EXPERIMENTAL COMPARISON OF SIGNAL SUBSPACE BASED NOISE REDUCTION METHODS

Peter S. K. Hansen, Per Christian Hansen, Steffen Duus Hansen and John Aasted Sørensen

Department of Mathematical Modelling, Section for Digital Signal Processing
Technical University of Denmark, Building 321, DK-2800 Lyngby, Denmark
E-mail: {pskh, pch, sdh, jaas}@imm.dtu.dk, URL: http://www.imm.dtu.dk

ABSTRACT

In this paper, the signal subspace approach for non-parametric speech enhancement is considered. Several algorithms have been proposed in the literature but only partly analyzed. Here, the different algorithms are compared, and the emphasis is put onto the limiting factors and practical behavior of the estimators. Experimental results show that the signal subspace approach may lead to a significant enhancement of the signal to noise ratio of the output signal.

1. INTRODUCTION

In single-microphone speech enhancement techniques such as the signal subspace approach [1, 2, 3, 4], the noise is attenuated outside the band of perceptual importance. Thus, the remaining noise is nonstationary (musical noise), and as shown in the comparison, some algorithms have partly met this problem, using modified weighting rules.

Now, let $\mathbf{x} = (x_1, x_2, \cdots, x_m)^T$ denote the noisy signal vector and assume that the noise component n_k is additive and uncorrelated with the speech signal s_k. A set of time shifted vectors can be organized in a data matrix $\mathbf{X} = \mathbf{S} + \mathbf{N} \in \mathbb{R}^{m \times n}$ with Toeplitz structure ($m \geq n$), where we assume broad-banded noise so rank($\mathbf{N}$) $= n$ and that the speech signal can be described by a low order model, giving a numerically rank deficient matrix $\mathbf{S}$ with rank($\mathbf{S}$) $= p < n$. This observation can be used to estimate the clean signal from the noisy signal in a signal subspace of dimension p. Traditionally, the SVD is used in frame-based methods to decompose the vector space as [1, 2, 3, 4]

$$\mathbf{X} = \begin{pmatrix} \mathbf{U}_{X1} & \mathbf{U}_{X2} \end{pmatrix} \begin{pmatrix} \mathbf{\Sigma}_{X1} & \mathbf{0} \\ \mathbf{0} & \mathbf{\Sigma}_{X2} \end{pmatrix} \begin{pmatrix} \mathbf{V}_{X1}^T \\ \mathbf{V}_{X2}^T \end{pmatrix} \quad (1)$$

which in recursive methods, can be efficiently approximated by the rank-revealing ULV decomposition (RRULVD) [4]

$$\mathbf{X} = \begin{pmatrix} \mathbf{U}_{X1} & \mathbf{U}_{X2} \end{pmatrix} \begin{pmatrix} \mathbf{L}_{X1} & \mathbf{0} \\ \mathbf{F}_X & \mathbf{G}_X \end{pmatrix} \begin{pmatrix} \mathbf{V}_{X1}^T \\ \mathbf{V}_{X2}^T \end{pmatrix} \quad (2)$$

where the decompositions are partitioned according to the signal subspace dimension. In practice, a clean speech frame also results in a matrix that span the total space, however, the quality of the speech is mainly associated with the formants, which are represented by the pairs of singular values in the first part of the singular spectrum or in the RRULVD case, by the lower triangular matrix $\mathbf{L}_{X1}$. The matrix dimensions (m, n) must be chosen so the 4-5 most important formants can be separated, e.g., $n = 20$ and $p = 12$ (see the complete analysis in [4]).

2. EXPERIMENTAL FRAMEWORK

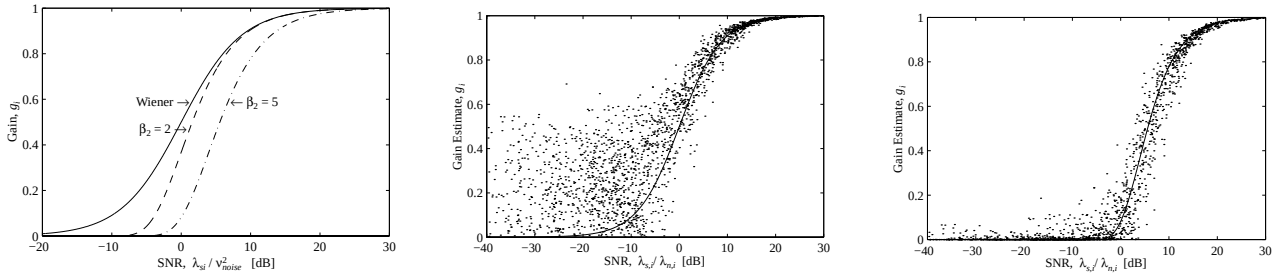
Non-parametric linear estimation of the clean signal using signal subspace methods can be represented by the general model

$$\hat{\mathbf{S}} = \mathbf{X}\mathbf{W} = \mathbf{X}\mathbf{V}_{X1}\mathbf{G}_1\mathbf{V}_{X1}^T \quad (3)$$

where the transformation $\mathbf{Y} = \mathbf{X}\mathbf{V}_X$ approximates the Karhunen-Loeve transform. Thus, the filter matrix $\mathbf{W}$ is applied to $\mathbf{X}$, and the enhanced vectors are combined using the overlap and add synthesis approach (averaging along the diagonals). The gain matrix $\mathbf{G}_1$ depends on the estimation method as shown in Table 1 for the Least Squares (LS), Minimum Variance (MV) [5], Time Domain Constrained (TDC) [2] and Spectral Domain Constrained (SDC) [2] case, where the last three are based on a white noise assumption, i.e., $\mathbf{N}^T\mathbf{N} = \sigma_{noise}^2\mathbf{I}_n$. Note, that the LS estimator results in the lowest possible signal distortion and in the highest possible residual noise level $(p/n)\sigma_{noise}^2$, while the MV estimator (Wiener gain function) is the optimal linear estimator, which gives the minimum total residual power. The TDC estimator keeps the residual noise power below some threshold while minimizing the signal distortion. Thus, this estimation criterion will control the musical noise component. The SDC estimator is a generalization of the TDC estimator which keep the residual noise power in each spectral component below some threshold α_i.

The gain g_i as function of the spectral SNR, i.e., $\sigma_{s,i}^2/\sigma_{noise}^2$, can be used to characterize the different estimation methods as shown in Fig. 1(a). Fig. 1(b) shows estimated Wiener gains obtained from a noisy sentence, where a large variance in the estimated gains is observed for small SNRs, which illustrates the importance of explicitly introducing a signal subspace. Thus, the performance depends on the estimation of σ_{noise}, i.e., the TDC and SDC based gain functions can be expected to perform better as illustrated in Fig. 1(c), since they are less sensitive to estimation errors.

The quality of the linear estimators are characterized by the residual matrix $\mathbf{R} = \mathbf{S}(\mathbf{W} - \mathbf{I}_n) + \mathbf{N}\mathbf{W} = \mathbf{R}_S + \mathbf{R}_N$, i.e., signal distortion denoted by the matrix $\mathbf{R}_S$ and residual noise denoted by the matrix $\mathbf{R}_N$, as shown in Fig. 2(a) for the noisy voiced speech frame in Fig. 2(c). The minimum residual power is obtained for the MV estimator ($\beta_2 \approx 2$) and is dominated by the residual noise, however, by choosing the perceptually more meaningful parameter $\beta_2 = 5$, the signal distortion will become dominant for the price of a slightly increase in the level of the total residual signal. Thus, in this case less noise accompany the low energy spectral components of the speech in accordance to the masking threshold of the auditory system (see Fig. 2(b)), i.e., both the quality and intelligibility of the noisy signal can be improved.



(a) (b) (c)

Figure 1 (a) Wiener and SDC gain functions for different choices of β_2. (b) Estimated Wiener gains $\{g_i\}_{i=1}^{12}$ of 165 speech frames ($\mathbf{X} \in \mathbb{R}^{141\times 20}$) obtained from a noisy speech sentence (white noise and SNR=10dB). (c) For the SDC estimator ($\beta_2 = 5$).

Method	SVD based gain matrix $\mathbf{G}_1$	RRULVD based gain matrix $\mathbf{G}_1$
LS	$\mathbf{I}_p$	$\mathbf{I}_p$
MV	$(\mathbf{I}_p - \sigma^2_{noise}\mathbf{\Sigma}^{-2}_{X1})$	$\mathbf{L}^{-1}_{X1}(\mathbf{L}_{X1} - \sigma^2_{noise}\mathbf{L}^{-T}_{X1})$
TDC	$(\mathbf{I}_p - \sigma^2_{noise}\mathbf{\Sigma}^{-2}_{X1})(\mathbf{I}_p - \sigma^2_{noise}(1-\gamma)\mathbf{\Sigma}^{-2}_{X1})^{-1}$	$(\mathbf{L}_{X1} - \sigma^2_{noise}(1-\gamma)\mathbf{L}^{-T}_{X1})^{-1}(\mathbf{L}_{X1} - \sigma^2_{noise}\mathbf{L}^{-T}_{X1})$
SDC	$\text{diag}(\sqrt{\alpha_1}, \ldots, \sqrt{\alpha_p}), \quad \alpha_i = \exp\left(\frac{-\beta_2\sigma^2_{noise}}{\sigma_{x,i} - \sigma^2_{noise}}\right)$	not available

Table 1 Gain matrix $\mathbf{G}_1 \in \mathbb{R}^{p\times p}$ for different estimators [4]. The TDC and SDC parameters must satisfy $\gamma \geq 0$ and $\beta_2 \geq 1$.

The consequence of prewhitening in signal subspace methods is illustrated in Fig. 2(c). Obviously, the magnitude of the prewhitened speech frame are rescaled in accordance with the noise spectrum, so in general, the effect of prewhitening is a (maybe large) bias of the signal subspace. This is a major limitation of subspace methods which is often overlooked.

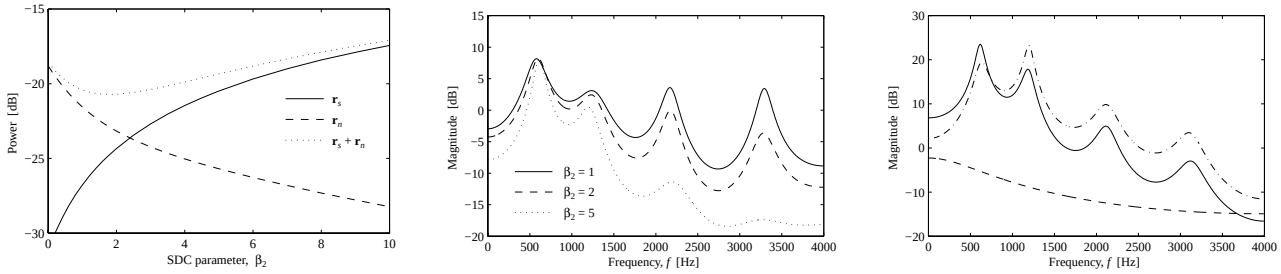
3. SIMULATIONS AND RESULTS

Comparisons of signal subspace based estimators are made on the basis of the improvement in segment SNR, tracking capability, and informal listening tests. The experiments illustrate the differences in speech enhancement that may arise from the use of different estimation strategies, decomposition methods (SVD or RRULVD) and window types (sliding or exponential). Also the effect of prewhitening is evaluated. The experiments have been performed by using a rectangular analysis window consisting of 160 samples, i.e., with data matrix dimensions $(m, n) = (141 \times 20)$, and by using a fixed signal subspace dimension $p = 12$. The noise matrix $\mathbf{N}$ (or σ_{noise}) is obtained from an initial noise-only segment.

A speech sentence contaminated by white noise is shown in Fig. 3(a), and Fig. 3(b) – 3(c) show the enhanced speech signals obtained by the MV and SDC estimator, respectively. From Fig. 4(a), it is seen that the segmental SNRs for the SDC estimator have been improved in most cases. Only frames with high SNR will not be enhanced due to the signal distortion obtained by introducing a signal subspace. Note also that the variations among the segmental SNRs are reduced, and that the SNRs of the enhanced signal are mainly above 0 dB. The latter observation rely on the actual gain function, which sets spectral components below 0 dB to zero (see Fig. 1(c)). Fig. 4(b) illustrates the improvements in segmental SNRs for the enhanced waveforms. The LS estimator gives a nearly constant improvement n/p as expected from the p-dimensional signal subspace, while the two other methods perform considerably better. At low SNRs, the improvements obtained by the SDC estimator are significantly larger than the ones obtained by the MV estimator, which can be explained by the practical behavior of the estimators (see Fig. 1). Fig. 5 shows the input-output relations of segmental SNRs for the MV and SDC estimators. Clearly, the improvement in output SNR increases for decreasing input SNR, and no improvement can be expected in frames with SNR close to 20 dB. Note again the 0 dB limit for the SDC estimator.

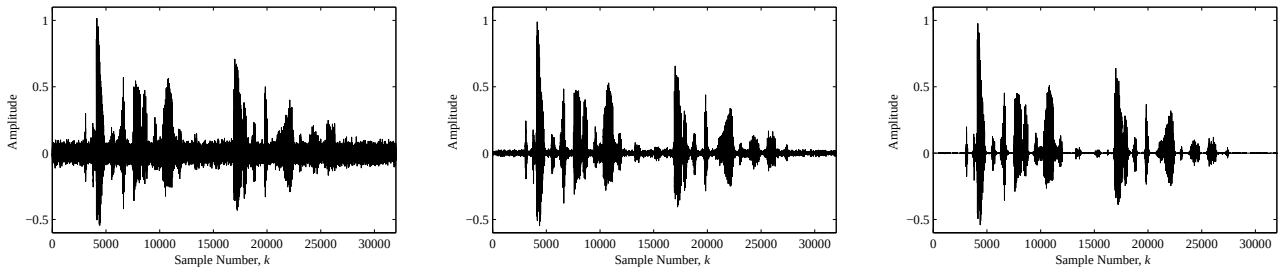
In the colored noise case, Fig. 6(a) illustrates the difference between SNR improvements obtained by estimators based on the QSVD and estimators based on the SVD. Thus, for most frames, the QSVD approach with integrated prewhitening delivers the best result, so in spite of the bias of the signal subspace as discussed previously, it is still better to use a signal subspace method with prewhitening, than without. However, the SNR improvement plots in Fig. 6(b) demonstrates a significant lower performance compared with the white noise case in Fig. 4(b). This can also be observed from Fig. 5(c) which shows the input-output relations of segmental SNRs for SDC estimator corresponding to the example in Fig. 5(b).

When the RRULVD-based estimators are used to enhance the noisy speech signal in Fig. 3(a), the improvements in segmental SNRs compared to the SVD approach is shown in Fig. 7(a) for the MV based estimation. Obviously, the recursive RRULVD method gives the best results, when there is a change in the dynamics of the signal, while the frame-based SVD approach is more accurate in stationary periods. However, as illustrated in Fig. 7(b), the variations between the two methods are larger in the colored noise case. The same observation is made in Fig. 7(c), when the RRULVD-based algorithm using a sliding window is compared with the one based on an exponential window (forgetting factor $\beta = 0.99$).



(a) (b) (c)

Figure 2 (a) Power of the residual noise $\mathbf{r}_n$ and the signal distortion $\mathbf{r}_s$ for the SDC estimator ($p = 12$) as function of β_2. The data matrix $\mathbf{X} \in \mathbb{R}^{141\times 20}$ represents a voiced speech frame of 160 samples added white noise (global SNR=5dB). (b) LPC-based magnitude spectra of the residual noise. (c) LPC-based magnitude spectra for a voiced speech frame (solid), AR(1,-0.7) noise process (dashed), and the speech prewhitened with the noise frame (dash-dot).



(a) (b) (c)

Figure 3 (a) Noisy speech sentence contaminated by white noise (SNR=5dB). (b) Enhanced speech signal obtained by the MV estimator. (c) Enhanced speech signal obtained by the SDC estimator ($\beta_2 = 5$).

4. INFORMAL LISTENING TESTS

Informal listening tests have been carried out for a number of speech sentences corrupted by white and colored noise. At higher noise levels (global SNR$<$10 dB), the enhanced speech signals obtained by the LS and MV methods are seriously affected by the musical noise. For the TDC and SDC estimators, the informal listening tests confirm that musical noise and/or audible distortions are still present in the processed speech. For example, the SDC estimator with $\beta_2 = 5$ results in enhanced speech almost free of musical noise, but with a significant distortion of the speech. In the case with colored noise, the audible speech distortion has increased and the musical noise is now dominated by low frequencies due to the bias of the signal subspace.

In the case of spatially uncorrelated noise, it is possible to eliminate the musical noise by using a multi-microphone solution, i.e., applying speech enhancement in each channel followed by summing of the outputs. Then the highly colored residual noise as shown in Fig. 2(b) will be whitened. Informal listening tests using four microphones confirm that the enhanced speech are almost free of both musical noise and speech distortions.

5. SUMMARY

The subspace-based noise reduction algorithms have been applied successfully to continuous speech embedded in white noise as well as colored broad-band noise. It has been demonstrated that the SVD-based signal subspace approach is able to achieve satisfactory improvements in the speech quality. Furthermore, arguments have been given for both introducing an recursive approach like the RRULVD, and for using a sliding window. In the colored noise case, the performance is highly dependent on the noise statistics. Thus, a noise process dominated by the same frequencies as the speech, will result in a less reliable algorithm. Furthermore, it should be emphasized that subspace techniques are a compromise between musical noise and signal distortion, and that this is less critical in combination with multi-channel solutions.

6. REFERENCES

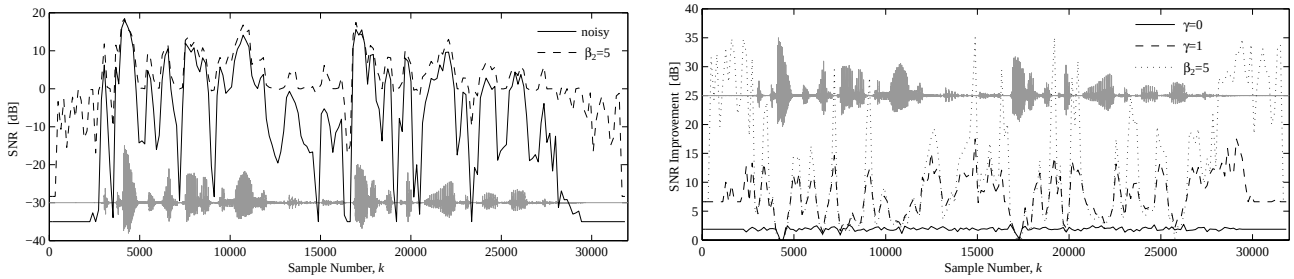
[1] M. Dendrinos, S. Bakamidis, and G. Carayannis. Speech Enhancement from Noise: A Regenerative Approach. *Speech Communication*, 10(1):45–57, February 1991.

[2] Yariv Ephraim and Harry L. Van Trees. A Signal Subspace Approach for Speech Enhancement. *IEEE Trans. on Speech and Audio Processing*, 3(4):251–266, July 1995.

[3] S. H. Jensen, P. C. Hansen, S. D. Hansen, and J. Aa. Sørensen. Reduction of Broad-Band Noise in Speech by Truncated QSVD. *IEEE Trans. on Speech and Audio Processing*, 3(6):439–448, November 1995.

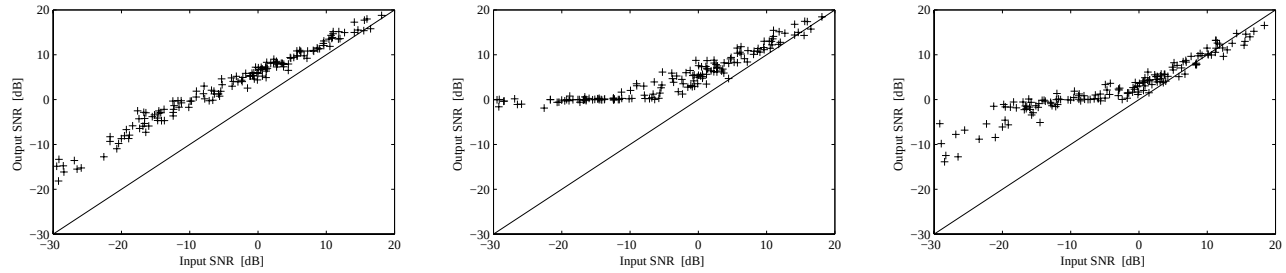
[4] P. S. K. Hansen. *Signal Subspace Methods for Speech Enhancement*. PhD thesis, Department of Mathematical Modelling, Technical University of Denmark, DK-2800 Lyngby, Denmark, September 1997.

[5] Bart De Moor. The Singular Value Decomposition and Long and Short Spaces of Noisy Matrices. *IEEE Trans. on Signal Processing*, 41(9):2826–2838, September 1993.



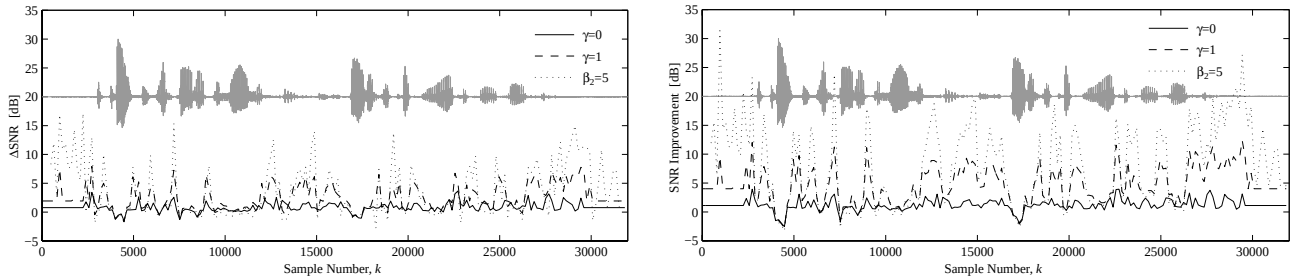
(a) (b)

Figure 4 (a) Segmental SNRs of the noisy signal and the SDC based enhanced waveform shown in Fig. 3. (b) Improvement in segmental SNRs for the enhanced waveforms obtained by the LS ($\gamma = 0$), MV ($\gamma = 1$) and SDC ($\beta_2 = 5$) estimators.



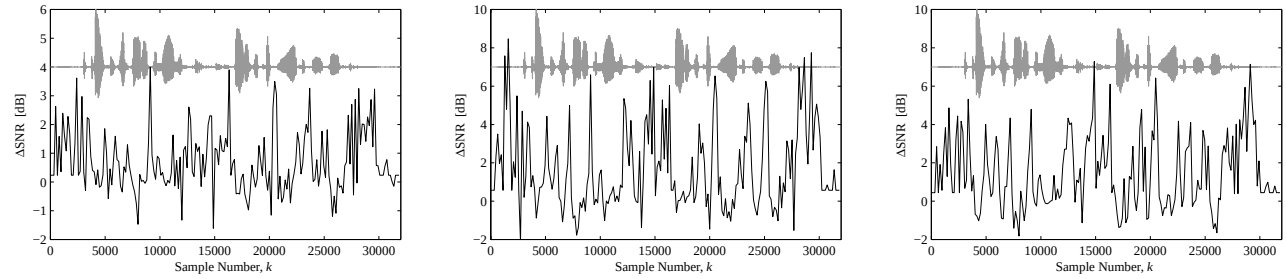
(a) (b) (c)

Figure 5 Segmental SNRs of the enhanced speech signal as function of the segmental SNRs of the noisy signal. (a) Using the MV estimator. (b) Using the SDC estimator ($\beta_2 = 5$). (c) As (b) but obtained by the QSVD in the colored noise case using an AR(1,-0.7) noise process (global SNR=5dB).



(a) (b)

Figure 6 (a) Difference between segmental SNRs of estimates obtained by using the QSVD and SVD algorithms, i.e., $\mathrm{SNR}_{QSVD}/\mathrm{SNR}_{SVD}$. The colored noise is an AR(1,-0.7) process (global SNR=5dB). (b) Improvement in segmental SNRs for the QSVD based estimators.



(a) (b) (c)

Figure 7 (a) Difference between segmental SNRs of enhanced speech (MV estimator) obtained by the RRULVD and the SVD, i.e., $\mathrm{SNR}_{RRULVD}/\mathrm{SNR}_{SVD}$. (b) As (a) but in the colored noise case obtained by the RRULLVD and the QSVD, i.e., $\mathrm{SNR}_{RRULLVD}/\mathrm{SNR}_{QSVD}$. (c) Difference between segmental SNRs of enhanced speech (MV estimator) obtained by the RRULVD using a sliding and exponential window ($\beta = 0.99$), respectively, i.e., $\mathrm{SNR}_{sli}/\mathrm{SNR}_{exp}$.