IMPROVED ITERATIVE WIENER FILTERING FOR NON-STATIONARY NOISE SPEECH ENHANCEMENT

Sharath Rao K., T.V. Sreenivas and A. Sreenivasa Murthy

Speech and Audio Laboratory
Department of Electrical Communication Engineering
Indian Institute of Science, Bangalore-560012, India
email: tvsree@ece.iisc.ernet.in

ABSTRACT

A clean speech VQ codebook has been shown to be effective in imposing intraframe constraints in Iterative Wiener Filtering (CCIWF) for speech enhancement. However, for time-varying noises, the performance is sub-optimum. We propose a smoothed noise update technique that uses the estimated signal spectrum for subsequent signal estimation. This leads to a more effective solution than the soft-decision based noise estimate found in literature. Further, the CCIWF performance is improved using codebook constraints in the LAR domain instead of LPC domain. Also, a new iteration initialization method is proposed which results in better enhancement in over 70% of the frames. Thus, we show that a combination of a robust parameter space, an effective initialization and continuous spectrum update significantly improves the performance of speech enhancement. Speech recognition results show that the new combination provides 10-20% increase in word recognition scores whereas simple spectral subtraction results in an actual decrease in recognition score.

1. INTRODUCTION

In [1], Lim and Oppenheim proposed the iterative Wiener filtering (IWF) technique for speech enhancement where the estimation of the all-pole parameters of speech in additive white gaussian noise was posed as a two step sequential MAP estimation problem. In [4], Hansen and Clements showed that constraints in the parameter estimation are essential in order to retain speech-like characteristics of enhanced speech. In [5], a clustering based approach namely the codebook constrainted iterative Wiener filtering scheme (referred henceforth as CCIWF) was proposed as an alternative method of imposing constraints. Here, the all-pole parameters are 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 types of speech constraints such as those between the formants and those due to speaker variability.

In all the above approaches only non-stationary noise is considered. However, in many practical applications the noise is time-varying and hence leads to sub-optimum results. Several techniques are found in literature [7]-[8] that address this problem. Most of them concentrate on avoiding explicit speech/non-speech classification and resort to measures of recursively estimating the noise psd. In [7]-[8], it is claimed that MCRA(minima controlled recursive averaging) and *aposteriori* SNR based recursive estimation are effective. Instead, we propose that an estimate of the noise spectrum could be obtained by using the estimated signal spectrum and signal subtraction (similar to noise subtraction). The noise spectral estimate can be appropriately smoothed either temporally or through model-fitting. This leads to an adaptive noise estimator with least assumptions about the signal and noise characteristics. In particular, since we are using an optimum signal estimator through CCIWF, the noise psd is sufficiently accurate for time-varying noise speech enhancement.

This paper (i) explores the adaptation of the CCIWF technique for non-stationary disturbances. (ii) proposes the spectral subtraction initialization (SSI) method which improves the enhancement performance with respect to parameter estimation and convergence. (iii) employs different LP parameters within CCIWF and identifies the best suited and most robust parameter domain through objective measures of enhancement and speech recognition tests.

2. NOISE UPDATE ALGORITHM (NUA)

Fig 1. shows the adaptive CCIWF scheme. We model the noisy signal as **x** = **s** + **d**, where **x**, **s** and **d** are noisy signal, speech and noise respectively. In the IWF technique, the speech signal **s** is modeled as a response of an all-pole system and the approach adopted is to solve for the MAP estimate of the signal, given **x**. In situations where background noise psd $P_d(\omega)$ is time-varying, the conventional method is to update the noise psd estimate in non-speech regions. This method has two major limitations : firstly, a speech/non-speech classification is required which in itself is challenging in noisy conditions ; secondly, this method is based on the assumption that sufficient non-speech duration is available to update the noise estimate which may not be the case. Moreover, the noise itself could be changing within a non-speech region. Thus, a poor estimate of $P_d(\omega)$ limits the performance of the wiener filter.

We have devised a simple and straightforward adaptive technique that tracks the changing noise characteristics. Since we are able to estimate the signal spectrum iteratively and the Wiener filter is optimum in estimating the signal, the noise spectrum in each frame can be estimated using signal subtraction (under the assumption that the speech and noise processes uncorrelated). This provides the means of estimating the time-varying spectrum. However, we can assume that noise is less time-varying than speech and hence, for each frame, the noise estimate is obtained by averaging the noise power spectrum of the last L frames as shown below. For

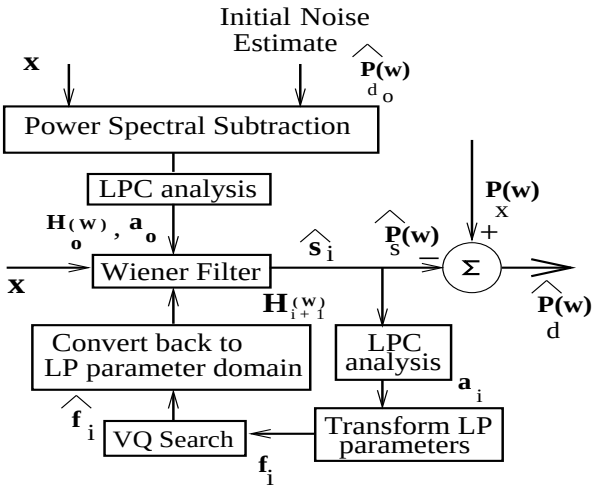

Fig. 1. *CCIWF adapted for non-stationary disturbances i - iteration index*

each frame m,

$$\widehat{P}_d(m;\omega) = \frac{1}{L}(\Sigma_{j=m-1}^{j=m-L}(F(j;\omega).W(m))) \quad (1)$$

where,

$$F(j;\omega) = P_x(j;\omega) - \widehat{P}_s(j;\omega); if P_x(j;\omega) > \widehat{P}_s(j;\omega) \quad (2)$$

$$F(j;\omega) = P_x(j;\omega); otherwise \quad (3)$$

Parameter ω is the frequency index, $\widehat{P}_d(\omega)$, $P_x(\omega)$, $\widehat{P}_s(\omega)$ are the noise psd estimate, noisy signal psd and the speech psd estimates of CCIWF respectively; and W(m) denotes the weighting function. To start the successive estimation, the noise psd estimate is obtained from an assumed initial non-speech duration of 0.2 seconds. The speech psd estimate is obtained with every iteration of CCIWF (Fig 1). The *smoothing parameter* L depends on the degree of non-stationarity of the noise. Ideally, the smaller the value of L, the better is the algorithm able to track rapidly varying noise. In addition, the weighting function W(m) chosen as a tapering window takes into account the higher correlation of the nearby frames rather than farther frames. Although, the algorithm makes no assumption regarding the type of noise, it is found to give robust performance for a variety of real world noises.

3. 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 **a** denoted as $\mathbf{a_o}$, based on which the speech vector $\widehat{\mathbf{s}}_\mathbf{1}$ is estimated through the Wiener filter. The current estimate $\widehat{\mathbf{s}}_\mathbf{1}$ is in turn used to calculate the next estimate of **a**. This procedure is continued until convergence is achieved. In [5], 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 method 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 issues. For each frame, power spectral subtraction [2] 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 CCIWF.

4. ROBUST PARAMETER DOMAIN SEARCH

The effectiveness of CCIWF consists in approximating the optimum filter through a codebook of clean speech vectors. Therefore, the parameter space used to represent these vectors has a significant bearing on the successive 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, each has been used with varied success in speech coding and recognition. In this study, we explore the different parameter spaces for CCIWF and identify the best performing parameter. The widely used IS distance measure is used for creating LPC codebooks. The Eucledean Distance (ED) is used for LAR and RC codebooks. For LSPs, we use the Eucledean Distance (ED) and also two other perception based weighted Eucledean distances - the Mel-Frequency Warping (MFW) based distance which is modeled on the auditory system and the Inverse Harmonic Mean (IHM) based distance. The IHM based distance is perceptually relevant since it weighs each LSF in the inverse proportion of its closeness to its neighbours due to the better chance of it representing formants[6].

5. EXPERIMENTAL EVALUATION CRITERIA

The speech data comprised of ten sentences by 6 male and 4 female speakers for a total of 170 sec. of speech sampled at 8 kHz. We reserved 4 sentences of 28 sec. 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 20 msec. 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[5]. In the Wiener filtering stage, non-overlapping frames of 20 msec duration are used.

The estimation of the all-pole parameters of the clean speech from degraded speech plays a key role in enhancement through IIWF. The performance, therefore can be evaluated in terms of both signal enhancement as well as robust parameter estimation. We used the average segmental SNR [5] and Log Likelihood ratio as the objective measures of CCIWF performance in our experiments. To quantify intelligibility, we used a HMM-based multispeaker Isolated Word Recognition (IWR) system on enhanced speech. The system used a MFCC front-end and a 10-word vocabulary. In all, 210 utterances of each word, uttered thrice by each of the 70 speakers used. Of these, 140 utterances were used for training HMMs on clean speech samples and remaining 70 for testing. During testing, clean speech samples were degraded with 10 dB and 5 dB SNR and recognition tests were performed after enhancement via different techniques.

6. RESULTS AND DISCUSSION

6.1. *NUA performance*

The adaptive behaviour of the Noise Update Algorithm (NUA) is shown in Fig 2(a) for fire engine noise. NUA is consistently able to track even variations of over 10 dB/sec. The performance of non-adaptive CCIWF and the adaptive CCIWF using NUA in terms of MSE betweeen the enhanced and the clean speech is shown in Fig 2(b). It is found that in all frames, the NUA performs better than the non-adaptive CCIWF with the MSE being upto 7 dB lower in some frames. As the noise deviates from the initial estimate, the performance of the CCIWF degrades, whereas the NUA adapts to the changing noise characteristics. Although NUA is not able to adapt fast enough in regions where noise is highly non-stationary due to the averaging effect, it still performs better than the non-adaptive CCIWF in all frames. We experimented with different smoothing parameters, weighting factors and non-stationary disturbances like waterflow, crowd babble, and door creak noise. It is found that L in the range of 4-7 and a uniform weighting across L past frames is adequate for a wide range of noise types. The upper-bound of L is dependent on the degree of non-stationarity of noise and is not a function of signal presence/SNR. However, if *apriori* knowledge of noise characteristics is available, further tuning of these parameters can result in a better performance.

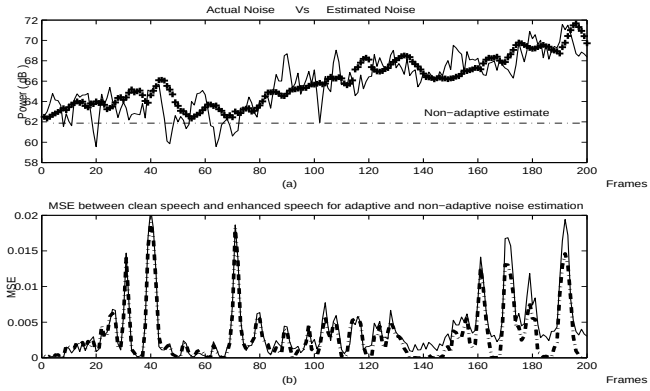

Fig. 2. *Fire Engine Noise (a) Noise tracking by NUA : — Actual Noise Power , + Estimated Noise through NUA (b) Enhancement performance : — Mean Square Error (MSE) between clean speech and Enhancement without adaptive noise estimate , - - MSE between clean speech and Enhancement with adaptive noise estimation through NUA showing consistent advantage.*

6.2. *SSI performance*

The purpose of spectral subtraction based initialization is to direct the course of iterations towards better convergence. Table 1 contrasts the results of SSI and $H_o = 1$ 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 is found that over 70 % frames converged to vectors in the codebook that provided a better match than that resulting from unity initialization. From Fig 3, 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. With the number of iterations decreasing by about 15-20% and both objective measures showing improved parameter estimation, SSI clearly aids better and quicker convergence.

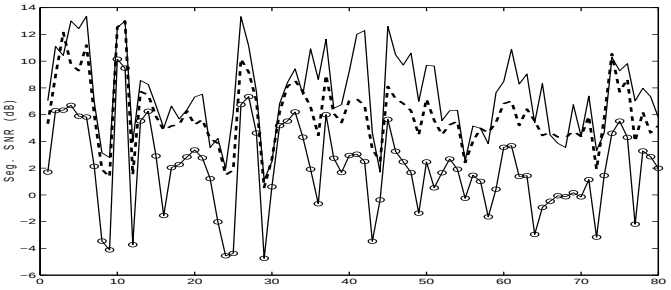

Fig. 3. *Input and Ouput Segmental SNR values for SSI and Unity filter Initialization; - o - Input Seg.SNR , - - Output Seg. SNR for CCIWF with Unity filter Initialization , — CCIWF after SSI*

Table 1.
Comparison between SSI and unity filter initialization for 0 dB input SNR

Speech Type	Avg.Seg.SNR		Avg. LLR	
	Unity	SSI	Unity	SSI
Degraded speech	3.862	3.862	0.557	0.557
Post-Iteration 1	7.822	9.682	0.337	0.313
Post-Convergence	9.614	9.977	0.329	0.310

6.3. *Optimum parameter space results*

Tables 2 and 3 summarize the performance of the various parameter sets for 2 different input SNRs. The average segmental SNR measures in Table 2 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 [3]. Moreover, LLR measures shown in Table 3 are least for LARs and are therefore consistent with corresponding highest segmental SNR values in Table 2. The theoritical limit for performance via MAP estimation obtained when original undistorted co-efficients are used in the wiener filter is shown in both Tables 2 and 3. It can be be seen that the performance of LAR based CCIWF approaches the theoritical limit. Further, even the 'worst' performing parameter set is found to be superior to the spectral subtraction [2] technique, both in terms of objective measures and artifacts like musical noise, which, unlike in spectral subtraction, are not found in enhancement through CCIWF. Thus, while spectral subtraction alone is not effective, in combination with CCIWF, results in an effective solution.

6.4. *IWR performance*

Since enhancement techniques are used as a preprocessor in several robust speech recognition systems, enhancement algorithms

Table 2.
Average Segmental SNR measure for CCIWF with different LP parameter sets for 0 dB and 5 dB input SNR

Parameter Set (Distance measure)	Avg. Seg.SNR *(0 dB SNR)*	Avg.Seg.SNR *(5 dB SNR)*
Noisy Speech	3.862	6.979
LAR (ED)	**9.723**	**12.194**
RC (ED)	9.203	11.735
LPC (IS)	9.614	12.184
LSF (IHM)	8.565	11.283
LSF (MFW)	8.627	11.328
LSF (ED)	8.435	11.060
Spectral Subtraction	7.185	9.916
True LPC	11.011	12.994

Table 3.
Avg. Log likelihood measures (LLR) for CCIWF with different LP parameter sets for 0 dB and 5 dB input SNR

Parameter Set (Distance measure)	LLR *(0 dB SNR)*	LLR *(5 dB SNR)*
Noisy Speech	.5568	.4321
LAR (ED)	**.320**	**.266**
RC (ED)	.324	.282
LPC (IS)	.330	.279
LSF (IHM)	.341	.284
LSF (MFW)	.327	.283
LSF (ED)	.342	.308
True LPC	.107	.089

can be evaluated by studying the recognition performance of the enhanced speech. Table 4 presents results for IWR after enhancement via different techniques and in parameter domains. Firstly, as is expected, it can be noted that with increasing noise, the performance falls from 89% for clean speech to as low as 46.28% at 5 dB SNR. On comparing CCIWF with spectral subtration, it is clear that CCIWF scores better in each and every parameter domain. Within CCIWF, the RCs and LARs perform better than all other parameters at 10dB with RC performing marginally better than LAR at 5 dB. Overall, the result is in agreement with SNR and LLR measures in Tables 3 and 4 which further reinforces the fact that the Log Area Ratio provides maximum robustness with respect to noise. Since the CCIWF technique iteratively searches a match for noisy vector in the clean speech codebook, it can be stated that the LAR based Eucledean distance performs best because it provides the best mapping between a given vector and its noisy counterpart. Interestingly, spectral subtraction performs worse than degraded speech, something that might be attributed to the artifacts like musical noise that the former introduces.

7. CONCLUSION

This study enhances the performance of the CCIWF technique for non-stationary disturbances. A new noise estimation algorithm is proposed that uses the optimum signal estimate of the CCIWF to calculate subsequent noise psd through signal subtraction. We also explored different parameter domains and found that the LARs

Table 4.
Recognition Scores post-CCIWF for different LP parameter sets for 10 dB and 5 dB input SNR

Parameter Set	Recog. Rate *(10 dB SNR)*	Recog. Rate *(5 dB SNR)*
Noisy Speech	62.7	46.2
LAR (ED)	75.4	65.4
RC (ED)	75.4	66.8
LPC (IS)	74.5	62.7
LSF (IHM)	70.2	53.2
LSF (MFW)	68.5	54.8
LSF (ED)	69.2	53.8
Spec. Sub.	56.8	43.2

best map the clean speech feature to its noisy counterpart and therefore are best suited for enhancement through CCIWF. Further, we also showed improvement in enhancement and convergence through a new initialization criterion. The scope for future work lies in incorporation of interframe constraints into the CCIWF framework which is currently under study.

We thank Tata Institute of Fundamental Research (TIFR), Mumbai for the speech database. We thank Siemens Communication Software Ltd. for providing the noise database.

8. 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] 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.

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

[4] 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.

[5] 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

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

[7] I. Cohen , B. Berdugo ; "Speech Enhancement for non-stationary noise environemnts,", *Signal Process.* , Vol 81,pp 2403-2418, Nov 2001.

[8] Lin, L., Holmes, W.H., Ambikairajah, E.; "Adaptive noise estimation algorithm for speech enhancement", *Electronics Letters* , Vol 39 ,no. 9, pp 754-755, May 2003.