# 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 output 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 $y(n)$

$$y(n) = z(n) + v(n) \tag{1}$$

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

$$z(n) = h(n) * s(n) \tag{2}$$

$$z(n) = \sum_{n'=0}^{T} h(n)s(n-n') \tag{3}$$

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

$$Y(k,m) = Z(k,m) + V(k,m) \tag{4}$$

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

where $m$ & $k$ are time frame and frequency bin index respectively. The first $N_e$ frames corresponds to the direct path which have strong peaks. The early & late reflection components are contained in the frame range of $[N_e\ N_T-1]$. $N_T$ 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

$$Z(k,m) = Z_d(k,m) + Z_l(k,m) \tag{6}$$

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:

$$H(k,m) = \begin{cases} H_d(k,m) & \text{for } 0 \le m \le N_e - 1 \\ H_l(k,m) & \text{for } N_e \le m \le N_T - 1 \end{cases} \quad (7)$$

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

$$Y(k,m) = Z_d(k,m) + Z_l(k,m) + V(k,m) \quad (8)$$

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

$$\hat{Z}(k,m) = \sum_{c=0}^{N_T-1} W_c^*(k,m) Y(k,m-c) \quad (9)$$

$$\hat{Z}(k,m) = \mathbf{w}^H(k,m)\mathbf{y}(k,m) \quad (10)$$

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

$$\hat{Z}(k,m) = \mathbf{w}^H(k,m)\mathbf{z}_d(k,m) + \mathbf{w}^H(k,m)\mathbf{z}_l(k,m) + \mathbf{w}^H(k,m)\mathbf{v}(k,m) \quad (11)$$

where $\mathbf{z}_d(k,m)$, $\mathbf{z}_l(k,m)$ and $\mathbf{v}(k,m)$ are defined in similar manner as $\mathbf{y}(k,m)$ is defined.

However as shown in [1], the vector $\mathbf{z}_d$ can be decomposed into desired signal $Z_d(k,m)$ at time frame $m$ and interference signal vector $\mathbf{z}'_d$ as shown below

$$\mathbf{z}_d(k,m) = Z_d(k,m)\gamma^*_{Z_d} + \mathbf{z}'_d(k,m) \quad (12)$$

$\mathbf{z}'_d(k,m) = [Z'_d(k,m) \ldots Z'_d(k,m-N_T+1)]^T$ is the interference signal vector where each element of this vector is defined in details in [1], and normalized inter frame correlation vector is defined below
$\gamma_{Z_d} = [\gamma_{Z_d}(k,m,0) \ldots \gamma_{Z_d}(k,m,N_T-1)]^T$

$$\gamma_{Z_d} = \frac{E[Z_d(k,m)\mathbf{z}^*_d(k,m)]}{E[|Z_d(k,m)|^2]} \quad (13)$$

Similarly, vector $\mathbf{z}_l(k,m)$ and vector $\mathbf{v}(k,m)$ can be decomposed into its desired signal and its corresponding interfering signal vector

$$\mathbf{z}_l(k,m) = Z_l(k,m)\gamma^*_{Z_l} + \mathbf{z}'_l(k,m) \quad (14)$$

$$\mathbf{v}(k,m) = V(k,m)\gamma^*_V + \mathbf{v}'(k,m) \quad (15)$$

where $\gamma_{Z_l}$ and $\gamma_V$ are defined in similar way to $\gamma_{Z_d}$. In similar fashion, $\mathbf{z}'_l(k,m)$ and $\mathbf{v}'(k,m)$ are also defined as $\mathbf{z}'_d(k,m)$.

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

$$\hat{Z}(k,m) = Z_f(k,m) + Z_{int}(k,m) + V_n(k,m) + Z_{lr}(k,m) \quad (16)$$

where $Z_f(k,m) = Z_d(k,m)\mathbf{w}^H(k,m)\gamma^*_{Z_d}$ is the filtered desired signal, $Z_{int}(k,m) = \mathbf{w}^H(k,m)\mathbf{z}'_d(k,m) + \mathbf{w}^H(k,m)\mathbf{z}'_l(k,m) + \mathbf{w}^H(k,m)\mathbf{v}'(k,m)$ is the residual interference signal, $V_n(k,m) = V(k,m)\mathbf{w}^H(k,m)\gamma^*_V$ is the residual noise signal and $Z_{lr}(k,m) = Z_l(k,m)\mathbf{w}^H(k,m)\gamma^*_{Z_l}$ 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 $Z_f(k,m)$ and remove all undesired signal terms (the last three terms of Equation 16). Thus, putting first set of constraint in matrix form, we obtain

$$\mathbf{P}_d^T\mathbf{w}_d(k,m) = \mathbf{I}_d \quad (17)$$

where, $\mathbf{P}_d^T = [\gamma_{Z_d}(k,m)\ \gamma_{V_d}(k,m)]$, $\mathbf{I}_d = [1\ 0]^T$ and $\mathbf{w}_d(k,m) = [W_0(k,m) \ldots W_{N_e-1}(k,m)]^T$. The dimension of constraint matrix $\mathbf{P}_d$ is $N_e \times 2$ and $\mathbf{I}_d$ is $2 \times 1$.

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

$$\mathbf{P}_l^T\mathbf{w}_l(k,m) = \mathbf{I}_l \quad (18)$$

where, $\mathbf{P}_l^T = [\gamma_{Z_l}(k,m)\ \gamma_{V_l}(k,m)]$, $\mathbf{I}_l = [\alpha\ 0]^T$ and $\mathbf{w}_l(k,m) = [W_{N_e}(k,m) \ldots W_{N_T-1}(k,m)]^T$. The dimension of constraint matrix $\mathbf{P}_l$ is $(N_T - N_e) \times 2$ and $\mathbf{I}_l$ is $2 \times 1$. Here $0 < \alpha < 1$, but $\alpha$=0.5 removes the early reflections which has most of the power of reverberation and few late reflections. The two constraint have been dealt separately in this paper to obtain weights ($\mathbf{w}_d(k,m)$ & $\mathbf{w}_l(k,m)$) using LCMV filter. The optimal filter obtained by minimising the energy at the filter output with the first set of constraint is given by [2] as shown below

$$\hat{\mathbf{w}}_d(k,m) = \min_{w_d(k,m)} \mathbf{w}_d^H(k,m)\mathbf{R}_{y_d}(k,m)\mathbf{w}_d(k,m) \quad (19)$$

subject to $\mathbf{P}_d^T\mathbf{w}_d(k,m) = \mathbf{I}_d$
The solution to the Equation 19 is given by [2] is

$$\hat{\mathbf{w}}_d(k,m) = \mathbf{R}^{-1}_{y_d}(k,m)\mathbf{P}_d[\mathbf{P}_d^T\mathbf{R}^{-1}_{y_d}(k,m)\mathbf{P}_d]^{-1}\mathbf{I}_d \quad (20)$$

where $\mathbf{R}_{y_d}(k,m) = E[\mathbf{y}_d(k,m)\mathbf{y}_d^H(k,m)]$ is the correlation matrix of $\mathbf{y}_d(k,m)$. $\mathbf{y}_d(k,m)$ can be obtained as $\mathbf{y}_d(k,m) = [Y(k,m) \ldots Y(k,m-N_e+1)]$, which is first $N_e$ frames of $\mathbf{y}(k,m)$.

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

$$\hat{\mathbf{w}}_l(k,m) = \mathbf{R}^{-1}_{y_l}(k,m)\mathbf{P}_l[\mathbf{P}_l^T\mathbf{R}^{-1}_{y_l}(k,m)\mathbf{P}_l]^{-1}\mathbf{I}_l \quad (21)$$

where $\mathbf{R}_{y_l}$ is defined similar to $\mathbf{R}_{y_d}$ and $\mathbf{y}_l(k,m)$ is defined as $\mathbf{y}_l(k,m) = [Y(k,m-N_e) \ldots Y(k,m-N_T+1)]$.

In general, $\gamma_{Z_d}$ in Equation 13 is found in terms of inter frame correlation vector of $\mathbf{y}_d(k,m)$ and $\mathbf{v}_d(k,m)$ as obtained in [1]. Here, noise vector $\mathbf{v}_d(k,m)$ is obtained from $\mathbf{v}(k,m)$ for first $N_e$ frames. Similarly $\gamma_{Z_l}$ can be obtained in terms of $\mathbf{y}_l(k,m)$ and $\mathbf{v}_l(k,m)$, where $\mathbf{v}_l(k,m)$ is obtained from $\mathbf{v}(k,m)$ for frame range $[N_e\ N_T - 1]$ . 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 $\hat{Z}(k,m)$=$[\hat{Z}_d(k,m)\quad \hat{Z}_l(k,m)]$. In other words, the enhanced

signal is produced by concatenating the desired signal $\hat{Z}_d(k,m)$ & $\hat{Z}_l(k,m)$ obtained from first and second constraint respectively as given below

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

$$\hat{Z}_l(k,m) = \hat{\mathbf{w}}_l^H(k,m)\mathbf{y}_l(k,m) \quad (23)$$

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 $\hat{Z}_1(k,m)$, $\hat{Z}_2(k,m)$, ...., $\hat{Z}_M(k,m)$. The output $\hat{Z}_1(k,m)$, $\hat{Z}_2(k,m)$, ...., $\hat{Z}_M(k,m)$ 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.

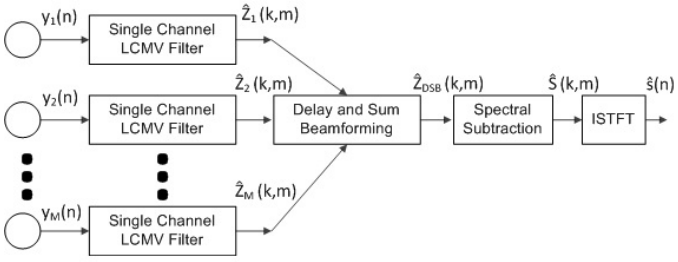


Figure 1: The block diagram for the proposed algorithm.

### 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

$$U(k,m) \approx \sum_{i=1}^{J} |\, \alpha_i(k) \,|^2 |\, \hat{Z}(k,m-i) \,|^2 \quad (25)$$

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

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

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.

### 4.2. Simulation Results for Joint Noise Cancellation & Dereverberation

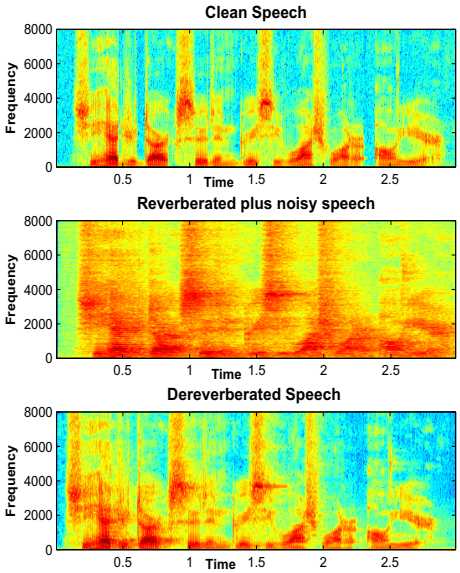


Figure 2: Figure illustrating the clean (Top), reverberant at DRR = -3dB (Middle), and the dereverberated speech signal from proposed algorithm (Bottom) spectrograms

The performance of the proposed M-LCMV algorithm is illustrated by considering a sentence from TIMIT database uttered by a male speaker sampled at 16 KHz. The AIR simulated here is done by image method [14]. Figure 2, illustrates the results of dereverberation using the proposed algorithm. It can be observed from dereverberated spectrogram (Figure 2) that there is a reasonable amount of noise & reverberation suppression.

# 5. Performance Evaluation

The performance of the proposed method is evaluated by conducting experiments on joint noise cancellation & dereverberation. The experiments on distant speech recognition are also conducted. The experiments are conducted on the TIMIT [15] database at various direct to reverberant ratios (DRR). The experiments on speech dereverberation evaluate subjective (mean opinion score (MOS)), and objective (log spectral distortion (LSD) and signal to reverberation ratios (SRR)) measures. The experiments on noise cancellation are conducted using segmental SNR ($SNR_{Seg}$) measure.

### 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 $10.4mX10.4mX6.2m$. 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.

Table 1: Experimental results on noise cancellation using segmental SNR measure for TIMIT database

| Methods | M-LCMV | ESI | STP | SE |
|---|---|---|---|---|
| DRR=-1dB | 14.42 | 10.07 | 9.13 | 7.72 |
| DRR=-3dB | 12.20 | 8.69 | 6.72 | 5.88 |
| DRR=-4dB | 10.11 | 7.25 | 5.32 | 4.79 |
| DRR=-5dB | 8.92 | 6.19 | 4.85 | 4.00 |

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 [17] 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.

Table 2: Experimental results on speech dereverberation for TIMIT database

| | DRR=-1dB | | | DRR=-5db | | |
|---|---|---|---|---|---|---|
| Methods | LSD | SRR | MOS | LSD | SRR | MOS |
| M-LCMV | 1.23 | 2.82 | 4.2 | 1.48 | 2.31 | 3.6 |
| ESI | 1.48 | 2.68 | 3.9 | 1.69 | 1.98 | 3.3 |
| STP | 1.66 | 2.36 | 3.5 | 1.95 | 1.82 | 2.8 |
| SE | 1.66 | 2.33 | 3.4 | 1.96 | 1.79 | 2.7 |

### 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

$$\text{WER} = \frac{S + D + I}{N} \tag{29}$$

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. In Figure 3, WER for all the methods increases as the distance between the source and microphone array increase (except for close talking microphone (CTM)). This is because as distance between source and microphone increases, the effect of reverberation increases. WER for CTM is constant & lowest because there is no reverberation effect during close talk. The proposed method has lower WER with respect to CTM indicating the higher recognition accuracy and least reverberation effect compared to ESI, STP and SE method. Table 3 illustrates the percentage increase in WER, when the distance between source & microphone array increases. The percentage increase in WER (obtained with respect to distance of one meter between source & microphone) is noted lowest for the proposed method compared to other methods.

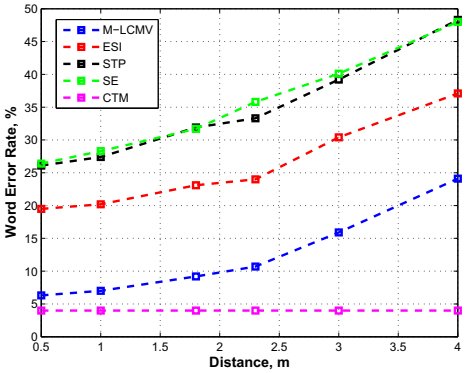


Figure 3: Variation in WER for all the methods with increase in distance between source & microphone array

Table 3: Percentage increase in WER for all methods

| Distance | M-LCMV | ESI | STP | SE |
|---|---|---|---|---|
| 2 meters | 3.0 | 4.0 | 4.8 | 4.0 |
| 3 meters | 8.9 | 10.2 | 11.8 | 11.8 |
| 4 meters | 17.0 | 17.0 | 20.9 | 19.8 |

## 6. Conclusion

In this paper, multi-channel speech enhancement algorithm is proposed by applying single channel linearly constrain minimum variance (LCMV) filter at the output of an array of microphones. The proposed method is based on orthogonal decomposition for the extraction of desired signal. The multi-channel output is combined by using DSB and spectral subtraction. This further improves the quality of reconstructed signal. The subjective and objective evaluation of proposed method shows reasonable improvement over other methods compared herein. Lower word error rate is noted from the experiments on distant speech recognition. The significance of the proposed method is also illustrated using segmental SNR, where higher SNR indicates higher intelligibility of reconstructed signal.

# 7. References


[1] J. Benesty and Y. Huang, "A single-channel noise reduction mvdr filter," in *Acoustics, Speech and Signal Processing (ICASSP), 2011 IEEE International Conference on*. IEEE, 2011, pp. 273–276.

[2] T. Long, J. Chen, J. Benesty, and Z. Zhang, "Single-channel noise reduction using optimal rectangular filtering matrices," *The Journal of the Acoustical Society of America*, vol. 133, no. 2, pp. 1090–1101, 2013. [Online]. Available: http://link.aip.org/link/?JAS/133/1090/1

[3] M. N. Murthi and B. D. Rao, "All-pole modeling of speech based on the minimum variance distortionless response spectrum," *Speech and Audio Processing, IEEE Transactions on*, vol. 8, no. 3, pp. 221–239, 2000.

[4] P. J. Sherman and K.-N. Lou, "On the family of ml spectral estimates for mixed spectrum identification," *Signal Processing, IEEE Transactions on*, vol. 39, no. 3, pp. 644–655, 1991.

[5] M. Wolfel, J. McDonough, and A. Waibel, "Minimum variance distortionless response on a warped frequency scale," in *Eurospeech 2003*, 2003.

[6] M. Wolfel and J. McDonough, "Minimum variance distortionless response spectral estimation," *Signal Processing Magazine, IEEE*, vol. 22, no. 5, pp. 117–126, 2005.

[7] S. Dharanipragada and B. D. Rao, "Mvdr based feature extraction for robust speech recognition," in *Acoustics, Speech, and Signal Processing, 2001. Proceedings.(ICASSP'01). 2001 IEEE International Conference on*, vol. 1. IEEE, 2001, pp. 309–312.

[8] M. Wolfel and J. McDonough, *Distant speech recognition*. Wiley, 2009.

[9] N. D. Gaubitch and P. A. Naylor, "Spatiotemporal averaging-method for enhancement of reverberant speech," in *Digital Signal Processing, 2007 15th International Conference on*. IEEE, 2007, pp. 607–610.

[10] B. Yegnanarayana, S. Prasanna, and K. S. Rao, "Speech enhancement using excitation source information," in *Acoustics, Speech, and Signal Processing (ICASSP), 2002 IEEE International Conference on*, vol. 1. IEEE, 2002, pp. I–541.

[11] E. A. Habets, "Towards multi-microphone speech dereverberation using spectral enhancement and statistical reverberation models," in *Signals, Systems and Computers, 2008 42nd Asilomar Conference on*. IEEE, 2008, pp. 806–810.

[12] J. Benesty, J. Chen, Y. Huang, and I. Cohen, *Noise reduction in speech processing*. Springer, 2009, vol. 2.

[13] J. Benesty, J. Chen, and Y. Huang, *Microphone array signal processing*. Springer, 2008, vol. 1.

[14] J. Allen and D. Berkley, "Image method for efficiently simulating small-room acoustics," *J. Acoust. Soc. Am*, vol. 65, no. 4, pp. 943–950, 1979.

[15] J. Garofolo, *TIMIT: Acoustic-phonetic Continuous Speech Corpus*. Linguistic Data Consortium, 1993.

[16] B. Grundlehner, J. Lecocq, R. Balan, and J. Rosca, "Performance assessment method for speech enhancement systems," in *Proc. 1st annu. IEEE BENELUX/DSP valley signal process. symp*, 2005.

[17] P. Naylor and N. Gaubitch, "Acoustic signal processing in noise: Its not getting any quieter," in *Acoustic Signal Enhancement; Proceedings of IWAENC 2012; International Workshop on*. VDE, 2012, pp. 1–6.

[18] T. Houtgast and H. J. Steeneken, "A review of the mtf concept in room acoustics and its use for estimating speech intelligibility in auditoria," *The Journal of the Acoustical Society of America*, vol. 77, p. 1069, 1985.