A SPECTRALLY-BASED SIGNAL SUBSPACE APPROACH FOR SPEECH ENHANCEMENT

Yariv Ephraim and Harry L. Van Trees

Department of Electrical and Computer Engineering
Center of Excellence in C^3I
George Mason University
Fairfax, VA 22030.

ABSTRACT

The signal subspace approach for enhancing speech signals degraded by uncorrelated additive noise is studied. The underlying principle is to decompose the vector space of the noisy signal into a signal plus noise subspace and a noise subspace. Enhancement is performed by removing the noise subspace and estimating the clean signal from the remaining signal subspace. The decomposition can theoretically be performed by applying the Karhunen-Loève transformto the noisy signal. Linear estimation of the clean signal is performed using a perceptually meaningful estimation criterion. The estimator is designed by minimizing signal distortion for a fixed desired spectrum of the residual noise. This criterion enables masking of the residual noise by the speech signal. The filter is implemented as a gain function which modifies the KLT components corresponding to the signal subspace. The gain function is solely dependent on the desired spectrum of the residual noise. Listening tests indicate that 14 out of 16 listeners strongly preferred the proposed approach over the spectral subtraction approach.

1. INTRODUCTION

The signal subspace approach for enhancing speech signals degraded by uncorrelated additive noise was introduced in [1]-[2]. It utilizes the fact that speech vectors lie in a subspace of the Euclidean space of the noisy signal while vectors of white noise occupy the entire space. Thus, the vector space of the noisy signal can be decomposed into a signal plus noise subspace and a noise subspace. Speech enhancement can be performed by eliminating the noise subspace and extracting the signal from the less noisy signal subspace. Decomposition of the vector space of the noisy signal is performed by applying the Karhunen-Loève transform(KLT) to the noisy signal. The approach is applicable to non-white noise sources using prewhitening.

A linear estimator which minimizes signal distortion while maintaining the residual noise below some prescribed level was developed in [1]. The estimator was shown to be a Wiener filter with adjustable input noise level. The estimator was implemented as a gain function which modifies the KLT components corresponding to the signal subspace. It was shown in [2] that the proposed estimator, and the spectral subtraction estimator which uses the Wiener gain function with adjustable input noise level, are asymptotically equivalent (in probability) under stationary and ergodic conditions when the frame length goes to infinity. Thus, optimality conditions were attributed to the heuristic spectral subtraction approach.

In this paper a second more perceptually meaningful linear estimator is developed and tested in speech enhancement. This estimator aims at minimizing signal distortion while controlling the spectrum of the residual noise. This strategy allows shaping the spectrum of the residual noise so as to minimize its effects. The new estimator was found significantly better that the old estimator and the spectral subtraction approach. It provides enhanced signals with clarity similar to that obtained with the spectral subtraction approach but with drastically reduced musical residual noise. The new estimator was judged significantly better than the spectral subtraction approach by 14 out of 16 listeners for 98% of the sentences.

2. SIGNAL SUBSPACE PRINCIPLES

In this section the principles of the signal subspace approach are reviewed. We assume that covariance matrices of speech vectors are positive semidefinite since most speech covariance matrices have some eigenvalues which are close to zero. The covariance matrix of the noise is assumed known. Thus, without loss of generality, the noise is assumed white, since knowledge of its covariance enables whitening the noise.

Let y, w and z denote, respectively, $K-$dimensional vectors of the clean, noise and noisy signals. Since the noise is assumed additive and uncorrelated with the signal, the covariance matrix of z is given by

$$R_z \triangleq E\{zz^*\} = R_y + R_w. \tag{1}$$

Let $R_z = U\Lambda_z U^*$ be the eigendecomposition of R_z. Here $U \triangleq [u_1, \ldots, u_K]$ denotes an orthogonal matrix of eigenvectors $\{u_k \in R^K\}$ of R_z, and $\Lambda_z \triangleq \mathrm{diag}(\lambda_z(1), \ldots, \lambda_z(K))$ denotes a diagonal matrix of eigenvalues of R_z. Since the noise is assumed white, the eigenvectors of R_z are also the eigenvectors of both R_y and R_w. Furthermore, all eigenvalues of R_w equal σ_w^2. Since $\mathrm{rank}\,(R_y) = M$, the matrix R_y has M positive eigenvalues and $K-M$ zero eigenvalues. Assume without loss of generality that the M positive eigenvalues of R_y are $\{\lambda_y(1), \ldots, \lambda_y(M)\}$ and the corresponding M eigenvectors are $\{u_1, \ldots, u_M\}$. For convenience, assume that $\{\lambda_y(1), \ldots, \lambda_y(M)\}$ are given in a descending order. Multiplying (1) by u_k, which is the common kth eigenvector of the three covariance matrices, we obtain

$$\lambda_z(k) = \begin{cases} \lambda_y(k) + \sigma_w^2 & \text{if } k = 1, \ldots, M \\ \sigma_w^2 & \text{if } k = M+1, \ldots, K. \end{cases} \tag{2}$$

Thus, the eigendecomposition of R_z is given by

$$R_z = U\Lambda_z U^* \tag{3}$$

$$\Lambda_z = \mathrm{diag}[\Lambda_{z,1}, \sigma_w^2 I] \tag{4}$$

$$\Lambda_{z,1} \triangleq \mathrm{diag}(\lambda_z(1), \ldots, \lambda_z(M)), \tag{5}$$

and the eigendecomposition of R_y is given by

$$R_y = U\Lambda_y U^* \tag{6}$$

0-7803-2431-5/95 $4.00 © 1995 IEEE

$$\Lambda_y = \text{diag}[\Lambda_{y,1}, 0I] \quad (7)$$

$$\begin{aligned}\Lambda_{y,1} &= \text{diag}(\lambda_y(1), \ldots, \lambda_y(M)) \\ &= \Lambda_{x,1} - \sigma_w^2 I. \end{aligned} \quad (8)$$

The eigenvalues in $\Lambda_{x,1}$ and their corresponding eigenvectors are referred to as the principal eigenvalues and eigenvectors of R_x, respectively.

Let $U = [U_1, U_2]$ where U_1 denotes the $K \times M$ matrix of principal eigenvectors of R_x, i.e.,

$$U_1 = \{u_k : \lambda_x(k) > \sigma_w^2\}. \quad (9)$$

Since U is orthogonal,

$$I = U_1 U_1^\# + U_2 U_2^\#. \quad (10)$$

The matrix $U_1 U_1^\#$ is idempotent and Hermitian. Hence, it is the orthogonal projector onto the subspace spanned by the columns of U_1. But span U_1 = span V [3, p. 454]. Hence, $U_1 U_1^\#$ is the orthogonal projector onto the signal subspace. The complementary orthogonal subspace is spanned by columns of U_2 and it constitutes the noise subspace. The matrix $U_2 U_2^\#$ is the orthogonal projector on that subspace. Thus, from (10), a vector z of the noisy signal can be decomposed as

$$z = U_1 U_1^\# z + U_2 U_2^\# z, \quad (11)$$

where $U_1 U_1^\# z$ is the projection of z onto the signal subspace and $U_2 U_2^\# z$ is the projection of z onto the noise subspace. The coefficient vectors of the two projections, $U_1^\# z$ and $U_2^\# z$, respectively, are obtained from $U^\# z$ which is the KLT of z. Note that since

$$E\{U^\# z\} = 0$$

$$\text{cov}(U^\# z) = \text{diag}[\Lambda_{y,1} + \sigma_w^2 I, \sigma_w^2 I], \quad (12)$$

$\text{cov}(U_2^\# z) = \sigma_w^2 I$, and the signal energy in the vector $U_2^\# z$ is zero. Hence, with probability one (w.p.1), this vector does not contain signal information, and can be nulled when estimating the clean signal.

3. LINEAR SIGNAL ESTIMATORS

Let $\hat{y} = Hz$ be a linear estimator of y where H is a $K \times K$ matrix. The residual signal obtained in this estimation is given by

$$\begin{aligned} r &= \hat{y} - y \\ &= (H - I)y + Hw \\ &\triangleq r_y + r_w, \end{aligned} \quad (13)$$

where $r_y \triangleq (H-I)y$ represents signal distortion and $r_w \triangleq Hw$ represents the residual noise. Let

$$\overline{\epsilon_y^2} \triangleq \text{tr}E\{r_y r_y^\#\} = \text{tr}\{(H-I)R_y(H-I)^\#\} \quad (14)$$

be the energy of the signal distortion vector r_y. Similarly, let

$$\overline{\epsilon_w^2} \triangleq \text{tr}E\{r_w r_w^\#\} = \sigma_w^2 \text{tr}\{HH^\#\} \quad (15)$$

denote the energy of the residual noise vector r_w. The linear estimator derived in [1] imposes a time domain constraint (tdc) on the residual noise and it was obtained from

$$\begin{aligned} &\min_H \overline{\epsilon_y^2} \\ &\text{subject to : } \tfrac{1}{K}\overline{\epsilon_w^2} \le \alpha\sigma_w^2, \end{aligned} \quad (16)$$

where $0 \le \alpha \le 1$. The estimator derived in this way minimizes the signal distortion over all linear filters which result in the permissible residual noise level $\alpha\sigma_w^2$. The optimal estimator in this sense is given by

$$H_{tdc} = R_y(R_y + \mu\sigma_w^2 I)^{-1}, \quad (17)$$

where μ is the Lagrange multiplier obtained from

$$\alpha = \frac{1}{K}\text{tr}\{R_y^2(R_y + \mu\sigma_w^2 I)^{-2}\}. \quad (18)$$

Hence, the optimal filter (17) is a Wiener filter with adjustable input noise level $\mu\sigma_w^2$.

Applying the eigendecomposition (6) of R_y to (17), we can re-write the optimal linear estimator as:

$$H_{tdc} = U_1 G_\mu U_1^\# \quad (19)$$

where

$$G_\mu \triangleq \Lambda_{y,1}(\Lambda_{y,1} + \mu\sigma_w^2 I)^{-1}. \quad (20)$$

In this paper a linear estimator which minimizes the signal distortion subject to spectral domain constraints (sdc) on the residual noise is derived. The spectrum of the residual noise is made similar to that of the speech, and thus the residual noise can be masked by the speech signal. The kth spectral component of the residual noise is given by $u_k^\# r_w$. For $k = 1, \ldots M$, we require that the energy in $u_k^\# r_w$ be smaller than or equal to $\alpha_k\sigma_w^2$, where $0 < \alpha_k \le 1$. For $k = M+1, \ldots K$, we require that the energy in $u_k^\# r_w$ be zero, since the signal energy in the noise subspace is zero. Hence, the filter H is designed by

$$\begin{aligned} &\min_H \overline{\epsilon_y^2} \\ &\text{subject to : } \begin{array}{lll} E\{|u_k^\# r_w|^2\} \le \alpha_k\sigma_w^2, & k = 1, \cdots, M \\ E\{|u_k^\# r_w|^2\} = 0, & k = M+1, \cdots, K. \end{array} \end{aligned} \quad (21)$$

Following an optimization procedure similar to that used in the time domain constrained problem, while taking into account that the matrix H may now have complex entries, it can be shown that the optimal H must satisfy the following gradient matrix equation

$$HR_y + \sigma_w^2 LH - R_y = 0, \quad (22)$$

where $L \triangleq U\Lambda_\mu U^\#$ and $\Lambda_\mu = \text{diag}(\mu_1, \cdots, \mu_K)$ is a diagonal matrix of Lagrange multipliers. Applying the eigendecomposition (6) of R_y to (22) we obtain,

$$(I - Q)\Lambda_y - \sigma_w^2 \Lambda_\mu Q = 0, \quad (23)$$

where $Q \triangleq U^\# HU$. A possible solution to (23) is obtained when Q is diagonal with elements given by

$$q_{kk} = \begin{cases} \frac{\lambda_y(k)}{\lambda_y(k) + \sigma_w^2\mu_k} & k = 1, \ldots, M \\ 0 & k = M+1, \ldots, K. \end{cases} \quad (24)$$

For this Q we have

$$E\{|u_k^\# r_w|^2\} = \begin{cases} \sigma_w^2 q_{kk}^2 & k = 1, \ldots, M \\ 0 & k = M+1, \ldots, K. \end{cases} \quad (25)$$

If the non-zero constraints in (21) are satisfied with equality then $\sigma_w^2 q_{kk}^2 = \alpha_k \sigma_w^2$ implies that

$$q_{kk} = (\alpha_k)^{1/2}, \quad k = 1, \ldots, M \quad (26)$$

and

$$\mu_k = \frac{\lambda_y(k)}{\sigma_w^2}[(1/\alpha_k)^{1/2} - 1], \quad k = 1, \ldots, M. \tag{27}$$

Since $\mu_k \geq 0$, the Kuhn-Tucker necessary conditions for the constrained minimization are satisfied by the proposed solution (24). Hence from (24) and (26) we conclude that the desired H is given by

$$\begin{aligned} H &= UQU^{\#} \\ Q &= \text{diag}(q_{11}, \cdots, q_{KK}) \\ q_{kk} &= \begin{cases} \alpha_k^{1/2} & k = 1, \ldots, M \\ 0 & k = M+1, \ldots, K. \end{cases} \end{aligned} \tag{28}$$

From (28) we see that the choice of $\{\alpha_k\}$ completely specifies the gains of the estimator. This is not surprising since the estimator is linear and the spectra of its input and output signals are known. The input noise is white with spectrum σ_w^2, and the non-zero spectrum of the output residual noise is $\alpha_k \sigma_w^2$. In theory $\{\alpha_k\}$ can be chosen independently of the statistics of the signal and noise. In this case, the second order statistics of the signal and noise affect the estimator through the KLT only. This is a dual situation to that exists in the spectral subtraction approach where the second order statistics of the signal and noise affect only the gain function of the estimator while the transform (i.e., the DFT) is signal independent.

A possible choice for α_k is

$$\alpha_k = \left(\frac{\lambda_y(k)}{\lambda_y(k) + \mu \sigma_w^2} \right)^{\gamma} \tag{29}$$

where $\gamma \geq 1$ and $\mu \geq 0$ are experimentally determined constant. This choice makes the spectrum of the residual noise look similar to that of the clean signal. It is interesting to note that this gain function has been commonly used in the spectral subtraction approach [4].

An alternative choice for α_k which results in a more aggressive noise suppression gain function is given by

$$\alpha_k = \exp\{-\nu \sigma_w^2 / \lambda_y(k)\} \tag{30}$$

where $\nu \geq 1$ is an experimentally chosen constant. The value of ν controls the suppression level of the noise as well as the resulting signal distortion. The motivation for choosing this gain function is that for $\nu = 2$, the first order Taylor approximation of $\alpha_k^{-1/2}$ is precisely the inverse of the Wiener gain function (20 with $\mu = 1$). Hence, we refer to this function as generalized Wiener gain function.

4. IMPLEMENTATION AND EVALUATION

4.1. Implementation

The sdc estimator (28) is significantly easier to implement than the tdc estimator since the Lagrange multipliers were analytically calculated. This estimator is also significantly more powerful than the tdc estimator since it allows constraints in the perceptually significant spectral domain. In addition, the tcd can be considered a particular case of the sdc [2]. For the given choice of gain function (30), the estimator (28) depends only on one fixed parameter ν whose value was experimentally chosen to be $\nu = 5$.

In implementing the linear signal estimators one must have good estimates of R_{x_t} and R_{w_t}, the covariance matrices of the vectors of the noisy signal and the noise process at time t, respectively. Furthermore, a good estimate of the dimension M_t of the signal subspace at time t is required. Other parameters which must be specified are those characterizing the analysis conditions, i.e., the frame length K, the overlap duration between adjacent frames, the type of analysis and synthesis windows, and the number $2T$ of non-overlapping $K-$dimensional vectors of the noisy signal from which R_{x_t} is estimated.

The sdc linear estimator was applied to frames of the noisy signal which overlapped each other by 50%. In order to preserve the whiteness of the input noise (i.e., $R_{w_t} = \sigma_w^2 I$), only a rectangular analysis window could be used. The enhanced vectors were Hanning windowed and combined using the overlap and add synthesis approach. An empirical Toeplitz covariance estimate R_{x_t} was used. This estimate was constructed from the first K samples of the biased autocorrelation function estimate. The latter estimate was obtained from $2TK$ samples of the noisy signal at instants $(t-T-1)K+1, \ldots, (t+T-1)K$. This estimator was efficiently implemented using the FFT algorithm. The covariance matrix R_{w_t} is obtained from a set of vectors of the noisy signal during which speech was absent.

In implementing the empirical Toeplitz covariance estimator, the values of K and T must be chosen. If the speech signal were strictly stationary, T and K would have been chosen to be as large as possible. The reasons are that large KT guarantees good estimate of R_{x_t}. In addition, the improvement in SNR obtained by the signal subspace approach is proportional to K/M, since the signal subspace dimension M is fixed, and the noise is evenly distributed in the entire K-dimensional space. Since speech signals are not strictly stationary, however, the values of K and T are restricted by the following constraints. First, $2TK$, the total number of speech samples used in estimating the Toeplitz covariance matrix must be smaller than the period during which the signal can be considered stationary. Typically, this number equals $300 - 400$ samples at 8kHz sampling rate. Second, $K > M$ should be chosen so that the SNR improvement expected from the existence of the signal subspace can be utilized. Third, the larger T is the more accurate $\hat{R}_{x_t}$ is, since on the average there are $2T$ samples of the noisy signal per each estimated autocorrelation sample. Finally, K should be chosen small to reduce computational complexity in performing eigendecomposition of the estimated covariance.

In this work we have obtained best results using $T = 5$ and $K = 40$. This amounts to estimating the covariance matrix from 400 samples of the noisy signal. For this value of K, only few frames resulted in estimated M which was equal to K. This is not a problem since for those frames the signal occupies the entire space and the noise subspace is null.

Estimating M_t was performed using the order estimation approach of Merhav, Gutman and Ziv (see references in [2]). This approach guarantees minimization of the probability of underestimation of the order, uniformly for all processes in the given class, while maintaining exponentially decaying probability of overestimation of the order. The approach was applied by Merhav to exponential probability density functions (pdf's), and it requires maximum likelihood (ML) estimation of the parameter set of the model [5]. Hence, when used here, it is assumed that the pdf of the noise is exponential, e.g., Gaussian. The estimator for M obtained in this way chooses the smallest dimension for which the energy of the noisy signal in the noise subspace is sufficiently close to the minimum possible energy of the noisy signal in that subspace [2].

4.2. Performance Evaluation

The sdc linear estimator and the spectral subtraction estimator were tested and compared in enhancing speech signals which have been degraded by computer generated additive white Gaussian noise at 10dB input SNR. Speech material which consists of two sentences spoken by three male speakers and three female speakers (total of 12 sentences)

was used. One of the sentences, "why were you away a year Roy?" contains vowels and glides only, and the other "his vicious father had seizures" contains fricatives only.

The evaluation was performed by a group of 16 listeners. Four subjects were individuals working on different aspects of speech coding and enhancement. These subjects were familiar with the sentences. The other 12 subjects were students (6) and professors (6) at George Mason University. The authors were obviously excluded from this test. These 12 subjects were not familiar with the sentences. All subjects claimed to have normal hearing. Their age ranged from 23 to 40 year old.

Each subject participated in two listening sessions. The goal of the first session was to compare the signal subspace approach with the plain noisy speech. The goal of the second session was to compare the signal subspace approach with the spectral subtraction approach. In each session, 12 pairs of sentences, each representing two different processing methods, were presented to the subjects through headphones. The subjects were asked to compare the two sentences and to vote for one of them. The order of the sentences in each pair was randomized. The comparison was subjective based on the perceived amount and nature of residual noise, possible distortion and nature of the processed speech, etc. No listening fatigue effects were taking into account since each session was relatively short. In comparing the two sentences in a pair the subjects could listen to the pair as many times as they wished. To minimize any bias, the subjects were not informed which versions of the speech material they will be comparing.

In the first session, 14 subjects preferred the speech material enhanced by the signal subspace approach over the non-processed noisy speech. On the average, the subjects in this group voted in favor of the signal subspace approach for 84% of the sentences with standard deviation of 16%. The general consensus was that the quality of the enhanced signal is far better than that of the raw noisy signal due to the reduction in the level of the input noise. For those sentences where the enhanced signal was preferred, the benefit of noise reduction was worth the slight distortion introduced by the noise removal algorithm. For the other sentences where the noisy signal was preferred, the distortion in the enhanced signal was more noticeable and/or the perception of the noise was tolerable. The 2 subjects who preferred the noisy signals over the enhanced signals did so for 67% of the sentences on the average, with standard deviation of 12%. These individuals preferred the "natural" sound of the raw signal and they were not bothered by the presence of the noise.

In the second session, the same 14 individuals who preferred the signal subspace processing over the raw noisy signal, also preferred the signal subspace processing over the spectral subtraction approach. On the average, the signal subspace approach was preferred for 98% of the sentences with standard deviation of 4%. The major complaint in this comparison was the noticeable annoying musical residual noise in the spectral subtraction approach. Such noise was not present in the sentences processed by the signal subspace approach. Nine subjects felt that the two approaches contribute comparable amount of distortion to the speech signals during the noise removal process. The other 5 subjects felt that some sentences processed by the signal subspace approach are slightly more muffled than the sentences processed by the spectral subtraction approach. One subject also indicated that the tonal residual noise in the spectral subtraction approach may be especially destructive over the telephone, since it can easily be confused with tonal signals used in the network. The 2 subjects who preferred the spectral subtraction processing over the signal subspace approach did so for 83% of the sentences with standard deviation of 12%. These individuals were not bothered by the musical noise and felt that the spectral subtraction approach provides crisper enhanced signals.

The improvement in SNR achieved by the spectral subtraction and the signal subspace approach was similar. The spectral subtraction approached elevated the SNR of the noisy signal by $4.21 - 6.02$dB, and the signal subspace approach improved the SNR of the noisy signal by $4.22 - 5.74$dB.

5. COMMENTS

A novel approach for non-parametric speech enhancement was developed. The basic principle is to decompose the vector space of the noisy signal into a signal plus noise subspace and a noise subspace. Enhancement is performed by removing the noise subspace and estimating the signal from the remaining subspace. Linear estimation is performed with the goal of minimizing signal distortion while masking the residual noise by the signal.

The proposed approach was found useful in interpreting many aspects of the popular spectral subtraction approach: 1) Asymptotic optimality of a version of the spectral subtraction approach in the linear minimum mean square error sense was proven in [2]. 2) The existence of the signal subspace was used to explain why nulling of weak spectral components of the noisy signal is necessary. 3) The estimation criterion proposed here confirms the intuition that subtracting more noise than actually exists balances signal distortion and residual noise level as conjectured in the spectral subtraction literature. Thus this work provides a theoretical basis for the heuristically derived spectral subtraction approach.

The proposed signal subspace approach was judged better than the spectral subtraction in our speech enhancement application where the noise was additive and white. It provided enhanced signals with comparable distortion to that obtained in the spectral subtraction approach but with essentially no musical residual noise.

The major difference between the spectral subtraction approach and the signal subspace approach is in the transform used to decompose the vector space of the noisy signal. The theoretically optimal transform is the KLT. The spectral subtraction approach uses the DFT while the signal subspace approach was implemented using an empirical estimate of the KLT. Future work should focus on studying other signal independent transforms which can well approximate the KLT of the speech signal, e.g., the DCT and the discrete wavelet transform.

REFERENCES

[1] Y. Ephraim and H. L. Van Trees, *A signal subspace approach for speech enhancement*, *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, pp. II-355-II-358, April 1993.

[2] Y. Ephraim and H. L. Van Trees, *A signal subspace approach for speech enhancement*, to appear in *IEEE Trans. Speech and Audio Processing.*

[3] S. Haykin, *Adaptive Filter Theory*, Prentice-Hall, Englewood Cliffs, NJ, 1991.

[4] J. S. Lim and A. V. Oppenheim, "Enhancement and bandwidth compression of noisy speech," *Proc. IEEE*, vol. 67, no. 12, pp. 1586-1604, Dec. 1979.

[5] N. Merhav, "The estimation of the model order in exponential families," *IEEE Trans. Inform. Theory*, vol. 35, no. 5, pp. 1109-1114, Sept. 1985.