International Journal of Engineering

Journal Homepage: www.ije.ir

A Novel Frequency Domain Linearly Constrained Minimum Variance Filter for Speech Enhancement

S. Kammi*

Faculty of Electrical and Computer Engineering, Babol Noshirvani University of Technology, Babol, Iran

PAPER INFO

Paper history:
Received 21 July 2019
Received in revised form 06 October 2019
Accepted 08 November 2019

Keywords:
Neighboring Correlation
Linearly Constrained Minimum Variance Filter
Reduced Rank Model
Speech Enhancement

ABSTRACT

A reliable speech enhancement method is important for speech applications as a pre-processing step to improve their overall performance. In this paper, we propose a novel frequency domain method for single channel speech enhancement. Conventional frequency domain methods usually neglect the correlation between neighboring time-frequency components of the signals. In the proposed method, we take this correlation into account via: 1) considering neighboring correlation for speech signals, we break down the clean speech into two uncorrelated components; 2) considering neighboring correlation for noise, we approximate the noise as a rank-1 component. Then, we design a linearly constrained minimum variance (LCMV) filter which aims at removing the dominant part of the noise, while keeping the speech signal undistorted. Performance of the proposed method is evaluated in terms of output signal to noise ratio (SNR) and speech distortion index under various noise environments. Evaluation results demonstrate that our method yields higher noise reduction and lower speech distortion compared to some recent methods.

***doi**: 10.5829/ije.2020.33.01a.08*

1. INTRODUCTION

Noise is ubiquitous and easily contaminates speech signals, and degrades signal quality. This affects performance of speech applications such as teleconferencing systems, speech coding, automatic speech recognition, and necessitates integrating speech enhancement module as a front-end processor to improve their overall performance. Therefore, due to the extent and diversity of its applications, speech enhancement has received a significant amount of research attention, and variety of methods have been proposed in the literature [1] Most of these methods work in the frequency domain, where the most common approach is to filter the short-time Fourier transform (STFT) of the noisy speech in order to reduce the effect of noise from the speech. In filtering techniques, since both clean speech and noise pass through the filter, speech distortion is inevitable. In fact, as proved in the literature, the more the noise is reduced, the more the speech is distorted. So, it is important to control the trade-off between noise reduction and speech distortion for optimal quality. Traditional methods like classical Wiener filter [2], spectral subtraction methods [3-5] and model based methods [6-8] have been developed with no explicit control on the amount of speech distortion. The coefficients of the Wiener filter are estimated by minimizing the mean square error between clean and estimated speeches. In spectral subtraction methods, the noise spectrum is obtained and updated during speech silent periods, and the clean speech spectrum is estimated by subtracting the noise spectrum from the spectrum of the noisy speech. Model based methods usually assume statistical models like Gaussian, Laplacian and Gama distributions for clean speech and noise, and then, recover the clean speech spectral amplitude by minimzing the mean square error or finding the maximum a posteriori estimate.

Recently some research efforts have been devoted to design optimal filters by considering noise reduction and speech distortion measures [9-13]. These filters can be divided into two categories: filtering vectors and rectangular filtering matrices. Using these filters, a sample or a vector of clean speech are respectively

*Corresponding Author Email: *sonaykammi@gmail.com* (S. Kammi)

Please cite this article as: S. Kammi, A Novel Frequency Domain Linearly Constrained Minimum Variance Filter for Speech Enhancement, International Journal of Engineering (IJE), IJE TRANSACTIONS A: Basics Vol. 33, No. 1, (January 2020) 68-74

estimated by passing a vector of noisy speech through a filtering vector or a rectangular filtering matrix. Each category has its own pros and cons, and different types of optimal filters e.g., minimum variance distortionless response (MVDR), maximum signal to noise ratio (SNR) and linearly constrained minimum variance (LCMV) filters, have been proposed in both categories [9]. MVDR filter reduces the noise as much as possible, while guarantees that no speech distortion is occurred. Maximum SNR filter yields the highest possible output SNR at the expense of tremendous speech distortion. In this paper, we focus on LCMV filter. The advantage of this approach is that, it can handle more than one constraint at the same time, which makes it possible to better manage the trade-off between noise reduction and speech distortion.

To simplify the problem of speech enhancement, conventional frequency domain methods usually neglect the correlation between neighboring STFT coefficients. So, filtering operation is performed by applying a gain function to the STFT of the noisy speech only in the current frequency bin and frame. Recently, the methods presented in papers [10] and [11] have respectively considered interframe and intraframe correlations of speech signals for speech enhancement. In this paper we take advantage of neighboring correlations [14] (which also contain both interframe and intraframe correlations) of speech and noise, and design an LCMV filter for speech enhancement. The main contributions of this paper are summarized as follows:

- We consider neighboring correlations for both clean speech and noise, and derive a low-rank model for noise based on this correlation.
- We design an LCMV filter based on two different decompositions for clean speech and noise: orthogonal decomposition for clean speech, and low-rank decomposition for noise.

We show effectiveness of the proposed method through various experiments.

The remainder of this paper is as follows: in Section 2, we describe neighboring correlation and model the signal in frequency domain. In Section 3, we derive an LCMV filter by taking neighboring correlations into account. We present experiments in Section 4 to show performance of the proposed method. The paper is concluded in Section 5.

2. SIGNAL MODEL

The aim of speech enhancement is to recover the clean speech $x(t)$ from the noisy speech $y(t)$, which is mathematically defined as:

$$y(t) = x(t) + v(t) \tag{1}$$

where t indicates the discrete time index and $v(t)$ is the additive noise, which is assumed to be uncorrelated with $x(t)$. Applying the STFT to the noisy speech $y(t)$ is done as follows:

$$Y(f,k) = X(f,k) + V(f,k) \tag{2}$$

where $f = 1,2,\dots,F$ denotes the frequency bin index and $k = 1,2,\dots,K$ denotes the time frame index. To take neighboring correlations into account, we stack neighboring STFT components of the noisy speech into a vector. This yields the following vector signal model:

$$\mathbf{y}(f,k) = \mathbf{x}(f,k) + \mathbf{v}(f,k) \tag{3}$$

with

$$\begin{aligned}\mathbf{y}(f,k) = [&Y(f-F_c,k-K_c)\,Y(f-F_c+1,k-K_c)\dots\\ &Y(f+F_c,k-K_c)\dots Y(f-F_c,k+K_c)\\ &Y(f-F_c+1,k+K_c)\dots Y(f+F_c,k+K_c)]^T\end{aligned} \tag{4}$$

which is a vector of size $L = (2F_c+1)\times(2K_c+1)$, the superscript T indicates the transpose operator, F_c and K_c respectively indicate the numbers of frequency bins and frames before and after frequency bin f and frame k, and $\mathbf{x}(f,k)$ and $\mathbf{v}(f,k)$ are defined similar to $\mathbf{y}(f,k)$. Based on the fact that clean speech and additive noise are assumed to be uncorrelated, the correlation matrix of noisy speech vector $\mathbf{y}(f,k)$ can be written as follows:

$$\begin{aligned}\mathbf{\Phi}_{\mathbf{y}}(f,k) = E[\mathbf{y}(f,k)\mathbf{y}(f,k)^{\mathrm{H}}] = \mathbf{\Phi}_{\mathbf{x}}(f,k) +\\ \mathbf{\Phi}_{\mathbf{v}}(f,k)\end{aligned} \tag{5}$$

where $E[\cdot]$ denotes mathematical expectation, the superscript H denotes complex transpose-conjugation, and $\mathbf{\Phi}_{\mathbf{x}}(f,k) = E[\mathbf{x}(f,k)\mathbf{x}(f,k)^{\mathbf{H}}]$ and $\mathbf{\Phi}_{\mathbf{v}}(f,k) = E[\mathbf{v}(f,k)\mathbf{v}(f,k)^{\mathbf{H}}]$ are respectively correlation matrices of $\mathbf{x}(f,k)$ and $\mathbf{v}(f,k)$.

3. PROPOSED LCMV FILTER

Our purpose in the proposed method is to estimate $X(f,k)$ from $\mathbf{y}(f,k)$. To do so, we use the linear filtering approach as follows:

$$\begin{aligned}\hat{X}(f,k) =&\\ &\textstyle\sum_{k'=-K_c}^{K_c}\sum_{f'=-F_c}^{F_c} H_{f',k'}^{*}(f,k)Y(f+f',k+k')\\ &= \mathbf{h}^{H}(f,k)\mathbf{y}(f,k),\end{aligned} \tag{6}$$

where $\hat{X}(f,k)$ is the estimate of $X(f,k)$, the superscript "$*$" denotes complex conjugation, and

$$\begin{aligned}\mathbf{h}(f,k) =&\\ &[H_{-F_c,-K_c}(f,k)\,H_{-F_c+1,-K_c}(f,k)\dots H_{F_c,-K_c}(f,k)\dots\\ &H_{-F_c,K_c}(f,k)\,H_{-F_c+1,K_c}(f,k)\dots H_{F_c,K_c}(f,k)]^T\end{aligned} \tag{7}$$

is a complex-valued filtering vector of size $L = (2F_c+1)\times(2K_c+1)$.

Orthogonal decomposition has recently been used in the literature to express the noisy speech vector as an explicit function of the signal of interest [9-11]. Using

this approach, we decompose $\mathbf{x}(f,k)$ into two orthogonal components:

$$\begin{aligned}\mathbf{x}(f,k) &= \boldsymbol{\rho}_X(f,k)X(f,k) + \mathbf{x}'(f,k)\\ &= \mathbf{x}_{\mathrm{d}}(f,k) + \mathbf{x}'(f,k),\end{aligned} \tag{8}$$

where

$$\boldsymbol{\rho}_X(f,k) = \frac{E[\mathbf{x}(f,k)X^*(f,k)]}{E[|X(f,k)|^2]} \tag{9}$$

is the normalized neighboring correlation vector, $\mathbf{x}_{\mathrm{d}}(f,k) = \boldsymbol{\rho}_X(f,k)X(f,k)$ is the desired signal vector, and $\mathbf{x}'(f,k) = \mathbf{x}(f,k) - \boldsymbol{\rho}_X(f,k)X(f,k)$ is the interference signal vector, which is uncorrelated with $X(f,k)$. Using (3) and (8), we have

$$\mathbf{y}(f,k) = \mathbf{x}_{\mathrm{d}}(f,k) + \mathbf{w}(f,k), \tag{10}$$

where $\mathbf{w}(f,k) = \mathbf{x}'(f,k) + \mathbf{v}(f,k)$ is the combined noise.

We also use the reduced rank technique, which is widely used in subspace based speech enhancement methods, to model the noise vector [15]. In subspace based methods, speech enhancement is performed in time domain and because of the high self-correlation of speech signal, it is assumed to have a low-rank linear model. So, the dimension of speech subspace is considered to be much smaller than that of the noisy speech space, and the bases of this subspace are obtained through eigenvalue decomposition of the correlation matrix or singular value decomposition of the data matrix. This technique has been used to model speech and noise in time domain [12, 13]. In STFT domain, noise matrix is assumed to lie near a rank-1 subspace because of the correlation within noise time frames [16]. We use the same assumption in the proposed method to model the noise vector because of the strong correlation between neighboring components of the noise in STFT domain. Using eigenvalue decomposition, we get,

$$\boldsymbol{\Phi}_{\mathbf{w}}(f,k) = \mathbf{Q}_{\mathbf{w}}(f,k)\boldsymbol{\Lambda}_{\mathbf{w}}(f,k)\mathbf{Q}_{\mathbf{w}}^{H}(f,k) \tag{11}$$

where $\mathbf{Q}_{\mathbf{w}}(f,k) = \left[\mathbf{q}_{\mathbf{w},1}(f,k)\ \mathbf{q}_{\mathbf{w},2}(f,k)\ \dots\ \mathbf{q}_{\mathbf{w},\mathrm{L}}(f,k)\right]$ is an orthogonal matrix holding eigenvectors and $\boldsymbol{\Lambda}_{\mathbf{w}}(f,k) = \mathrm{diag}\left(\lambda_{\mathbf{w},1}(f,k), \lambda_{\mathbf{w},2}(f,k), \dots, \lambda_{\mathbf{w},\mathrm{L}}(f,k)\right)$ is a diagonal matrix holding eigenvalues of the noise correlation matrix. Rank-1 approximation for noise subspace yields to the assumption: $\lambda_{\mathbf{w},1}(f,k) \gg \lambda_{\mathbf{w},2}(f,k) \geq \cdots \geq \lambda_{\mathbf{w},\mathrm{L}}(f,k)$. Based on this assumption, we model the noise vector as follows:

$$\begin{aligned}\mathbf{w}(f,k) &= \mathbf{q}_{\mathbf{w},1}(f,k)\mathbf{q}_{\mathbf{w},1}^{\mathrm{H}}(f,k)\mathbf{w}(f,k) + \mathbf{w}'(f,k)\\ &= \mathbf{w}_{\mathrm{d}}(f,k) + \mathbf{w}'(f,k)\end{aligned} \tag{12}$$

where $\mathbf{q}_{\mathbf{w},1}(f,k)$ is the eigenvector corresponding to the largest eigenvalue $\lambda_{\mathbf{w},1}(f,k)$, $\mathbf{q}_{\mathbf{w},1}(f,k)\mathbf{q}_{\mathbf{w},1}^{\mathrm{H}}(f,k)$ is the orthogonal projection matrix which projects the noise vector to a rank-1 subspace that is assumed to concentrate most of the energy of the noise, so we call $\mathbf{w}_{\mathrm{d}}(f,k) = \mathbf{q}_{\mathbf{w},1}(f,k)\mathbf{q}_{\mathbf{w},1}^{\mathrm{H}}(f,k)\mathbf{w}(f,k)$ as the dominant noise vector, and $\mathbf{w}'(f,k) = \sum_{l=2}^{L}\mathbf{q}_{\mathbf{w},l}(f,k)\mathbf{q}_{\mathbf{w},l}^{\mathrm{H}}(f,k)\mathbf{w}(f,k)$ is the remaining noise. Using (6), (10) and (12) we have:

$$\begin{aligned}\hat{X}(f,k) &= \mathbf{h}^H(f,k)\mathbf{x}_{\mathrm{d}}(f,k) + \mathbf{h}^H(f,k)\mathbf{w}_{\mathrm{d}}(f,k)\\ &\quad + \mathbf{h}^H(f,k)\mathbf{w}'(f,k)\\ &= X_{fd}(f,k) + W_{rd}(f,k) + W'_{rr}(f,k)\end{aligned} \tag{13}$$

where $X_{fd}(f,k) = \mathbf{h}^H(f,k)\mathbf{x}_{\mathrm{d}}(f,k)$ is the filtered desired signal, $W_{rd}(f,k) = \mathbf{h}^H(f,k)\mathbf{w}_{\mathrm{d}}(f,k)$ is the residual dominant noise, and $W'_{rr}(f,k) = \mathbf{h}^H(f,k)\mathbf{w}'(f,k)$ is the residual remaining noise. The error signal between the estimated signal $\hat{X}(f,k)$ and the desired signal $X(f,k)$ is defined as:

$$\begin{aligned}e(f,k) &= \hat{X}(f,k) - X(f,k)\\ &= X_{fd}(f,k) - X(f,k) + W_{rd}(f,k) + W'_{rr}(f,k)\\ &= e_d(f,k) + e_{\mathbf{q}_{\mathrm{v},1}}(f,k) + e_r(f,k),\end{aligned} \tag{14}$$

where $e_d(f,k) = X_{fd}(f,k) - X(f,k) = (\mathbf{h}^H(f,k)\,\boldsymbol{\rho}_X(f,k) - 1)\,X(f,k)$ is the speech distortion due to the complex filter vector, $e_{\mathbf{q}_{\mathrm{v},1}}(f,k) = W_{rd}(f,k)$ is the residual dominant noise and $e_r(f,k) = W'_{rr}(f,k)$ is the residual remaining noise.

In the proposed method, we derive the LCMV filter by minimizing the energy at the filter output with the constraints that: 1) the speech is not distorted ($E\{e_d^2(f,k)\} = 0$), and 2) the residual dominant noise is cancelled ($E\{e_{\mathbf{q}_{\mathbf{w},1}}^2(f,k)\} = 0$). This is mathematically equivalent to:

$$\begin{aligned}&\mathbf{h}_{LCMV}(f,k) = \arg\min_{\mathbf{h}} \mathbf{h}^H(f,k)\boldsymbol{\Phi}_{\mathbf{y}}(f,k)\mathbf{h}(f,k)\\ &s.t.\quad \mathbf{h}^H(f,k)\boldsymbol{\rho}_X(f,k) = 1\,,\ \ \mathbf{h}^H(f,k)\mathbf{q}_{\mathbf{w},1}(f,k) = 0\end{aligned} \tag{15}$$

which can be rewritten as

$$\begin{aligned}&\mathbf{h}_{LCMV}(f,k) = \arg\min_{\mathbf{h}} \mathbf{h}^H(f,k)\boldsymbol{\Phi}_{\mathbf{y}}(f,k)\mathbf{h}(f,k)\\ &\qquad s.t.\ \ \mathbf{h}^H(f,k)\boldsymbol{\Gamma}(f,k) = [1\ \ 0],\end{aligned} \tag{16}$$

where $\boldsymbol{\Gamma}(f,k) = \left[\boldsymbol{\rho}_X(f,k)\ \ \mathbf{q}_{\mathbf{w},1}(f,k)\right]$. To solve this optimization problem, we first adjoin the constraint to the objective function using a Lagrange multiplier. Then, we derivate the objective function with respect to $\mathbf{h}(f,k)$ and set it to zero. So, the solution is obtained as:

$$\begin{aligned}&\mathbf{h}_{LCMV}(f,k) =\\ &\boldsymbol{\Phi}_{\mathbf{y}}^{-1}(f,k)\boldsymbol{\Gamma}(f,k)\left(\boldsymbol{\Gamma}^H(f,k)\boldsymbol{\Phi}_{\mathbf{y}}^{-1}(f,k)\boldsymbol{\Gamma}(f,k)\right)^{-1}\begin{bmatrix}1\\0\end{bmatrix}.\end{aligned} \tag{17}$$

Once the clean speech STFT components are estimated by applying the proposed LCMV filter to the noisy speech vectors, the enhanced speech is obtained in time domain by performing inverse STFT followed by the overlap-add method.

4. EXPERIMENTS AND RESULTS

In this section, we carry out a number of simulations to evaluate and compare the proposed method for speech enhancement.

4. 1. Simulation Conditions In our experiments, we use 10 randomly selected speech files from the TIMIT database [17]. Six different types of noises from the NOISEX-92 database [18] including white, car, babble, factory, f16 and hfchannel are added to clean speech signals at SNR levels of -5 dB, 0 dB, 5 dB and 10 dB to generate noisy speech signals. The sampling rates of the noise and speech signals are adjusted to 8 kHz. The STFT is implemented with a hamming window of length 64 samples, and a hop of 48 samples. To design the filter, we empirically set $F_c = K_c = 2$ in our experiments, which yields $L = 25$.

To implement the proposed LCMV filter, we need to estimate correlation matrices $\mathbf{\Phi_y}(f,k)$ and $\mathbf{\Phi_v}(f,k)$, and the normalized neighboring correlation vector $\boldsymbol{\rho}_X(f,k)$. Because the noisy speech is accessible, we can easily compute $\mathbf{\Phi_y}(f,k)$. But to compute $\mathbf{\Phi_v}(f,k)$ we would need a noise estimator like the one proposed in [19]. However, we skip the noise estimation process and directly estimate noise statistics from the noise signal as was done in papers [10-13]. In this paper, we initially estimate $\mathbf{\Phi_y}(f,k)$ and $\mathbf{\Phi_v}(f,k)$ using 200 frames from their corresponding signals. Then, we use the rest frames of the signals for performance evaluations, where $\mathbf{\Phi_y}(f,k)$ and $\mathbf{\Phi_v}(f,k)$ are recursively updated as [10]:

$$\mathbf{\Phi_y}(f,k) = \lambda_y \mathbf{\Phi_y}(f,k-1) + (1-\lambda_y)\mathbf{y}(f,k)\mathbf{y}(f,k)^H \quad (18)$$

$$\mathbf{\Phi_v}(f,k) = \lambda_v \mathbf{\Phi_v}(f,k-1) + (1-\lambda_v)\mathbf{v}(f,k)\mathbf{v}(f,k)^H, \quad (19)$$

where $0 < \lambda_y < 1$ and $0 < \lambda_v < 1$ are forgetting factors. We set $\lambda_y = \lambda_v = 0.8$ in our experiments. After $\mathbf{\Phi_y}(f,k)$ and $\mathbf{\Phi_v}(f,k)$ are estimated, based on (5), we obtain estimate of $\mathbf{\Phi_x}(f,k)$ according to $\mathbf{\Phi_x}(f,k) = \mathbf{\Phi_y}(f,k) - \mathbf{\Phi_v}(f,k)$. Then, we take $\boldsymbol{\rho}_X(f,k)$ as the $((L+1)/2)$th vector of $\mathbf{\Phi_x}$ normalized by its $((L+1)/2)$th element.

4. 2. Performance Evaluations In Figure 1, we present a visual example of speech enhancement using the proposed method. Figures 1(a) and (b) show waveform and spectrogram of the noisy speech (with SNR=0 dB), Figures 1(c) and (d) show wavform and spectrogram of noise, Figures 1(e) and (f) show wavform and spectrogram of clean speech, Figures 1(g) and (h) show wavform and spectrogram of enhanced speech (which is a combination of filtered desired signal, residual dominant noise and residual remaining noise), Figures 1(i) and (j) show wavform and spectrogram of filtered desired signal, Figures 1(k) and (l) show wavform and spectrogram of residual dominant noise, Figures 1(m) and (n) show wavform and spectrogram of residual remaining noise. As the figure shows, the filtered desired signal highly resembles the clean speech signal, and the residual dominant noise is highly mitigated. This example shows effectiveness of the proposed method. For further evaluations, we compare performance of our proposed method with those of the following methods:

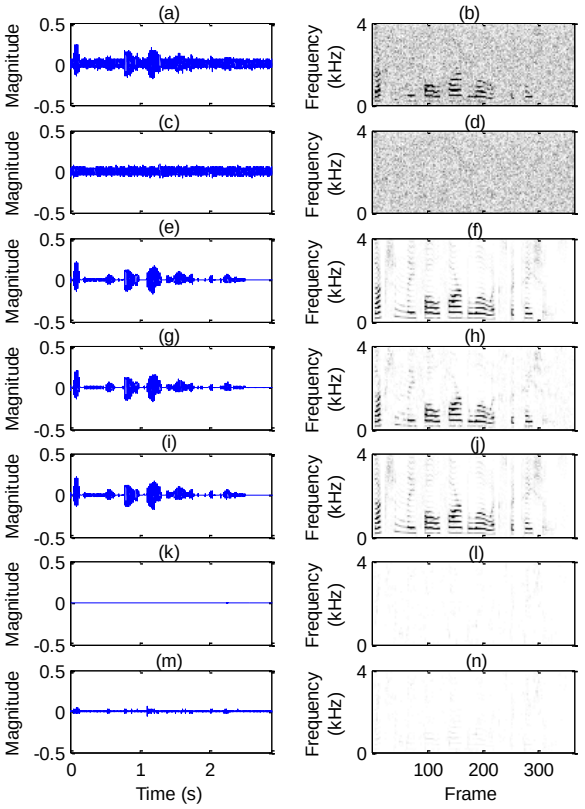

Figure 1. Signal waveforms and spectrograms for the proposed LCMV filter: (a) and (b) noisy speech with SNR=0 dB, (c) and (d) white noise, (e) and (f) clean speech, (g) and (h) enhanced speech, (i) and (j) filtered desired signal, (k) and (l) residual dominant noise, (m) and (n) residual remaining noise

1. Frequency domain minimum variance distortionless response filter, which takes interframe correlation into account [10].
2. Frequency domain minimum variance distortionless response filter, which takes intraframe correlation into account [11].

To compare performance of these methods, we calculate output SNR and speech distortion index on the enhanced speechs obtained using these methods. Output SNR which quantifies the level of noise remaining at the output of the filter is defined as [11]:

$$\text{oSNR} = 10\log_{10}\frac{\mathrm{E}\{x_{fd}^2(t)\}}{\mathrm{E}\left\{\left(w_{rd}(t)+w'_{rr}(t)\right)^2\right\}} \quad (20)$$

where $x_{fd}(t)$, $w_{rd}(t)$ and $w'_{rr}(t)$ are respectively the time domain signals reconstructed from $X_{fd}(f,k)$, $W_{rd}(f,k)$ and $W'_{rr}(f,k)$. Higher oSNR means lower residual noise, which indicates a better performance of the method.

Speech distortion index which quantifies the distortion level of the desired signal due to the filtering operation is defined as [11]:

$$v_{sd} = 10\log_{10}\frac{\mathrm{E}\left\{\left(x_{fd}(t)-x(t)\right)^2\right\}}{\mathrm{E}\{x^2(t)\}} \quad (21)$$

Lower v_{sd} indicates lower speech distortion, which means better performance of the method.

Tables 1 and 2 present comparison results in terms of output SNR and speech distortion index, respectively, at different noise conditions.

TABLE 1. Performance comparisons in terms of output SNR

Noise type	Method	-5 dB	0 dB	5 dB	10 dB
White	Proposed	**8.07**	**10.90**	**13.86**	**17.01**
	[11]	7.56	10.44	13.48	16.69
	[10]	6.96	10.06	13.33	16.67
Car	Proposed	**8.16**	**10.92**	**13.83**	**17.06**
	[11]	7.67	10.55	13.53	16.67
	[10]	7.31	10.33	13.53	16.97
Babble	Proposed	**6.58**	**9.22**	**12.13**	**15.32**
	[11]	5.93	8.64	11.48	14.55
	[10]	5.78	8.52	11.62	15.04
Factory	Proposed	**7.60**	**10.39**	**13.37**	**16.63**
	[11]	7.12	10.02	13.04	16.27
	[10]	6.70	9.73	13.02	16.56
F16	Proposed	**7.32**	**10.11**	**13.07**	**16.30**
	[11]	6.84	9.76	12.81	16.01
	[10]	6.40	9.40	12.66	16.13
Hfchannel	Proposed	**8.15**	**10.82**	**13.65**	**16.74**
	[11]	7.81	10.44	13.35	16.55
	[10]	7.10	10.05	13.14	16.40
Average	Proposed	**7.64**	**10.39**	**13.31**	**16.51**
	[11]	7.15	9.97	12.94	16.12
	[10]	6.70	9.68	12.88	16.29

TABLE 2. Performance comparisons in terms of speech distortion index

Noise type	Method	-5 dB	0 dB	5 dB	10 dB
White	Proposed	**-22.28**	**-30.85**	**-43.95**	**-51.40**
	[11]	-14.28	-21.54	-30.92	-43.30
	[10]	-16.44	-22.32	-29.66	-38.10
Car	Proposed	**-22.73**	**-30.11**	**-39.89**	-51.00
	[11]	-17.09	-23.92	-33.54	**-51.58**
	[10]	-16.38	-22.46	-29.63	-37.71
Babble	Proposed	**-19.15**	**-25.29**	**-34.71**	**-47.86**
	[11]	-13.98	-20.33	-30.56	-47.05
	[10]	-12.85	-18.08	-24.64	-32.27
Factory	Proposed	**-22.38**	**-30.78**	**-44.25**	-52.82
	[11]	-15.48	-22.50	-32.77	**-53.20**
	[10]	-15.82	-21.98	-29.56	-38.33
F16	Proposed	**-20.65**	**-28.34**	**-39.64**	**-48.83**
	[11]	-14.07	-21.19	-31.21	-46.84
	[10]	-14.08	-20.17	-27.53	-35.74
Hfchannel	Proposed	**-25.19**	**-35.75**	**-46.65**	**-53.12**
	[11]	-16.04	-23.11	-32.20	-45.55
	[10]	-17.42	-23.28	-30.68	-39.79
Average	Proposed	**-22.06**	**-30.18**	**-41.51**	**-50.83**
	[11]	-15.15	-22.09	-31.86	-47.92
	[10]	-15.49	-21.38	-28.61	-36.99

As the results show, compared to the competing methods, our proposed method achieves the highest oSNR at all noise conditions while achieving the lowest speech distortion index at 22 out of 24 noise conditions, where the only exceptions are car and factory noises at SNR level of 10 dB. Also, averaged results over all noise types are presented at the last row of the tables to give general comparisons at different SNR levels. Based on these results, our method yields higher output SNR and lower speech distortion than the competing methods at all SNRs, which confirms superiority of the proposed method.

Here we explain the reasons for the superiority of the proposed method:

1. Neighboring correlation is a stronger correlation than interframe or intraframe correlations, which means Equation (8) is a better model for speech signal in our proposed method. This explains why the proposed method causes less speech distortion.

2. By considering a low rank model for noise, because of its neighboring correlation in STFT domain, we manage the proposed filter to cancel the dominant part of the noise. This explains why the proposed method achieves higher noise reduction.

To further evaluate performace of the proposed method, we apply the well-known perceptual evaluation of speech quality (PESQ) metric [20]. PESQ reflects perceptual quality of enhanced speech and has a high correlation with subjective judgements of speech quality [21]. Higher PESQ indicates higher quality of the enhanced speech. Table 3 presents PESQ results of the proposed method and the following competing methods:

TABLE 3. Performance comparisons in terms of PESQ

Noise type	Method	-5 dB	0 dB	5 dB	10 dB
White	Proposed	2.42	**2.76**	**3.09**	**3.40**
	[11]	2.32	2.59	2.87	3.17
	[10]	1.96	2.34	2.74	3.09
	[14]	**2.43**	2.67	2.93	3.17
Car	Proposed	**2.55**	**2.87**	**3.17**	**3.46**
	[11]	2.42	2.68	2.92	3.22
	[10]	2.27	2.65	2.98	3.28
	[14]	2.50	2.73	2.97	3.20
Babble	Proposed	**2.40**	**2.67**	**2.98**	**3.33**
	[11]	2.30	2.52	2.77	3.05
	[10]	2.16	2.46	2.79	3.14
	[14]	2.34	2.54	2.77	3.01
Factory	Proposed	2.75	**3.06**	**3.39**	**3.68**
	[11]	2.60	2.87	3.19	3.50
	[10]	2.57	2.89	3.25	3.56
	[14]	**2.79**	3.02	3.25	3.44
F16	Proposed	2.44	**2.75**	**3.08**	**3.38**
	[11]	2.36	2.60	2.87	3.16
	[10]	2.11	2.48	2.82	3.19
	[14]	**2.45**	2.65	2.87	3.11
Hfchannel	Proposed	**2.39**	**2.71**	**3.03**	**3.34**
	[11]	2.31	2.58	2.87	3.12
	[10]	1.97	2.33	2.72	3.07
	[14]	2.36	2.63	2.87	3.09
Average	Proposed	**2.49**	**2.80**	**3.12**	**3.43**
	[11]	2.38	2.64	2.91	3.20
	[10]	2.17	2.52	2.88	3.22
	[14]	2.47	2.70	2.94	3.17

1. Frequency domain minimum variance distortionless response filtering vector, which takes interframe correlation into account [10].
2. Frequency domain minimum variance distortionless response filtering vector, which takes intraframe correlation into account [11].
3. Frequency domain minimum variance distortionless response rectangular filtering matrix, which takes neighboring correlation into account [14].

Comparison of the results show that our method outperforms the methods in [10] and [11] at all noise conditions, and outperforms the method in [14] at 21 out of 24 noise conditions, where the only exceptions are white, factory and f16 noises at SNR level of -5 dB. General comparison results presented in the last row of the table confirms superiority of the proposed method at all noise levels in terms of PESQ.

5. CONCLUSION

This paper deals with the problem of single channel speech enhancement in frequency domain. Unlike conventional frequency domain methods that assume the neighboring STFT coefficients are independent, this neighboring correlation is considered in the proposed method. Also, noise is considered to lie near a rank-1 subspace. Then, an LCMV filter is derived to preserve the speech and remove the dominant part of the noise. The evaluation using output SNR and speech distortion index metrics showed that the proposed method outperforms the two recently developed methods.

6. REFERENCES

1. Loizou, P.C., Speech enhancement: theory and practice, CRC press, (2013).
2. Vaseghi, S.V., Advanced signal processing and digital noise reduction, John Willey & Sons LTD. And B. G Teubner, (1996).
3. Udrea, R.M., Vizireanu, N.D. and Ciochina, S., "An improved spectral subtraction method for speech enhancement using a perceptual weighting filter", ***Digital Signal Processing***, Vol. 18, No. 4, (2008), 581–587.
4. Upadhyay, N. and Karmakar, A., "An improved multi-band spectral subtraction algorithm for enhancing speech in various noise environments", ***Procedia Engineering***, Vol. 64, (2013), 312–321.
5. Zhang, R. and Liu, J., "An Improved Multi-band Spectral Subtraction using Mel-scale", ***Procedia computer science***, Vol. 131, (2018), 779–785.
6. Chen, B. and Loizou, P. C., "A Laplacian-based MMSE estimator for speech enhancement", ***Speech Communication***, Vol. 49, No. 2, (2007), 134–143.
7. Wang, H.Y., Zhao, X.H. and Gu, H. J., "Speech enhancement using super gauss mixture model of speech spectral amplitude", ***The Journal of China Universities of Posts and Telecommunications***, Vol. 18, (2011), 13–18.

8. Abutalebi, H.R. and Rashidinejad, M., "Speech enhancement based on β-order MMSE estimation of Short Time Spectral Amplitude and Laplacian speech modeling", ***Speech Communication***, Vol. 67, (2015), 92–101.
9. Benesty, J. and Chen, J., Optimal time-domain noise reduction filters, Springer, (2011).
10. Huang, Y.A. and Benesty, J., "A multi-frame approach to the frequency-domain single-channel noise reduction problem", ***IEEE Transactions on Audio, Speech, and Language Processing***, Vol. 20, No. 4, (2011), 1256–1269.
11. Huang, H., Zhao, L., Chen, J. and Benesty, J., "A minimum variance distortionless response filter based on the bifrequency spectrum for single-channel noise reduction", ***Digital Signal Processing***, Vol. 33, (2014), 169–179.
12. Jensen, J.R., Benesty, J., Christensen, M.G. and Chen, J., "An LCMV filter for single-channel noise cancellation and reduction in the time domain", In 2013 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics, IEEE, (2013), 1–4.
13. Jensen, J.R., Benesty, J., Christensen, M.G. and Chen, J., "A class of optimal rectangular filtering matrices for single-channel signal enhancement in the time domain", ***IEEE Transactions on Audio, Speech, and Language Processing***, Vol. 21, No. 12, (2103), 2595–2606.
14. Kammi, S., "Single channel speech enhancement using an MVDR filter in the frequency domain", ***International Journal of Speech Technology***, Vol. 22, No. 2, (2019), 383–389.
15. Hansen, P.C. and Jensen, S. H., "Subspace-based noise reduction for speech signals via diagonal and triangular matrix decompositions: Survey and analysis", ***EURASIP Journal on Advances in Signal Processing***, Vol. 2007, No. 1, (2007), 1–24.
16. Sun, C., Zhu, Q. and Wan, M., "A novel speech enhancement method based on constrained low-rank and sparse matrix decomposition", ***Speech Communication***, Vol. 60, (2014), 44–55.
17. Garofolo, J.S., Lamel, L.F., Fisher, W.M., Fiscus, J.G. and Pallett, D.S., "Getting started with the DARPA TIMIT CD-ROM: An acoustic phonetic continuous speech database," National Institute of Standards and Technology (NIST), Gaithersburgh, (1988).
18. Varga, A. and Steeneken, H. J., "Assessment for automatic speech recognition: II. NOISEX-92: A database and an experiment to study the effect of additive noise on speech recognition systems", ***Speech Communication***, Vol. 12, No. 3, (1993), 247–251.
19. Cohen, I., "Noise spectrum estimation in adverse environments: Improved minima controlled recursive averaging", ***IEEE Transactions on Speech and Audio Processing***, Vol. 11, No. 5, (2003), 466–475.
20. ITU-T P, Perceptual evaluation of speech quality (PESQ): An objective method for end-to-end speech quality assessment of narrow-band telephone networks and speech codecs, Rec. ITU-T P. 862., (2001).
21. Hu, Y. and Loizou, P. C., "Evaluation of objective quality measures for speech enhancement", ***IEEE Transactions on Audio, Speech, and Language Processing***, Vol. 16, No. 1, (2007), 229–238.

A Novel Frequency Domain Linearly Constrained Minimum Variance Filter for Speech Enhancement

S. Kammi

Faculty of Electrical and Computer Engineering, Babol Noshirvani University of Technology, Babol, Iran

P A P E R I N F O

Paper history:
Received 21 July 2019
Received in revised form 06 October 2019
Accepted 08 November 2019

Keywords:
Neighboring Correlation
Linearly Constrained Minimum Variance Filter
Reduced Rank Model
Speech Enhancement

چکیده

استفاده از روش‌های قابل اتکای بهسازی گفتار به عنوان مرحله‌ی پیش پردازش برای بهبود عملکرد نهایی سیستم‌های پردازش گفتار حائز اهمیت می‌باشد. در این مقاله روشی جدید برای بهسازی سیگنال گفتار در حوزه‌ی فرکانس ارائه شده است. روش‌های متداول حوزه‌ی فرکانس معمولا از همبستگی بین مولفه‌های مجاور زمان–فرکانس سیگنال صرف نظر می‌کنند. در روش پیشنهادی، این همبستگی در نظر گرفته می‌شود: ۱) با در نظر گرفتن این همبستگی در سیگنال گفتار، سیگنال تمیز به دو مولفه ناهمبسته تجزیه می‌شود؛ ۲) با در نظر گرفتن این همبستگی در نویز، سیگنال نویز بصورت مولفه‌ی رتبه–۱ تقریب زده می‌شود. سپس یک فیلتر LCMV طراحی می‌شود تا قسمت غالب نویز حذف شده و سیگنال گفتار بدون اعوجاج باقی بماند. عملکرد روش پیشنهادی بر حسب نسبت سیگنال به نویز خروجی و شاخص اعوجاج سیگنال گفتار تحت شرایط نویزی مختلف ارزیابی می‌شود. نتایج ارزیابی نشان می‌دهند که روش پیشنهادی در مقایسه با چند روش اخیر منجر به کاهش بیشتر نویز و اعوجاج کمتر سیگنال گفتار می‌گردد.

***doi**: 10.5829/ije.2020.33.01a.08*