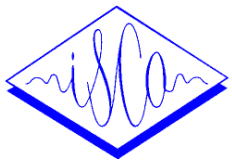


7th International Conference on Spoken Language Processing [ICSLP2002]
Denver, Colorado, USA
September 16-20, 2002

SUBSPACE SPEECH ENHANCEMENT USING SUBBAND WHITENING FILTER

Jong Uk Kim and Chang D. Yoo

Korea Advanced Institute of Science and Technology
Department of Electrical Engineering and Computer Science
373-1, Guseong-dong, Yuseong-gu, Daejon, Republic of Korea, 305-701
oribros@mail.kaist.ac.kr , cdyoo@ee.kaist.ac.kr

ABSTRACT

A novel subspace approach for speech enhancement using a subband whitening filter is proposed. Previous subspace approaches for enhancement either assumed white noise or used a fullband pre-whitening filter before enhancement for colored noise. The previous approaches were successful only in reducing the upper bound of the signal distortion while reducing noise. However, the proposed method minimizes the overall signal distortion while reducing noise. Experimental results show that the proposed method attains higher segmental signal-to-noise ratio (seg_SNR) than that attained by Ephraim et al. and also by the Wiener filter algorithm. In addition, the proposed algorithm requires less computational load than previous subspace approaches.

1. INTRODUCTION

Many speech enhancement algorithms including spectral subtraction [1] and Wiener filtering [2] function in the Fourier domain where the exponential family of the form $e^{-jn\omega}$ forms the bases. In a subspace approach suggested by Ephraim and Van Trees [3], the enhancement algorithm functions in Khrunen-Loeve transform (KLT) domain where the eigenvectors of the covariance matrix of a given signal forms the bases. These bases are obtained by solving

$$Ru = \lambda u, \tag{1}$$

where R, u, and λ are covariance matrix, the eigenvector and eigenvalue. The eigenvalues indicate the strength of the eigenvectors in describing the signal. The bases in the KLT domain are optimal in the terms of energy compaction, and for this reason, subspace enhancement methods have been known to offer results better than those obtained in the Fourier domain.

A number of enhancement algorithms based on subspace have been published; however, most of them either assume white noise in the formulation or use a whitening filter $R_w^{-1/2}$ to whiten the colored noise that does not minimize the signal distortion energy. In this paper, a subband whitening filter is used which minimizes the signal distortion by lowering the upper bound of the signal distortion.

This work was supported by grant No. R01-2000-00259 from the Korea Science & Engineering Foundation.

2. OUTLINE OF SUBSPACE APPROACH

Let the clean speech y is contaminated by additive white noise w to produce noisy speech vector z

$$z = y + w. \tag{2}$$

Clean speech y can be estimated using linear estimator H as $\hat{y} = Hz$. Then the residual error r is given by

$$\begin{aligned} r &= \hat{y} - y \\ &= (H - I)y + Hw \\ &\triangleq r_y + r_w, \end{aligned} \tag{3}$$

where $r_y = (H-I)y$ and $r_w = Hw$ represent signal distortion and residual noise respectively, and I denotes identity matrix. From (2), the covariance matrix of z is given by

$$R_z \triangleq E\{zz^T\} = R_y + R_w = R_y + \sigma^2 I, \tag{4}$$

where $(\cdot)^T$ denotes matrix transpose. Let $R_z = Q_z D_z Q_z^T$ and $R_y = Q_y D_y Q_y^T$ be the eigendecomposition of z and y respectively. Matrices Q_z, Q_y are orthogonal matrices whose column vectors are eigenvectors of R_z, R_y. Matrices D_z, D_y are diagonal matrices whose diagonal elements are eigenvalues of R_z, R_y. Then (4) can be rewritten as

$$\begin{aligned} R_z &= Q_z D_z Q_z^T \\ &= Q_y (D_y + \sigma^2 I) Q_y^T. \end{aligned} \tag{5}$$

Let $Q_z = Q_y \triangleq [Q, \bar{Q}]$, where

$$Q = [q_k : \lambda_z(k) > \sigma^2] \tag{6}$$

$$\bar{Q} = [q_k : \lambda_z(k) = \sigma^2], \tag{7}$$

and q_k is kth eigenvector of R_z associated with kth eigenvalue $\lambda_z(k)$. Denoting

$$Q = [q_1, q_2, \cdots, q_M], \tag{8}$$

the linear filter

$$H = QGQ^T \tag{9}$$

with $G = \text{diag}\Big\{\alpha_1^{1/2}, \alpha_2^{1/2}, \cdots, \alpha_M^{1/2}\Big\}$ is obtained by minimizing

$$\epsilon_y^2 = \text{tr}\Big[E\Big\{r_y r_y^T\Big\}\Big] \tag{10}$$

subject to

$$E\{|q_k^T r_w|^2\} \leq \alpha_k \sigma^2, \tag{11}$$

where diag$\{\cdot\}$ and tr$[\cdot]$ denote diagonal matrix composed of given elements and trace of a matrix respectively. From the minimization (10) and (11), α_k is given by generalized Wiener filter

$$\alpha_k = \exp\Big\{-\frac{\nu\sigma^2}{\lambda_z(k) - \sigma^2}\Big\}, k = 1, 2, \cdots, M \tag{12}$$

with $\nu \geq 1$ being an experimentally determined constant.

3. SUBBAND WHITENING

In comparison to spectral subtraction and Wiener filtering, the subspace approach reduces more noise while minimizing the distortion of speech; however, the algorithm requires the additive noise to be white. For colored noise, a whitening filter $R_w^{-1/2}$ can be applied to noisy speech z such that the filtered output of z is given by

$$\begin{aligned} \tilde{z} &= R_w^{-1/2}(y + w) \\ &= R_w^{-1/2} y + R_w^{-1/2} w \\ &= \tilde{y} + \tilde{w}, \end{aligned} \tag{13}$$

where $\tilde{y} = R^{-1/2} y$ and $\tilde{w} = R^{-1/2} w$. Define

$$\epsilon_{\tilde{y}}^2 = \text{tr}\Big[E\Big\{r_{\tilde{y}} r_{\tilde{y}}^T\Big\}\Big], \tag{14}$$

where $r_{\tilde{y}} = (\tilde{H} - I)\tilde{y}$, and $\tilde{H}$ is the linear estimator applied to (13). Then as pointed out in [4], minimizing $\epsilon_{\tilde{y}}^2$ does not minimize ϵ_y^2 but minimizes the upper bound of it. But if we decompose z into K bands, the upper bound of each band can be reduced, thus allowing ϵ_y^2 to be reduced.

3.1. Algorithm description

3.1.1. Subband whitening filter

When the additive noise w is white with variance σ^2, the covariance matrix R_w has the form $\sigma^2 I$, so the eigendecomposition of R_w is

$$R_w = Q(\sigma^2 I)Q^T, \tag{15}$$

where Q is arbitrary orthogonal matrix. But when the noise is colored, the eigendecomposition of R_w is given by the general form as

$$R_w = QDQ^T, \tag{16}$$

where

$$Q = \Big[q_1, q_2, \cdots, q_{(NK)}\Big] \tag{17}$$

$$D = \text{diag}\Big\{\sigma_1^2, \sigma_2^2, \cdots, \sigma_{(NK)}^2\Big\}, \tag{18}$$

and (NK) is the analysis frame size. The eigenvectors $\{q_i\}_{i=1}^{(NK)}$ and eigenvalues $\{\sigma_i^2\}_{i=1}^{(NK)}$ can be grouped into N subband matrices $\{Q_n\}_{n=1}^N$ and $\{D_n\}_{n=1}^N$ of dimension K such that

$$Q_n \triangleq \Big[q_{((n-1)K+1)}, \cdots, q_{(nK)}\Big] \tag{19}$$

$$D_n \triangleq \text{diag}\Big\{\sigma_{((n-1)K+1)}^2, \cdots, \sigma_{(nK)}^2\Big\}. \tag{20}$$

3.1.2. Signal distortion

The nth subband whitening filter is given by

$$F_n = D_n^{-1/2} Q_n^T. \tag{21}$$

Similar to (13), the output of F_n is given by

$$\tilde{z}_n = F_n y + F_n w = \tilde{y}_n + \tilde{w}_n, \tag{22}$$

where $\tilde{y}_n = F_n y$ and $\tilde{w}_n = F_n w$. Let $\|X\|_F$ represents the Frobenius norm of a matrix X. Using

$$\epsilon_y^2 = \Big\|QD^{1/2} E\Big\{r_{\tilde{y}} r_{\tilde{y}}^T\Big\}^{1/2}\Big\|_F^2, \tag{23}$$

and

$$\|XY\|_F^2 \leq \|X\|_F^2 \cdot \|Y\|_F^2, \tag{24}$$

the signal distortion for fullband signal is given by

$$\begin{aligned} \epsilon_y^2 &\leq \text{tr}[D] \cdot \epsilon_{\tilde{y}}^2 \\ &\leq \text{tr}[D] \cdot \text{tr}[(\tilde{H} - I)^2] \cdot \text{tr}[R_{\tilde{y}}] \\ &= \text{tr}[D] \cdot \text{tr}[(\tilde{G} - I)^2] \cdot \text{tr}[R_{\tilde{y}}] \\ &\triangleq \gamma \end{aligned} \tag{25}$$

where $\tilde{H} = \tilde{Q}\tilde{G}\tilde{Q}^T$ is a linear estimator for $\tilde{y}$ which has the same form as (9). From the relationship

$$\tilde{y} = \Big[\tilde{y}_1^T, \cdots, \tilde{y}_N^T\Big]^T, \tag{26}$$

the following is derived

$$\text{tr}[R_{\tilde{y}}] = \text{tr}[R_{\tilde{y}_1}] + \cdots + \text{tr}[R_{\tilde{y}_N}]. \tag{27}$$

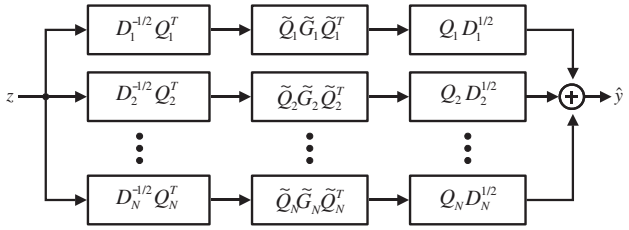


Fig. 1. Overall system of subspace speech enhancement using N subband whitening filter.

For nth subband signal, the signal distortion is given similarly by

$$\begin{aligned}\epsilon_{y_n}^2 &\leq \text{tr}[D_n]\cdot\text{tr}[(\tilde{G}_n - I)^2]\cdot\text{tr}[R_{\tilde{y}_n}] \\ &\triangleq \gamma_n \end{aligned} \tag{28}$$

where $\tilde{G}_n$ is defined in $\tilde{H}_n = \tilde{Q}_n\tilde{G}_n\tilde{Q}_n^T$ which is a linear estimator for nth subband signal $\tilde{y}_n$ as in (25).

From (25), (27) and (28), we note that

- $\sum_{n=1}^{N}\gamma_n < \gamma$: upper bound on the total signal distortion is smaller in subband than in fullband structure,
- $\epsilon_{y_n}^2 < \text{tr}[D_n]\cdot\text{tr}[(\tilde{G}_n - I)^2]\cdot\text{tr}[R_{\tilde{y}_n}]$: upper bound on each subband signal distortion is smaller in subband than in fullband structure.

3.1.3. Computational complexity

The computational load of the proposed subband whitening filter algorithm is reduced with increase in number of subbands. Referring to [5], computational complexities for eigendecomposition vary from $O\big((NK)^2\big)$ to $O\big((NK)^3\big)$, where (NK) is the analysis frame size. When we use N band structure, then the computational complexities vary from $O\big(NK^2\big)$ to $O\big(NK^3\big)$.

3.2. Overall system

Overall system of subspace speech enhancement algorithm using N subband whitening filter is shown in Figure 1. Noisy signal is first sent through a set of N whitening filter

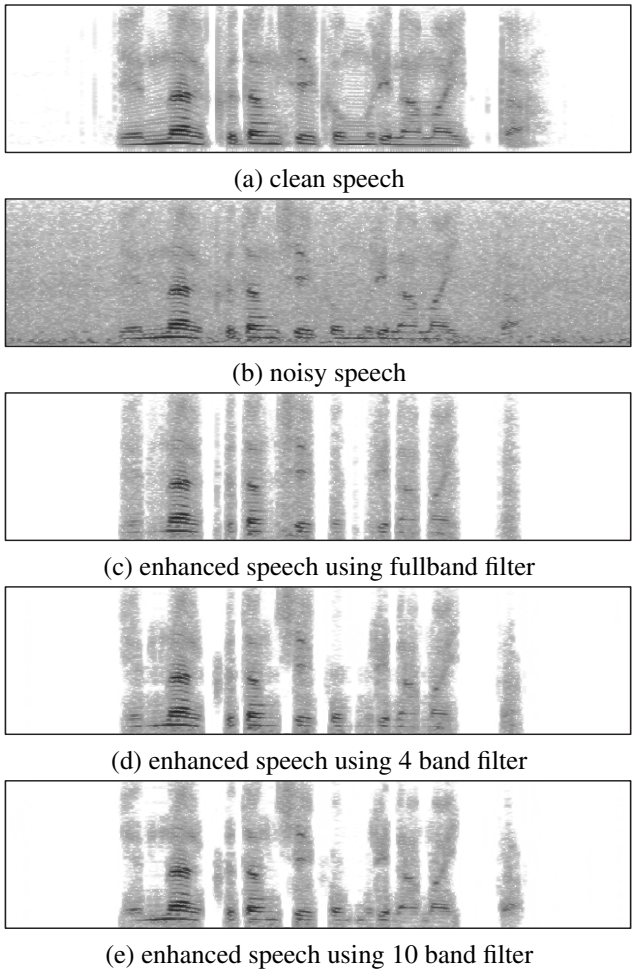
$$\{F_n = D_n^{-1/2}Q_n^T\}_{n=1}^N. \tag{29}$$

The output of each is enhanced by

$$\{\tilde{H}_n = \tilde{Q}_n\tilde{G}_n\tilde{Q}_n^T\}_{n=1}^N \tag{30}$$

to produce $\hat{\tilde{y}}_n$ as shown by

$$\hat{\tilde{y}}_n = \tilde{H}_n\tilde{z}_n. \tag{31}$$



(a) clean speech

(b) noisy speech

(c) enhanced speech using fullband filter

(d) enhanced speech using 4 band filter

(e) enhanced speech using 10 band filter

Fig. 2. Spectrograms of (a) clean speech, (b) noisy speech contaminated by stationary colored noise with SNR 5 dB , and enhancement results using (c) fullband whitening filter with $(NK) = 40$, (d) 4 band whitening filter with $(NK) = 40$ and (e) 10 band whitening filter with $(NK) = 40$.

Finally the nth subband estimate $\hat{\tilde{y}}_n$ is passed through an inverse filter $F_n^* = Q_nD_n^{1/2}$, then all filtered estimates are summed to get $\hat{y}$ as

$$\hat{y} = \sum_{n=1}^{N} F_n^*\hat{\tilde{y}}_n. \tag{32}$$

It should be noted that $F_n^*\hat{\tilde{y}}_n$ is an (NK)- dimensional vector.

4. EXPERIMENTAL RESULTS

The spectrograms of clean, noisy and enhanced speech are given in Figure 2. Clean speech of a Korean male speaker sampled at 8 KHz and pronounced as /*yet-naal goo-da:ng-si-eui guhn-mu-ri na-ma it-dda*/ was used. Noisy speech

is produced by adding clean speech with stationary noise with SNR 5 dB. (c) $\sim$ (e) is the enhanced speech using noise whitening filter in (c) fullband with $(NK) = 40, N = 1, K = 40$, (d) 4 bands with $(NK) = 40, N = 4, K = 10$ and (e) 10 bands with $(NK) = 40, N = 10, K = 4$. Note that (e) shows better performance than (c) and (d).

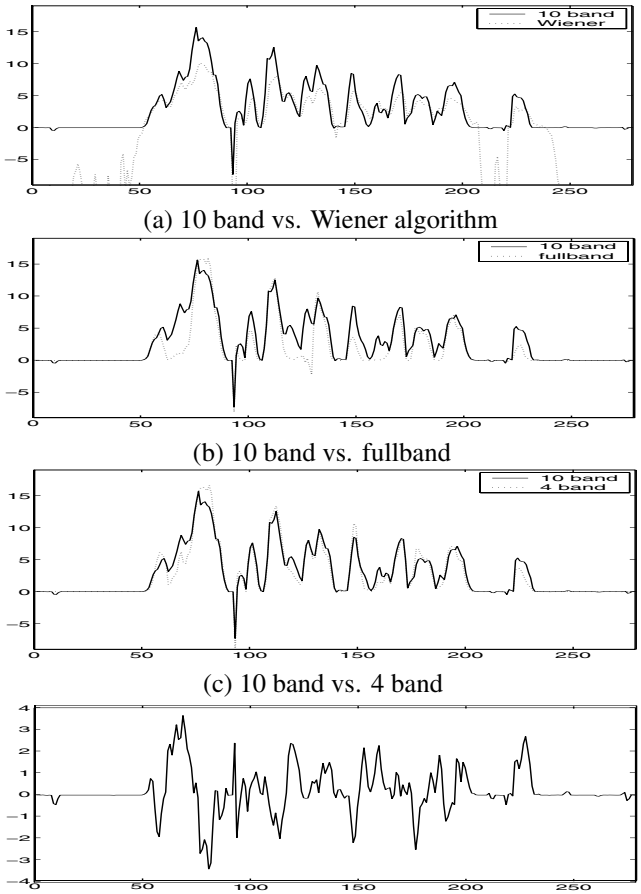
Output seg_SNRs as an objective quality measure on every 10ms frames are given in Figure 3. The ith output seg_SNR(i) is given by

$$\text{seg\_SNR}(i) \triangleq 10 \log_{10} \left(\frac{y_i^T y_i}{e_i^T e_i} \right), \tag{33}$$

where y_i, $\hat{y}_i$ are ith frames of clean speech and estimated speech respectively, and $e_i = y_i - \hat{y}_i$. In the figure, extremely low seg_SNR (< -10dB) regions are considered to be silence region. Ignoring this extremely low seg_SNR, we observe that the performance of 10 band structure is superior to 4 band, fullband and Wiener filtering algorithm. In the figure, (d) shows the SNR differences between 10 band and 4 band structures (seg_SNR of 10 band - seg_SNR of 4 band) with average differences of 0.18dB including non-speech region.

5. CONCLUSIONS

A subspace approach for speech enhancement using a subband whitening filter is proposed. While the previous approaches using a fullband whitening filter were only successful in reducing the upper bound of the signal distortion, the proposed method using subband whitening filter minimizes the overall signal distortion while reducing the noise. Experimental results show that the proposed method attains higher segmental signal-to-noise ratio (seg_SNR) than that obtained by fullband structure and by Wiener filtering algorithm. In addition to the advance in seg_SNR, the computational complexity is also reduced, because we reduce the dimension to be processed from (NK) to K. This means that eigendecomposition can be performed efficiently.



(a) 10 band vs. Wiener algorithm

(b) 10 band vs. fullband

(c) 10 band vs. 4 band

(d) SNR difference between 10 band and 4 band structure

Fig. 3. Output seg_SNRs for (a) 10 band structure vs. Wiener algorithm, (b) 10 band vs. fullband, and (c) 10 band vs. 4 band, and (d) the SNR difference between 10 band and 4 band structure.

6. REFERENCES

[1] Steven F. Boll, "Suppression of Acoustic Noise in Speech Using Spectral Subtraction," *IEEE Trans. Acoust., Speech, and Signal Proc.,* vol. ASSP-29, pp. 113-120, Apr. 1979.

[2] Jae S. Lim, Alan V. Oppenheim, "Enahncement and Bandwidth Compression of Noisy Speech," *Proc. IEEE,* vol. 67, No. 12, pp. 1586-1979, Dec. 1979.

[3] Yariv Ephraim, Harry L. Van Trees, "A Signal Subspace Approach for Speech Enhancement," *IEEE Trans. Speech, Audio Proc.,* vol. 3, No. 4, pp. 251-266, Jul. 1995.

[4] Udar Mittal, Nam Phamdo, "Signal/Noise KLT Based Approach for Enhancing Speech Degraded by Colored Noise," *IEEE Trans. Speech, Audio Proc.,* vol. 8, No. 2, pp. 159-167, Mar. 2000.

[5] Bin Yang, "Projection Approximation Subspace Tracking," *IEEE Trans., Signal Proc.,* vol. 43, No. 1, pp. 95-107, Jan. 1995.