Multi-channel Iterative Dereverberation based on Codebook Constrained Iterative Multi-channel Wiener Filter

Ajay S.[1], and *T. V. Sreenivas*[2]

Department of Electrical Communication Engineering,
Indian Institute of Science, Bangalore-560012. INDIA
[1]ajays@ece.iisc.ernet.in, [2]tvsree@ece.iisc.ernet.in

Abstract

A novel Multi-channel Iterative Dereverberation (MID) algorithm based on Codebook Constrained Iterative Multi-channel Wiener Filter (CCIMWF) is proposed. We extend the classical iterative wiener filter (IWF) to the multi-channel dereverberation case. The late reverberations are estimated using Long-term Multi-step Linear Prediction (LTMLP). This estimate is used in CCIMWF framework through a doubly iterative formulation. A clean speech VQ codebook is effective for inducing intra-frame constraints and improve the convergence of IWF, thus, a joint-CCIMWF algorithm is proposed for the multi-channel case. The signal to reverberation ratio (SRR) and log spectral distortion (LSD) measures improve through the double-iterations, showing that the algorithm suppresses the effect of late reverberations and improves speech quality and intelligibility. The algorithm also has fair convergence properties through the iterations.

Index Terms: Dereverberation, Multi-channel wiener filter, Long term prediction

1. Introduction

Hands free speech recorded with distant microphones is affected by the acoustic characteristics of the room, mainly reflections. This degradation of speech due to room reverberation adversely affects speech intelligibility, as well as automatic speech recognition (ASR) performance. Hence it is desired to enhance speech quality by reducing the effect of reverberation. Though reverberant speech enhancement has been addressed widely in the literature, it is still a challenging problem because of the difficulty in characterizing reverberation. Several methods have been proposed for dereverberation of speech [1] and multi-channel approaches provide an advantage, since theoretically we can obtain perfect dereverberation under certain assumptions on the room impulse response [2].

Room reverberation is modeled effectively using the Room Impulse Response (RIR) which consists of a direct path, early reflections and late reflections [3]. The early reflections correspond to the reflections which arrive within ∼30ms of the direct component. The late reflections are the main cause for the smearing of phonemes which affects human intelligibility as well as ASR performance. The early reflections can be handled during the feature extraction of ASR systems by cepstral mean subtraction.

In this paper, we develop a doubly iterative dereverberation algorithm to suppress the effect of late reverberations. The algorithm estimates the effect of late reverberation using the approach presented in [4]. We extend the Codebook Constrained Iterative Wiener filter [5] to the multi-channel case and use it to suppress the late reverberations.

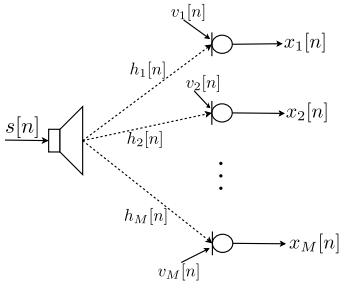

Figure 1: Multi-channel Reverberant Speech model

2. Multi-channel Reverberation Model

Let the speech signal $s[n]$ be recorded at M distant microphones, with the reverberant speech signal at the microphones being $x_i[n]$, $i = 1, 2, \cdots, M$ as shown in figure 1. Using the RIR $h_i[n]$, of length L ($\simeq$ 0.5s), from the source to the i^{th} microphone, we can write,

$$x_i[n] = \sum_{k=0}^{L-1} h_i[k]\, s[n-k] \;+\; v_i[n] \tag{1}$$

$v_i[n]$ indicates the additive noise at the microphone; for the present, we consider $v_i[n] = 0$. The RIR for each channel is assumed to be time invariant. From the statistical room acoustics theory [3, 6], we can assume that the direct signal component at the receiver and the reverberant components are uncorrelated. Further, if the early reflections in the room impulse response have sufficient energy, even they can be assumed to be uncorrelated with the late reverberant components (see appendix of [4]). The RIR from the source to the i^{th} mic, $\mathbf{h}_i$, is decomposed into the early reflections $\mathbf{h}_{i,e}$ and late reflections $\mathbf{h}_{i,l}$, with the parameter D delineating the boundary. D is typically chosen in the range of 20-30 ms.

$$\mathbf{h}_i = [\underbrace{h_i(0)\, h_i(1) \;\ldots\; h_i(D)}_{\mathbf{h}_{i,e}}\; \underbrace{h_i(D+1) \;\ldots\; h_i(L-1)}_{\mathbf{h}_{i,l}}] \tag{2}$$

Thus, we can simplify Eq.(1) by separating the early and late reverberant speech as,

$$\begin{aligned} x_i[n] &= \mathbf{h}_{i,e} * s[n] + \mathbf{h}_{i,l} * s[n] \\ x_i[n] &\triangleq y_i[n] + \eta_i[n] \end{aligned} \tag{3}$$

We consider $y_i[n]$ as the target speech signal which includes the direct path and early reverberant speech component, and $\eta_i[n]$ is the late reverberations, assumed to be uncorrelated with $y_i[n]$. The above is an additive model of late reverberations, which permits optimum statistical estimation.

3. Multi-channel iterative dereverberation

Using the additive model of reverberation, we develop a doubly iterative dereverberation algorithm using the LTMLP [4] to estimate $\eta_i[n]$. In [4], they use spectral subtraction to suppress late reverberations using this estimate. However, spectral subtraction has the difficulties of musical noise, over subtraction and flooring. Instead, we explore the IWF framework to achieve better dereverberation. We extend the IWF to the multi-channel case (IMWF) and to address the convergence problem, we use the codebook constrained approach, proposed by us earlier [5]. The late reverberations estimate is obtained using multi-channel long term prediction. In this, we expect different channels to contribute to the long-term predictability, which is expected to improve after enhancing the speech through wiener filtering. Therefore, we propose an outer iteration between multi-channel wiener filtered speech and late reverberation estimation; this is referred to as Multi-channel Iterative Dereverberation (MID) iteration, shown in figure 2.

3.1. Long Term Multi-step Linear Prediction (LTMLP)

LTMLP uses high order long term prediction over a long speech utterance from several channels of recording to obtain an estimate of late reverberation signal. The long term prediction estimation here is interpreted as a MISO system for blind channel identification as in [2]. For multi-channel data, LTMLP of order P and step-size D can be formulated as,

$$x_i[n] \;=\; \sum_{m=1}^{M}\sum_{p=0}^{P-1} w_{m,i}(p)\, x_m(n-p-D) \;+\; e_i[n] \quad (4)$$

for each of the channels $i = 1, 2, \ldots, M$ and $w_{m,i}(p)$ are the P^{th} order prediction coefficients for each channel-i, using the speech from m^{th} channel (MISO). The prediction coefficients $w_{m,i}(p)$ are estimated by minimizing the mean squared energy in the residual error $e_i(n)$. If we define,

$$\mathbf{w}_{m,i} \triangleq [w_{m,i}(0),\, w_{m,i}(1),\, \cdots,\, w_{m,i}(P-1)]^T$$
$$\mathbf{w}_i = \left[\mathbf{w}_{1,i}^T,\, \mathbf{w}_{2,i}^T,\, \cdots,\, \mathbf{w}_{M,i}^T\right]^T$$
$$\mathbf{x}_i[n] = [x_i[n],\, x_i[n-1],\, \cdots,\, x_i[n-P+1]\,]^T \text{ and}$$
$$\mathbf{x}[n] = \left[\mathbf{x}_1^T[n],\, \mathbf{x}_2^T[n],\, \cdots,\, \mathbf{x}_M^T[n]\right]^T$$

The MMSE solution for $\mathbf{w}_i$ in Eq.(4) is got then by solving the linear system,

$$\left(\mathbf{E}\left\{\mathbf{x}[n-D]\mathbf{x}^T[n-D]\right\}\right)\mathbf{w}_i = \mathbf{E}\left\{\mathbf{x}[n-D]x_i[n]\right\} \quad (5)$$

$\mathbf{E}(.)$ is the time average operator, computed over a long reverberant speech segment ($\sim$1s). The $\mathbf{w}_i$ are computed only once for a speech utterance of several seconds long, since the RIR is time invariant. For the coefficients calculated as above, we consider,

$$\hat{\eta}_i[n] \triangleq \sum_{m=1}^{M}\sum_{p=0}^{P-1} w_{m,i}(p)\, x_m[n-p] \quad (6)$$

$\hat{\eta}_i[n]$ in Eq.(6) is an estimate of late reverberations for each channel i, assuming the source $s[n]$ is white. However, if the source is not white (which is the case with speech), we use a whitening filter of a short order for each of the channels to reduce the short term autocorrelation as suggested in [4]. We could also increase the delay D, which reduces the correlation between early and late reverberation components.

3.2. Multi-channel Wiener Filter

The codebook constrained iterative wiener filter uses a clean speech VQ codebook to impose intra-frame constraints for the Iterative Wiener Filter (IWF) [7] leading to better convergence and improved intelligibility. Here, we extend the single channel formulation of IWF to the multi-channel case and incorporate codebook constraints. In the frequency domain, we form the vector signal, with multi-channel components stacked as,

$$\mathbf{X}(\omega) = \mathbf{Y}(\omega) \;+\; \mathbf{N}(\omega) \quad (7)$$

where,

$$\mathbf{X}(\omega) = [X_1(\omega)\; X_2(\omega)\; \ldots X_M(\omega)]^T$$
$$\mathbf{Y}(\omega) = [Y_1(\omega)\; Y_2(\omega)\; \ldots Y_M(\omega)]^T$$
$$\mathbf{N}(\omega) = [N_1(\omega)\; N_2(\omega)\; \ldots N_M(\omega)]^T$$

where $(.)^T$ is the transpose operator and $N_i(\omega)$ is the late reverberation component of $X_i(\omega)$. The MWF is a MIMO formulation in which all channels are enhanced using all the available channels. Let $\mathbf{G}_i(\omega) = [G_{i,1}(\omega)\; G_{i,2}(\omega)\; \ldots G_{i,M}(\omega)]^T$; $i = 1, 2, \ldots, M$ and let $(.)^H$ be the conjugate transpose operator and $\mathbf{G}$ be an $M^2 \times 1$ stacked vector defined as,

$$\mathbf{G}^H(\omega) = [\mathbf{G}_1^H(\omega)\; \mathbf{G}_2^H(\omega)\; \ldots \mathbf{G}_M^H(\omega)]$$

In the frequency domain, wiener filtering is performed at each frequency independently; hence we can drop the variable ω for convenience. For each frame of the signal, at each ω, for an optimum MWF, we consider the following modified cost function [5], which is the weighted sum of the residual noise energy and the speech distortion energy:

$$J(\mathbf{G}) = \mathcal{E}\left\{\left\|\mathbf{Y}^T - \left[\mathbf{G}_1^H\mathbf{Y}\;\; \mathbf{G}_2^H\mathbf{Y} \ldots \mathbf{G}_M^H\mathbf{Y}\right]\right\|_2^{\mathbf{2}}\right\}$$
$$+\, \mu\, \mathcal{E}\left\{\left\|\left[\mathbf{G}_1^H\mathbf{N}\;\; \mathbf{G}_2^H\mathbf{N} \ldots \mathbf{G}_M^H\mathbf{N}\right]\right\|_2^{\mathbf{2}}\right\} \quad (8)$$

The parameter μ can be controlled to give different weightage to the speech distortion vs residual noise. Minimizing the cost function in Eq.(8) assuming that the early reverberations are uncorrelated with the late reverberations, and that the late reverberations in all the channels are uncorrelated with each other, we obtain the expression for the optimum MMSE wiener filter coefficient vector $\mathbf{G}_i^{opt}$ as:

$$\mathbf{G}_i^{opt} = \left(R_{\mathbf{YY}} + \mu R_{\mathbf{NN}}\right)^{-1} R_{\mathbf{Y}Y_i} \quad (9)$$

with $R_{\mathbf{YY}} = \mathcal{E}\left\{\mathbf{Y}\mathbf{Y}^H\right\}$, $R_{\mathbf{NN}} = \mathcal{E}\left\{\mathbf{N}\mathbf{N}^H\right\}$ and

$$R_{\mathbf{Y}Y_i} = \left[\mathcal{E}\left\{Y_1Y_i^H\right\}\; \mathcal{E}\left\{Y_2Y_i^H\right\}\; \ldots\; \mathcal{E}\left\{Y_MY_i^H\right\}\right]^T$$

Since the late reverberations in each channel is uncorrelated with the rest of the channels, $R_{\mathbf{NN}}$ is assumed to be diagonal. The target clean speech estimate is obtained by performing MWF as:

$$\hat{Y}_i = (\mathbf{G}_i^{opt})^H\mathbf{X} \quad (10)$$

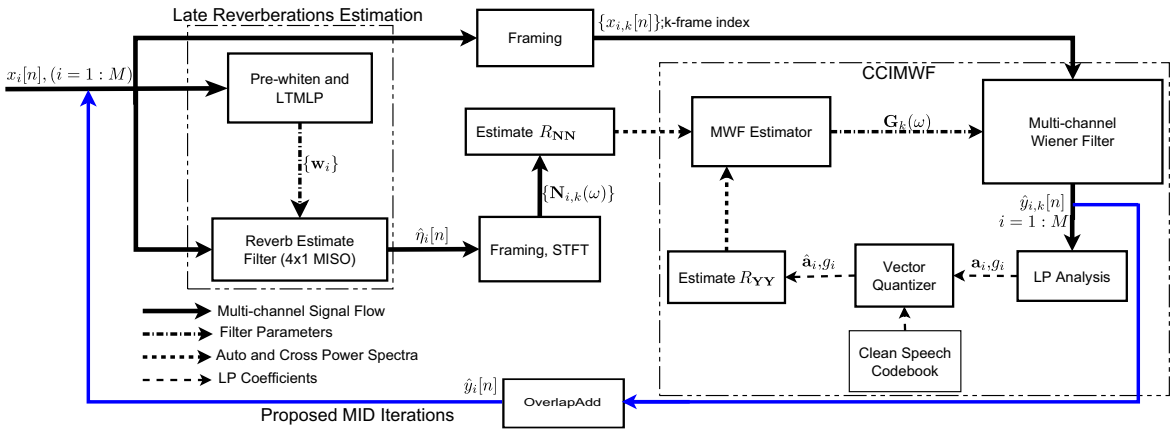

Figure 2: Block Diagram of the doubly iterative MID algorithm

3.3. Codebook constrained iterative MWF

Iterative wiener filtering (IWF) has been extensively studied in literature and in particular codebook constrained IWF is shown to be quite successful for various types of stationary and non-stationary distortions in signal. We extend this to the multi-channel case by considering the late reverberation component as the additive noise in each channel. Further, multi-channel measurement improves the performance of the MWF because of its MIMO nature: a high SRR in one of the channels helps to obtain a better estimate in a different channel with low SRR. The iterations provide successive improvement through both MWF as well as codebook constrained short-time LP of each channel.

The CCIMWF is performed over short frames of 20ms duration, while the late reverberation estimation is over the entire utterance of several seconds. In the IMWF iterations, we use the power spectrum estimate from the parametric LP analysis rather than that from the reverberant signal. Let iterations through the CCIMWF be referred to as the IMWF iterations.

For the codebook constraints, the short-time LP coefficient vector of each channel is estimated and quantized to the nearest clean speech codebook vector for the target signal $y_i[n]$, which is the clean speech with early reverberation. Thus, the codebook is expected to be that of clean speech with early reflections included. However, since RIR is an all-zero model and the LP analysis is an all-pole analysis, we have a premise that the LP analysis of clean speech and speech with early reverberation produce nearly identical LP spectra [6]. Hence, we use the clean speech codebook itself, instead of an early reverberant speech codebook. For effective quantization, we use equivalent LSF (Line Spectral Frequency) representation in the LP codebook.

For this multi-channel case, since all the channels of speech are recorded from a single source, ideally we expect the LP coefficients computed at each iteration of CCIMWF to be equal for all channels. So, we also examine a variation to CCIMWF, called the joint-CCIMWF (j-CCIMWF), where we use a joint distance measure and constrain all the channels in the current frame to quantize to the same codebook vector. A replicated codebook is used and all the M channel LP coefficients are jointly quantized minimizing a single distance measure to a single codebook vector.

3.4. Multi-channel Iterative Dereverberation (MID) algorithm

The CCIMWF as formulated above enhances the signal against the additive model of late reverberation. The LTMLP formulation will get benefited by the increased SRR through the CCIMWF. Hence, any residual reverberation present in the target speech after CCIMWF iterations can be re-estimated using LTMLP, over the entire utterance. Thus, we form an outer iteration loop of MID using successive LTMLP and CCIMWF, resulting in a doubly iterative procedure. The SRR is expected to improve through both the iterations. Though **we do not have an explicit convergence criterion** for the MID and CCIMWF iterations, we experimentally find the number of MID and IWF iterations for best performance. Since the algorithm is doubly iterative, the late reverberation estimation need not be most accurate. The parameter μ of the wiener filter is also not critical. The MID scheme is shown in figure 2.

4. Experiments and Results

We consider 4 channel reverberant speech sampled at 8kHz in all the experiments. The room impulse response of a reverberant room (5m $\times$ 6m $\times$ 4m) is computed using the image method [8], for the source at (1.9,1.9,3.5) and mics at (2,2,1.5), (1,2,1.5), (3,2,1.5), (2,3,1.5) (All in m). The reverberation time (RT_{60}) for the simulated room environment is about 0.5 sec.

The long-term predictor order $P = 1000$ was chosen with the step-size parameter $D = 160$ samples (=20ms). The late reverberation signal is estimated over the entire speech utterance of $\sim$ 5 sec. For pre-whitening, a 20^{th} order short-time predictor is used as suggested in [4]. For CCIMWF, a 256 size, 10^{th} order LSF VQ codebook was generated using the clean speech from TIMIT database. We have used about 38 min of clean speech from dialect region 2 of TIMIT to generate the codebook. The feature vectors of clean speech are calculated over 20ms frames. Wiener filter parameter, $\mu = 2$ is used.

Dereverberation is performed using CCIMWF and j-CCIMWF and IMWF (without codebook constraints), under identical conditions of LTMLP. The dereverberation performance is measured using the known average Segmental SRR $(\overline{SSRR})$ measure defined between the clean speech and reverberant speech. Average Segmental Log Spectral Distortion $(\overline{SLSD})$, defined between the LP power spectra of clean and

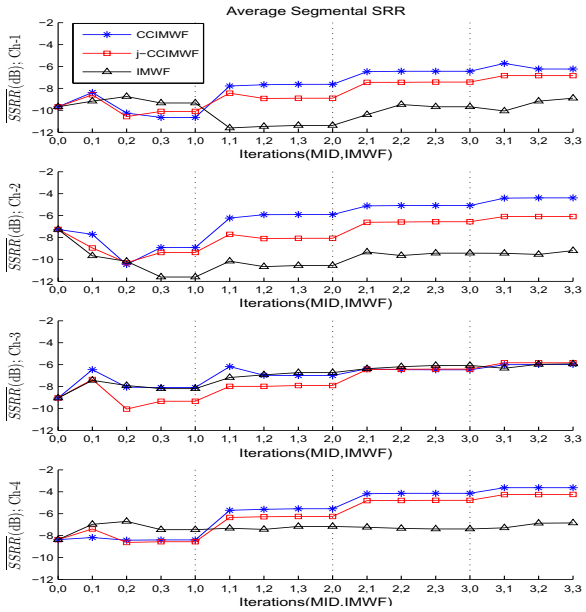

Figure 3: $\overline{SSRR}$ through MID and IMWF iterations (MID iteration boundaries are demarked using a vertical dotted line)

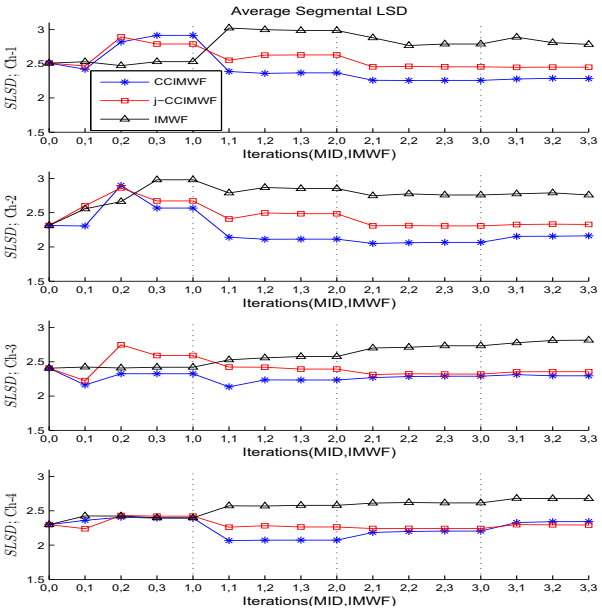

Figure 4: $\overline{SLSD}$ through MID and IMWF iterations

reverberant speech is used as a measure of intelligibility. The $(\overline{SSRR})$ and $(\overline{SLSD})$ are traced through the MID and IMWF iterations.

4.1. Results and Discussion

The performance of the algorithm over both the CCIMWF and MID iterations is shown for all the channels in figure 3 and figure 4. The x-axis is marked using ordered pairs indicating the MID and IMWF iteration count. A value of 0 for IMWF indicates the end of LTMLP and the initialization of CCIMWF.

We found that more than 70% of the signal frames converge in their segmental SRR within 3 iterations of CCIMWF. Hence, we fix 3 iterations of CCIMWF for each MID iteration. Since the residual late reverberations reduce through MID iterations, the late reverberations estimate provided by LTMLP also gets smaller through the MID iterations. The improvement in $\overline{SSRR}$ is marginal after the 3^{rd} iteration of MID as shown.

From figure 3, we see that the $\overline{SSRR}$ improves over iterations, except during the first MID iteration. However, the improvement over MID iterations is monotonic after 1^{st} iteration. We also see that CCIMWF performs always better than j-CCIMWF. This suggests that the individual channels get benefited from independent quantization rather than joint quantization. Since the target speech has early reverberation included, all channels may not be quantized to the same codebook vector. Both CCIMWF and j-CCIMWF have better convergence than without codebook constraints (IMWF), emphasizing the need for codebook constraints. From figure 4, we see that $\overline{SLSD}$ decreases marginally through the iterations for CCIMWF and j-CCIMWF, indicating better intelligibility. For IMWF though, $\overline{SLSD}$ increases for a majority of channels, reinforcing that codebook constraints are crucial for the IMWF. The improvement over all the channels is not similar, with channels closer to source having a better performance than farther channels. Listening the enhanced speech supports some of these observations and additional artifacts.

5. Conclusions

The proposed MID algorithm provides an iterative method to suppress late reverberations. The double-iterations reduce the residual reverberation through the iterations. The joint-CCIMWF is sub-optimal compared to CCIMWF. Though the CCIMWF formulation is used for dereverberation, it could be used in other multi-channel speech enhancement applications also.

6. References

[1] K. Eneman, M. Moonen, "Multimicrophone Speech Dereverberation: Experimental Validation," *EURASIP Journal on Audio, Speech, and Music Processing*, Vol 2007.

[2] D. Gesbert, P. Duhamel, "Robust Blind Channel Identification and Equalization based on Multi-Step Predictors," in *Proc. Intl. Conf. Acoustics, Speech and Signal Processing*, 1997, vol 26(5), pp. 3621-3624.

[3] E.A.P. Habets, "Single and Multi-Microphone Speech Dereverberation using Spectral Enhancement," *Ph.D. Thesis*, Technische Universiteit Eindhoven, June 2007

[4] K. Kinoshita, M. Delcroix, T. Nakatani, "Suppression of Late Reverberation Effect on Speech Signal Using Long-Term Multiple-step Linear Prediction," *IEEE Trans. on Audio, Speech and Language Processing*, Vol. 17, No. 2, Feb 2009.

[5] N. Cazi, T V Sreenivas, "Enhancement of Binaural Speech Using Codebook Constrained Iterative Binaural Wiener Filter," in *Proc. INTERSPEECH*, pp. 1335 - 1338, 2009.

[6] N.D. Gaubitch, D.B. Ward, and P.A. Naylor, "Statistical analysis of the autoregressive modeling of reverberant speech," *Journal of the Acoustical Society of America*, 120(6), December 2006

[7] P. C. Loizou, *"Speech Enhancement: Theory and Practice,"* pp. 163-177, Boca Raton:CRC Press, 2007

[8] J.B. Allen and D.A. Berkley, "Image Method for Efficiently Simulating Small Room Acoustics," *Journal of the Acoustical Society of America*, vol. 65, no. 4, pp. 943950, 1979.