Joint Noise Cancellation and Dereverberation using Multi-Channel Linearly Constrained Minimum Variance Filter

Karan Nathwani and Rajesh M. Hegde

Department of Electrical Engineering, Indian Institute of Technology Kanpur, India

{nathwani,rhegde}@iitk.ac.in

Abstract

Speech acquired from an array of distant microphones is affected by ambient noise and reverberation. Single channel linearly constrained minimum variance (LCMV) filters have been proposed to remove ambient noise. In this paper, an algorithm for joint noise cancellation and dereverberation using a multi channel LCMV filter in the frequency domain is proposed. A single channel LCMV filter which accounts for the inter frame correlation is applied on each channel to remove the early reverberation component. A modified spectral subtraction method is also proposed to remove the late reverberation component present in the speech signal. Experimental results on joint noise cancellation and dereverberation indicate a reasonable improvement over conventional speech enhancement methods. Additional experiments on distant speech recognition are also conducted to illustrate the significance of the method.

Index Terms: Multi-channel, Noise Cancellation, Speech dereverberation, linearly constrained minimum variance filter, modified spectral subtraction, delay and sum beamformer.

1. Introduction

Hands free audio source acquisition from distant microphones are often smeared by reverberation. Reverberation is defined as multiple delayed and attenuated versions of a signal added to signal itself due to multiple reflections from the surrounding walls and other objects. The reverberation results in degradation of fidelity, intelligibility of speech signal and distant speech recognition performance.

In [1], single channel minimum variance distortionless response (MVDR) filter has been proposed for noise cancellation. This can improve both narrowband and fullband output SNRs and is extended to LCMV filter in [2]. The MVDR has been widely used in spectrum estimation [3], [4], [5], [6] and feature extraction [7], [8]. In [9], spatiotemporal averaging method is defined which operates on the linear prediction residual (LPR) of spatially averaged multi-microphone observations for enhancement of reverberant speech. Speech enhancement using source information by computing the coherently added signal from LPR of degraded speech from different microphones is discussed in [10]. In [11], multi-channel speech dereverberation algorithm is proposed to suppress late reverberation components which employ a MVDR and a single channel MMSE estimator operates on the beamformers output signal.

In this paper, the single channel linearly constrained minimum variance (LCMV) filter has been defined by defining the constraints separately for direct path and late reflection path required to obtain the distortionless response. This single channel LCMV filter is then applied to each of the microphone outputs to suppress the noise & early part of reverberation (reflection) components from each multi-channel. The delay and sum beamformer (DSB) is used to combine multi-channel LCMV filter outputs to obtain the enhanced speech. The late reverberant components are then eliminated by post processing at output of DSB using modified spectral subtraction method. The inter frame correlation defined in this work was introduced in [12].

The rest of the paper is organized as follows. Section 2 gives the problem formulation. Section 3, explains single channel LCMV filter for noise and early reverberation suppression. In Section 4, the multi-channel speech dereverberation & noise cancellation is described. The performance evaluation for proposed method is discussed in Section 5. Section 6 presents brief conclusion.

2. Problem Formulation

The audio signal recorded by single microphone is often smeared by reverberation and surrounding noise. In general, the acoustic impulse response (AIR) $h(n)$ is assumed to be time-invariant. The microphone output is given by $y(n)$

$$y(n) = z(n) + v(n) \quad (1)$$

where $n$ is time samples, $v(n)$ is additive noise and $z(n)$ is the reverberated signal given by

$$z(n) = h(n) \ast s(n) \quad (2)$$

$$z(n) = \sum_{n'=-T}^{T} h(n')s(n-n') \quad (3)$$

where $s(n)$ is the clean speech signal of $T$ time samples. In the short time Fourier transform (STFT) domain, the microphone signal output $y(n)$ is given by

$$Y(k,m) = Z(k,m) + V(k,m) \quad (4)$$

$$Z(k,m) = \sum_{m'=0}^{N_f-1} H(k,m)S(k,m-m') + \sum_{m'=N_f}^{N_T-1} H(k,m)S(k,m-m') \quad (5)$$

where $m$ & $k$ are time frame and frequency bin index respectively. The first $N_s$ frames corresponds to the direct path which have strong peaks. The early & late reflection components are contained in the frame range of $[N_s, N_T - 1]$. $N_T$ corresponds to the total number of consecutive frames obtained after STFT of $y(n)$. $V(k,m)$ is STFT of additive noise $v(n)$. The Equation 5 can be written as

$$Z(k,m) = Z_d(k,m) + Z_l(k,m) \quad (6)$$
where \( Z_d(k, m) \) is direct spectral components and \( Z_l(k, m) \) denote early & late spectral speech components. In Equation 5, \( H(k, m) \) (AIR in frequency domain) is split in two parts:

\[
H(k, m) = \begin{cases} 
H_d(k, m) & \text{for } 0 < m < N_e - 1 \\
H_l(k, m) & \text{for } N_e \leq m \leq N_r - 1 
\end{cases}
\]  
(7)

where \( H_d(k, m) \) models the direct path and \( H_l(k, m) \) models all early & late reflections path. Substituting Equation 6 in Equation 4, the microphone output is obtained as

\[
Y(k, m) = Z_d(k, m) + Z_l(k, m) + V(k, m)
\]  
(8)

A single channel LCMV filter which accounts for the inter frame correlation is discussed in ensuing Section. The single channel LCMV filter suppress noise and early reverberation.

### 3. Suppression of Noise & Early Reverberation in Single Channel LCMV Filter

In this work, the inter frame correlation is considered which is due to the highly correlated nature of speech signal. The complex gain required to remove noise & early reverberation components using single channel LCMV filter is given by

\[
\hat{Z}(k, m) = \sum_{c=0}^{N_T-1} W_c^H(k, m) Y(k, m - c)
\]  
(9)

\[
\hat{Z}(k, m) = w^H(k, m) y(k, m)
\]  
(10)

where \( w^H \) and \( * \) denote transpose conjugate and complex conjugate operation respectively, \( w(k, m) = [W_0(k, m) ... W_{N_T-1}(k, m)]^T \), \( y(k, m) = [Y(k, m) ... Y(k, m - N_T + 1)]^T \) are vectors of length \( N_T \) and \( T \) denotes transpose operation. The Equation 10 can also be written as

\[
\hat{Z}(k, m) = w^H(k, m) z_d(k, m) + w^H(k, m) z_l(k, m) + w^H(k, m) v(k, m)
\]  
(11)

where \( z_d(k, m) \), \( z_l(k, m) \) and \( v(k, m) \) are defined in similar manner as \( y(k, m) \) is defined. However as shown in [1], the vector \( z_d \) can be decomposed into desired signal \( Z_d(k, m) \) at time frame \( m \) and interference signal vector \( z^*_d \) as shown below:

\[
z_d(k, m) = Z_d(k, m) \gamma_d + z^*_d(k, m)
\]  
(12)

\[
z^*_d(k, m) = [Z^*_d(k, m) ... Z^*_d(k, m - N_T + 1)]^T
\]

\( \gamma_d = [\gamma_d(k, m, 0) ... \gamma_d(k, m, N_T - 1)]^T \)

\[
\gamma_d = E[Z_d(k, m) z^*_d(k, m)] / [E[Z_d(k, m)]^2]
\]  
(13)

Similarly, vector \( z_l(k, m) \) and vector \( v(k, m) \) can be decomposed into its desired signal and its corresponding interfering signal vector

\[
z_l(k, m) = Z_l(k, m) \gamma_l + z^*_l(k, m)
\]  
(14)

\[
v(k, m) = V(k, m) \gamma_v + v^*(k, m)
\]  
(15)

where \( \gamma_d \) and \( \gamma_l \) are defined in similar way to \( \gamma_d \). In similar fashion, \( z^*_d(k, m) \) and \( v^*(k, m) \) are also defined as \( z^*_d(k, m) \).

Substituting Equation 12, 14 and 15 into Equation 11, we obtain

\[
\hat{Z}(k, m) = Z_l(k, m) + Z_{int}(k, m) + V(k, m) + Z_r(k, m)
\]  
(16)

where \( Z_l(k, m) = Z_d(k, m) w^H(k, m) \gamma_d \) is the filtered desired signal, \( Z_{int}(k, m) = w^H(k, m) z^*_d(k, m) + w^H(k, m) z^*_l(k, m) + w^H(k, m) y(k, m) \gamma_v \) is the residual interference signal, \( V(k, m) = V_k(k, m) w^H(k, m) \gamma_v \) is the residual noise signal and \( Z_r(k, m) = Z_l(k, m) w^H(k, m) \gamma_d \) is the filtered early & late reverberant signal. The estimated desired signal \( \hat{Z}(k, m) \) (Equation 16) is the sum of four mutually uncorrelated terms.

The objective of single channel LCMV filter is to recover the filtered desired signal \( \hat{Z}(k, m) \) and remove all undesired signal terms (the last three terms of Equation 16). Thus, putting first set of constraint in matrix form, we obtain

\[
olvep T \owe_d(k, m) = I_t
\]  
(17)

where, \( \olvep = [\gamma_d(k, m) \gamma_l(k, m)], \ I_t = [1 0]^T \) and \( \owe_d(k, m) = [\owe_0(k, m) ... \owe_{N_r-1}(k, m)]^T \). The dimension of constraint matrix \( \olvep \) is \( N_r \times 2 \) and \( I_t \) is \( 2 \times 1 \).

The second set of constraint can be put in matrix form as

\[
olvep T \owe_d(k, m) = I_l
\]  
(18)

subject to \( \olvep T \owe_d(k, m) = I_l \)

The solution to the Equation 19 is given by [2] is

\[
\owe_d(k, m) = R^{-1}_{ue}(k, m) \olvep \owe_d(k, m) \olvep^{-1}(k, m) \owe_d(k, m)
\]  
(20)

where \( R_{ue}(k, m) = E[y_d(k, m) y^H_d(k, m)] \) is the correlation matrix of \( y_d(k, m) \). \( y_l(k, m) \) can be obtained as \( y_d(k, m) = [Y(k, m) ... Y(k, m - N_e + 1)] \), which is first \( N_e \) frames of \( y(k, m) \).

Similarly weights required to minimize the energy at the filter output with second set of constraint is obtained as

\[
\owe(k, m) = R^{-1}_{ue}(k, m) \olvep \owe(k, m) \olvep^{-1}(k, m) \owe(k, m)
\]  
(21)

where \( \olvep \) is defined similar to \( R_{ue} \) and \( y_l(k, m) \) is defined as \( y_l(k, m) = [Y(k, m) ... Y(k, m - N_r + 1)] \). In general, \( \gamma_{zd} \) in Equation 13 is found in terms of inter frame correlation vector of \( y_d(k, m) \) and \( y_l(k, m) \) as obtained in [1]. Here, noise vector \( v_o(k, m) \) is obtained from \( v(k, m) \) for first \( N_e \) frames. Similarly \( \gamma_{zl} \) can be obtained in terms of \( y_l(k, m) \) and \( v_l(k, m) \), where \( v_l(k, m) \) is obtained from \( v(k, m) \) for frame range \( [N_e N_r - 1] \). The statistics of the noise signal is computed during silences as other noise reduction algorithm.

The enhanced signal is denoted by \( \hat{Z}(k, m) \) obtained as \( \hat{Z}(k, m) = [Z_d(k, m) \hat{Z}_l(k, m)] \). In other words, the enhanced
signal is produced by concatenating the desired signal \( \hat{Z}_d(k, m) \) & \( Z_i(k, m) \) obtained from first and second constraint respectively as given below

\[
\hat{Z}_d(k, m) = \mathbf{w}_d^H (k, m) \mathbf{y}_d(k, m) \quad (22)
\]

\[
\hat{Z}_i(k, m) = \mathbf{w}_i^H (k, m) \mathbf{y}_i(k, m) \quad (23)
\]

The single channel LCMV filter model described herein is applied to an array of microphones in next Section.

4. Multi-Channel Noise Cancellation & Speech Dereverberation

The output of LCMV filter at each channel of an array of M microphones is denoted by \( \hat{Z}_1(k, m), \hat{Z}_2(k, m), \ldots, \hat{Z}_M(k, m) \).

The output \( \hat{Z}_1(k, m), \hat{Z}_2(k, m), \ldots, \hat{Z}_M(k, m) \) are generated individually by using notion explained in Section 3. The block diagram of proposed multi-channel LCMV (M-LCMV) algorithm is shown in Figure 1. The application of LCMV filter to each microphone signal output results in removal of noise and all early reflection components. The delay and sum beamformer [13] is used to combine multi-channel LCMV filter outputs which results in an enhanced speech denoted by \( \hat{Z}_{DSB}(k, m) \).

The late reverberant components are then eliminated by post processing at output of DSB using modified spectral subtraction method.

If we assume that reverberation components are approximately uncorrelated between frames, the coefficients of the late reverberation are estimated by

\[
\alpha_i(k) = E \left[ \frac{\hat{Z}_{DSB}(k, m) \hat{Z}^*_i(k, m - i)}{|\hat{Z}_{DSB}(k, m - i)|^2} \right] \quad (26)
\]

Spectral subtraction is now employed to obtain the dereverberated signal:

\[
\hat{S}(k, m) = \hat{Z}_{DSB}(k, m) G(k, m) \quad (27)
\]

where \( \hat{S}(k, m) \) is the STFT of the recovered speech \( \hat{s}(n) \)

\[
G(k, m) = \left[ \frac{|\hat{Z}_{DSB}(k, m - i)|^2 - U(k, m)}{|\hat{Z}_{DSB}(k, m - i)|^2} \right]^{\frac{1}{2}} \quad (28)
\]

The dereverberated signal \( \hat{s}(n) \) is reconstructed from the estimated STFT \( \hat{S}(k, m) \), through the inverse-STFT (ISTFT) and overlap add techniques.

4.1. Modified spectral subtraction for removal of Late Reverberation

The post processing of the enhanced speech signal using a modified spectral subtraction method is described in this Section. This method helps in removing the late reverberation component present in the enhanced speech signal. The enhanced speech signal \( \hat{Z}_{DSB}(k, m) \) obtained after DSB output is a linear combination of the STFT \( S(k, m) \) of original speech \( s(n) \), that is

\[
\hat{Z}_{DSB}(k, m) = S(k, m) + \sum_{i=1}^{J} \alpha_i(k) S(k, m - i) \quad (24)
\]

where indexes \( k \) and \( m \) refer to frequency index and time frame, respectively, \( \alpha_i(k) \) is the coefficient of the late reverberation for previous \( i \) frames, and \( J \) is the duration of the reverberation. Here, \( \alpha_i(k) \ll 1 \) because the early reflection components that has most of the power of reverberation is reduced from each microphone response. Therefore, the power spectrum of late reverberation can be approximated by

\[
U(k, m) \approx \sum_{i=1}^{J} |\alpha_i(k)|^2 |\hat{Z}(k, m - i)|^2 \quad (25)
\]
5.1. Experimental Conditions

The TIMIT database are used in the experiments. An array of eight microphones are used in experiments. Sentences from the database are reverberated at various DRR and noise has been added into reverberated signal at SNR equal to 15 dB. The room dimension used for simulation is \(10.4m \times 10.4m \times 6.2m\). The subjective evaluation (MOS) was done by 25 listeners in the age group of 21 to 25 years on dereverberated sentences from the TIMIT database.

5.2. Experimental Results on Noise Cancellation & Speech Dereverberation

The noise cancellation results are presented in terms of segmental SNR [16], which is defined as average of SNR measured over short frames (good frames). The good frames are defined as those frames which are having SNR above a lower threshold (for example -10 dB) and saturated at an upper threshold (in our application +30 dB). The length of the short frames is generally 10-25 ms. The proposed method has higher segmental SNR values at different DRR, when compared to excitation source information (ESI) method of speech enhancement [10], spatio temporal processing (STP) method [9] and spectral enhancement (SE) method [11]. The \(SNR_{seg}\) values decreases with decrease in DRR for all methods and this decrement in \(SNR_{seg}\) is lowest for the proposed method. Higher the SNR values, better is the method in terms of speech intelligibility.

The speech Dereverberation results are listed as both objective and subjective measures in Table 2. The LSD is speech distortion measure obtained by root mean square (RMS) value of the difference between log spectra of clean speech signal and dereverberated signal [17], [18]. The \(SRR_{seg}\) is a measure of reverberation which depends on the signal before and after processing. The increment in LSD values and decrement in SRR & MOS values with decrease in DRR is noted for all the methods in Table 2. But, it is observed that proposed algorithm has lower LSD values and higher SRR & MOS values at different DRR indicating better reverberation suppression compared to other method used herein. In general, lower the LSD values and higher the SRR & MOS values, better is the method for reverberation suppression.

5.3. Experimental Results on Distant Speech Recognition

Spatialized version TIMIT (S-TIMIT) database is generated by acquiring TIMIT data over a microphone array and used for performing the experiments on distant speech recognition. The recognition results are presented in terms of word error rate (WER). 15 states, 3 mixtures triphone HMM with 39 MFCC with delta and acceleration coefficients have been used in the speech recognition experiments. In order to train the baseline triphone models of the recognition system, the clean speech data from database are used. The reconstructed signals from all these methods at different DRR are used for testing the recognition system. WER is defined as

\[
WER = \frac{S + D + I}{N}
\]

where \(S\) is the number of substitutions, \(D\) the number of deletions, \(I\) the number of insertions and \(N\) is the number of words in the reference.

Figure 3: Variation in WER for all the methods with increase in distance between source & microphone array.
7. References


