop an enhancement preprocessor that will produce speech or its recognition features that are less sensitive to background noise. Speech enhancement methods attempt to improve either the *subjective quality* of speech to reduce listener fatigue or improve the *intelligibility*. In many applications, the intelligibility of speech is of central importance, and it would be generally acceptable to sacrifice quality if the intelligibility could be improved. There are several techniques of speech enhancement [3]. The present work is an extension of iterative Wiener filtering technique proposed by Lim and Oppenheim [5]. This technique is known for its theoretical appeal. Its performance, however, is limited because of the difficulty in convergence of the iterative algorithm. Hansen and Clements [2] have proposed a constrained iterative algorithm that is shown to have better convergence properties. The present work is a new approach to incorporating constraints, which results in much faster convergence. In addition, since the constraints are derived directly from the clean speech in an unsupervised learning mode, the new algorithm leads to even better speech enhancement.

II. Iterative Wiener Filter (IWF)

This technique is a sequential maximization of the *a posteriori* probability (MAP estimation) of the speech signal and its all-pole parameters as originally formulated by Lim and Oppenheim [5]. In this method, the speech signal is modeled as the response of an all-pole system, and the approach is to solve for the MAP estimate of the signal $\mathbf{s}$, given the noisy signal $\mathbf{y} = \mathbf{s} + \mathbf{d}$. However, the resulting equations for the joint MAP estimate of the all-pole speech parameter vector $\mathbf{a}$, gain G, and noise-free speech vector $\mathbf{s}$ are found to be nonlinear. In order to simplify the solution, Lim and Oppenheim proposed a suboptimal iterative solution using sequential estimation of $\mathbf{a}$ and G, given $\mathbf{s}_i$, where $\mathbf{s}_i$ is the estimated signal at the ith iteration. The sequential estimation procedure is linear at each iteration and would continue until a criterion of convergence is satisfied. They assumed the *a posteriori* pdf $p(\mathbf{s} \mid \mathbf{a}_i, \mathbf{y})$ to be Gaussian. Consequently, the MAP estimate of $\mathbf{s}$ based on maximizing the pdf is equivalent to the minimum mean square error (MMSE) estimation of $\mathbf{s}$. Further, as the number of data samples available for estimation increases, the procedure for obtaining the MMSE of $\mathbf{s}$ approaches a noncausal Wiener filter, i.e., $\mathbf{s}$ is estimated by filtering $\mathbf{y}$ through a noncausal Wiener filter. It is shown [5] that this technique increases the joint likelihood of $\mathbf{a}$ and $\mathbf{s}_i$ with each iteration and leads to the joint MAP estimate. Fig. 1 gives a block diagram of the iterative Wiener filter scheme. The transfer function of the noncausal Wiener filter [9] is given by

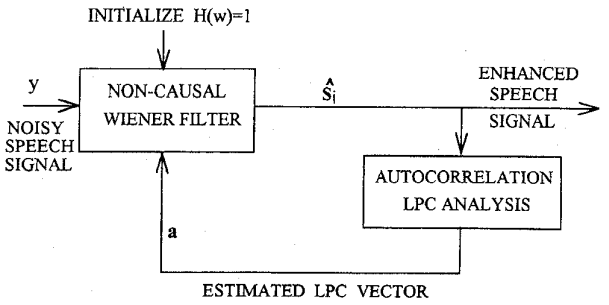

Fig. 1. Linearized MAP algorithm for estimation of all-pole parameters and speech vector from noisy speech vector.

$$H(\omega) = \frac{P_s(\omega)}{P_s(\omega) + P_d(\omega)}. \tag{1}$$

$P_d(\omega)$ is the power spectral density (psd) of the additive noise, and $P_s(\omega)$ is the psd of the clean speech signal. For the given model parameter values of $\mathbf{a}_i$ and G

$$P_s(\omega) = G^2 \Big/ \left| 1 + \sum_{m=1}^{p} a_i(m) \cdot e^{-j\omega m} \right|^2. \tag{2}$$

It may be noted that G is estimated from the energy of the input signal. Thus

$$\frac{1}{2\pi} \int_{-\pi}^{+\pi} \left[G^2 \Big/ \left| 1 + \sum_{m=1}^{p} a_i(m) e^{-j\omega m} \right|^2 \right] d\omega = \frac{1}{N} \sum_{n=0}^{N-1} y^2(n) - \sigma_d^2 \tag{3}$$

where σ_d^2 is the variance of noise assumed to be stationary.

III. Constrained IWF

Hansen and Clements [2] performed an extensive investigation of the IWF technique for speech with additive white Gaussian noise and found some anomalies:

i) With increasing number of iterations, the individual formants of speech consistently decrease in bandwidth. In addition, the formant frequencies drift from the correct values. This results in unnatural sounding reconstructed speech.

ii) Pole frequency jitter in the all-pole model is observed between successive frames, causing artificial discontinuities in formant contours.

iii) Although the sequential MAP estimation technique should theoretically increase the joint likelihood of the speech wave-

Manuscript received June 29, 1994; revised March 3, 1996. The associate editor coordinating the review of this paper and approving it for publication was Prof. John H. L. Hansen.

T. V. Sreenivas is with the Department of Electrical Communication Engineering, Indian Institute of Science, Bangalore 560012, India (e-mail: tvsree@ece.iisc.ernet.in).

P. Kirnapure is with Hughes Software Systems, New Delhi, India.

Publisher Item Identifier S 1063-6676(96)06715-6.

form and its all-pole parameters, proper convergence criteria is necessary for optimum results.

In the original algorithm [5], the convergence criteria is not specified because, according to the MAP formulation, the probability would improve monotonically with iterations. However, the true measure of speech enhancement is through the perception of $\hat{s}$. Using an objective measure of speech perception quality, viz., Itakura–Saito (IS) likelihood measure [8], Hansen and Clements [2] experimentally determined the optimum number of iterations for best perception of enhanced speech. It is found that different classes of speech sounds require a different number of iterations for optimum performance. The obvious problem with such a criterion is that outside of simulation, the clean speech is unavailable, and hence, comparative evaluation is not possible. However, simulation results indicated that the number of iterations leading to maximum objective quality measure is consistent for a specific speech class.

Hansen and Clements also proposed to incorporate constraints in estimating the all-pole model. They suggested the use of speech specific constraints in either the LPC based representation or line spectral pair (LSP) representation of the all-pole model. Constraints are imposed on the estimated vocal tract spectrum at each iteration. The constraints applied to $\mathbf{a}$ ensure the following:

i) The all-pole model is stable.
ii) It possesses speech-like characteristics (poles are in reasonable positions with respect to each other and the unit circle),
iii) The vocal tract characteristics do not vary too much between successive frames.

The procedure for determining G remains the same. The constraints have been shown to result in a consistently superior objective measure of performance at convergence. Surprisingly, in addition, the convergence occurred at the *seventh* iteration for *all* sound classes. Experiments showed that the "optimum" iteration number is also satisfactory over a range of SNR's. However, the same experiments also indicate that the performance degrades before or beyond the "optimum" number of iterations.

Considering the above experiments, there is clearly a need for a better criteria of convergence. In addition, better convergence should lead to a better estimate of $\mathbf{a}$ and, thus, better enhancement performance.

IV. Codebook Constrained IWF

Speech signals are highly dynamic in nature, and accurate perception of phonemes, such as semivowels, stop consonants, nasals, etc., depends on the precise changes in the signal spectra. Hence, dynamic characteristics should be retained in the enhanced signal. Application of interframe constraints to LPC or LSP parameters, such as simple smoothing [2], can be detrimental to intelligibility of enhanced speech. Further, the intraframe formant frequency constraints derived from acoustic-phonetic knowledge of speech is limited in scope because of nonunique mapping of poles to formants and large variability among talkers. These difficulties of applying explicit human derived knowledge of speech spectra can be overcome using unsupervised learning techniques to derive the requisite knowledge. A simple approach to this is through pattern clustering of clean speech spectra. The time-varying nature of formant frequency contours and the strong correlation between formant frequencies can be effectively captured using a codebook of formant contour segments [10]. In the present work, we explore the effectiveness of applying only intraframe constraint using a codebook approach.

Let $\{\mathbf{a}\}$ be a set of LPC vectors derived from clean speech data of several continuously spoken sentences. The perceptual difference between two LPC vectors is well correlated to the IS distortion

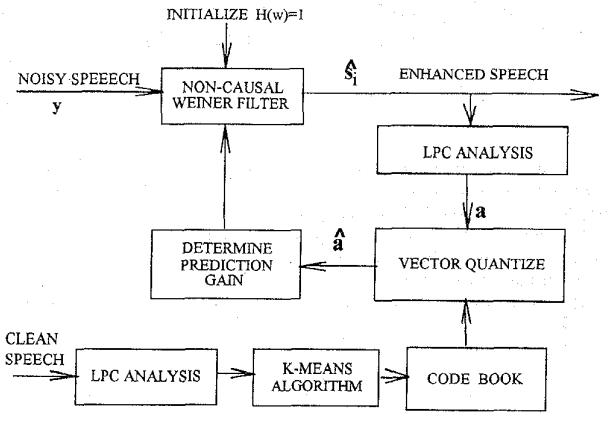

Fig. 2. Iterative speech enhancement based on vector quantization of LPC parameters.

measure. A close approximation to the ML measure has been widely used in vector quantization [6], [7], viz.

$$d(\mathbf{a}_x, \mathbf{a}_y) = (\mathbf{a}_x - \mathbf{a}_y)^T \mathbf{R}_x (\mathbf{a}_x - \mathbf{a}_y) \tag{4}$$

where $\mathbf{a}_x$ and $\mathbf{a}_y$ are any two LPC vectors of the same order, i.e., $[a_1, a_2, \ldots, a_p]^T$, and $\mathbf{R}_x$ is the normalized autocorrelation matrix corresponding to the LPC vector $\mathbf{a}_x$.

$$\mathbf{R}_x(i,k) = \{\mathbf{E}[x(n-i)x(n-k)]\}/\{\mathbf{E}[x^2(n)]\} \tag{5}$$

Using the IS distortion measure, the problem of pattern clustering can be stated as follows: Let the variety of valid speech spectra be represented by a codebook of LPC vectors $\{\mathbf{c}_k; 1 \le k \le K\}$, where K is of the order of 1024 (since such a codebook size has been found sufficient in speech coding applications). Each $\mathbf{c}_k$ is a representative of a cluster of vectors $S_k = \{\mathbf{a}_i, 1 \le i \le N_k\}$ such that $\frac{1}{N_k} \sum_{\forall \mathbf{a} \in S_k} d(\mathbf{a}, \mathbf{c}_k)$ is minimum. Now, clustering implies finding $\mathbf{c}_k$ such that

$$\bar{D} = \frac{1}{K} \sum_{k=1}^{K} \frac{1}{N_k} \sum_{\forall \mathbf{a} \in S_k} d(\mathbf{a}, \mathbf{c}_k) \tag{6}$$

is minimum. This is effectively solved using the iterative K-means algorithm [6]. Over the iterations, the centroid of a cluster is found by [6]

$$\mathbf{c}_k = \left[\sum_{i:\ \mathbf{a}_i \in S_k} \mathbf{R}_i \right]^{-1} \left[\sum_{i:\ \mathbf{a}_i \in S_k} \mathbf{R}_i \mathbf{a}_i \right]. \tag{7}$$

The matrix $\sum \mathbf{R}_i$ is positive definite and Toeplitz. Hence, Levinson's algorithm [8] can be used to calculate the centroid. The centroids, thus found, represent the most often occurring spectra of speech. In addition, the error caused in speech due to quantization of $\mathbf{a}$ by the nearest neighbor c_j, i.e.,

$$d(\mathbf{a}, \mathbf{c}_j) \le d(\mathbf{a}, \mathbf{c}_k) \quad \forall k \tag{8}$$

is small for a large size codebook and can be assumed to be perceptually not significant.

In the iterative Wiener filtering algorithm at each iteration the LPC vector $\mathbf{a}$ is estimated from the estimated speech $\hat{\mathbf{s}}$. The random errors in formants of $\mathbf{a}$ due to noise in $\hat{\mathbf{s}}$ can be corrected in a perceptual sense by choosing the codebook vector $\mathbf{c}_i$ closest to $\mathbf{a}$, i.e.,

$$d(\mathbf{a}, \mathbf{c}_i) \le d(\mathbf{a}, \mathbf{c}_k) \quad \forall k. \tag{9}$$

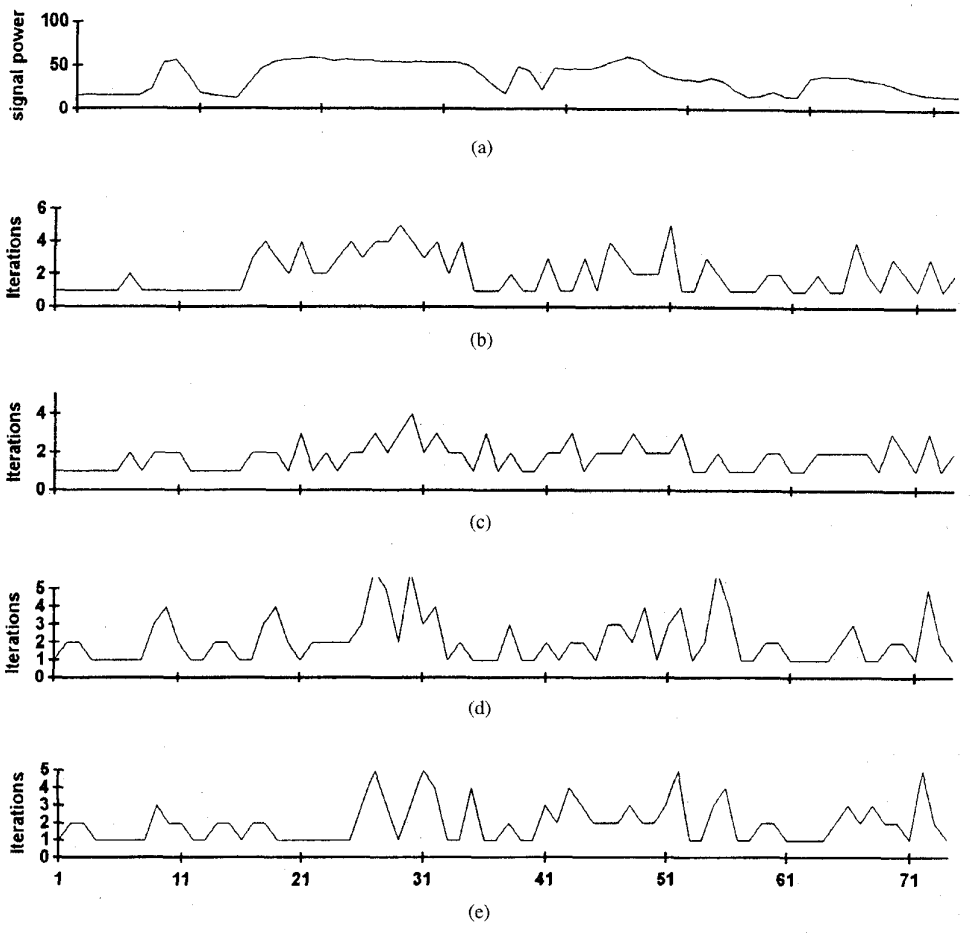

Fig. 3. Convergence of *Codebook constrained iterative Wiener filtering* algorithm at different frames of the utterance "a tube tent" (male voice): (a) Short-time energy contour. (b) $K = 128$, SNR= 0 dB. (c) $K = 128$, SNR= 10 dB. (d) $K = 1024$, SNR= 0 dB. (e) $K = 1024$, SNR= 10 dB.

Since $\mathbf{c}_i$ has been estimated from clean speech, it satisfies the assumption of clean signal power spectrum for the Wiener filter in (2), i.e.,

$$P_s(\omega) = G^2 \Big/ \left|1 + \sum_{m=1}^{p} c_i(m) e^{-j\omega m}\right|^2. \tag{10}$$

The rest of the iterative algorithm can be continued until $\hat{\mathbf{s}}$ reaches a stationary point as in the original IWF algorithm.

V. Experiments

The design of the LPC codebook needs a large number of vectors for training. The data used for codebook design is recorded speech sentences of eight male and four female speakers for a duration of about 170 and 135 s, respectively. The LPC parameters are extracted using a quasistationary analysis with consecutive frames overlapping. These LPC parameters are directly used for training the quantizer using the distortion measure of (4). The K-means algorithm [6] with a splitting codebook initialization approach is used for the codebook design. Some of the parameters in the preprocessing are sampling frequency = 8 kHz, LPC model order $p = 10$, frame window width $N_w = 20$ ms, and successive frame overlap = 15 ms.

For the purpose of simulation, the pseudo-random numbers with Gaussian distribution are used as the interfering noise. The Box–Muller method [9] is used to generate the random number sequence. Stationary white Gaussian noise is added to the clean speech signal to get the noisy signal. The variance of the noise to be added determines the SNR of the signal, i.e., $y(n) = s(n) + \sigma w(n)$ and

$$\text{SNR} = 10\log_{10}\left[\left(\frac{1}{N}\sum_{n=1}^{N} s^2(n)\right) \Big/ \sigma^2\right] \text{dB} \tag{11}$$

where N is the total number of samples in the test utterance. The noisy signal $\mathbf{y}$ is processed as successive quasistationary frames of duration 20 ms. Codebook constrained iterative Wiener filtering (Fig. 2) is applied to each frame independently to obtain a succession of frames of $\hat{\mathbf{s}}$. The Wiener filtering is based on MMSE criterion, which suggests the use of MSE between the estimated signal and the clean speech signal (which is known in the experiment) as a measure of performance of the enhancement algorithm. The MSE in relation to signal power can be defined as the segmental SNR, which is often used to measure the performance of speech coders.

$$\text{SNR\_seg} = 10\log_{10}\left[\frac{\mathbf{S}^T\mathbf{S}}{(\mathbf{S}-\hat{\mathbf{S}})^T(\mathbf{S}-\hat{\mathbf{S}})}\right] \text{dB}. \tag{12}$$

For the overall speech signal, an average performance is defined as

$$\text{SNR\_seg\_avg} = \mathbf{E}\{\text{SNR\_seg}\} \tag{13}$$

where the expectation is over all the speech frames. The SNR_seg_avg is a good measure of the perceptual speech quality.

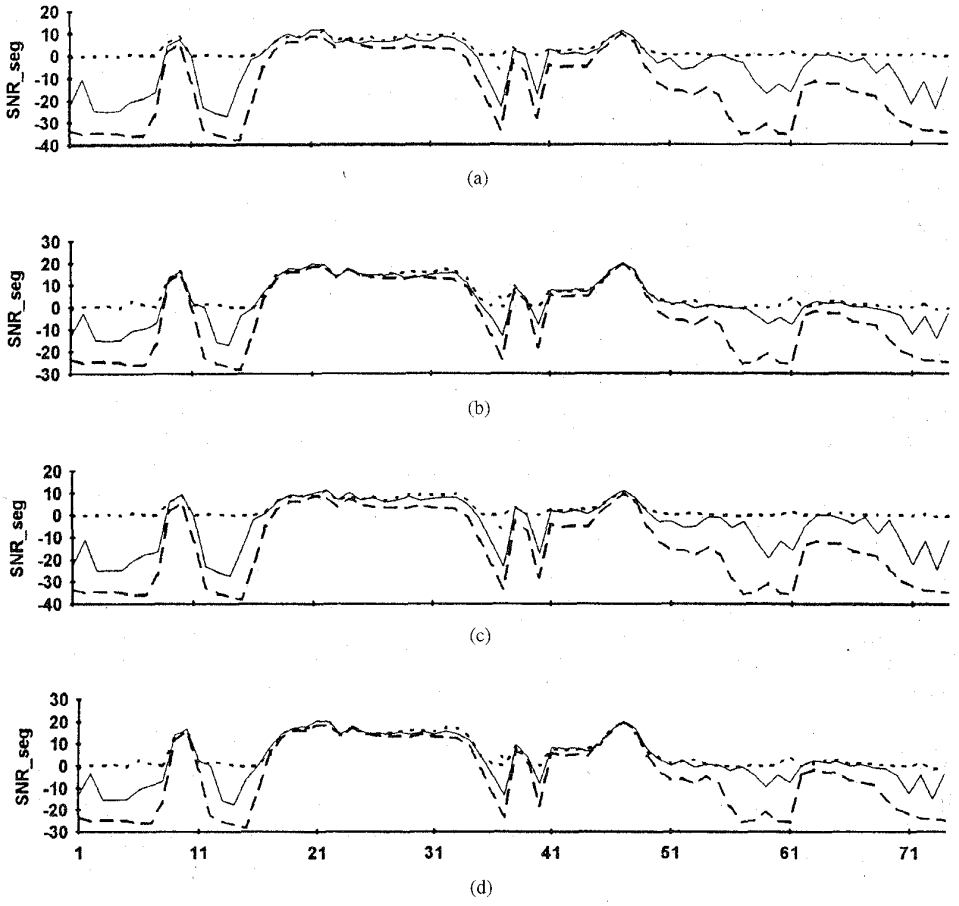

Fig. 4. Speech enhancement performance of the *Codebook constrained iterative Weiner filtering (CCIWF)* algorithm; utterance: "a tube tent" (male voice). - - - noisy input signal, ··· performance of CSLWF, —— performance of CCIWF.

TABLE I
WHITE NOISE PERFORMANCE: SNR-seg-avg, AVERAGED OVER NINE SETENCES, OBTAINED FOR DIFFERENT SPEECH ENHANCEMENT ALGORITHMS AT THREE LEVELS OF INPUT SNR

	Input SNR=10 dB (SNR_seg_avg = 1.00 dB)	*Input SNR=5dB (SNR_seg_avg = -3.3 dB)*	*Input SNR=0dB (SNR_seg_avg = -6.1 dB)*
Clean speech LPC Wiener filter	9.00 dB	6.55 dB	4.43 dB
CCIWF (CB=1024)	7.07 dB	3.40 dB	-0.31 dB
CCIWF (CB=512)	7.14 dB	3.39 dB	-0.26 dB
CCIWF (CB=256)	7.06 dB	3.37 dB	-0.32 dB
CCIWF (CB=128)	7.00 dB	3.30 dB	-0.37 dB
Spectral Subtraction.	0.02 dB	-1.79 dB	-3.96 dB

The codebook constrained iterative enhancement algorithm was run on nine sentences totaling about 10 s of speech comprised of two male and two female speakers with a British and an Indian accent. The only parameter varied in the algorithm is the codebook size K. While designing the codebook using the splitting approach, optimum code books of size 128, 256, 512, and 1024 are all saved. The nine sentences are corrupted with various levels of noise corresponding to SNR = 0, 5, and 10 dB.

A. White Noise Results

Fig. 3 shows the performance of the new algorithm with regard to convergence. Interestingly, in about 75% of cases the algorithm

TABLE II
PINK NOISE PERFORMANCE: SNR-seg-avg, AVERAGED OVER ONE SETENCE, "IT NEVER RAINS, BUT IT POURS IN BOMBAY, AT LEAST," OBTAINED FOR DIFFERENT SPEECH ENHANCEMENT ALGORITHMS AT THREE LEVELS OF INPUT SNR

	Input SNR=20 dB (SNR_seg_avg = 9.04 dB)	*Input SNR=10 dB (SNR_seg_avg = -1.0 dB)*	*Input SNR=0 dB (SNR_seg_avg = -11 dB)*
Clean speech LPC Wiener filter	12.84 dB	7.57 dB	3.28 dB
CCIWF (CB=1024)	12.53 dB	4.93 dB	-2.72 dB
CCIWF (CB=512)	12.44 dB	4.91 dB	-2.71 dB
CCIWF (CB=256)	12.48 dB	4.95 dB	-2.79 dB
CCIWF (CB=128)	12.53 dB	4.97 dB	-2.67 dB
Spectral Subtraction.	10.9 dB	1.61 dB	-7.79 dB

converged within three iterations and, in most cases, within five iterations. In addition, the number of iterations at convergence is found to be not related to speech type or input segmental SNR, unlike in the Hansen–Clements algorithm [2]. The enhancement performance of the new algorithm is shown in Fig. 4. It can be seen that the low level regions of the speech signal have very poor segmental SNR (<-10 dB), and high-level regions have good segmental SNR (>10 dB), averaging to the required SNR level. The improvement in SNR_seg is large at low level regions than at high level regions. The 1–2 dB improvement at high-level regions may not be perceptually significant. However, in the 0–5 dB SNR_seg range, an improvement of 2–3 dB contributes significantly to the perceptual quality. These regions usually correspond to the transitional parts of speech comprising consonants, hence, leading to improved intelligibility of enhanced speech. The low-level regions (<0 dB SNR_seg) correspond to regions of closure in stop consonants or weak fricatives. While there may not be much hope of estimating the correct signal spectrum in these regions, one can only expect improvement by way of attenuating the noise to the original signal level. This is evident from the theoretical performance limit shown in the figure, which is obtained using the clean speech model parameters a and G in the Wiener filter. For speech segments in the 0–5 dB SNR_seg range, the slightly higher values of the theoretical limit than the performance obtained using the CCIWF algorithm indicates the further scope for improvement.

Table I summarizes the relative performance of CCIWF over all the utterances compared with the spectral subtraction algorithm [1]. Considering the improvement in SNR_seg_avg between input and output, we can see that the CCIWF algorithm provides about 6-dB improvement even for the 0-dB SNR signal. For the 10-dB case, the spectral subtraction method results in a *reduction* in SNR_seg_avg. The reduction in SNR_seg_avg is a good indicator of the lack of improvement in intelligibility using that method. The uniform 6-dB improvement in SNR_ seg_avg using CCIWF is better compared with the 4–5-dB SNR enhancement results of the recent HMM-based algorithm [11]. The HMM-based algorithm performed poorly for SNR less than 10 dB, whereas the performance of CCIWF is satisfactory even at 0 dB SNR. The performance of CCIWF for different code books indicates that the algorithm is not sensitive to the size of the codebook. However, the enhancement due to the 1024 codebook is found to be perceptually better than that due to the 128 size codebook, particularly at low SNR. The SNR_seg_avg achieved using the unquantized clean speech LPC vector Wiener filtering (CSLWF) is also shown in the table. This sets the upper limit to the enhancement performance of the Wiener filter formulation. It can be seen that there is scope for improving the SNR_seg_avg by another 5 dB through better estimation of the LPC vector from the noisy data. In addition, perceptually, the CSLWF-enhanced speech shows very good intelligibility and listenability.

B. Pink Noise Results

The stationary white noise considered in the previous section is quite restrictive, and in many applications, the noise is nonwhite as well as nonstationary. To determine the effectiveness of CIWF algorithm for practical applications, Hansen and Clements [2] have tested for nonwhite noise of an aircraft. In this experiment, pink noise generated using an analog noise generator and digitized using an A/D converter is used. The power spectral density (psd) of noise is found to have a first-order AR characteristic. One speech sentence of about 3.5 s from the previous set of nine test sentences is chosen. Pink noise is added at SNR levels of 0, 10, and 20 dB. A higher SNR value is included because pink noise is harder for enhancement than white noise. For the Wiener filter in CCIWF, the *a priori* computed psd of pink noise is used. For comparison, the noisy signal is also enhanced using the spectral subtraction method as well as the CSLWF method.

Table II shows the average segmental SNR for the single sentence. In comparison with Table I, it can be seen that the overall performance of CCIWF is as good as that in the case of white noise corruption. The input-to-output improvement in segmental SNR is at least 6 dB in all cases. Even at 0 dB SNR, there is an enhancement of about 8 dB, which is surprisingly better than the white noise case. As before, even a small codebook of size 128 provides a performance comparable with that of 1024. In contrast, the spectral subtraction method provides only about 2–3 dB improvement. The number of iterations for convergence of the CCIWF algorithm for the case of pink noise is only slightly more than that for white noise: about 85% of the frames within five iterations and 99% within eight iterations.

Fig. 5 shows a comparison of the SNR_seg plots for the different enhancement methods, revealing some of the characteristics of the new algorithm. The clean signal power in Fig. 5(a) fluctuates more than 40 dB, and the SNR_seg of the input signal in Fig. 5(b) follows the same trend. The CCIWF output consistently enhances the mid-SNR regions, which helps to improve stop consonant perception. The low-SNR valleys also get enhanced, increasing the SNR_seg_avg. The spectral subtraction algorithm, instead, shows a moderate improvement, uniformly retaining the SNR profile of the input signal. In the case of the CSLWF method, all the low-SNR valleys are pulled up

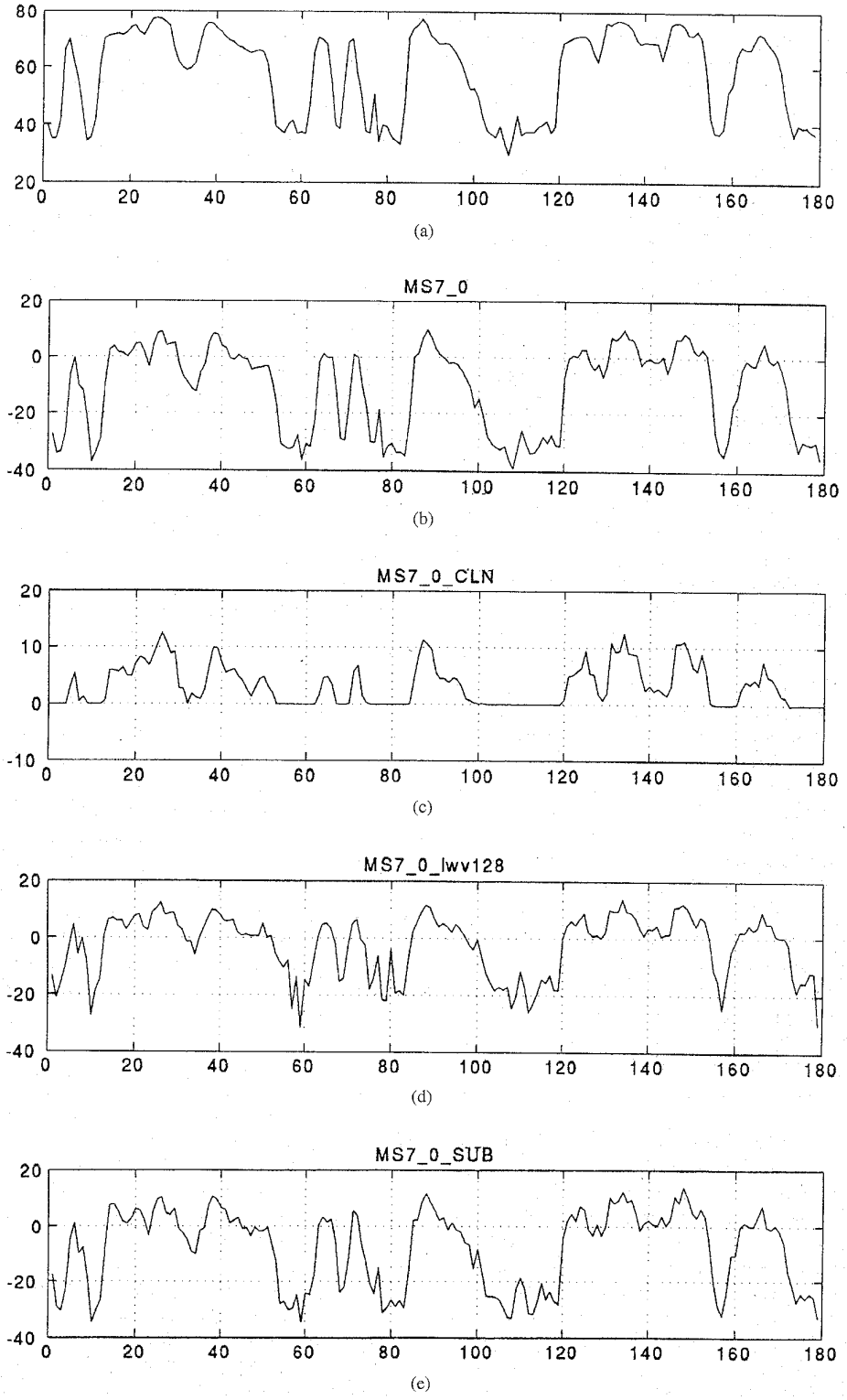

Fig. 5. Input and output signal SNR_seg for the utterance "it never rains, but it pours in Bombay at least," at 0 dB SNR: (a) Clean signal segmental power. (b) Input noisy signal SNR_seg. (c) Clean signal LPC Wiener filter output. (d) 128 size CCIWF output. (e) Spectral subtraction output.

to a uniform level of 0 dB. This is an interesting case showing the upper limit to enhancement for very low SNR segments. The 0-dB upper limit can be explained as shown in the Appendix. However, for mid-SNR and high-SNR segments, the performance is similar to that of the CCIWF algorithm.

VI. Conclusion

Lack of proper convergence criteria is not a limitation to the Lim and Oppenheim iterative Wiener filtering algorithm for speech enhancement. Proper constraints have to be applied to the parameter

estimation from the noisy signal. As shown in this correspondence, the clean speech derived codebook constrained approach is very effective. In addition, we have shown the enhancement performance limit using the clean speech parameters and accordingly, very low SNR segments can be improved only upto 0 dB SNR. Other segments with moderate and high SNR can only be slightly improved using further improvement in parameter estimation.

APPENDIX

The iterative Wiener filtering can be expressed as

$$\mathbf{S}_i(\omega) = \mathbf{Y}(\omega)\frac{\hat{\mathbf{P}}_s(\omega)}{\hat{\mathbf{P}}_s(\omega) + g^2\mathbf{P}_d(\omega)}; \quad i = 1, 2, \ldots$$

where

$$\hat{\mathbf{P}}_s(\omega) = \frac{G_{i-1}^2}{|A_{i-1}(\omega)|^2}$$

and

$$A_{i-1}(z) = \left[1 + \sum_{k=1}^{p} a_k z^{-k}\right]$$

is estimated from $\hat{\mathbf{s}}_{i-1} \Leftrightarrow \hat{\mathbf{S}}_{i-1}(\omega)$ (Fourier transform pair). In addition, $\mathbf{Y}(\omega) \Leftrightarrow \mathbf{y} = \mathbf{s} + g\mathbf{d}$, where $\sigma_d^2 = 1$ and SNR $= \sigma_s^2/g^2$. For very low SNR, $g^2 \gg \sigma_s^2$.

Considering such a low SNR case and clean speech LPC vector Wiener filtering where iterations become redundant, we can write

$$\hat{\mathbf{S}}_i(\omega) = \hat{\mathbf{S}}(\omega) = \mathbf{Y}(\omega)\frac{\hat{\mathbf{P}}_s(\omega)}{\hat{\mathbf{P}}_s(\omega) + g^2\mathbf{P}_d(\omega)} \cong \mathbf{Y}(\omega)\frac{\hat{\mathbf{P}}_s(\omega)}{g^2\mathbf{P}_d(\omega)}.$$

Defining the estimation error signal spectrum as $\mathbf{E}(\omega) = \mathbf{S}(\omega) - \hat{\mathbf{S}}(\omega)$ and substituting for $\mathbf{Y}(\omega) = \mathbf{S}(\omega) + g\mathbf{D}(\omega)$, we get

$$\mathbf{E}(\omega) = \mathbf{S}(\omega)\left[1 - \frac{\hat{\mathbf{P}}_s(\omega)}{g^2\mathbf{P}_d(\omega)}\right] - \frac{g\mathbf{D}(\omega)\hat{\mathbf{P}}_s(\omega)}{g^2\mathbf{P}_d(\omega)} \cong \mathbf{S}(\omega) - \frac{\mathbf{D}(\omega)\hat{\mathbf{P}}_s(\omega)}{g\mathbf{P}_d(\omega)}.$$

Now, let us consider the error signal power. Substituting for $\hat{\mathbf{P}}_s(\omega)$ and $\mathbf{P}_d(\omega)$, we get

$$|\mathbf{E}(\omega)|^2 \approx \left||\mathbf{S}(\omega)|\exp\{\theta_s(\omega)\} - \frac{|\mathbf{S}(\omega)|^2\mathbf{D}(\omega)\exp\{\theta_d(\omega)\}}{g|\mathbf{D}(\omega)|^2}\right|^2$$

where $\theta(\omega)$ are the respective phase spectra. Further simplifying, we get

$$|\mathbf{E}(\omega)|^2 \approx |\mathbf{S}(\omega)|^2\left|\exp\{\theta_s(\omega)\} - \frac{|\mathbf{S}(\omega)|\exp\{\theta_d(\omega)\}}{g|\mathbf{D}(\omega)|}\right|^2.$$

The second term within the modulus is negligible because $g \gg 1$ for the case of very low SNR considered here. Thus, $|\mathbf{E}(\omega)|^2 \approx |\mathbf{S}(\omega)|^2$, which implies a 0 dB segmental SNR of the output enhanced signal.

ACKNOWLEDGMENT

The authors are thankful to S. Sunil, Project Assistant, IISc, for helping to modify the software for colored noise experiments. They also thank the Rice University DSP Group for providing the various noise data through the Internet.

REFERENCES

[1] S. F. Boll, "Suppression of acoustic noise in speech using spectral subtraction," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 27, no. 2, pp 113–120, Apr. 1979.
[2] J. H. L. Hansen and M. A. Clements, "Constrained iterative speech enhancement with application to speech recognition," *IEEE Trans. Signal Processing*, vol. 39, no. 4, pp 795–805, Apr. 1991.
[3] J. S. Lim, Ed., *Speech Enhancement*. Englewood Cliffs, NJ: Prentice-Hall, 1983.
[4] J. S. Lim and A. V. Oppenheim, "All-pole modeling of degraded speech," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-26, no. 3, pp. 197–210, June 1978.
[5] J. S. Lim and A. V. Oppenheim, "Enhancement and bandwidth compression of noisy speech," *Proc. IEEE*, vol. 67, no. 12, pp. 1586–1604, Dec. 1979.
[6] Y. Linde, A. Buzo, and R. Gray, "An algorithm for vector quantizer design," *IEEE Trans. Commun.*, vol. COM-28, no. 1, pp. 84–94, Jan. 1980.
[7] J. Makhoul, "Vector quantization in speech coding," *Proc. IEEE*, vol. 73, no. 11, Nov. 1985.
[8] J. D. Markel and A. H. Gray, *Linear Prediction of Speech*. New York: Springer Verlag, 1976.
[9] A. Papoulis, *Probability, Random Variables, and Stochastic Processes*. New York: McGraw-Hill, 1965.
[10] N. Sunder, "VQ of formant contours for speech enhancement," M.E. Project Rep., Dept. of Elect. Commun. Eng., Indian Inst. of Sci., Bangalore, June 1992.
[11] Y. Ephraim, "Statistical model based speech enhancement systems," in *Proc. IEEE*, vol. 80, no. 10, pp. 1526–1555, Oct. 1992.

Modeling Acoustic Transitions in Speech by Modified Hidden Markov Models with State Duration and State Duration-Dependent Observation Probabilities

Y. K. Park, C. K. Un, *Fellow, IEEE*, and O. W. Kwon

***Abstract*—We propose a modified hidden Markov model (MHMM) that incorporates nonparametric state duration and state duration-dependent observation probabilities to reflect state transitions and to have accurate temporal structures in the HMM.**

In addition, to cope with the problem that results from the use of insufficient amount of training data, we propose to use the modified continuous density hidden Markov model (MCDHMM) with a different number of mixtures for the probabilities of state duration-independent and state duration-dependent observation. We show that this proposed method yields improvement in recognition accuracy in comparison with the conventional CDHMM.

I. INTRODUCTION

It is well known that one major weakness of the conventional hidden Markov model (HMM) is its inaccurate modeling of state durations and state transitions [1]. The conventional HMM treats the spectral modeling and the duration modeling as being separate or loosely connected [1]. However, we note that transient portions, as well as steady portions of speech signal, play an important role in

Manuscript received March 23, 1994; revised March 3, 1996. The associate editor coordinating the review of this paper and approving it for publication was Prof. John H. L. Hansen.

The authors are with the Communication Research Laboratory, Department of Electrical Engineering, Korea Advanced Institute of Science and Technology, 373-1 Kusong-Dong, Yusong-Ku, Taejon 305-701, Korea (e-mail: ckun@eekaist.kaist.ac.kr).

Publisher Item Identifier S 1063-6676(96)06717-X.