

# A SUBSPACE APPROACH FOR ENHANCING SPEECH CORRUPTED BY COLORED NOISE

*Yi Hu and Philipos C. Loizou*

Department of Electrical Engineering  
University of Texas at Dallas  
Richardson, Texas 75083-0688  
{yihuyxy, loizou}@utdallas.edu

## ABSTRACT

A generalized subspace approach is proposed for enhancement of speech corrupted by colored noise. The proposed approach is based on the simultaneous diagonalization of the clean speech and noise covariance matrices, which is shown to be a generalization of the approach proposed by Ephraim and Van Trees for white noise. Objective and subjective measures demonstrated significant improvements over other subspace-based methods when tested with sentences corrupted with speech-shaped noise and multi-talker babble.

## 1. INTRODUCTION

Several subspace-based speech enhancement methods (e.g., [1], [2]) have been proposed recently, following the pioneering work of Ephraim and Van Trees [3]. The original paper [3] dealt primarily with white noise and suggested using prewhitening to handle colored noise. However, as proved in [2], this only guarantees minimizing the upper bound of the variance of the speech distortion. The spectral domain constrained approach proposed by Ephraim and Van Trees [3] assumed that the solution had a special form, which does not hold if the noise is colored. Rezayee and Gazor [1] extended Ephraim and Van Trees' time domain constrained method to deal with colored noise by approximating the covariance matrix of the KLT-transformed noise vectors with a diagonal matrix. Their approximation, however, led to a sub-optimal estimator. In this paper, we derive two optimal linear estimators based on time domain and frequency domain constraints for subspace speech enhancement. The proposed method makes no assumption about the nature of the noise (white or colored).

This paper is organized as follows. In section II, the proposed approach using time-domain constraints is explained. In section III, the proposed approach using

frequency-domain constraints is described. Experimental results are discussed in Section IV, and the conclusions are given in Section V.

## 2. SUBSPACE APPROACH BASED ON TIME DOMAIN CONSTRAINTS

### 2.1. Principles

The model used in the subspace approach assumes that the noise signal is additive and uncorrelated with the speech signal, i.e.,

$$\mathbf{y} = \mathbf{x} + \mathbf{n} \quad (1)$$

where  $\mathbf{y}$ ,  $\mathbf{x}$  and  $\mathbf{n}$  are the  $K$ -dimensional noisy speech, clean speech and noise vectors respectively. Let  $\hat{\mathbf{x}} = H \cdot \mathbf{y}$  be a linear estimator of the clean speech  $\mathbf{x}$ , where  $H$  is a  $K \times K$  matrix. The error signal  $\varepsilon$  obtained by this estimation is given by:

$$\varepsilon = \hat{\mathbf{x}} - \mathbf{x} = (H - I) \cdot \mathbf{x} + H \cdot \mathbf{n} = \varepsilon_{\mathbf{x}} + \varepsilon_{\mathbf{n}} \quad (2)$$

where  $\varepsilon_{\mathbf{x}}$  represents the speech distortion and  $\varepsilon_{\mathbf{n}}$  represents the residual noise [3]. Defining the energies of the signal distortion  $\overline{\varepsilon_{\mathbf{x}}^2}$  and the energies of the residual noise  $\overline{\varepsilon_{\mathbf{n}}^2}$  as

$$\overline{\varepsilon_{\mathbf{x}}^2} = \text{tr} (E [\varepsilon_{\mathbf{x}} \varepsilon_{\mathbf{x}}^T]) \quad (3)$$

$$\overline{\varepsilon_{\mathbf{n}}^2} = \text{tr} (E [\varepsilon_{\mathbf{n}} \varepsilon_{\mathbf{n}}^T]) \quad (4)$$

we can obtain the optimum linear estimator by solving the following time-domain constrained optimization problem [3]:

$$\begin{aligned} & \min_H \overline{\varepsilon_{\mathbf{x}}^2} \\ & \text{subject to : } \frac{1}{K} \overline{\varepsilon_{\mathbf{n}}^2} \leq \sigma^2 \end{aligned} \quad (5)$$

where  $\sigma^2$  is a positive constant. The solution to Eq. 5 is given by [1]:

$$H_{opt} = R_{\mathbf{x}} (R_{\mathbf{x}} + \mu R_{\mathbf{n}})^{-1} \quad (6)$$

Research supported in part by Grant No. R01 DC03421 from NICHD/NIH.

where  $R_{\mathbf{x}}$  and  $R_{\mathbf{n}}$  are the covariance matrices of the clean speech and noise respectively, and  $\mu$  is the Lagrangian multiplier. Eq. 6 can be simplified using the eigen-decomposition of  $R_{\mathbf{x}} = U\Delta_{\mathbf{x}}U^T$  to:

$$H_{opt} = U\Delta_{\mathbf{x}}(\Delta_{\mathbf{x}} + \mu U^T R_{\mathbf{n}} U)^{-1} U^T \quad (7)$$

where  $U$  is the (unitary) eigenvector matrix and  $\Delta_{\mathbf{x}}$  is the diagonal eigenvalue matrix of  $R_{\mathbf{x}}$ . Note that for white noise with variance  $\sigma_{\mathbf{n}}^2$ ,  $R_{\mathbf{n}} = \sigma_{\mathbf{n}}^2 I$  and the above estimator reduces to the Ephraim and Van Trees' estimator. In [1], the matrix  $U^T R_{\mathbf{n}} U$  was approximated by the diagonal matrix  $\Delta_{\mathbf{n}}$ :

$$\Delta_{\mathbf{n}} = \text{diag}(|\mathbf{u}_1^T \mathbf{n}|^2, |\mathbf{u}_2^T \mathbf{n}|^2, \dots, |\mathbf{u}_K^T \mathbf{n}|^2) \quad (8)$$

where  $\mathbf{u}_k$  is the  $k$ -th eigenvector of  $R_{\mathbf{x}}$ , and  $\mathbf{n}$  is the noise vector estimated from the speech-absent segments of speech. The above approximation yielded the following sub-optimal estimator [1]:

$$H_{opt} \approx U\Delta_{\mathbf{x}}(\Delta_{\mathbf{x}} + \mu\Delta_{\mathbf{n}})^{-1} U^T \quad (9)$$

Because of the approximation in Eq. 8, the estimator used in [1] was sub-optimal. Next, we present an optimal (in the sense of Eq. 6) estimator suited for colored noise.

Computer simulations indicated that the matrix  $U^T R_{\mathbf{n}} U$  in Eq. 7 was not diagonal, although in some cases it was nearly diagonal. This was not surprising, since  $U$ , being the eigenvector matrix of the symmetric matrix  $R_{\mathbf{x}}$ , diagonalizes  $R_{\mathbf{x}}$  and not  $R_{\mathbf{n}}$ . Rather than trying to approximate  $U^T R_{\mathbf{n}} U$ , we looked for a matrix that would simultaneously diagonalize  $R_{\mathbf{x}}$  and  $R_{\mathbf{n}}$ . It was shown in [4] that such a matrix exists and can simultaneously diagonalize the two matrices in the following way:

$$\begin{aligned} V^T R_{\mathbf{x}} V &= \Lambda \\ V^T R_{\mathbf{n}} V &= I \end{aligned} \quad (10)$$

where  $\Lambda$  and  $V$  are the eigenvalue matrix and eigenvector matrix respectively of  $\Sigma = R_{\mathbf{n}}^{-1} R_{\mathbf{x}}$ , i.e.,

$$\Sigma V = V \Lambda \quad (11)$$

Applying the above eigendecomposition of  $\Sigma$  to Eq. 7, and using Eq. 10, we can rewrite the optimal linear estimator as:

$$\begin{aligned} H_{opt} &= R_{\mathbf{n}} V \Lambda (\Lambda + \mu I)^{-1} V^T \\ &= V^{-T} \Lambda (\Lambda + \mu I)^{-1} V^T \end{aligned} \quad (12)$$

The enhanced signal  $\hat{\mathbf{x}}$  is obtained by applying the transform  $V^T$  to the noisy signal, appropriately modifying the components of  $V^T \mathbf{y}$  by a gain function, and by taking the inverse transform ( $V^{-T}$ ) of the modified components. Note that in our case,  $V^T \mathbf{y}$  is not the Karhunen-Loeve transform

(KLT) of  $\mathbf{y}$ . However, as we show below, if the noise is white,  $V^T \mathbf{y}$  becomes the KLT of  $\mathbf{y}$ .

Comparing the above estimator given in Eq. 12 with the corresponding linear estimator obtained for white noise in [3], we can see that both estimators have the same form. In fact, the Ephraim and Van Trees estimator [3] is a special case of the proposed estimator in Eq. 12. For white noise  $R_{\mathbf{n}} = \sigma_{\mathbf{n}}^2 I$ , and  $V$  becomes the unitary eigenvector matrix ( $U$ ) of  $R_{\mathbf{x}}$ , since  $\Sigma = \frac{1}{\sigma_{\mathbf{n}}^2} R_{\mathbf{x}}$ , and the diagonal matrix  $\Lambda$  becomes  $\frac{1}{\sigma_{\mathbf{n}}^2} \Delta_{\mathbf{x}}$ , where  $\Delta_{\mathbf{x}}$  is the diagonal eigenvalue matrix of  $R_{\mathbf{x}}$ . Therefore, for white noise Eq. 12 reduces to:

$$H_{opt} = U\Delta_{\mathbf{x}}(\Delta_{\mathbf{x}} + \mu\sigma_{\mathbf{n}}^2 I)^{-1} U^T \quad (13)$$

which is the Ephraim and Van Trees estimator [3]. The proposed approach is therefore the generalization of the subspace approach developed in [3] and can be used for both white and colored noise. The proposed approach makes no assumptions about the nature of the noise.

For the above proposed estimator, we need an estimate of the matrix  $\Sigma$ . Since we have no access to the covariance matrix of the clean speech signal, we can estimate  $\Sigma$  from the noisy speech signal as follows. Assuming that speech is uncorrelated with noise, we have:

$$R_{\mathbf{y}} = R_{\mathbf{x}} + R_{\mathbf{n}} \quad (14)$$

and so:

$$\Sigma = R_{\mathbf{n}}^{-1} R_{\mathbf{x}} = R_{\mathbf{n}}^{-1} (R_{\mathbf{y}} - R_{\mathbf{n}}) = R_{\mathbf{n}}^{-1} R_{\mathbf{y}} - I \quad (15)$$

## 2.2. Implementation

The proposed approach can be formulated in the following four steps. For each speech frame:

*Step 1:* Compute the covariance matrix  $R_{\mathbf{y}}$  of the noisy signal, and estimate the matrix  $\Sigma = R_{\mathbf{n}}^{-1} R_{\mathbf{x}}$ . The noise covariance matrix  $R_{\mathbf{n}}$  can be computed using noise samples collected during speech-absent frames.

*Step 2:* Perform the eigendecomposition of  $\Sigma$  :

$$\Sigma V = V \Lambda \quad (16)$$

*Step 3:* Assuming that the eigenvalues of  $\Sigma$  are ordered as  $\lambda_{\Sigma}(1) \geq \lambda_{\Sigma}(2) \geq \dots \geq \lambda_{\Sigma}(K)$ , estimate the dimension of the speech signal subspace as follows:

$$M = \arg \max_{1 \leq m \leq K} \{\lambda_{\Sigma}(m) > 0\} \quad (17)$$

The optimal linear estimator is computed as follows:

$$q_{\Sigma}(m) = \frac{\lambda_{\Sigma}(m)}{\lambda_{\Sigma}(m) + \mu} \quad 1 \leq m \leq M \quad (18)$$

$$Q_1 = \text{diag}\{q_{\Sigma}(1), \dots, q_{\Sigma}(M)\} \quad (19)$$

$$\begin{aligned}
H_{opt} &= V^{-T} \begin{bmatrix} Q_1 & 0 \\ 0 & 0 \end{bmatrix} V^T \\
&= R_n V \begin{bmatrix} Q_1 & 0 \\ 0 & 0 \end{bmatrix} V^T
\end{aligned} \quad (20)$$

*Step 4:* Estimate the enhanced speech signal by:  $\hat{\mathbf{x}} = H_{opt} \cdot \mathbf{y}$

The above estimator was applied to 4-ms duration frames of the noisy signal, which overlapped each other by 50%. A rectangular analysis window was used for the estimation of the covariance matrix  $R_y$ . The enhanced speech vectors were Hamming windowed and combined using the overlap and add approach. The parameter  $\mu$  was set to 5. No voice activity detection algorithm was used in our approach to update the noise covariance matrix. The noise covariance matrix was estimated using speech vectors from the initial silence frames of the sentences.

### 3. SUBSPACE APPROACH BASED ON FREQUENCY DOMAIN CONSTRAINTS

#### 3.1. Principles

The idea behind the frequency domain constrained linear optimal estimator is that the signal distortion can be minimized subject to constraints on the spectrum of the residual noise [3]. Specifically, suppose that the  $k$ -th spectral component of the residual noise is given by  $v_k^T \boldsymbol{\varepsilon}_n$ , where  $v_k$  is the  $k$ -th column vector of the eigenvector matrix of  $\Sigma = R_n^{-1} R_x$ . For  $k = 1, \dots, M$ , we require the energy in  $v_k^T \boldsymbol{\varepsilon}_n$  to be smaller than or equal to  $\alpha_k$ , where  $\alpha_k > 0$ , while for  $k = M+1, \dots, K$ , we require the energy in  $v_k^T \boldsymbol{\varepsilon}_n$  to be zero, since the signal energy in the noise subspace is zero. Therefore, the filter  $H$  is designed by [3]:

$$\begin{aligned}
&\min_H \overline{\boldsymbol{\varepsilon}_x^2} \\
\text{subject to: } &E\{|v_k^T \boldsymbol{\varepsilon}_n|^2\} \leq \alpha_k, k=1, \dots, M \\
&E\{|v_k^T \boldsymbol{\varepsilon}_n|^2\} = 0, k=M+1, \dots, K
\end{aligned} \quad (21)$$

The optimal estimator in the sense of Eq. 21 can be found using the method of Lagrange multipliers. It can be shown that the optimal  $H$  must satisfy the following matrix equation:

$$H R_x + L H R_n - R_x = 0 \quad (22)$$

where  $L = V \Lambda_\mu V^T$ , and  $\Lambda_\mu = \text{diag}(\mu_1, \dots, \mu_K)$  is a diagonal matrix of Lagrange multipliers. Using Eq. 10, we can rewrite Eq. 22 as:

$$V^T H V^{-T} \Lambda_x + V^T V \Lambda_\mu V^T H V^{-T} - \Lambda_x = 0 \quad (23)$$

which can be further reduced to the following equation:

$$Q \Lambda_x + V^T V \Lambda_\mu Q = \Lambda_x \quad (24)$$

where  $Q = V^T H V^{-T}$ .

The above equation is the well known Lyapunov equation encountered in control theory. Unfortunately, this equation does not have a closed-form solution, however, techniques developed in [5] can be used to obtain a numerical solution. After solving for  $Q$  in Eq. 24, we can compute the optimal  $H$  by:  $H_{opt} = V^{-T} Q V^T$ . Note that for white noise, Eq. 24 reduces to the same equation given in [3] (pp. 255) for the spectral domain estimator.

#### 3.2. Implementation

The proposed approach can be formulated in the following five steps. For each speech frame:

*Step 1:* Compute the covariance matrix  $R_y$  of the noisy signal, and estimate the matrix  $\Sigma = R_n^{-1} R_x$ .

*Step 2:* Perform the eigendecomposition of  $\Sigma$  :

$$\Sigma V = V \Lambda \quad (25)$$

*Step 3:* Assuming that the eigenvalues of  $\Sigma$  are ordered as  $\lambda_\Sigma(1) \geq \lambda_\Sigma(2) \geq \dots \geq \lambda_\Sigma(K)$ , estimate the dimension of the speech signal subspace as follows:

$$M = \arg \max_{1 \leq m \leq K} \{\lambda_\Sigma(m) > 0\} \quad (26)$$

*Step 4:* For a given  $\Lambda_\mu$  matrix, solve the Lyapunov equation in Eq. 24 to obtain  $Q$ , and from that obtain  $H_{opt}$  as:

$$H_{opt} = V^{-T} Q V^T$$

*Step 5:* Estimate the enhanced speech signal by:  $\hat{\mathbf{x}} = H_{opt} \cdot \mathbf{y}$ .

It is generally difficult to estimate the exact values of the Lagrange multiplier  $\mu_k$  due to the fact that the Lyapunov equation has no closed-form solution, however, for consistency with the time domain constrained approach, we set all the  $\mu_k$  values in  $\Lambda_\mu$  to 5.

### 4. EXPERIMENTAL RESULTS

For evaluation purposes, we used 20 sentences from the HINT (Hearing in Noise Test) database [6]. The HINT database is commonly used for speech intelligibility studies and it contains various lists of sentences, which were designed to be equally intelligible in noise. We randomly chose two of the sentence lists for testing, with each list containing ten sentences.

For colored noise, we used speech-shaped noise and multi-talker babble (seven talkers) added at an SNR of 0 and 5 dB. The speech-shaped noise (included in the HINT database) was computed by filtering white noise through an FIR filter with frequency response that matched the long-term spectrum of all the sentences in the database.

We used the Itakura-Saito (IS) distance measure [7] as the objective measure for evaluation of the proposed algorithms. The highest 5% of the IS distance values were discarded, as suggested in [8], to exclude unrealistically high spectral distance values. This method ensures a reasonable overall measure of performance. For comparative purposes, we also implemented and evaluated the approach in [1]. Table 1 presents the results for the two types of noise: speech shaped and multitalker babble in terms of the mean IS measure for 20 sentences.

Input SNR	0dB		5dB	
Noise type	S-S	M-B	S-S	M-B
Approach in [1]	9.28	2.84	5.87	1.78
TDC approach	1.31	1.48	0.93	1.07
SCD approach	1.87	1.70	1.35	1.25

**Table 1.** Comparative performance, in terms of mean Itakura-Saito distance measure for 20 sentences, obtained by the KLT approach in [1], and the proposed subspace methods based on time-domain constraints (TDC) and spectral domain constraints (SDC).

As can be seen from Table 1, our proposed approach outperformed Rezayee and Gazor’s approach [1] for both types of noise and for both SNRs. Informal listening tests also confirmed that the proposed subspace method preserved more speech information and had less speech distortion than the approach in [1].

Although we expected the spectral domain constrained (SDC) approach to have better speech quality than the time domain constrained (TDC) approach, we did not find that to be the case. We suspect that this is due to the difficulty in choosing the right  $\mu_k$  values for the matrix  $\Lambda_\mu$ . Overall, both TDC and SDC methods yielded comparable speech quality.

## 5. CONCLUSIONS

A new type of signal subspace approach for enhancing speech degraded by colored additive noise was proposed. The proposed approach outperformed the adaptive KLT approach proposed in [1] for colored noise. We showed that the linear estimator derived in this paper is a generalization of the Ephraim and Van Trees’ estimator [3]. Unlike the subspace method proposed by Ephraim and Van Trees for white noise, our method makes no assumptions about the nature of the noise (white or colored).

## 6. REFERENCES

- [1] A. Rezayee and S. Gazor, “An adaptive KLT approach for speech enhancement,” *IEEE Trans. Speech Audio Processing*, vol. 9, pp. 87–95, Feb. 2001.
- [2] U. Mittal and N. Phamdo, “Signal/noise KLT based approach for enhancing speech degraded by colored noise,” *IEEE Trans. Speech Audio Processing*, vol. 8, pp. 159–167, Mar. 2000.
- [3] Y. Ephraim and H. L. Van Trees, “A signal subspace approach for speech enhancement,” *IEEE Trans. Speech Audio Processing*, vol. 3, pp. 251–266, 1995.
- [4] S. B. Searle, *Matrix Algebra Useful for Statistics*, John Wiley & Sons, 1982.
- [5] R. H. Bartels and G. W. Stewart, “Solution of the matrix equation  $AX+XB=C$ ,” *Communication of the ACM*, vol. 15, no. 9, pp. 820–822, 1972.
- [6] M. Nilsson, S. Soli, and J. Sullivan, “Development of the Hearing in Noise Test for the measurement of speech reception thresholds in quiet and in noise,” *J. Acoust. Soc. Am.*, vol. 95, pp. 1085–1099, 1994.
- [7] Jr. J. R. Deller, J. Hansen, and J. G. Proakis, *Discrete-Time Processing of Speech Signals*, IEEE Press, NY, 2000.
- [8] J. Hansen and B. Pellom, “An effective quality evaluation protocol for speech enhancement algorithms,” in *Int. Conf. Spoken Language Processing*, Dec. 1998, pp. 2819–2822, Sydney, Australia.