SPEECH ENHANCEMENT USING PCA AND VARIANCE OF THE RECONSTRUCTION ERROR IN DISTRIBUTED SPEECH RECOGNITION

Amin Haji Abolhassani[1], *Sid-Ahmed Selouani*[2], *Douglas O'Shaughnessy*[1]

[1]INRS-Energie-Matériaux-Télécommunications, Montréal, Canada
[2]Université de Moncton, Campus de Shippagan, Canada

ABSTRACT

We present in this paper a signal subspace-based approach for enhancing a noisy signal. This algorithm is based on a principal component analysis (PCA) in which the optimal subspace selection is provided by a variance of the reconstruction error (VRE) criterion. This choice overcomes many limitations encountered with other selection criteria, like overestimation of the signal subspace or the need for empirical parameters. We have also extended our subspace algorithm to take into account the case of colored and babble noise. The performance evaluation, which is made on the Aurora database, measures improvements in the distributed speech recognition of noisy signals corrupted by different types of additive noises. Our algorithm succeeds in improving the recognition of noisy speech in all noisy conditions.

Index Terms— Speech enhancement, speech recognition, model identification, signal subspace, principal component analysis, colored noise

1. INTRODUCTION

Among all classes of speech enhancement techniques, signal subspace filtering has gained a lot of attention. Using this approach, we obtain a nonparametric linear estimate of the unknown clean-speech signal, based on a decomposition of the observed noisy signal into mutually orthogonal signal and noise subspaces. This decomposition is possible under the assumption of a low-rank linear model for clean speech and an uncorrelated additive noise interference. Assuming these conditions, the energy of less correlated noise spreads over all dimensions of the observation space while the energy of the correlated speech components is concentrated in a subspace thereof. This so-called signal subspace can be recovered consistently from the noisy data. Generally speaking, enhancement is obtained by removing the noise subspace and optimally weighting the signal subspace to remove noise energies from this subspace.

In [1] we had reported a novel signal subspace-based model identification approach for single channel speech enhancement in noisy environments based on the Karhunen-Loève Transform (KLT), and implemented it via Principal Component Analysis (PCA) [2] [3]. The motivation to choose KLT is its optimality in compression of information, while the DFT and the DCT are suboptimal. The main problem in subspace approaches is the optimal choice of signal dimension. In [1] we introduced a novel approach for the optimal subspace partitioning using the Variance of the Reconstruction Error (VRE) criterion [4]. This criterion provides consistent parameter estimates and allows us to implement an automatic noise reduction algorithm that can be simply applied to the observed data. In this paper we apply this method to a distributed speech recognition task carried out on $TESTA$ of the Aurora database. This test set includes four types of noise (subway, babble, car and exhibition hall noise) artificially added to clean signals provided in the same set. In the end, we prove the method to be merited the best to ameliorate the quality of a noisy signal as well as the recognition accuracy.

The organization of the paper is given as follows. Section 2 of this paper describes the proposed subspace approach in enhancement of the noisy signal. Performance evaluation is made in section 3, and in section 4 the paper is concluded.

2. PROPOSED SUBSPACE APPROACH

In this section we first present fundamental relations and notations of PCA and then introduce the proposed method for model identification. The last part of this section explains how to reconstruct the clean signal from the observation, using the identified model.

2.1. Principal component analysis

We define a real-valued observation vector $x(t) \in \mathbb{R}^K$ to be the sum of the signal vector $s(t) \in \mathbb{R}^K$ and noise vector $n(t) \in \mathbb{R}^K$, i.e.,

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

where

$$x(t) = [x_1, x_2, \ldots, x_K]^T, \tag{2}$$

where K is chosen such that *Wide Sense Ergodicity* is satisfied, and $s(t)$ and $n(t)$ are defined similar to $x(t)$. We arrange

a K-dimensional observation vector in a $M \times N$ Hankelstructed (i.e., constant across the anti-diagonals) observation matrix $X_{M\times N}(t)$, where $K = M + N - 1$, i.e.,

$$X_{M\times N}(t) = \begin{pmatrix} x_1 & x_2 & \dots & x_N \\ x_2 & x_3 & \dots & x_{N+1} \\ \vdots & \vdots & \ddots & \vdots \\ x_M & x_{M+1} & \dots & x_K \end{pmatrix}. \quad (3)$$

The time-variable notation is from now on considered implicit and will therefore be left out in the remainder of the paper. From x we can calculate the covariance matrix R_{xx} which we define to be the expectation value of the outer product of the observation vector with itself, i.e.,

$$R_{xx} = E\{xx^T\}. \quad (4)$$

Due to the ergodicity assumption made in (2), we can estimate the covariance matrix R_{xx} using the zero-mean-scaled version of (3) as

$$\hat{R}_{xx} = \frac{1}{M-1} X^T X \in \mathbb{R}^{N\times N}. \quad (5)$$

The covariance matrix $\hat{R}_{xx} \in \mathbb{R}^{N\times N}$ can be examined by its eigenvalues and corresponding eigenvectors. Let $q_1, q_2, \dots, q_N$ be eigenvectors corresponding to the eigenvalues $\lambda_1, \lambda_2, \dots, \lambda_N$ of the covariance matrix $\hat{R}_{xx}$. We define the matrix Q as

$$Q = [q_1, q_2, \dots, q_N] \in \mathbb{R}^{N\times N} \quad (6)$$

where, due to the symmetry in $\hat{R}_{xx}$, the elements (eigenvectors) are orthonormal. If we arrange the eigenvalues in decreasing order in a diagonal matrix

$$\Lambda = diag(\lambda_1, \lambda_2, \dots, \lambda_N) \in \mathbb{R}^{N\times N}, \quad (7)$$

where

$$\lambda_1 \geq \lambda_2 \geq \dots \geq 0 \quad (8)$$

for positive-definite covariance matrices, we can decompose $\hat{R}_{xx}$ into its eigenvalue decomposition (EVD), i.e.,

$$\hat{R}_{xx} = Q\Lambda Q^T. \quad (9)$$

Since noise is assumed to be spread in the whole space, and noise and clean signal are assumed uncorrelated, the eigenvectors of $\hat{R}_{xx}$, $\hat{R}_{ss}$ and $\hat{R}_{nn}$ are the same and

$$\hat{R}_{xx} = \hat{R}_{ss} + \hat{R}_{nn}. \quad (10)$$

In all subspace signal enhancement algorithms it is assumed that every short-time speech vector $s = [s_1, s_2, \dots, s_N]$ can be written as a linear combination of $k < N$ linearly independent basic functions $m_i, i = 1, 2, \dots, k$ where

$$s = My, \quad (11)$$

where M is a $(N \times k)$ matrix containing the basis functions in columns and y is a $(k \times 1)$ weight vector. Since $rank(R_{ss}) = k$, there are k positive and $N - k$ zero eigenvalues in EVD of R_{ss}.

The speech enhancement procedure can now be summarized as follows:

- Separate the signal (signal + noise) subspace from the noise-only subspace.
- Remove the noise-only subspace.
- Remove the noise components in the signal subspace.

The first operation needs a prior knowledge of signal dimension to correctly define the signal subspace. Theoretically the signal dimension is defined by the order of the linear signal model in (11) where in practice due to the strong variation in speech contents (e.g., voiced versus unvoiced segments), the entire signal will never exactly obey one model. Numerous approaches for estimating the order of a model were reported in the literature, but in most of them the noise is assumed to be white noise and sometimes further assumptions are made. As an example a popular approach is the minimum description length (MDL) of Rissanen [5]. Schwarz [6] has shown that MDL is optimal in the minimum probability of error sense in detecting the order of the model, assuming that it has an underlying exponential (Koopman-Darmois) probability distribution.

2.2. Model identification using VRE

In [4] Valle et al. have estimated the Variance of the Reconstruction Error (VRE) to detect the faulty sensor. In this part we apply this technique to the speech enhancement domain and define the rank of the speech signal and enhance the observation signal by removing the remaining noise-only subspace. The minimum of the VRE consistently corresponds to the best reconstruction. When reconstruction of the noisy signal is based on the PCA model, the error is a function of the number of PCs and the minimum found in the VRE calculation directly determines the number of PCs. This is because the VRE is decomposed into the principal components subspace and a residual subspace. The portion in the principal components subspace has a tendency to increase with the number of PCs, and that in the residual subspace has a tendency to decrease, resulting in a minimum in VRE.

Imagine that our signal is corrupted with a noise along a direction ξ_i

$$x = s + n_i\xi_i \quad (12)$$

where s is the clean portion, n_i is the noise magnitude and $\xi_i \in \mathbb{R}^N$ where $\|\xi_i\| = 1$. The reconstruction of the signal is given by correction along the noise direction, that is,

$$\hat{s} = x - n_i\xi_i, \quad (13)$$

so that $\hat{s}$ is most consistent with the PCA model. The difference $s - \hat{s}$ is known as the *reconstruction error*. In [7] Qin and Dunia define the variance of the reconstruction error along each dimension as

$$u_i(l) \equiv var\{\xi_i^T(x - \hat{s})\} = \frac{\zeta_i^T(l)\hat{R}_{xx}\zeta_i(l)}{(\zeta_i^T(l)\zeta_i(l))^2} \qquad (14)$$

where

$$\zeta_i(l) = (I - \hat{Q}(l)\hat{Q}^T(l))\xi_i. \qquad (15)$$

In (14) and (15), l is an assumption for the rank of clean speech signal (k) and $\hat{Q}(l)$ is obtained from Q in (9) by keeping only the first l columns as the PCs. In order to find the number of PCs, we have to minimize $u_i(l)$ with respect to the number of PCs. Considering different noise directions, we propose the VRE to be minimized as

$$VRE(l) = \sum_{i=1}^{N} \frac{u_i(l)}{var\{\xi_i^T x\}} = \sum_{i=1}^{N} \frac{u_i(l)}{\xi_i^T \hat{R} \xi_i}. \qquad (16)$$

In this equation we calculate the VRE by summing $u_i(l)$ in all dimensions $(i = 1, 2, \ldots, N)$. In order to equalize the importance of each variable, variance-based weighting factors are applied.

In summary, in order to select the rank of the signal we have to go through the following steps:

- Build a PCA model for the original noisy data.
- Calculate the u_i and VRE using (14) and (16).
- The minimum VRE occurs in a specific number of PCs, which corresponds to the best reconstruction.

In a particular dimension ξ_i that is highly uncorrelated to the others, it is possible that $u_i(l) \geq var\{\xi_i^T x\}$, which means the model gives a worse prediction than the mean of the data (put $\hat{s} = 0$ in (14)). In this case we should drop the uncorrelated variable from the model.

2.3. Signal reconstruction

In order to reconstruct a signal from an observation, after identifying the speech-signal model, we should remove the noise-only subspace and modify the signal subspace to eliminate the effect of noise from this subspace.

Ephraim and Trees [8] developed two estimators of the clean signal using two perceptually meaningful estimation criteria. In the first estimator, signal distortion is minimized while the residual noise energy is maintained below some given threshold. This criterion results in a Wiener filter with adjustable input noise level. In the second one, signal distortion is minimized for a fixed spectrum of the residual noise. In this estimator the speech signal masks the residual noise and results in a filter whose structure is similar to that obtained in the first case, except that now the gain function which modifies the KLT coefficients is solely dependent on the desired spectrum of the residual noise. Generally the first estimator allows a time domain constraint (TDC) on the residual noise, while the second is designed for noise shaping using spectral domain constraints (SDC). In this paper we use a modified version of a TDC estimator.

The original TDC estimator in [8] is as

$$\hat{S} = X\hat{Q}G_\mu\hat{Q}^T \qquad (17)$$

where $\hat{Q}$ is the truncated Q. The truncation is made by cutting the last $N - l$ columns of Q (l is the rank of signal which minimizes (16)). As in [8] only white noise is considered, G_μ is a diagonal matrix containing l diagonal elements as

$$g_\mu(m) = \frac{\lambda_s(m)}{\lambda_s(m) + \mu\sigma_\omega^2}, \qquad (18)$$

where σ_ω^2 is the variance of the white noise and $\lambda_s(m)$ is the clean signal's variance in the m^{th} dimension. In our case as we are dealing with real world noise signals, we associate different variances of noise to each space dimension:

$$g_\mu(m) = \frac{\lambda_s(m)}{\lambda_s(m) + \mu\sigma_m^2}. \qquad (19)$$

In (18) and (19), μ is the Lagrange multiplier in [8].

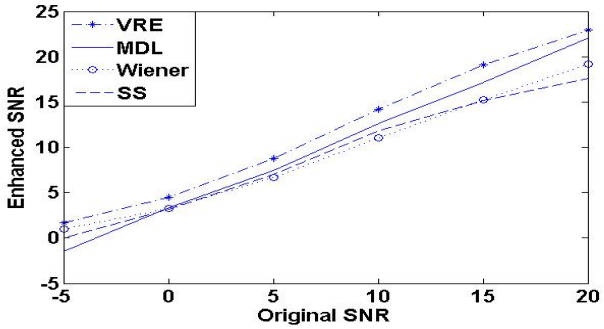
After estimating $\hat{S}$ using the modified TDC estimator, we can simply estimate the clean signal $(\hat{s})$ by averaging the anti-diagonal values of $\hat{S}$.

3. EXPERIMENTS

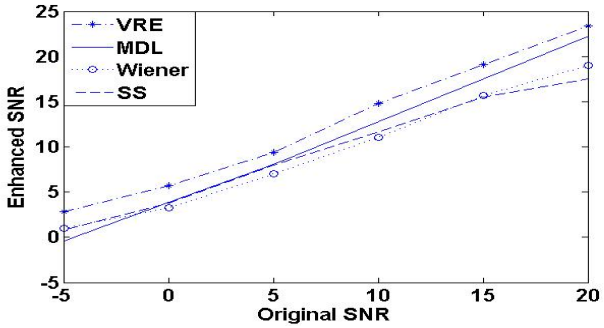
In order to perform the evaluation, we have chosen three enhancement methods, each from different categories of single channel enhancement algorithms. The methods to be compared with VRE are as follows:

- Minimum description length (MDL) [5]: A subspace approach using the KLT transform and MDL model identification.
- Wiener [9]: A well-known minimum mean-square error (MMSE) algorithm using mean-square error criterion to enhance a noisy signal in the discrete fourier transform (DFT) domain.
- Spectral subtraction (SS) [10] [11]: A Maximum likelihood (ML) approach using spectral subtraction to remove noise from the speech signal.

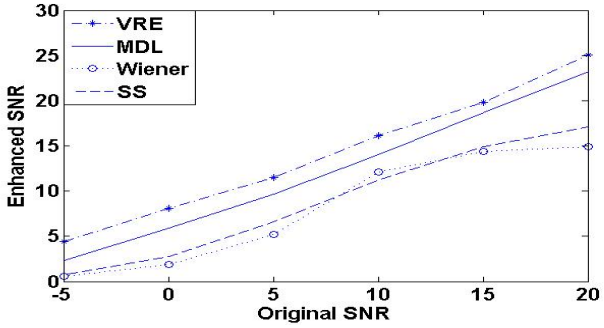
In this section we analyze the performance of our method comparing to other methods in terms of recognition rate and global signal-to-noise ratio (SNR) improvement as two main objective quality measurements. To evaluate and to compare the performance of the different estimators, we carried out



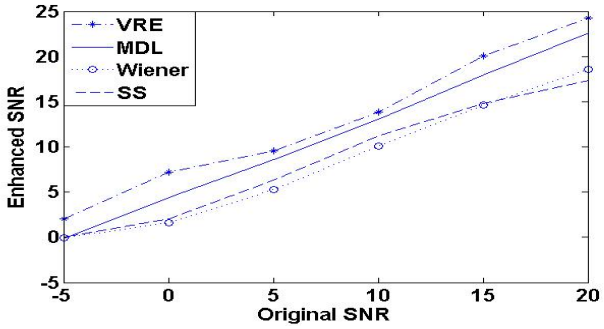
(a)



(b)



(c)



(d)

Fig. 1. SNR improvement using different enhancing methods in (a) N1: Subway noise, (b) N2: Babble noise (c) N3: Car noise (d) N4: Exhibition hall.

Table 1. Word accuracy (%) in different noisy conditions

(a) $N1$: Subway noise

SNR	Noisy	VRE	MDL	Wiener	SS
-5	13.39	27.01	25.07	21.05	19.89
0	27.68	40.45	38.64	35.33	32.69
5	52.95	63.42	59.68	55.72	56.45
10	77.97	82.30	82.21	81.24	80.32
15	93.11	95.20	94.11	93.23	93.02
20	97.09	97.95	96.96	95.62	95.59

(b) $N2$: Babble noise

SNR	Noisy	VRE	MDL	Wiener	SS
-5	5.12	21.35	18.18	16.97	14.38
0	12.20	24.95	20.24	18.74	16.98
5	26.89	37.40	35.34	30.49	30.02
10	48.96	55.65	54.53	51.86	51.32
15	74.02	76.75	75.96	72.50	72.43
20	88.97	90.92	90.54	77.83	75.23

(c) $N3$: Car noise

SNR	Noisy	VRE	MDL	Wiener	SS
-5	9.51	23.97	20.65	17.28	16.29
0	15.03	25.63	23.37	18.42	17.38
5	33.50	45.35	42.48	38.94	37.45
10	66.61	71.93	70.29	69.47	68.49
15	88.42	91.93	91.32	89.51	90.03
20	95.11	96.94	96.41	95.19	95.10

(d) $N4$: Exhibition hall noise

SNR	Noisy	VRE	MDL	Wiener	SS
-5	8.36	26.31	21.38	18.55	16.03
0	17.19	30.74	27.23	22.83	20.73
5	44.25	54.74	51.30	44.33	41.47
10	76.31	82.24	79.63	78.63	77.53
15	90.18	95.73	93.33	92.35	92.54
20	95.86	97.85	96.97	94.80	94.89

computer simulations with the $TESTA$ database of Aurora [12] ($f_s = 8$ kHz). These speech signals were corrupted with four types of noise at different global SNR levels. These types of noises are as follows:

- $N1$: Subway noise.
- $N2$: Babble noise.
- $N3$: Car noise.
- $N4$: Exhibition hall noise.

In the segmentation process, a frame length of 30 milliseconds, 40% overlap and a hamming window is applied. In VRE and MDL, $N = 21$, $\mu = 2$ and σ_m^2 is obtained by averaging noise power in each dimension over the initial non-speech segments. In Wiener, α (the smoothing factor for the

Decision-Directed method for estimation of A Priori SNR) equals 0.99 and the smoothing factor for the noise updating is 9. In SS, c (the scaling factor in silence periods) is set to 0.03.

In Fig. 1, SNR improvements using different methods as well as the SNRs of the original noisy signals are shown. Each figure represents different noisy conditions ($N1$ to $N4$). As we can see, in all situations VRE outperforms other methods.

Thus far we have proven VRE to be better than other methods regarding SNR improvement. Now we apply our enhancement algorithm as a preprocessing stage to distributed speech recognition in noise. We have used the speech recognizer provided in Aurora [12], which has been designed and trained on clean speech for digit recognition. The recognizer has been designed with the HTK HMM toolkit version 3.4. The features for speech recognition are the 12 MFCC and the energy, together with the first and second order derivatives of these 13 parameters which constitute a 39-dimensional vector. The general model for the isolated digit recognition consists of a model for silence between the digits (3 emitting states). The testing database contains $TESTA$ of Aurora database.

Table 1 gives the recognition results in terms of correctness for the compared algorithms. These results underline that our method allows an extraction of the relevant features of speech even in highly noisy conditions.

4. CONCLUSIONS

We have presented in this paper a promising enhancement method based on subspace approach for distributed speech recognition in noisy environments. This approach is based on PCA and an associated VRE subspace selection. The performance evaluation based on recognition accuracy shows clearly that our algorithm makes the front-end more robust to noise than other existing enhancement methods based on MMSE, ML or even PCA, even in the existence of colored and babble noise. A prominent point in our method is that it does not require any empirical parameter.

5. ACKNOWLEDGEMENTS

The authors would like to express their appreciation to Nuance Communications, Inc.

6. REFERENCES

[1] Amin Haji Abolhassani, Sid-Ahmed Selouani, Douglas O'Shaughnessy, and Mohamed-Faouzi Harkat, "Speech enhancement using pca and variance of the reconstruction error model identification," INTERSPEECH, 2007.

[2] S. Haykin, "Adaptive filter theory," Printice Hall, 1991.

[3] Kris Hermus, Patrick Wambacqa, and Hugo Van Hamme, "A review of signal subspace speech enhancement and its application to noise robust speech recognition," *Journal on Advances in Signal Processing*, vol. 2007, pp. 15, EURASIP, 2007.

[4] S. Valle, W. Li, and S.J. Qin, "Selection of the number of principal components: the variance of the reconstruction error criterion with a comparison to other methods," vol. 38, pp. 4389–4401, Industrial and Engineering Chemistry Research, 1999.

[5] J. Rissanen, "Modeling by shortest data description," vol. 14, pp. 465–471, Automatica, 1978.

[6] G. Schwarz, "Estimating the dimension of a model," vol. 6, pp. 461–464, Ann. Statist., 1978.

[7] S.J. Qin and R. Dunia, "Determining the number of principal components for best reconstruction," *Journal of Process Control*, vol. 10, pp. 245–250, 2000.

[8] Y. Ephraim and V. Trees, "A signal subspace approach for speech enhancement," *Transactions on Speech and Audio Processing*, vol. 3, No.4, IEEE, July 1995.

[9] R. Martin, "Speech enhancement using mmse short time spectral estimation with gamma distributed speech priors," IEEE ICASSP'02, May 2002.

[10] S. Boll, "Suppression of acoustic noise in speech using spectral subtraction," *IEEE Transactions on Acoustics Speech and Signal Processing*, vol. 27(2), pp. 113–120, ASSP, 1979.

[11] Douglas O'Shaughnessy, "Speech communications: Human and machine," Second Edition, IEEE, 2000.

[12] H. Hirsch and D. Pearce, "The aurora experimental framework for the performance evaluation of speech recognition systems under noisy conditions," ISCA ITRW ASR, September 2000.