ULV-BASED SIGNAL SUBSPACE METHODS FOR SPEECH ENHANCEMENT

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@imm.dtu.dk, pch@imm.dtu.dk, sdh@imm.dtu.dk and jaas@imm.dtu.dk

ABSTRACT

In this paper the signal subspace approach for nonparametric speech enhancement is considered. Traditionally, the SVD (or the eigendecomposition) is used in frame-based methods to decompose the vector space of the noisy signal into a signal- and noise subspace [1, 2, 5]. Linear estimation of the clean signal from the information in the signal subspace is then performed using a set of nonparametric estimation criteria. In this paper, the rank-revealing ULV decomposition is used instead of the SVD, and we use recursive updating of the estimate instead of working in frames. An ULV formulation of three different estimation strategies is considered: Least Squares, Minimum Variance and Time Domain Constrained. Experiments indicate that the ULV-based algorithm is able to achieve the same quality of the reconstructed speech signal as the SVD-based method.

1 SIGNAL AND NOISE MODEL

Let $\mathbf{x} = (x_1, x_2, \cdots, x_m)^T$ denote the noisy signal vector of m samples and assume that the noise component $\mathbf{n}$ is additive and uncorrelated with the speech signal $\mathbf{s}$, i.e., $\mathbf{x} = \mathbf{s} + \mathbf{n}$.

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 where $m \geq n$. We assume that the noise is broad-banded so $\text{rank}(\mathbf{X}) = \text{rank}(\mathbf{N}) = n$, and that the speech signal can be described by a low order model, giving a rank deficient matrix $\mathbf{S}$ with $\text{rank}(\mathbf{S}) = p < n$. This formulation includes, for example, the *damped complex sinusoid* model, which has often been attributed to speech signals.

2 ULV BASED SIGNAL ESTIMATION

One approach for *nonparametric* speech enhancement is linear estimation of the clean signal from the noisy signal using signal subspace methods, which are based on the rank-revealing ULV decomposition (RRULVD) introduced by Stewart [8].

Assume that the singular values of $\mathbf{X}$ satisfy

$$\sigma_1 \geq \cdots \geq \sigma_p \geq \tau \gg \sigma_{p+1} \geq \cdots \geq \sigma_n \qquad (1)$$

then there exists a matrix $\mathbf{U}_X \in \mathbb{R}^{m \times n}$ with orthogonal columns and an orthogonal matrix $\mathbf{V}_X \in \mathbb{R}^{n \times n}$ such that

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

where $\mathbf{L}_{X1} \in \mathbb{R}^{p \times p}$, $\mathbf{G}_X \in \mathbb{R}^{(n-p) \times (n-p)}$ and $\mathbf{L}_X \in \mathbb{R}^{n \times n}$ are lower triangular, and

$$\sigma_{min}(\mathbf{L}_{X1}) \approx \sigma_p \qquad (3)$$

$$\|\mathbf{F}_X\|_F^2 + \|\mathbf{G}_X\|_F^2 \approx \sigma_{p+1}^2 + \cdots + \sigma_n^2 \qquad (4)$$

Thus, the signal- and noise subspaces defined by the gap in the singular values can be estimated using the RRULVD, where the quality depends on $\|\mathbf{F}_X\|_2$.

An *approximate* LS estimate $\hat{\mathbf{S}}_{ALS}$ of the signal matrix $\mathbf{S}$ can be computed by essentially substituting the RRULVD for the SVD based estimate [3], thus replacing one problem with a similar, nearby problem that can be solved more efficiently, i.e.,

$$\hat{\mathbf{S}}_{ALS} = \mathbf{X}\mathbf{V}_{X1}\mathbf{V}_{X1}^T \qquad (5)$$

The estimate converges to the true LS solution, if the following condition is satisfied

- The off-diagonal matrix $\mathbf{F}_X$ is zero.

Assume now that the estimator $\hat{\mathbf{S}}$ of the pure signal matrix $\mathbf{S}$ is constrained to be a *linear function* of the data matrix $\mathbf{X}$, i.e., $\hat{\mathbf{S}} = \mathbf{X}\mathbf{W}$ where $\mathbf{W} \in \mathbb{R}^{n \times n}$ is a filter matrix, then the *Minimum Variance* (MV) estimator problem [7] is to find the matrix $\mathbf{W}$ that minimizes

$$\min_{\mathbf{W}} \text{tr}\left((\mathbf{X}\mathbf{W} - \mathbf{S})^T(\mathbf{X}\mathbf{W} - \mathbf{S})\right) \Rightarrow \qquad (6)$$

$$\mathbf{W}_{MV} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{S} \qquad (7)$$

Note, that under stationary and ergodic conditions, the MV estimator converges asymptotically to the *Linear Minimum Mean-Squared Error* (LMMSE) estimator as the number of rows $m \to \infty$ [7].

To obtain the RRULVD based MV estimate proposed in [4], i.e.,

$$\hat{\mathbf{S}}_{MV} = \mathbf{X}\mathbf{V}_{X1}\mathbf{L}_{X1}^{-1}(\mathbf{L}_{X1} - \sigma_{noise}^2\mathbf{L}_{X1}^{-T})\mathbf{V}_{X1}^T \quad (8)$$

we need the additional conditions

- The signal is orthogonal to the noise: $\mathbf{S}^T\mathbf{N} = \mathbf{0}$.
- The matrix $\mathbf{N}$ satisfy: $\mathbf{N}^T\mathbf{N} = \sigma_{noise}^2\mathbf{I}_n$.
- There is a distinct gap in the singular values of the matrix $\mathbf{X}$: $\sigma_p > \sigma_{p+1}$.
- $\mathbf{G}_X = \sigma_{noise}\mathbf{I}_{n-p}$ is a diagonal matrix.

The residual matrix $\mathbf{R} = \mathbf{S}(\mathbf{W} - \mathbf{I}_n) + \mathbf{N}\mathbf{W} = \mathbf{R}_S + \mathbf{R}_N$ minimized in the above method represents signal distortion $\mathbf{R}_S$ and residual noise $\mathbf{R}_N$. Since both terms can not be simultaneously minimized, a *Time Domain Constrained* (TDC) estimator is proposed in [2] which keep the residual noise energy $\epsilon_n^2 = \mathrm{tr}(\mathbf{R}_N^T\mathbf{R}_N)$ below some threshold while minimizing the signal distortion energy $\epsilon_s^2 = \mathrm{tr}(\mathbf{R}_S^T\mathbf{R}_S)$

$$\min_{\mathbf{W}} \epsilon_s^2 \quad \text{subject to} \quad \epsilon_n^2 \leq \alpha n \sigma_{noise}^2 \quad \Rightarrow \quad (9)$$

$$\mathbf{W}_{TDC} = (\mathbf{S}^T\mathbf{S} + \gamma\sigma_{noise}^2\mathbf{I}_n)^{-1}\mathbf{S}^T\mathbf{S} \quad (10)$$

where α is a fixed or SNR-dependent parameter ($0 \leq \alpha \leq 1$), and γ is the Lagrange multiplier in (9). In a practical implementation, γ is actually used as the parameter.

Given the above conditions, we propose a RRULVD based TDC estimate, which can be obtained by using the following RRULVD formulations for $\mathbf{S}$ and $\mathbf{X}$

$$\mathbf{S} = \begin{pmatrix} \mathbf{U}_{S1} & \mathbf{U}_{S2} \end{pmatrix} \begin{pmatrix} \mathbf{L}_{S1} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{pmatrix} \begin{pmatrix} \mathbf{V}_{S1}^T \\ \mathbf{V}_{S2}^T \end{pmatrix} \quad (11)$$

$$\mathbf{X} = \begin{pmatrix} (\mathbf{U}_{S1}\mathbf{L}_{S1} + \mathbf{N}\mathbf{V}_{S1})\mathbf{L}_{X1}^{-1} & \mathbf{N}\mathbf{V}_{S2}\sigma_{noise}^{-1} \end{pmatrix} \cdot \begin{pmatrix} \mathbf{L}_{X1} & \mathbf{0} \\ \mathbf{0} & \sigma_{noise}\mathbf{I}_{n-p} \end{pmatrix} \begin{pmatrix} \mathbf{V}_{S1}^T \\ \mathbf{V}_{S2}^T \end{pmatrix} \quad (12)$$

and the relation $\mathbf{L}_{X1}^T\mathbf{L}_{X1} = \mathbf{L}_{S1}^T\mathbf{L}_{S1} + \sigma_{noise}^2\mathbf{I}_p$, i.e.,

$$\hat{\mathbf{S}}_{TDC} = \mathbf{X}\mathbf{V}_{X1}(\mathbf{L}_{X1} - \sigma_{noise}^2(1-\gamma)\mathbf{L}_{X1}^{-T})^{-1} \cdot (\mathbf{L}_{X1} - \sigma_{noise}^2\mathbf{L}_{X1}^{-T})\mathbf{V}_{X1}^T \quad (13)$$

Note that for $\gamma = 0$ we obtain (5) and for $\gamma = 1$ we obtain (8). For speech signals, the TDC estimation criterion will control the nonstationary residual noise with annoying noticeable tonal characteristics, referred to as *musical noise*, since this noise component decreases as $\gamma \to \infty$.

In practice, the above mentioned conditions are never satisfied exactly, but the RRULVD is robust with respect to mild violations of these conditions.

If the additive noise matrix $\mathbf{N}$ is colored, $\mathbf{N}^T\mathbf{N} \neq \sigma_{noise}^2\mathbf{I}_n$, then a prewhitening transformation can be applied to the data matrix using the QR decomposition of $\mathbf{N} = \mathbf{Q}\mathbf{R}$

$$\mathbf{X}\mathbf{R}^{-1} = \mathbf{S}\mathbf{R}^{-1} + \mathbf{N}\mathbf{R}^{-1} = \mathbf{S}\mathbf{R}^{-1} + \mathbf{Q} \quad (14)$$

One problem concerning the prewhitening transformation is the complicated update of the matrix $\mathbf{X}\mathbf{R}^{-1}$ when $\mathbf{X}$ and $\mathbf{N}$ are updated. This can be avoided by using the rank-revealing ULLV decomposition [6] of the matrix pair $(\mathbf{X}, \mathbf{N})$, which allows each matrix to be updated individually and delivers the required factorizations without forming the quotients and products.

3 IMPLEMENTATION

The transformation $\mathbf{y} = \mathbf{V}_X^T\mathbf{x}$ approximates the *Karhunen-Loeve* transform (KLT) of $\mathbf{x}$. Hence, all the above mentioned linear signal estimates are obtained by the following steps (see Figure 1)

- KLT of the noisy signal onto the signal subspace.
- Modify the components of the KLT by a gain filter matrix $\mathbf{G}_1$.
- Inverse KLT of the modified components to reconstruct the signal in the signal subspace.

This scheme results in a generalized formulation of the optimal linear estimator

$$\hat{\mathbf{s}} = \mathbf{W}\mathbf{x} = \mathbf{V}_{X1}\mathbf{G}_1\mathbf{V}_{X1}^T\mathbf{x} \quad (15)$$

where the matrix $\mathbf{G}_1 \in \mathbb{R}^{p \times p}$ depends on the estimation method as shown in the last section.

The two matrices $\mathbf{L}_X$ and $\mathbf{V}_X$ necessary for computing $\mathbf{W}$ are updated for each new sample x_k corresponding to a new row in the data matrix $\mathbf{X}$. A new row is processed in the following four steps.

- Updating: The new row of $\mathbf{X}$ is incorporated into the decomposition.
- Downdating: The oldest row of $\mathbf{X}$ is isolated and removed in the decomposition.
- Deflation: Establishes and maintains the rank-revealing nature of the decomposition.
- Refinement: The norm of $\mathbf{F}_X$ is reduced to improve the subspace quality.

Obviously, the filter matrix $\mathbf{W}$ is estimated in an analysis window of width $(m+n-1)$, centered around

the middle row of $\mathbf{X}$. The linear estimator is applied to this row, giving a n-sample synthesis window. Finally, the enhanced vectors are combined using the overlap and add synthesis approach, which corresponds to the LS esimate of the noise-free signal s_k from the enhanced vectors [9].

4 EXPERIMENTS

A recursive RRULLV algorithm has been developed based on the methods given in [8, 6], and was applied to speech signals contaminated by an AR(1, -0.7) noise process. The noise matrix $\mathbf{N}$ was only updated in periods without speech, and the matrix dimension was $m = 141$ and $n = 20$. In all simulations, a TDC estimator is used.

The typical average SNR of a reconstructed speech segment (voiced) using 100 independent noise realizations and SNR = 5 dB is illustrated in Fig. 2 as a function of the signal subspace dimension p and the parameter γ. Clearly, the MV estimate ($\gamma = 1$) gives the best SNR improvement and is less sensitive to the choice of p compared with the other estimates. However, if γ is chosen in the neighbourhood of 1, the variations are minimal. Thus, using a fixed value of $p = 14$ as in the following results, we are able to achieve a satisfactory quality of the reconstructed speech. An informal listening test gave $\gamma \approx 2$ as the best fixed value, but a better choice is to make γ dependent on the local SNR.

The RRULLV algorithm using a sliding window was applied to the speech signal in Fig. 3 with added broad-band noise (global SNR of 5 dB). Observe from Fig. 4 that their is a SNR improvement using the TDC estimate and that the variations among the local SNRs of the various segments are reduced.

In the RRULLV algorithm computations can be saved by omitting the refinement step, i.e., accepting a larger $\|\mathbf{F}_X\|_2$, but then the singular values of $\mathbf{L}_{X1}$ will underestimate the first p values $\sigma_i(\mathbf{X}\mathbf{R}^{-1})$ by a larger factor. Similarily, the singular values of $\mathbf{G}_X$ will in general overestimate the corresponding last $n - p$ values $\sigma_i(\mathbf{X}\mathbf{R}^{-1})$.

The graphs in Fig. 5 and 6 illustrates this problem. Here, the average singular values of a prewhitened voiced speech frame are compared with the one obtained from $\mathbf{L}_{X1}$ and $\mathbf{G}_X$ with $p = 14$ and no refinement. Note, that $\sigma_i(\mathbf{L}_{X1})$ are plotted against the first p indices, and $\sigma_i(\mathbf{G}_X)$ are plotted from index $p + 1$ to n. It is seen that the largest and smallest singular values and thereby the dominant range and null space are well determined by the RRULLVD, while the subspaces are blurred together near the rank-revealing point p due to the off-diagonal block $\mathbf{F}_X$ and the small gap in the singular spectrum. As shown in Fig. 6, the quality can be increased by applying a number of refinement steps.

In Fig. 7, the canonical angles between the QSVD and RRULLVD based signal subspaces are plotted against their indices, where the example corresponds to the one in Fig. 5. As expected, there is a group of large angles due to the mix of signal and noise subspace. However, since the singular spectrum of speech signals is relative constant at the rank-revealing point, this has no practical effect in a noise reduction algorithm as shown in Fig. 8. Here, four different speech segments all result in a reconstructed average SNR, which is nearly independent of the number of refinement steps. This is also why these results closely match the QSVD based method.

Another issue is that the conditions for the RRULLV based estimates are typically not satisfied. However, as demonstrated in Fig. 8, the method is very robust concerning this.

5 SUMMARY

A recursive signal subspace approach for noise reduction of speech signals is presented. The algorithm is formulated by means of the RRULVD using a proposed set of estimators. The method is demonstrated to be comparable with SVD-based methods.

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] Ricardo D. Fierro and Per Christian Hansen. Accuracy of TSVD Solutions Computed from Rank-Revealing Decompositions. *Numerische Mathematik*, 70:453–471, 1995.

[4] P. S. K. Hansen, P. C. Hansen, S. D. Hansen, and J. Aa. Sørensen. Noise Reduction of Speech Signals using the Rank-Revealing ULLV Decomposition. In *Signal Processing VIII Theories and Applications*, volume 2, pages 967–970. EUSIPCO-96, 1996.

[5] 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.

[6] F. T. Luk and S. Qiao. A New Matrix Decomposition for Signal Processing. In M. S. Moonen et al., editor, *Linear Algebra for Large Scale and Real-Time Applications*, pages 241–247. Kluwer Academic Publishers, 1993.

[7] 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.

[8] G. W. Stewart. Updating a Rank-Revealing ULV Decomposition. *SIAM Journal on Matrix Analysis and Applications*, 14(2):494–499, April 1993.

[9] Donald W. Tufts and Abhijit A. Shah. Estimation of a Signal Waveform from Noisy Data Using Low-Rank Approximation to a Data Matrix. *IEEE Trans. on Signal Processing*, 41(4):1716–1721, April 1993.

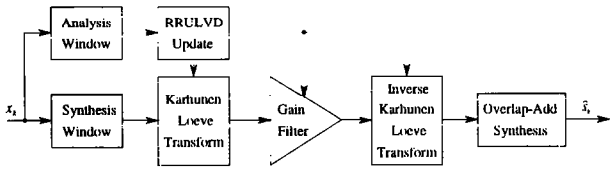

Figure 1 Filter structure.

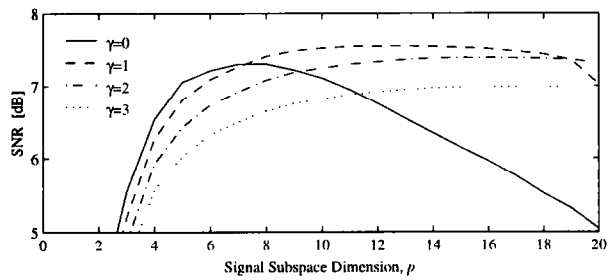

Figure 2 Average SNR of a reconstructed noisy (voiced) speech segment using a TDC estimator with the listed γ values, and SNR=5dB. The average is taken over 100 independent noise realizations.

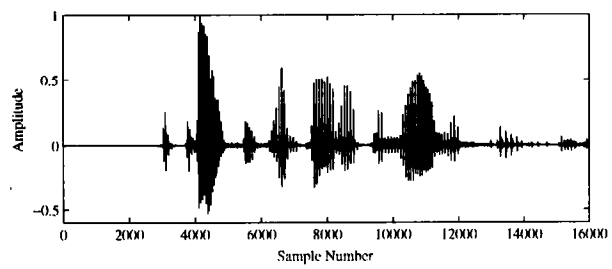

Figure 3 Noise-free speech signal.

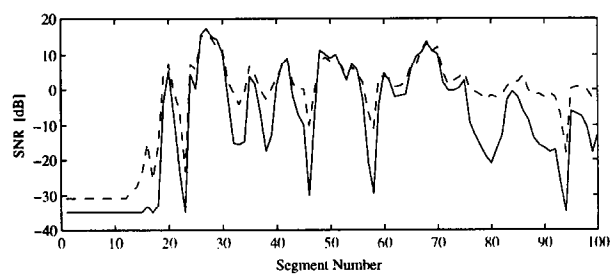

Figure 4 The local SNR of noisy speech signal (global SNR=5dB) and a TDC estimate with p=14 and γ=2.

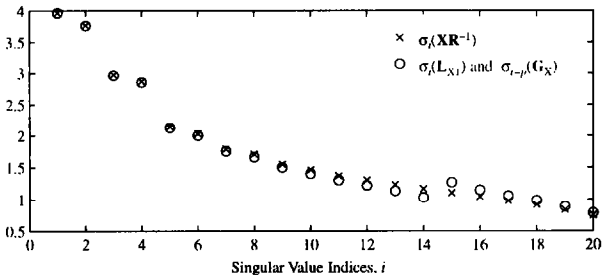

Figure 5 Average singular values of prewhitened (voiced) speech segment using 100 independent noise realizations and SNR=5dB. The rank revealed in $\mathbf{L}_X$ is $p = 14$ (without refinement).

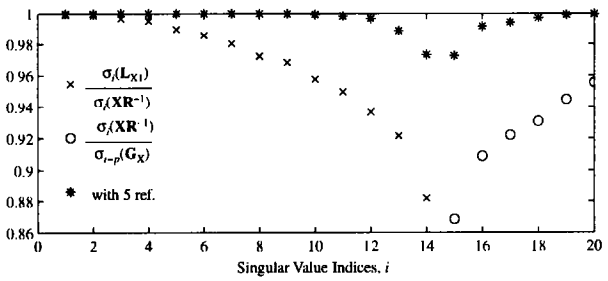

Figure 6 The ratios corresponds to the example in Fig. 5 without refinement, and a case with 5 refinement steps (*).

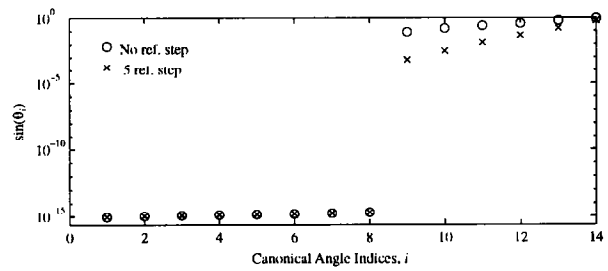

Figure 7 The average canonical angles $\sin(\theta_i)$ between the 14 dimensional signal subspaces obtained from the QSVD and the RRULLVD, respectively. The signals correspond to the example in Fig. 6 without refinement (o) and with 5 steps (*).

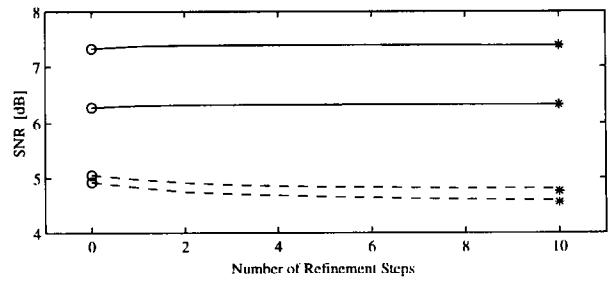

Figure 8 Average SNR corresponding to four different reconstructed noisy (voiced) speech segments using a TDC estimator with $\gamma = 2$, $p = 14$ and SNR=5dB. The (*) marks are the QSVD based estimates and voiced/unvoiced frames are given by solid/dashed lines. The average is taken over 100 independent noise realizations.