Two Stage Iterative Wiener Filtering for Speech Enhancement

Krishna Nand. K., T.V. Sreenivas

Department of Electrical Communication Engineering
Indian Institute of Science, Bangalore, India 560012

krishnaanandk@gmail.com, tvsree@ece.iisc.ernet.in

Abstract
We formulate a two-stage Iterative Wiener filtering (IWF) approach to speech enhancement, bettering the performance of constrained IWF, reported in literature. The codebook constrained IWF (CCIWF) has been shown to be effective in achieving convergence of IWF in the presence of both stationary and non-stationary noise. To this, we include a second stage of unconstrained IWF and show that the speech enhancement performance can be improved in terms of average segmental SNR (SSNR), Itakura-Saito (IS) distance and Linear Prediction Co-efficients (LPC) parameter coincidence. We also explore the tradeoff between the number of CCIWF iterations and the second stage IWF iterations.

Index Terms: Two stage Wiener filtering, Codebook constraint, Voronoi region, LAR parameters, IS distance.

1. Introduction
The technique of Iterative Wiener Filtering (IWF), introduced by Lim and Oppenheim [1], for speech enhancement, is a sequential maximization of the a posteriori probability (MAP) of the speech signal and its all-pole parameters. However, lack of inter-frame and intra-frame constraints has been shown to affect the performance of Iterative Wiener Filtering (IWF) [2]. Intra-frame constraints are applied during the iterations of IWF to ensure that the poles of the estimated speech spectra (formant frequencies) are at reasonable positions with respect to each other and the unit circle. On the other hand, inter-frame constraints are also important to ensure that the vocal tract characteristic of the estimated speech does not vary randomly between successive speech frames.

One of the successful techniques of imposing intra-frame constraints, introduced by Sreenivas and Kirmapure [3], is Codebook Constrained Iterative Wiener Filtering (CCIWF) which uses a vector Quantization (VQ) codebook of Linear Prediction Co-efficients (LPC) derived from clean speech. It yields faster and assured convergence of the IWF process and results in better speech enhancement. Later CCIWF has been extended to incorporate inter-frame constraints using matrix quantization [5].

We recognize that CCIWF is a parametric approach to speech enhancement, and the codebook constraint aids the process of convergence. Speech parameter vectors form a continuum in the parameter space and hence it is important to maintain smooth continuity between successive frames. Towards this, we explore the idea of carrying out unconstrained IWF further to CCIWF enhanced speech. Because of the MAP formulation, we can expect further improvement to the parameter vector through the second stage IWF and thus the enhanced speech to be further close to the clean speech. This approach will also aid in exploiting the inter-frame constraints [5], which will be explored subsequently. In addition to the performance measures of improved average segmental SNR and Itakura-Saito (IS) distance measure, we also evaluate the closeness of enhanced speech with clean speech by comparing the Log area ratio (LAR) vector of enhanced speech with that of clean speech and see if they belong to the same VQ cluster. We show that the two-stage algorithm indeed improves on all the three performance measures.

2. Codebook Constrained Iterative Wiener Filtering
The strong correlations between the formant frequencies of a speech segment are effectively captured by the vector quantization (VQ) codebook of LPC vectors derived from clean speech training data. Exploiting this idea, CCIWF avoids explicit application of human derived knowledge about speech spectra for imposing the intra-frame constraints. The constraints are utilized cleverly through the VQ codebook in an un-supervised manner. However, each frame is independently enhanced and the correlation between consecutive frames is not exploited. The inter-frame constraints are later exploited through matrix quantization [5] or stochastic modeling [6]. In this work, we focus the analysis to VQ codebook constrained IWF.

The general noisy signal model is:

$$y[n] = x[n] + d[n]$$

where $y[n]$, $x[n]$, and $d[n]$ are noisy speech, clean speech and noise respectively. The block diagram of CCIWF is shown in Figure 1.

To enhance each frame of speech, $H^0(\omega)$ is initialized to identity and Wiener filtering is performed on $y[n]$ to get $\hat{x}[n]$. At iteration $t_1$, using the Wiener filter output $\hat{x}^{t_1-1}[n]$, the LPC vector $a_{t_1}^{t_1-1}$ is estimated from the enhanced signal and vector quantized using the clean speech codebook, to form a new wiener filter $H^{t_1} (\omega)$; The updated $H^{t_1} (\omega)$ is used for the next iteration. The iterations are repeated until the LP vector $a$ converges to the same code vector from the codebook, for successive iterations. The performance measures of speech enhancement are (1) Average segmental SNR and (2) Itakura-Saito (IS) distance, or spectral distortion (SD) measure. The IS measure is an important measure of intelligibility. The intelligibility of enhanced speech is considered to be improved through the reduction of IS distance of enhanced speech with respect to the clean speech LPC vector. The LP Co-efficients have several non-linear transformations, such as Log area ratio (LAR), Linear spectral frequency (LSF), Linear prediction based cepstral Co-efficients (LPCC) etc, and it has been shown [4] that log area ratio (LAR) parameters provide the best map between the clean speech and enhanced speech.
speech feature to the noisy counterpart. Hence, LAR domain is best suited for enhancement through CCIWF and we evaluate the performance of the two stage technique also in the LAR space.

3. Parameter Space Representation of CCIWF

The progress of CCIWF iterations in the LAR space is illustrated in Figure 2(a). Let the size of LAR codebook be \( N \). Let \( C_i \) represent the centroid of the \( i^{th} \) cluster (\( 1 \leq i \leq N \)). Let \( X_m \) and \( Y_m \) denote the clean speech and noisy speech LAR vectors corresponding to the \( m^{th} \) frame of \( x[n] \) and \( y[n] \) respectively. Let \( \hat{Y}_m^{t_1} \) denote the LAR vector after \( t_1 \) iteration of CCIWF corresponding to the \( m^{th} \) frame of \( y[n] \). Let \( \hat{Y}_m^{t_2} = c \) denote the CCIWF converged LAR vector corresponding to the \( m^{th} \) frame where \( c \) is the iteration count when convergence is achieved. Typical movement of LAR vector caused by CCIWF is shown by broken line and the path of IWF after CCIWF convergence is shown by the solid line.

To start with, let the clean LAR vector \( X_m \) reside in cluster (say) 4. Addition of noise has caused the movement of \( X_m \) into the cluster (say) 1 resulting in \( Y_m \). Upon vector quantization, \( Y_m \) moves to the centroid \( C_1 \), which is the closest to it among all the code vectors of the codebook. Wiener filtering is now applied on the vector \( C_1 \) (which is the starting point now). This results in vector \( \hat{Y}_m^{t_1} \) in cluster (say) 2. This completes the first iteration of CCIWF. Vector quantization is now again applied on \( \hat{Y}_m^{t_1} \), which moves it to \( C_2 \) (which is the centroid of cluster 2). Wiener filtering is now applied on \( C_2 \) because of which it moves to cluster (say) 3, resulting in \( \hat{Y}_m^{t_2} \) and also completing the second iteration of CCIWF. This goes on until the converged enhanced LAR vector \( \hat{Y}_m^{t_2} = c \) is reached. Both Codebook constraint and Wiener Filtering jointly influence the movement of enhanced speech LAR vector towards the clean speech LAR vector; Wiener filtering will reduce the Mean Square Error and codebook constraint brings in intra-frame parameter structure.

4. Unconstrained IWF post CCIWF

IWF is known to fail in the absence of any constraints. Here, we show that IWF after CCIWF enhancement, is successful in moving the enhanced speech much closer to the clean speech.

During every iteration of CCIWF, there exists two possibilities:

- a. Wiener filter is successful in moving the parameter vector out of the cluster, in which case, CCIWF algorithm proceeds into the next iteration in the new cluster.

- b. Wiener filter is unsuccessful in moving the parameter vector out of the cluster. In this case, during successive iterations, the parameter vector is constrained to the same cluster and convergence is said to be achieved.

CCIWF performance would be best when the convergence is in the same cluster as that of the associated clean speech parameter vector. We note that in CCIWF, the convergence achieved is to the centroid of the cluster and not to the clean speech parameter vector itself (see Figure 2(a)); i.e., \( \hat{Y}_m^{t_1} = c \) is the converged centroid of CCIWF enhanced speech and \( X_m \) is the clean speech LAR vector. Through CCIWF iterations, it may also happen that convergence is reached to a neighboring cluster and not the cluster in which the associated clean speech parameter vector is present. This is illustrated in Figure 2(b), where \( \hat{Y}_m^{t_2} = c \) is the converged centroid of CCIWF enhanced speech in cluster \( C_4 \), whereas the associated clean speech parameter vector \( X_m \) is present in the neighboring cluster \( C_5 \). In this case, the CCIWF performance is sub-optimum.

To take the enhanced speech further close to the clean speech, we propose to carry out unconstrained IWF post CCIWF convergence. This will exploit the fact that CCIWF has already lessened the gap between clean speech and noisy speech and further Wiener filtering can make the enhanced speech better by taking it further close to the clean speech. In Figure 2(a), the idea is to move the enhanced speech parameter vector from \( \hat{Y}_m^{t_1} = c \) to \( X_m \), inside the same cluster. In Figure 2(b), the idea is to move the enhanced speech parameter vector from \( \hat{Y}_m^{t_2} = c \) in cluster \( C_4 \), to \( X_m \) in cluster \( C_5 \), across the cluster boundary. The second-stage IWF is shown in Figure 1, as an extension of CCIWF. We carry out IWF on the CCIWF enhanced speech in such a way that after each iteration of IWF, the predicted signal \( \hat{x}[n] \), is taken as the input for constructing the second stage wiener filter \( \hat{H}_2^{t_2}(\omega) \), for the next iteration \( t_2 \) of IWF post CCIWF convergence.

5. Experiments and results

The speech database used for the experiments comprised of 60 sentences by 30 male and 30 female speakers, for a total of around 2400 secs of speech, sampled at 8 kHz. We reserved
Figure 2: Two examples of feature space traversing through CCIWF+IWF iterations. (a) CCIWF convergence is shown in the same cluster as that of the clean speech parameter vector; (b) IWF iterations leading to neighboring cluster where the clean speech parameter vector is present.

10 sentences totaling around 600 secs, spoken by 5 male and 5 female speakers for testing and the rest for training. Degraded speech is generated with 5, 0 and -5 dB SNR by digitally adding various types of noises to clean speech. Noise files were taken from NOISEX-92 database. Clean speech codebooks are generated using a 10th order LPC model to extract features by quasi-stationary analysis with 75% overlap between consecutive frames of length 20 msec. Clustering is performed using the LBG algorithm using Euclidean distance measure in the LAR domain. Codebooks of size 64-1024 are generated and 128 size codebook is used in most of the investigations [3]. For Wiener filtering in both stages, non-overlapping frames of 20 msec is used.

The average segmental SNR of enhanced speech resulting from IWF post CCIWF, have been tabulated in Table 1 for varying number of iterations of IWF, at 5, 0 and -5 dB input SNR, for white Gaussian noise. It can be seen that at all input SNRs, CCIWF provides 4-5 dB avg.segmental SNR improvement. This is also consistent with improved intelligibility, in terms of avg.IS distance, reducing by more than 50%. After CCIWF convergence, IWF iterations are performed without the VQ constraint. It is seen that initially IWF improves both the IS distance as well as segmental SNR, but starts to diverge later. The noise performance is seen to be best optimized at 2 iterations of unconstrained IWF post CCIWF, as shown highlighted. The higher average segmental SNR and lower IS distance show that the technique has been successful in achieving better enhancement in addition to retaining speech intelligibility. Further iterations of IWF are seen to give sub-optimal results, because of the fact that unconstrained IWF is taking the enhanced speech away from the clean speech when the number of iterations are more by taking the LPC vector out of the cluster in which CCIWF convergence was achieved.

However, we can see from the results of Table 2 that the percentage of speech frames, whose centroids match that of the associated clean speech centroid, has improved by about 10% over the CCIWF only performance. We see that after 2 iterations of IWF post CCIWF, the enhanced speech is closest to the clean speech with maximum number of centroids of enhanced speech frames matching that of centroids of associated clean speech frames. This is also consistent with the segmental SNR and IS distance results, further strengthening the advantages of the second stage IWF.

Table 2: Percentage of speech frames for which the centroids of enhanced speech and that of associated clean speech match, at 5 and 0 dB input SNR, for increasing iterations of IWF, post CCIWF convergence. (CCIWF codebook size = 128)

<table>
<thead>
<tr>
<th>Speech type</th>
<th>Percentage at 5 dB</th>
<th>Percentage at 0 dB</th>
</tr>
</thead>
<tbody>
<tr>
<td>CCIWF</td>
<td>20.056%</td>
<td>14.192%</td>
</tr>
<tr>
<td>t_2 = 1 Iteration</td>
<td>27.64%</td>
<td>20.42%</td>
</tr>
<tr>
<td>t_2 = 2 Iterations</td>
<td>30.35%</td>
<td>21.42%</td>
</tr>
<tr>
<td>t_2 = 3 Iterations</td>
<td>29.49%</td>
<td>21.39%</td>
</tr>
<tr>
<td>t_2 = 4 Iterations</td>
<td>27.91%</td>
<td>20.93%</td>
</tr>
<tr>
<td>t_2 = 5 Iterations</td>
<td>26.47%</td>
<td>20.15%</td>
</tr>
</tbody>
</table>

Table 3 compares the performance of the two-stage technique with single stage CCIWF of increasing codebook sizes. Two stage is seen to perform better in terms of higher segmental SNR, even when compared to CCIWF with codebook size of 1024. This again demonstrates the fact that, in CCIWF technique, codebook constraint has more to do with aiding convergence along with retaining perceptual quality of speech, whereas IWF improves noise performance, contributing more towards achieving increased segmental SNR. IWF post CCIWF is performed with codebook of size 64 and was also observed to be best optimized at t_2=2 iterations.

From Table 4, we see that the two-stage technique performs well even under different noise conditions.
iterations for convergence when compared to 64 and 128 sized codebooks, which is expected. The tradeoff involved between the codebook size, \( t_1 \) and \( t_2 \), in the two-stage technique, is in choosing the right sized codebook, which requires fewer iterations for CCIWF convergence but yet achieves better performance than CCIWF only with larger codebook size, through \( t_2 \) iterations of IWF post CCIWF. This is illustrated in Figure 4, where \( t_{1,64} + t_2 \) has resulted in higher Avg.seg SNR even when compared to \( t_{1,128} \), thus reducing the computation significantly.

### 6. Conclusions

We have explored unconstrained IWF further to CCIWF enhanced speech. We found that this can take the CCIWF enhanced speech further close to the clean speech, resulting in improved segmental SNR which is indicative of improved speech quality and improved IS distance which is indicative of speech intelligibility. The performance is shown to be best optimized with 2 iterations of IWF post CCIWF. We have also demonstrated this through the LPC parameter coincidence measure. Finally we have shown that the technique performs better than CCIWF only, even with increased codebook sizes, and also the tradeoff involved between codebook size, number of CCIWF iterations and second stage IWF iterations, in achieving improved performance.

### 7. References


