AN ENERGY-CONSTRAINED SIGNAL SUBSPACE METHOD FOR SPEECH ENHANCEMENT AND RECOGNITION IN COLORED NOISE

Jun Huang *Yunxin Zhao*

Beckman Institute and Department of Electrical and Computer Engineering
University of Illinois at Urbana-Champaign, Urbana, Illinois 61801, USA
jhuang@ifp.uiuc.edu *yxz@ifp.uiuc.edu*

ABSTRACT

An energy-constrained signal subspace (ECSS) method is proposed for speech enhancement and recognition under an additive colored noise condition. The key idea is to match the short-time energy of the enhanced speech signal to the unbiased estimate of the short-time energy of the clean speech, which is proven very effective for improving the estimation of the noise-like, low-energy segments in speech signal. The colored noise is modelled by an autoregressive (AR) process. A modified covariance method is used to estimate the AR parameters of the colored noise and a prewhitening filter is constructed based on the estimated parameters. The performance of the proposed algorithm was evaluated using the TI46 digit database and the TIMIT continuous speech database. It was found that the ECSS method can significantly improve the signal-to-noise ratio (SNR) and word recognition accuracy (WRA) for isolated digits and continuous speech under various SNR conditions.

1. INTRODUCTION

It is well known that when speech signals are degraded by background noises, the performance of many voice communication and recognition systems become unacceptable. An important problem is to enhance speech degraded by colored noise which represent most acoustic ambient noise. In this paper, we introduce an energy constrained signal subspace (ECSS) method to enhance the speech signal contaminated by additive colored noise. The enhanced speech is passed into an existing speech recognition system which was trained in a noise-free setting to evaluate the quality of the enhanced speech signal. In this paper, the noisy signal is modelled as:

$$x(m) = s(m) + n(m) \tag{1}$$

where m is the discrete time index, $s(m)$ is the clean speech signal, $x(m)$ is the noisy speech signal, and $n(m)$ is the additive noise.

The signal subspace principle was proposed by Ephraim and Van Trees in 1995 [4]. The key idea is to decompose the vector space of the noisy signal into a signal-plus-noise subspace and a noise subspace under the assumption that the additive noise is white. The enhancement is performed by removing the noise subspace and estimating the clean speech from the remaining signal-plus-noise subspace. However, we observed that the signal subspace (SS) method didn't work well for the noise-like, low-energy speech units in continuous speech such as consonants. In this work, we introduce an energy constraint that uses the unbiased estimate of the short-time energy of clean speech to adjust the speech enhanced by the SS method. It was found that the energy-constrained signal subspace (ECSS) method is very effective for recovering the low-energy segments in continuous speech. As a result, the recognition accuracy on the enhanced speech was significantly improved. Furthermore, the colored noise is modelled by an autoregressive (AR) process. A modified covariance method is used to estimated the AR parameters of the colored noise process and a prewhitening filter is constructed based upon the estimated AR parameters. The noisy speech signal filtered by the prewhitening filter is then enhanced by the ECSS method and an inverse filter is used in the post-processing stage in order to remove the distortion to the speech signal caused by the prewhitening filter.

This paper is organized as follows. The design of the prewhitening filter is discussed in Section 2. Two energy-constrained signal subspace (ECSS) algorithmss for the colored noise case are proposed in Section 3. The experiment results are presented in Section 4 and a conclusion is given in Section 5.

2. PREWHITENING OF THE COLORED NOISE

For colored noise, the assumption that the correlation matrix of the noise being positive definite no longer holds. The approach taken in this work is to use a prewhitening filter to whiten the noise before enhancing the noisy speech signal and use an inverse filter after the signal subspace based processing to remove the effect of prewhitening filter on the clean speech signal.

Suppose that the discrete time signal $n(m)$ can be modeled by an AR process of order p,

$$n(m) = -\sum_{i=1}^{p} a(i)n(m-i) + v(m) \tag{2}$$

where $v(m)$ is a white Gaussian process with variance λ_v^2. The problem is to estimate the AR coefficients $a(1), \cdots, a(p)$ and the white Gaussian noise variance λ_v^2 given the colored noise observation data $n(1), \cdots n(L)$.

The resulted optimal AR coefficients $\hat{a}_1, \cdots, \hat{a}_p$ is given by [5]:

$$\begin{bmatrix} \hat{a}_1 \\ \hat{a}_2 \\ \cdots \\ \hat{a}_p \end{bmatrix} = -C_n^{-1} \begin{bmatrix} c_{nn}(1,0) \\ c_{nn}(2,0) \\ \cdots \\ c_{nn}(p,0) \end{bmatrix} \quad (3)$$

where:

$$\begin{aligned} c_{nn}(j,k) &= \frac{1}{2(L-p)}\Big(\sum_{m=p}^{L-1} x(m-j)x(m-k) \\ &+ \sum_{m=0}^{L-1-P} x(m+j)x(m+k)\Big) \\ & 0 \le j \le p, 0 \le k \le p \end{aligned} \quad (4)$$

$$C_n = \begin{bmatrix} c_{nn}(1,1) & c_{nn}(1,2) & \cdots & c_{nn}(1,p) \\ c_{nn}(2,1) & c_{nn}(2,2) & \cdots & c_{nn}(2,p) \\ \vdots & \vdots & \ddots & \vdots \\ c_{nn}(p,1) & c_{nn}(p,2) & \cdots & c_{nn}(p,p) \end{bmatrix} \quad (5)$$

The prewhitening filter $h(m)$ and its inverse filter $h^{-1}(m)$ are constructed based on the estimated AR parameters:

$$h(m) = \begin{cases} \hat{a}(m), & \text{if } 0 \le m \le p, \\ 0, & \text{otherwise} \end{cases} \quad (6)$$

where $\hat{a}(0) = 1$, and

$$h^{-1}(m) = -\sum_{i=1}^{p} \hat{a}(i)h(m-i) + \delta(m), \quad m = 0, 1, \cdots \quad (7)$$

3. THE ECSS ALGORITHM FOR COLORED NOISE

In the current work, two methods of using the energy constraint are proposed. For each frame t, the ECSS algorithms are formulated in seven steps. The first method, referred to as ECSS1 method, is based on the rescaling of the magnitude of the estimated signal in step 6 (setting $\alpha = 1$ in step 5). The second one, referred to as ECSS2 method, is based on the modification of the transformation matrix $H^{(t)}$ at step 5 where α in Eq. (17) is determined by a line search to satisfy the energy constraint of Eq. (21) with step 6 being skipped. The criterion for the estimation of the dimension of the signal subspace in step 4 was first introduced by Merhav in [6] and also used in [4].

Step 1 Prewhitening the colored noise.
Use the prewhitening filter in Eq. (6) to filter the noisy speech signal.

Step 2 Estimate the correlation matrices for the noisy and clean speech signal vectors.

$$\hat{r}_{xx}^{(t)}(k) = \frac{1}{2TK} \sum_{n=(t-T-1)K+1}^{(t+T-1)K-k} x(n)x(n+k), \quad 0 \le k \le K-1 \quad (8)$$

$$\hat{R_x}^{(t)} = \begin{bmatrix} \hat{r}_{xx}^{(t)}(0) & \hat{r}_{xx}^{(t)}(1) & ... & \hat{r}_{xx}^{(t)}(K-1) \\ \hat{r}_{xx}^{(t)}(1) & \hat{r}_{xx}^{(t)}(0) & ... & \hat{r}_{xx}^{(t)}(K-2) \\ ... & ... & ... & ... \\ \hat{r}_{xx}^{(t)}(K-1) & \hat{r}_{xx}^{(t)}(K-2) & ... & \hat{r}_{xx}^{(t)}(0) \end{bmatrix} \quad (9)$$

$$\hat{R_s}^{(t)} = \hat{R_x}^{(t)} - \hat{R}_n \quad (10)$$

where $\hat{R}_n$ is calculated using Eq. (8) and Eq. (9) during a non-speech period.

Step 3 Perform eigen decomposition for the estimated correlation matrices.

$$\hat{R_x}^{(t)} = \hat{U_x}^{(t)} \Lambda_x{}^{(t)} (\hat{U_x}^{(t)})^T \quad (11)$$

$$\hat{R_s}^{(t)} = \hat{U_s}^{(t)} \Lambda_s{}^{(t)} (\hat{U_s}^{(t)})^T \quad (12)$$

Step 4 Assuming that the eigenvalues of the estimated correlation matrix $\hat{R}_s^{(t)}$ are $\lambda_s^{(t)}(1) \ge \lambda_s^{(t)}(2) \ge \cdots \ge \lambda_s^{(t)}(N)$, estimate the dimension of the signal subspace.

$$\bar{M}^{(t)} = arg \max_{1 \le m \le K} \{\lambda_s^{(t)}(m) > 0\} \quad (13)$$

$$\hat{\sigma_n^2}(m) = \frac{1}{K} \|\hat{U}_{x2,K-m} \hat{U}_{x2,K-m}^T X^{(t)}\|^2 \quad (14)$$

$$\hat{M}^{(t)} = arg \ \min\{\frac{1}{2} log\hat{\sigma_n^2}(m) - \frac{1}{2} log\hat{\sigma_n^2}(\bar{M}^{(t)}) < \delta\} - 1 \quad (15)$$

where $\hat{\sigma_n^2}(m)$ represents the energy of the noisy signal in the noise subspace assuming that its dimension is $K - m$. $U_{x2,K-m}^{(t)}$ is a $K \times (K-m)$ matrix of the eigenvectors $\{u_{x,m+1}^{(t)}, \cdots, u_{x,K}^{(t)}\}$ of $R_x^{(t)}$.

Step 5 Compute the MMSE estimates.

$$\sigma_n^2 = \hat{r}_{nn}(0) \quad (16)$$

$$q_k^{(t)} = \frac{\lambda_s^{(t)}(k)}{\lambda_s^{(t)}(k) + \alpha\sigma_n^2} \quad 1 \le k \le \hat{M}^{(t)} \quad (17)$$

$$Q_1^{(t)} = diag\{q_1^{(t)}, \cdots, q_{M^{(t)}}^{(t)}\} \quad (18)$$

$$H^{(t)} = \hat{U}_{s1}^{(t)} Q_1^{(t)} (\hat{U}_{s1}^{(t)})^T \quad (19)$$

$$\hat{S}^{(t)} = H^{(t)} X^{(t)} \quad (20)$$

where $\hat{U}_s^{(t)} = [\hat{U}_{s1}^{(t)}, \ \hat{U}_{s2}^{(t)}]$ and $\hat{U}_{s1}^{(t)}$ is the principal eigenvector matrix of $\hat{R}_s^{(t)}$.

Step 6 Rescale the magnitude of the enhanced signal.

$$\widehat{\|S^{(t)}\|}^2 = max(\|X^{(t)}\|^2 - \sigma_n^2, 0) \quad (21)$$

$$r = max(1, \frac{\widehat{\|S^{(t)}\|}}{\|\hat{S}^{(t)}\|}) \quad (22)$$

$$\tilde{S}^{(t)} = r\hat{S}^{(t)} \quad (23)$$

Step 7 Use inverse filter to recover the clean speech signal.

Use the inverse filter in Eq. (7) to filter the enhanced speech signal in Eq. (23).

4. EXPERIMENT RESULTS

The test materials in our experiment consist of two sets of speech data: isolated digits and continuous speech. The clean speech signals were degraded by computer generated additive colored noise at different SNR levels. The SNR of the noisy signal is defined as:

$$SNR = 10\log_{10}(\frac{\sum_{k=1}^{L} s^2(k)}{\sum_{k=1}^{L}(x(k)-s(k))^2}) \qquad (24)$$

where L is the number of samples; $x(k)$, $s(k)$ are the k^{th} samples of the noisy and clean speech signals, respectively.

The proposed algorithms were tested in two ways. First, the SNR improvement was evaluated to quantify the overall quality of the enhanced speech signal. Second, the enhanced speech signals were recognized by existing speech recognition systems. The WRAs on the enhanced speech from two test data sets were evaluated and compared with those of the noisy speech signals under different SNR conditions.

4.1. Experiment on speech enhancement

4.1.1. Evaluation on isolated digits

In this experiment, we chose 10 isolated digits 0 through 9 from the TI46 database. Each digit was spoken by a male and a female speaker and was repeated twice. So there were totally 40 isolated digits in this test set. The speech data rate was down-sampled to 8 kHz. The colored noise was generated from an AR(2) model with parameters $a(1) = -0.45, a(2) = 0.55, \lambda_u = 1.0$. The colored noise data were added to the clean digits to generate the noisy digits under SNR conditions of 0 dB, 5 dB, 10 dB and 20 dB.

Table 1 shows the SNR improvement for the isolated digits under the condition of colored noise. It can be seen from this table that the ECSS1 and ECSS2 methods can improve the digit SNR from 2.2 dB to 7.6 dB for the colored noise case. The SNR improvements for the input SNR conditions of 0 dB, 5 dB, 10 dB and 20 dB are around 7 dB, 6 dB, 5 dB and 2 dB. This means that the lower the input SNR, the higher the SNR improvement. We can also see from Table 1 that the SNR improvement for the male speaker is slightly higher than that for the female speaker.

Table 1. SNR improvement for isolated digits

Input SNR	0 dB	5 dB	10 dB	20 dB
ECSS1 method				
Male speaker	7.6 dB	6.4 dB	5.2 dB	2.9 dB
Female speaker	7.5 dB	6.3 dB	5.1 dB	2.2 dB
ECSS2 method				
Male speaker	7.0 dB	5.8 dB	4.8 dB	2.9 dB
Female speaker	6.9 dB	5.8 dB	4.7 dB	2.3 dB

4.1.2. Evaluation on continuous speech sentences

The test data set in this experiment consisted of 16 continuous speech sentences from the TIMIT database. The colored noise was generated from the same AR(2) process as described in previous section. The colored noise data were added to the clean speech data to generate noisy speech data under the SNR conditions of 5 dB, 10 dB and 20 dB.

Tables 2 shows the SNR improvement using the ECSS1 and ECSS2 methods in the case of colored noise. It can be seen from this table that the ECSS methods can improve the SNR by approximately 6 dB, 5 dB and 2 dB under the input SNR conditions of 5 dB, 10 dB and 20 dB, respectively. It seems that the SNR improvement by ECSS methods are nearly the same for the isolate digits and continuous speech sentences.

Table 2. SNR improvement for continuous speech sentences

Input SNR	5 dB	10 dB	20 dB
ECSS1 Method	6.3 dB	4.8 dB	2.4 dB
ECSS2 Method	5.9 dB	4.5 dB	2.3 dB

4.2. Experiment on speech recognition

4.2.1. Evaluation on isolated digits

In this experiment, the digit utterances 0 through 9 spoken by a male and a female speaker were used. Five utterances of each digit were taken as training data and two utterance of each digit were used as test data. The endpoints of both the training and test data were hand-labelled by phone units. The analysis window size was 25 msec and the window shift was 12.5 msec.The mel frequency cepstral coefficients (MFCC) of order 16 and their temporal regression coefficients were used as the feature vector, where the regression were made over five adjacent frames (52.5 msec). The speech recognizer was a simple speaker-dependent Viterbi decoder based on the acoustic models of the digit-dependent phone units.

Table 3 shows the recognition accuracy of the noisy isolated digits and the digits enhanced by the ECSS methods under the SNR conditions of 0 dB, 5 dB 10 dB and 20 dB. It can be seen from Table 3 that the ECSS methods yield very high digit recognition accuracy for both male and female speakers under various input SNR conditions. The recognition accuracy for the enhanced digits is 90% for the male speaker and 95% for the female speaker under the input SNR condition of 0 dB and 100% for both speakers under the rest SNR conditions.

4.2.2. Evaluation on continuous speech sentences

In this part, the enhanced speech signals were recognized by a HMM-based speaker-independent continuous speech recognition (SICSR) system. The cepstrum coefficients of the PLP analysis (of order 8) and log energy were taken as instantaneous features and their first-order 50 msec temporal regression coefficients as dynamic features [3]. The task vocabulary size was 853 and the grammar perplexity was 64. The input SNR conditions were 5 dB, 10 dB and 20 dB. The number of test sentences was 16 and the WRA for the clean speech was 98.0%, where the WRA is defined by:

Table 3. Word recognition accuracy (WRA) for the noisy digits enhanced by the ECSS methods

Input SNR	0 dB	5 dB	10 dB	20 dB
Baseline				
Male speaker	40.0%	70.0%	90.0%	100.0%
Fmale speaker	75.0%	90.0%	95.0%	100.0%
ECSS1 method				
Male speaker	90.0%	100.0%	100.0%	100.0%
Female speaker	95.0%	100.0%	100.0%	100.0%
ECSS2 method				
Male speaker	90.0%	100.0%	100.0%	100.0%
Female speaker	95.0%	100.0%	100.0%	100.0%

$$WRA = (1 - \frac{\sharp\ of\ \ Sub + \sharp\ of\ \ Ins + \sharp\ of\ \ Del}{\sharp\ of\ \ reference\ \ words}) \times 100\% \quad (25)$$

where Sub means substitutions, Ins means insertions, Del means deletions and $\sharp$ denotes number.

The WRAs for the original noisy speech and the speech enhanced by the ECSS methods are listed in Table 4.

Table 4. Word recognition accuracy (WRA) for noisy continuous speech enhanced by the ECSS methods

Input SNR	5 dB	10 dB	20 dB
Baseline	-2.0%	33.3%	74.5%
WRA using ECSS1 method	53.9%	81.4%	93.1%
WRA using ECSS2 method	53.9%	80.4%	96.1%

It can be seen from Table 4 that the ECSS1 and ECSS2 methods improved the WRA by 55.9%, 47.1%–48.1% and 18.6%–21.6% under the SNR conditions of 5 dB, 10 dB and 20 dB. The WRA improvement by ECSS1 method is slightly higher than improvement by ECSS2 method under the SNR condition of 10 dB and the ECSS2 method perform slightly better than the ECSS1 method under the SNR condition of 20 dB.

5. CONCLUSION

In this paper, an energy-constrained signal subspace method is proposed for speech enhancement and recognition under colored noise conditions. The key idea is to match the short-time energy of the enhanced speech signals to the estimated short-time energy of the clean speech. The colored nosie was modelled by an AR process and a modified covariance method was used to estimate the AR parameters. A prewhitening filter and its inverse were constructed from the estimated AR parameters and were used before and after the ECSS speech enhancement system formulated for the white noise condition.

Both SNR and recognition accuracy improvements were evaluated to verify the effectiveness of the proposed methods. It was observed that the ECSS methods provided SNR gains around 6 dB, 5 dB and 2 dB for both isolated digits and continuous speech sentences under the input SNR conditions of 5 dB, 10 dB and 20 dB, respectively. The ECSS methods also led to very high accuracy for isolated digit recognition and provided significant WRA improvements for continuous speech under various input SNR conditions. Future work will be focused on further WRA improvement for continuous speech recognition under low SNR conditions.

ACKNOWLEDGMENT

This material is based upon work supported by the National Science Foundation under Grant No. IRI-95-02074, and by a grant from the Whitaker Foundation. The gift fund from AT&T/Lucent is also acknowledged.

REFERENCES

[1] S. F. Boll, "Suppression of acoustic noise in speech using spectral substraction", *IEEE Trans. Acoust. Speech Signal Process.*, vol. 27, pp. 113–120, 1979

[2] R. Cole et al. (1995), "The challenge of spoken language systems: Research directions for the nineties", *IEEE Trans. Speech and Audio Process.*, vol. 3, pp. 1–20, 1995.

[3] Y. Zhao, "A Speaker-independent continuous speech recognition system using continuous mixture Gaussian density HMM of phoneme-sized units", *IEEE Trans. on Speech and Audio Process.* vol. 1, pp. 345–361, 1993.

[4] Y. Ephraim and H. L. Van Trees, "A signal subspace approach for speech enhancement", *IEEE Trans. on Speech and Audio Process.*, vol. 3, pp. 251–266, 1995.

[5] S. M. Kay, *Modern spectral estimation: Theory and application* Englewood Cliffs: Prentice-Hall, 1988.

[6] N. Merhav, "The estimation of the model order in exponential families", *IEEE Trans. Inform. Theory*, vol. 35, pp. 1109–1114, 1985.

[7] E. Oja, *Subspace methods of pattern recognition* Letchworth: Research Studies Press, 1983.

[8] M. B. Priestley, *Spectral analysis and time series* Orlando: Academic Press, 1981.

[9] J. Huang and Y. Zhao, "Energy constrained signal subspace method for speech enhancement and recognition", *IEEE Signal Processing Letter*, vol. 10, pp. 283–285, 1997.

[10] L. Rabiner and B. H. Juang, *Fundamentals of speech recognition* Englewood Cliffs: Prentice-Hall, 1993.