VQ Based Techniques in Speech Enhancement

Sharath Rao K. , *Student Member, IEEE,*
SP 1.08, Dept of ECE,
Indian Institute of Science
email : sharath@udukku.ece.iisc.ernet.in

***Abstract*— The Wiener filter has been a popular technique for single and dual channel speech enhancement and has been widely studied. In this paper, we study its performance in the iterative mode. The Iterative wiener filter is found to perform better when spectral constraints are imposed across iterations and time. We explore the use of quantization techniques in different LP parameter domains in imposing constraints. The role of initialization of iterations is illustrated via an initialization technique which results in improved all-pole model estimation and fewer iterations. Results in terms of enhancement performance and convergence are presented. Through objective measures and informal listening tests, we show that operating in a robust parameter domain and suitably initializing the iterations can significantly improve the performance of the Iterative Wiener filtering (IWF) technique.**

I. Introduction

WHEN speaker and listener are close to each other in a quiet environment, communication is generally easy and accurate. However, at a distance or in a noisy background, the ability to understand speech decreases. This apart, when speech is sent electronically, the conversion and transmission media introduce distortions yielding a noisy speech signal. Such degradation can lower the quality and intelligibility of speech. Thus, enhancement techniques play a vital role in making voice communication a viable option. In addition, the study of enhancement techniques is of interest in a wider context. It is well known that that the performance of these speech coders, recognition systems etc. tends to degrade if they are operating in environments that they are not designed for. Speech enhancement techniques, when used as a front-end preprocessors to these systems, help make these systems robust to noise. Several techniques for speech enhancement have been proposed. A survey can be found in [2].

In [1], Lim and Oppenheim proposed the iterative wiener filtering (IWF) technique for speech enhancement. In this technique, the estimation of the all-pole speech parameters of speech in additive white gaussian noise was posed as a two step sequential MAP estimation problem. Although theoritically appealing, this method suffered from the following drawbacks [3] : (i) it was found that increasing iterations caused the formant bandwidths to decrease and formant locations to shift. (ii) Frame to frame pole frequency jitter caused artifical discontinuities in the formant contours. (iii) No consistent convergence criterion was defined. These effects contributed to unnatural sounding speech and arose out of the fact that IWF merely optimizes a mathematical criterion, which may not always be well-correlated with perceptual aspects. Hansen and Clements addressed these issues by incorporating constraints in the all-pole model estimation which retained speech-like characteristics of the enhanced utterance. They suggested the use of the LSP parameters which, owing to their excellent interpolation properties, lend themselves well to imposing constraints.

Codebook constrainted iterative wiener filtering scheme (referred as CCIWF), a clustering based approach was proposed as an alternative technique of imposing constraints [4]. Here, the all-pole parameters were constrained to belong to a codebook of clean speech vectors. Apart from successfully defining a convergence criterion, this approach was quite effective in taking care of several type of speech constraints such as those between formants and speaker variability.

In this paper, we propose an initialization criterion to address the issue of faster and better convergence of IWF. Further, we also explore various parameter domain codebooks for noise robustness and therefore, their effectiveness in imposing intraframe constraints.

II. Codebook Based Constraints

The effectiveness of the VQ based method consists in successfully approximating the optimum filter through the codebook of clean speech vectors. Therefore, the parameter space used to represent these vectors has a significant bearing on these approximations. Line Spectral Frequencies (LSF), Reflection Coeffecients (RC) and Log Area Ratios (LAR), though share a one-to-one mapping, have different clustering properties due to the non-linear relationships between them. Hence, they have been used with varied success in speech coding and recognition [5]. In this study, we study of the behaviour of these different spaces in the VQ based IWF scheme.

The codebook based approach can be considered to be a two-stage problem - (i) Codebook Generation, where clean speech data of sufficient duration is clustered so as to be representative of a large number of speakers and acoustic-phoentic classes and (ii) VQ based IWF where all-pole model parameters are estimated using the above codebooks.

A. Codebook Generation

Let $\{\mathbf{a}\}$ be a set of LPC vectors derived from clean speech data. In order to create codebooks in different parameter domains, the vector a is first converted to the appropriate domain. Perceptually relevant and computationally affordable distance measures have to defined for each of these parameter spaces. The LBG algorithm [9] is then used to create a codebook of a fixed pre-determined size. In the present study, we used the Itakura Saito distance measure [4] to cluster LP coefficients and the Euclidean Distance (ED) to cluster data in the LAR and RC space. We have used the above since these have been widely accepted as effective distance measures and have been used in Vector Quantization and Speech Recognition. For LSFs, we used two perceptually motivated weighted Eucledean distances (WED) ; the Mel-Frequency Warping (MFW) based WED and the Inverse Harmonic Mean (IHM) based WED [6]. The inverse harmonic mean measure between any test LSF vector Ft and reference vector ,Fr is given as below :

$$d(f_t, f_r) = w_i * (f_t(i) - f_r(i))^2 \tag{1}$$

where the weights w_i are defined as :

$$w_i = s_i^2 * (\frac{1}{f_i - f_{i-1}} + \frac{1}{f_{i+1} - f_i}) \tag{2}$$

where $w_0 = 0$ and $w_{p+1} = \pi$
and f_t and f_r are the test vector and reference vector respectively.

The IHM based WED is perceptually relevant in the sense that it weighs each LSF in the inverse proportion of its closeness to its neighbours due to the better chance of it representing formants.

The MFW based weights are defined as :

$$w_i = (1 + \frac{2}{\omega_i} tan^{-1} \frac{a * sin\omega_i}{1 - acos\omega_i})^2 \tag{3}$$

It is an auditory motivated measure and has been successfully used in speech recognition [6].

B. VQ based Iterative Wiener Filtering

Fig 1. shows the VQ based IWF scheme. The non-causal wiener filter is defined as :

$$H(\omega) = \frac{P_s(\omega)}{(P_s(\omega) + P_d(\omega))} \tag{4}$$

where $P_s(\omega)$ and $P_d(\omega)$ are the speech and noise psd respectively. However, since the actual psd is not available, we use the psd estimates $\widehat{P}_s(\omega)$ and $\widehat{P}_d(\omega)$ in place of $P_s(\omega)$ and $P_d(\omega)$ respecteively.

The Wiener filter defined above is an estimator that minimizes the Mean Square Error between the actual signal and the estimated signal. The input to the wiener filter is the noisy speech frame. For each frame, the LPC parameters obtained through Levinson-Durbin recursion. These LPC parameters are converted into the approriate parameter domain and the closest match from the corresponding codebook is obtained through minimizing the associated distance measure between the estimated vector and clean speech codebook vector. The codebook vector is then converted back to LP coefficient representation to be denoted as $\mathbf{a_c}$.

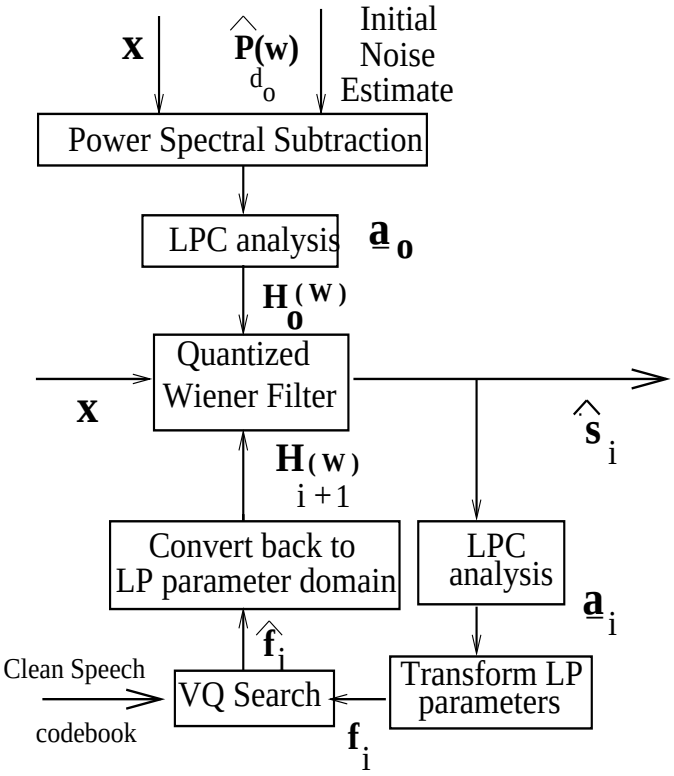

Fig. 1. VQ based Constraints in Iterative Wiener Filtering ; *i : Iteration index*

The speech psd estimate to be used in (1) is obtained by :

$$\widehat{P}_s(\omega) = \frac{G^2}{(|1 - \Sigma_{k=1}^{k=p} \mathbf{a_c}(k) * \epsilon(-j * \omega * k)|)^2} \tag{5}$$

The parameter G is obtained as below:

$$G^2 = R_o - \Sigma_{k=1}^{k=p} \mathbf{a_c}(k) * R(k) \tag{6}$$

where R(k) is the k^{th} autocorrelation lag of the noisy speech.

The parameter $\widehat{P}_d(\omega)$ is obtained by averaging the noise only portions of the speech.

III. SPECTRAL SUBTRACTION BASED INITIALIZATION (SSI)

The sequential MAP estimation implies that for each frame we begin with an assumed set of initial values for vector $\mathbf{a}$ denoted as $\mathbf{a_o}$, based on which the speech vector $\widehat{\mathbf{s}}_\mathbf{1}$ is estimated through MAP. The current estimate $\widehat{\mathbf{s}}_\mathbf{1}$ is in turn used to calculate the next estimate of $\mathbf{a}$. This procedure is continued until convergence is achieved. In the present formulation, H(ω) is started as unity which is highly suboptimum. This gives rise to two possibilities. Firstly, the iterations might converge such that the resulting filter is not perceptually the best. Secondly, even if they do converge to an optimum filter, the number of iterations

will be large. Therefore, an initialization criteria which can direct the course of iterations towards better and quicker convergence is required. We propose a spectral subtraction based initialization (SSI) method to addresses the above issue.

For each frame, power spectral subtraction [7] is performed to obtain the enhanced speech estimate. Following LPC analysis, the above estimate gives $\mathbf{a}_o$ which determines $H_o(\omega)$. Clearly, this $H_o(\omega)$ is better than starting with a unity WF and therefore, leads to better convergence properties of VQ based IWF.

IV. EXPERIMENTS

The speech data comprised ten sentences by 6 male and 4 female speakers for a total of 170 seconds sampled at 8 khz. We reserved 4 sentences of 28 seconds spoken by 2 male and 2 female speakers for testing and the rest for training. Degraded speech with different SNRs was generated by digitally adding noise to clean speech. For codebook generation, a 10^{th} order LPC model was used to extract features by quasi-stationary analysis with 75% overlap between consecutive frames of length 20msec. Clustering was performed using the LBG algorithm for the various parameter spaces with the above mentioned distance measures. Codebooks of size 128 were used since they were found to be adequate in earlier investigations of CCIWF.[4]. For enhancement through IWF, non-overlapping frames of 20 msec duration were used.

The estimation of the all-pole parameters of the clean speech from degraded speech plays a key role in enhancement. The performance, therefore can be evaluated in terms of signal enhancement as well as robust parameter estimation. We used the average segmental SNR [4] and Log Likelihood ratio as the objective measures of enhancement in our experiments.

V. RESULTS

A. Alternate parameter space VQ results

Table I and II summarize the performance of the various parameter sets for 2 different input SNRs. The segmental SNR measures in Table II show that LAR yields the best performance for both 0 dB and 5 dB input SNR. This result is consistent with the higher correlation that LAR based Eucledean distance has with the Diagnostic Acceptability Measure (DAM) in comparison with other LP measures [8]. Moreover, LLR measures shown in Table I are least for LARs and are therefore consistent corresponding highest segmental SNR values in Table II.

The theoritical limit for performance via MAP estimation obtained when original undistorted co-efficients were used in parameter estimation is shown in Table I and II. It can be be seen that the performance of LAR based VQ approaches the theoritical limit. Further, even the 'worst' performing parameter set is found to be superior to the spectral subtractionBol79 technique, both in terms of objective measures and artifacts like musical noise, which, unlike in spectral subtraction, are not found in enhancement through VQ based constraints [7].

TABLE I

LOG LIKELIHOOD MEASURES (LLR) FOR VQ BASED IWF FOR DIFFERENT LP PARAMETER SETS
INPUT SNR : 0 dB AND 5 dB

Parameter Set	LLR *(0 dB SNR)*	LLR *(5 dB SNR)*
Noisy Speech	.5568	.4321
LPC (IS)	.3295	.2781
LSF (IHM)	.3418	.2840
LSF (MFW)	.3271	.2831
LAR (ED)	.3205	.2669
RC (ED)	.3241	.2824
True LPC	.1072	.0886

In terms of convergence, the IHM based WED converges in remarkably least number of iterations as shown in Table III. However, in terms of LLR and segmental SNR measures, it is not as effective as the LAR space. It might therefore be inferred that although converging faster, IHM based WED is unable to find the closest match. This might be attributed to the fact that in presence of noise, the weighting of the degraded LSFs is not effective.

TABLE II

AVERAGE SEGMENTAL SNR VALUES FOR VQ BASED IWF FOR DIFFERENT LP PARAMETER SETS
INPUT SNR : 0 dB AND 5 dB

Parameter Set	Avg. Seg.SNR *(0 dB SNR)*	Avg.Seg.SNR *(5 dB SNR)*
Noisy Speech	3.862	6.979
LPC (IS)	9.614	12.184
LSF (IHM)	8.565	11.283
LSF (MFW)	8.627	11.328
RC (ED)	9.203	11.735
Spectral Subtraction	7.185	9.916
True LPC	11.011	12.994

TABLE III

PERFORMANCE OF VQ BASED CONSTRAINTS IN TERMS OF NUMBER OF ITERATIONS AT 0 dB SNR
TOTAL FRAMES : 326

Iterations	LPC	LAR	RC	LSF (IHM)	LSF (MFW)
2	20	30	60	176	26
3	74	100	110	54	20
4	122	88	84	48	48
5	72	32	38	32	90
$= 6$	38	76	34	16	142

B. SSI performance

The purpose of spectral subtraction based initialization is to direct the course of iterations towards better convergence. Table IV contrasts the performances of SSI and unity filter initialization. Interestingly, on comparing the results at the end of the first iteration to those at convergence, it is observed that SSI nearly obviates the need for iterations.

It was found that over 70 % frames converged to vectors in the codebook that provided a better match than that resulting

from unity initialization. From Fig 2, it is clear that in over 90 % of the cases, SSI does better than unity filter initialization. As expected, in our experiments we found that the relative improvement of SSI over unity initialization increases with increasing noise levels. We also found that the number of iterations decreased by about 15-20% thus indicating that the course of iterations have been positively influenced after SSI.

TABLE IV

COMPARISON BETWEEN SSI AND UNITY FILTER INITIALIZATION
INPUT SNR : 0 dB

Speech Type	Avg.Seg.SNR		Avg. LLR	
	Unity	SSI	Unity	SSI
Degraded speech	3.862	3.862	.5568	.5568
Post-Iteration 1	7.822	9.682	.3371	.3134
Post-Convergence	9.614	9.97	.3295	.3100

Fig 3 shows the segmental SNR histograms of the noisy speech and the enhanced utterances with and without SSI. Firstly, comparing Fig 3(a) and Fig 3(b) and 3(c), one can notice that the significant enhancement has resulted through VQ based IWF. Further, Fig 3(c) is peakier in the sense that more number of frames are now in higher ranges of segmental SNR compared to Fig 3(b). This result is particularly important because as more and more frames have higher segmental SNRs, the quality and intellegibility improves significantly. Thus, SSI clearly aids better and quicker convergence.

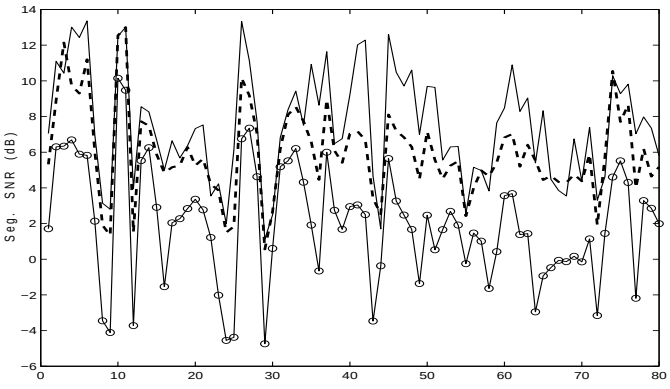

Fig. 2. *Input and Ouput Segmental SNR values for SSI and Unity filter Initialization; - o - Input Seg.SNR , - - Output Seg. SNR for VQ based IWF with Unity filter Initialization , — VQ based IWF after SSI*

VI. Conclusion

This study investigates the effectiveness of the different LP parameter sets with respect to intraframe constraints in VQ based IWF. Better initialization of the iterations is shown to result in better performance and faster convergence. The scope for future work lies in incorporation of interframe constraints into the present framework.

Acknowledgment

The author would like to thank Dr. T.V.Sreenivas, Department of Electrical Communication, Indian Institute of Science (IISc), Bangalore, India. Thanks also to Mr. Sreenivasa Murthy A., Doctoral candidate, IISc, Bangalore.

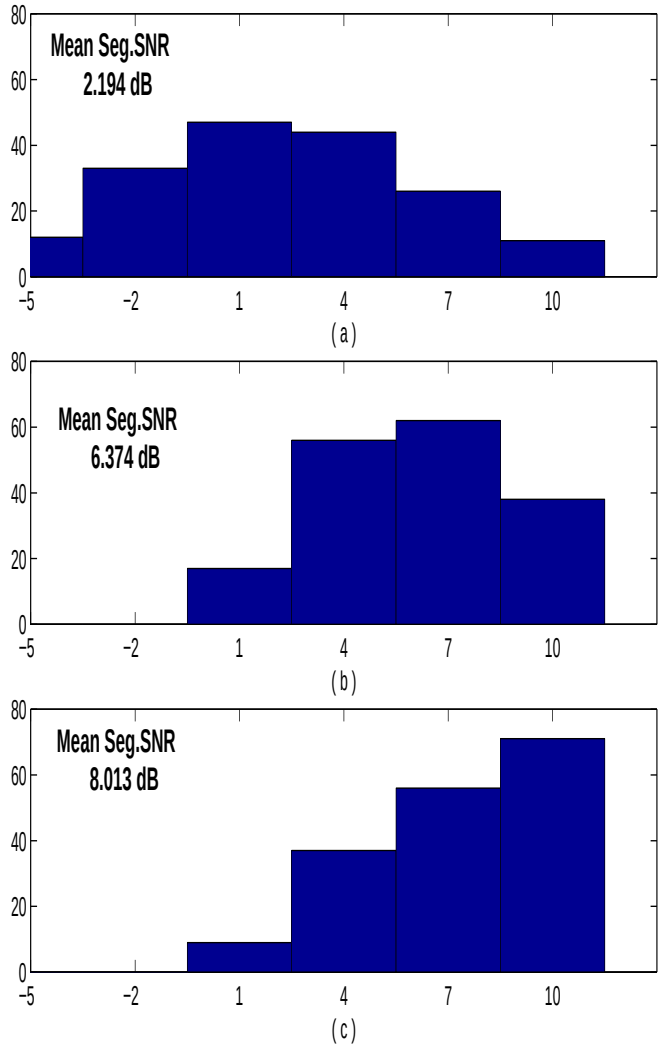

Fig. 3. *Segmental SNR histograms ; (a) Noisy Speech (b) Enhanced with unity filter initialization (c) Enhancement with SSI*

References

[1] J. S. Lim and Alan Oppenheim,"All-pole Modeling of Degraded Speech", *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol ASSP-6, no 3,pp. 197-220,June 1978.

[2] Y. Ephrahim,"Statistical model-based speech enhancement systems," *Proc. IEEE*, vol. 80, pp. 1526-1555,1992

[3] J.H.L Hansen and M.A.Clements,"Constrained Iterative Speech Enhancement with applicaiton to Speech Recognition.",*IEEE Transactions on Acoustics, Speech and Signal Processing,*vol 39, no. 4, pp 795-805, Apr 1991.

[4] T.V. Sreenivas and Pradeep Kirnapure,"Codebook Constrained Wiener Filtering for Speech Enhancement",*IEEE Transactions on Acoustics, Speech and Signal Processing,*vol.4, no.5, pp 383-389, Sep. 1996

[5] K.K.Paliwal and P.V.S.Rao,"Evaluation of various Prediction parametric representations in vowel recognition", *Signal Processing,*Vol 4. , no.4, July 1982

[6] Seung Ho Choi, Hong Kook Kim and HwangSoo Lee, "Speech Recognition using quantized LSF parameters and their transformations in digital communications", *Speech Communication*, pp 223-233, 1999

[7] S. F. Boll,"Supression of acoustic noise in speech using spectral subtraction", *IEEE Transactions on Acoustics, Speech and Signal Processing*, Vol. ASSP-27, pp. 113-120, April 1979.

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

[9] Y.Linde, A.Buzo and R.Gray, "An Algorithm for vector quantizer design", *IEEE Tran.Commun.* vol. COM-28, no. 1, pp 84-94, Jan 1980