16th European Signal Processing Conference (EUSIPCO 2008), Lausanne, Switzerland, August 25-29, 2008, copyright by EURASIP

SUBSPACE-BASED SPEECH ENHANCEMENT BY UPDATING NOISE CHARACTERISTICS IN THE PRESENCE OF SPEECH

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. In our previous works we have developed an algorithm based on principal component analysis (PCA) in which the optimal subspace selection is provided by a variance of the reconstruction error (VRE) criterion. In this work we will improve our previous technique by applying an updating noise variance algorithm. The performance evaluations show that our method provides a higher noise reduction and a lower signal distortion than our previous one.

1. INTRODUCTION

In all single channel speech enhancement methods, there has always been a challenge to define the statistical characteristics of non-stationary additive noise. Almost in all of these techniques, noise statistics are updated during the non-speech intervals [1]. This fact makes the process of updating noise features highly dependent on the accuracy of the existing Voice Activity Detector (VAD). Besides, in some utterances there is not any silence between the speech signals, which makes the initial statistics obtained from the beginning of the signals quite useless due to the non-stationarity of the additive noise.

In this paper we have exploited our previously developed subspace-based speech enhancement technique [2], towards an online, unsupervised algorithm that updates noise statistics during the enhancement process. Especially when we are to enhance a highly noisy corrupted speech signal, due to the difficulty of the task of distinguishing non-speech intervals from the speech-present ones, we can see the virtue of our new method over all other VAD-based enhancement methods. Moreover, the simplicity of this method introduces only a small computational complexity overhead.

In our method we apply a subspace model identification approach for single channel speech enhancement in noisy environments based on the Karhunen-Loève Transform (KLT), and implement it via Principal Component Analysis (PCA) [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 [2] we introduced therefore a novel approach for the optimal subspace partitioning using the Variance of the Reconstruction Error (VRE) criterion [4].

In this paper we improve our previous work by equipping it with a noise statistical updating algorithm provided here. In the end, we prove the new method to have a good performance in ameliorating the quality of a noisy signal especially in lower SNRs.

The organization of the paper is given as follows. Section two describes the basics of our subspace approach on which we will later base our novel noise parameter updating algorithm in section three. Performance evaluation is made in section four, and in section five the paper is concluded.

2. SUBSPACE-BASED SPEECH ENHANCEMENT

In this section we first present our recently developed VRE-based speech enhancement method and then improve it in the next section.

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

$$x(t) = [x_{t+0}, x_{t+1}, \ldots, x_{t+K-1}]^T, \tag{2}$$

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 Hankel-structed (i.e., constant across the anti-diagonals) observation matrix of arbitrary dimension $M \times N$ (i.e., $X_{M\times N}(t)$), where $K = M + N - 1$, i.e.,

$$X_{M\times N}(t) = \begin{pmatrix} x_{t+0} & x_{t+1} & \cdots & x_{t+N-1} \\ x_{t+1} & x_{t+2} & \cdots & x_{t+N} \\ \vdots & \vdots & \ddots & \vdots \\ x_{t+M-1} & x_{t+M} & \cdots & x_{t+K-1} \end{pmatrix}. \tag{3}$$

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.,

$$R_{xx} = E\{xx^T\}. \tag{4}$$

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

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

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

$$Q = [q_1, q_2, \ldots, q_N] \in \mathbb{R}^{N\times N}. \tag{6}$$

If we arrange the eigenvalues in decreasing order in a diagonal matrix,

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

where

$$\lambda_1 \geq \lambda_2 \geq \ldots \geq 0 \tag{8}$$

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

$$\hat{R}_{xx} = Q\Lambda Q^T. \tag{9}$$

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

$$s = My. \tag{10}$$

In this equation, 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 the EVD of R_{ss}.

In summary, in order to enhance a noisy signal we should first separate the signal (signal + noise) subspace from the noise-only subspace, then remove the noise-only subspace and finally remove the noise components in the signal subspace. The first operation needs a prior knowledge of the signal dimension to correctly define the signal subspace. In the following subsection we introduce our VRE-based method to tackle this problem.

2.2. Model identification using VRE

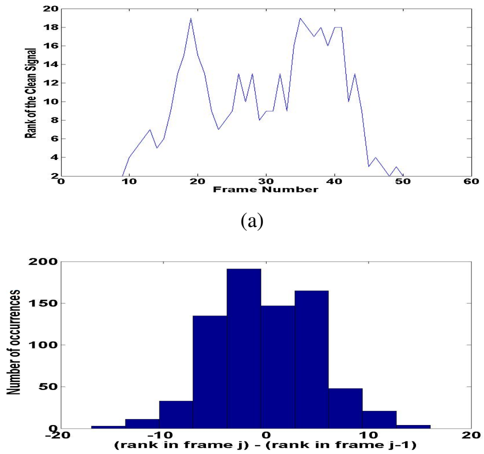
In [2] we have developed a VRE method to enhance the speech signal by defining the rank of the speech signal and removing the remaining noise-only subspace. The minimum of the VRE consistently corresponds to the best reconstruction. When reconstruction of the noisy signal is based on the PCA model, the 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,

$$x = s + n_i\xi_i, \tag{11}$$

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 a correction along the noise direction, that is,

$$\hat{s} = x - n_i\xi_i, \tag{12}$$



(a)

(b)

Fig. 1. (a) Rank of the clean signal in an utterance, decided by VRE. (b) Histogram of the inter-frame rank rising in 25 signals ($mean = 0$).

so that $\hat{s}$ is most consistent with the PCA model ($\hat{s}$ is the reconstructed signal obtained by de-noising the noisy signal). The difference $s - \hat{s}$ is known as the *reconstruction error*. In [4], Qin and Dunia define the variance of the reconstruction error along each dimension as

$$u_i(l) \equiv var\{\xi_i^T(x - \hat{s})\} = \frac{\zeta_i^T(l)\hat{R}_{xx}\zeta_i(l)}{(\zeta_i^T(l)\zeta_i(l))^2} \tag{13}$$

where

$$\zeta_i(l) = (I - \hat{Q}(l)\hat{Q}^T(l))\xi_i. \tag{14}$$

In (13) and (14), 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

$$VRE(l) = \sum_{i=1}^{N} \frac{u_i(l)}{var\{\xi_i^T x\}} = \sum_{i=1}^{N} \frac{u_i(l)}{\xi_i^T \hat{R} \xi_i}. \tag{15}$$

In this equation, in order to equalize the importance of each variable, variance-based weighting factors are applied.

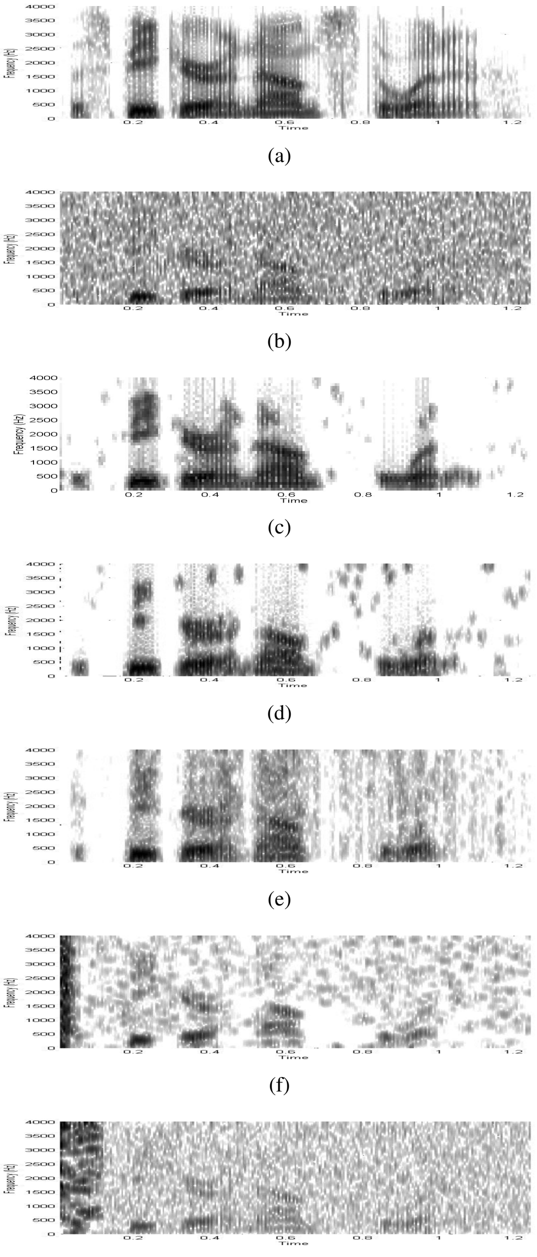
2.3. Signal reconstruction

In [2] we have developed a modified time domain constraint based technique (TDC) [5], in order to reconstruct the clean signal from an observation matrix X. The equation generated for reconstruction is:

$$\hat{S} = X\hat{Q}(k)G_\mu\hat{Q}^T(k), \tag{16}$$

where G_μ is a diagonal matrix containing l diagonal elements as

$$g_\mu(m) = \frac{\lambda_s(m)}{\lambda_s(m) + \mu\sigma_m^2}. \tag{17}$$



(a)

(b)

(c)

(d)

(e)

(f)

(g)

Fig. 2. The spectrograms of (a) the clean signal of utterance: "The speaker announced the winner", (b) noisy signal (SNR = -3 dB) and enhanced signal using (c) VRE-U, (d) VRE, (e) MDL, (f) Wiener and (g) SS.

In (17) σ_m^2 and $\lambda_s(m)$ are the variances of the colored noise and clean signal along the m^{th} dimension, respectively. Also in (17), μ is the Lagrange multiplier in [5].

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. UPDATING THE NOISE STATISTICS

In (17), in order to estimate the variance of the clean signal in each dimension ($\lambda_s(m)$), we assume that the clean signal and the noise signal have Gaussian distributions and are uncorrelated. Using Maximum Likelihood (ML) estimation, we obtain the variance of the clean signal by subtracting the variance of the noise from the variance of the noisy signal in each dimension, i.e.,

$$\lambda_s(m) = \lambda_m - \sigma_m^2 \tag{18}$$

Therefore, in (18), the role of the noise variance (σ_m^2) in each dimension of the eigenspace seems to be critical. However, defining the statistics of noise normally depends on detecting non-speech intervals in the signal, which seems to be a hard task in low SNRs.

In this part we suggest updating the noise variance from the noise subspace. The idea is that by minimizing (15) with regard to l, we can divide the eigenspace into signal-plus-noise and noise sub-areas. In the noise subspace, only noise components exist and as a result we can simply find its variance in that dimension. On the other hand the value of l changes frame by frame, meaning that, in an arbitrary frame j, one dimension can be a component of the noise subspace, while in frame $j+1$ that specific dimension can be in a signal-plus-noise sub-area. Considering this idea, we can update the noise statistics during the speech intervals by making use of the transformed statistics (along the new eigenvector, which could probably be only slightly changed) from the previous frame.

A note should be made here regarding the clean speech rank risings and fallings. Since normally the initial and final frames of a speech signal are silent, the rank of signal is the same in these intervals. Thus, the number of risings and fallings should be the same for all utterances leading to a useful update of the noise statistics (rising order), 50% of the times, when there is a change in the estimated order of the clean signal. In Fig. 1(a) the rank of the signal in each frame defined by the VRE algorithm is shown. Also in Fig. 1(b) we can see the histogram of difference between the rank of the clean signal in the j^{th} frame and the $(j-1)^{th}$ frame estimated using 25 signals. The mean of this histogram is zero, meaning that there is always the same number of risings and fallings.

Normally σ_m^2, which is obtained by averaging noise power in each dimension during the initial non-speech segments, is only updated in non-speech intervals decided by VAD. In our method though, for the noise updating VRE estimator, first we define the rank of signal using (15) and then update the variance of noise in each frame and in each dimension of noise subspace (defined by VRE) by a recursive relation with a forgetting factor α, i.e.,

$$\sigma_{m,new}^2 = \alpha\sigma_{m,old}^2 + (1-\alpha)\lambda_m cos^2(\beta), \tag{19}$$

where β is the angle between the m^{th} eigenvectors in two consecutive frames and λ_m can be obtained from (7). We also update our noise statistics in non-speech intervals by applying (19) in all dimensions.

This algorithm is especially useful when the SNR is low. In lower SNRs removing noise even in one dimension can

Table 1. SNRs obtained using different enhancement methods

(a) $N1$: Subway noise

SNR	SS	Wiener	MDL	VRE	VRE-U
-5	0.00	0.96	-1.44	1.02	**3.26**
0	3.17	3.28	3.31	3.93	**5.85**
5	7.01	6.68	7.49	8.12	**9.61**
10	11.81	11.02	12.65	13.12	**14.04**
15	15.20	15.22	17.20	17.62	**17.89**
20	17.62	19.16	22.05	22.30	**22.41**

(b) $N2$: Babble noise

SNR	SS	Wiener	MDL	VRE	VRE-U
-5	0.80	0.95	-0.40	1.17	**3.21**
0	3.82	3.24	3.85	4.55	**6.42**
5	7.95	7.03	8.06	8.67	**9.12**
10	11.70	11.04	12.80	13.37	**14.01**
15	15.50	15.66	17.55	17.99	**18.00**
20	17.55	19.04	22.24	**22.50**	22.48

(c) $N3$: Car noise

SNR	SS	Wiener	MDL	VRE	VRE-U
-5	0.73	0.53	2.32	3.04	**5.91**
0	2.74	1.89	5.92	6.89	**8.56**
5	6.61	5.24	9.66	9.92	**11.05**
10	11.24	12.12	14.04	14.21	**15.30**
15	14.90	14.42	18.67	18.73	**19.21**
20	17.11	18.29	23.24	23.31	**23.57**

(d) $N4$: Exhibition hall noise

SNR	SS	Wiener	MDL	VRE	VRE-U
-5	0.05	-0.06	-0.13	1.69	**3.98**
0	2.07	1.57	4.39	5.21	**7.29**
5	6.35	5.27	8.57	8.98	**10.11**
10	11.25	10.10	13.04	13.34	**14.87**
15	14.83	14.61	17.98	18.33	**19.05**
20	17.38	18.54	22.61	**22.82**	22.59

help us considerably in improving the quality of speech. Additionally, distinguishing between speech and non-speech intervals is difficult in low SNR signals.

4. EXPERIMENTS

In order to perform the evaluation, we compare five different enhancement methods, gathered from different categories of single channel enhancement algorithms. The methods to be compared are as follows:

- VRE-U: The method presented in this paper which benefits from a noise variance updating algorithm.
- VRE [2]: A subspace-based approach using VRE model selection criteria.
- Minimum description length (MDL) [6]: A subspace-based approach using the KLT transform and MDL model selection criteria.
- Wiener [7]: 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) [1], [8]: A Maximum likelihood (ML) approach using spectral subtraction to remove noise from the speech signal.

We should mention here that in order to make a fair comparison between these methods, we have used the same noise estimation algorithm for all of them which comprises an energy-based voice activity detector (VAD).

As a subjective test, spectrograms of signals are illustrated in Fig. 2. In these figures the spectrograms of the original clean and noisy signals as well as the output of different methods are illustrated. This illustration is carried out on the sentence “The speaker announced the winner” uttered by a male speaker and corrupted by white Gaussian noise at an input SNR = -3 dB.

As an objective validation of our algorithm we have also analyzed the performance in terms of global signal-to-noise ratio (SNR). To evaluate and to compare the performance of these techniques, we carried out the simulations with the $TESTA$ database of Aurora [9]. 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 are applied (40% shows the best result when changing the overlap from 0% to 70%). Moreover, we have chosen $N = 21$ and $\mu = 2$. In (19) we have heuristically chosen $\alpha = 0.9$ and since there exists an overlap of 40% between two adjacent frames, β approaches zero and as a result $cos^2(\beta)$ tends to one and can be omitted.

In Table 1, SNRs achieved by each method in different noisy conditions as well as the SNRs of the original noisy signals are shown. As we can see, both in Fig. 2 and Table 1, the noise updating VRE estimator (VRE-U) outperforms other estimators.

5. CONCLUSIONS

We have improved our previously generated PCA-based enhancement technique by applying a novel idea which involves updating the statistics of noise even during speech present intervals of a noisy signal. This approach is based on PCA, an associated VRE subspace selection and a newly developed technique based on a recursive noise variance updating algorithm. The performance evaluations based on spectrogram and SNR show clearly that our improved algorithm seems to be very promising in enhancing signals corrupted by stationary or nonstationary noises.

6. REFERENCES

[1] D. O’Shaughnessy, “Speech communications: Human and machine,” IEEE press, Piscataway, NJ, USA, 2nd edition, 2000.

[2] A. Haji Abolhassani, S.A. Selouani, D. O'Shaughnessy, and M.F. Harkat, "Speech enhancement using pca and variance of the reconstruction error model identification," *INTERSPEECH*, pp. 974 – 977, 2007.

[3] K. Hermus, P. Wambacq, and H.V. Hamme, "A review of signal subspace speech enhancement and its application to noise robust speech recognition," *EURASIP Journal on Advances in Signal Processing*, vol. 2007, pp. Article ID 45821, 15 pages, 2007.

[4] S.J. Qin and R. Dunia, "Determining the number of principal components for best reconstruction," *Journal of Process Control*, vol. 10, pp. 245–250(6), April 2000.

[5] Y. Ephraim and V. Trees, "A signal subspace approach for speech enhancement," *Transactions on Speech and Audio Processing*, vol. 3, No.4, pp. 251–266, IEEE, 1995.

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

[7] R. Martin, "Speech enhancement using mmse short time spectral estimation with gamma distributed speech priors," pp. 253–256, IEEE ICASSP'02, May 2002.

[8] 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.

[9] 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.