ASSESSMENT OF SIGNAL SUBSPACE BASED SPEECH ENHANCEMENT FOR NOISE ROBUST SPEECH RECOGNITION

Kris Hermus and Patrick Wambacq

Katholieke Universiteit Leuven – Dept. ESAT
Kasteelpark Arenberg 10, B-3001 Heverlee, Belgium
{hermus,wambacq}@esat.kuleuven.ac.be

ABSTRACT

Subspace filtering is an extensively studied technique that has been proven very effective in the area of speech enhancement to improve the speech intelligibility. In this paper, we review different subspace estimation techniques (Minimum Variance, Least Squares, Singular Value Adaptation, Time Domain Constrained and Spectral Domain Constrained) in a modified singular value decomposition (SVD) framework, and investigate their capability to improve the noise robustness of speech recognisers. An extensive set of recognition experiments with the Resource Management (RM) database showed that significant reductions in WER can be obtained, both for the white noise and the coloured noise case. Unlike for speech enhancement approaches, we found that no truncation of the noisy signal subspace should be done to optimise the recognition accuracy.

1. INTRODUCTION

One solution to mitigate the performance degradation of automatic speech recognition systems in noisy operating environments, is to enhance the observed speech prior to the recogniser's preprocessing and decoding operations. At first sight, the enhancement step can be performed independently of the recognition process by one of the manifold available (single-microphone) speech enhancement algorithms. However, the main objective of noise reduction in speech enhancement and in automatic speech recognition differs. While the first tries to improve the intelligibility of the speech and/or to ease listener's fatigue, the latter is effective if it succeeds in closing the gap between noisy and clean speech recognition accuracy. Nevertheless, a correlation can be expected between the improvements in speech quality on the one hand, and the improvement in recognition accuracy on the other hand.

A particular class of speech enhancement techniques, that has gained a lot of attention, is signal subspace filtering. In this approach, a non-parametric linear estimate of the unknown clean speech signal is obtained, based on a decomposition of the observed noisy signal into mutually orthogonal speech and noise subspaces. This decomposition is possible under the assumption of a low-rank linear model for speech and an uncorrelated additive (white) noise interference. Noise reduction is then obtained by nulling the noise subspace (Least Squares (LS) estimation), combined with a suppression of the noise threshold in the speech subspace (Minimum Variance (MV) or Singular Value Adaptation (SVA) estimators). Although this theory is developed for white noise, it can easily be extended to general coloured noise if the noise covariance matrix is known (or estimated). A theoretical analysis of the underlying principles of subspace filtering can e.g. be found in [1, 2]. In [2] a subspace-based speech enhancement with noise shaping is proposed. Based on the observation that signal distortion and residual noise can not be minimised simultaneously, two new linear estimators are designed – Time Domain Constrained (TDC) and Spectral Domain Constrained (SDC) – that keep the level of the residual noise below a chosen threshold while minimising signal distortion. Parameters of the algorithm control the trade-off between residual noise and signal distortion. In perceptual subspace speech enhancement, the residual noise is shaped according to an estimate of the clean signal masking threshold, as discussed in recent papers [3, 4]. The excellent noise reduction capabilities of subspace filtering techniques are confirmed by several studies, both with the basic LS estimate [5] and with the more advanced optimisation criteria [2, 6, 7]. Especially for the MV and SDC estimators, a speech quality improvement that outperforms the Spectral Subtraction approach is revealed by listening tests.

Very few papers discuss the application of signal subspace methods to robust speech recognition. In [8] an energy-constrained signal subspace (ECSS) method is proposed based on the MV estimator. For the recognition of LV-CS corrupted by additive white noise, a relative reduction in WER of 70% is reported. In [9], MV subspace filtering is applied on a LV-CSR task distorted with white and coloured noise. Significant WER reductions are reported that outperform spectral subtraction. To our knowledge, no results with other signal subspace estimators than MV were reported.

In this paper we investigate the impact of five different subspace estimation techniques on the noise robustness of LV-CSR, and compare their performance, both for white and coloured noise removal. Some crucial parameters, such as the analysis window length, the Hankel matrix dimensions, the signal subspace dimension, and method-specific design parameters will be discussed.

2. SUBSPACE FILTERING

2.1. Basic Theory

Let $s(k), k = 1 \ldots N$ represent a N-dimensional vector of clean speech samples and $n(k), k = 1 \ldots N$ be the zero-mean, additive white noise distortion that is assumed to be uncorrelated with the clean speech. The observed noisy speech $x(k), k = 1 \ldots N$ is then given by

$$x(k) = s(k) + n(k)$$

Further, let R_x, R_s, and R_n be $N \times N$ covariance matrices from $x(k)$, $s(k)$ and $n(k)$ respectively. It is clear that

$$R_x = R_s + R_n$$

The eigenvalue decomposition (EVD) of R_s, R_n and R_x can be written as follows

$$R_s = U\,E\,U^T \quad (1)$$
$$R_n = U\,(\sigma_w^2 I)\,U^T \quad (2)$$
$$R_x = U\,(E+\sigma_w^2 I)\,U^T \quad (3)$$

with σ_w^2 the white noise variance and I the identity matrix. Observe that the eigenvectors of the noise are identical to the clean speech eigenvectors due to the white noise assumption.
Speech and noise are separated based on the assumption that the clean speech is confined to a $p < N$ dimensional subspace, the so-called signal subspace, whereas the noise occupies the N-dimensional observation space. This implies that E has only p non-zero eigenvalues e_i, such that the EVD of R_x can be rewritten as :

$$R_x = \left[U_p U_{N-p}\right]\left(\left[E_p E_{N-p}\right]+\sigma_w^2\left[I_p I_{N-p}\right]\right)\left[U_p^T U_{N-p}^T\right]$$

if we assume that the elements e_i of E are in descending order.
Regardless of the specific optimisation criterion, speech enhancement is now obtained by

(1) restricting the enhanced speech to occupy solely the signal subspace by nulling its components in the noise subspace
(2) changing (i.e. lowering) the energy of the eigenvalues that correspond to the signal subspace

Mathematically this enhancement procedure can be written as a linear filtering operation on the noisy speech $x(k)$:

$$\hat{s} = Fx$$

with the filter matrix F given by :

$$F = U_p\,G_p\,U_p^T$$

in which the $p \times p$ diagonal matrix G_p contains the weighting factors $g(i)$ for the first p eigenvalues of R_x, while U^T and U are known as the KLT (Karhunen Loeve transform) matrix and inverse KLT matrix, respectively.
In many implementations the above covariance matrices are estimated as $R_x = H_x H_x^T$ with H_x a Hankel/Toeplitz signal observation matrix. In that case an equivalent speech enhancement can be obtained via the SVD of H_x. A commonly used modified SVD based speech enhancement procedure proceeds as follows :
Let H_x $(= H_s + H_n)$ be a $m \times q$ $(m+q = N+1$ and $m > q)$ noisy Hankel/Toeplitz matrix constructed from x, with SVD

$$H_x = U\Sigma V^T$$

The enhanced matrix $\hat{H}_s$ is then obtained as

$$\hat{H}_s = U_p\,G_p\,\Sigma_p\,V_p^T$$

or

$$\hat{H}_s = \sum_{i=1}^{p} g(i)\,\sigma_i\,u_i\,v_i^T$$

from which the enhanced signal $\hat{s}(k)$ is recovered by averaging along the diagonals (Toeplitz) or anti-diagonals (Hankel) of $\hat{H}_s$. An equivalent FIR filter implementation of this overall procedure is described in [10].
The main advantage of working with the SVD, instead of the EVD, is that the first scheme requires less computations since no explicit estimation of the covariance matrix is needed. Therefore, our speech enhancement algorithm is implemented in terms of the SVD, and this paper will further focus on the SVD description.

2.2. Optimisation criteria

By applying a specific estimation criterion, the elements of the weighting matrix G_p can be found. In this section the most common of these criteria are briefly reviewed.

Least Squares (LS) The LS estimate $\hat{H}_{\mathrm{LS}}$ is defined as the best rank-p approximation of H_x,

$$\min_{rk(\hat{H}_{\mathrm{LS}})=p} ||H_x - \hat{H}_{\mathrm{LS}}||_F^2$$

and is obtained by truncating the SVD $U\Sigma V^T$ of H_x to rank p

$$\hat{H}_{LS} = U_p\,\Sigma_p\,V_p^T$$

Observe that this estimate removes the noise subspace, but keeps the noisy signal unaltered in the signal subspace. This estimate yields an enhanced signal with the highest residual noise level but with the lowest signal distortion.

Minimum Variance (MV) Given the rank p of the clean speech, the MV estimate $\hat{H}_{\mathrm{MV}}$ is the best approximation of the original matrix H_s that can be obtained by making linear combinations of the columns of H_x

$$\hat{H}_{\mathrm{MV}} = \min_{T \in \mathbb{R}^{q \times q}} ||H_x T - H_s||_F^2$$

In algebraic terms, $\hat{H}_{\mathrm{MV}}$ is the geometric projection of H_s onto the column space of H_x, and is obtained by setting

$$g_{MV}(i) = 1 - \frac{\sigma_w^2}{\sigma_i^2}$$

The MV estimate is the linear estimator with the lowest residual noise level [1, 11].

Singular Value Adaptation (SVA) In the SVA method the singular values of H_x are mapped onto the estimated original (clean) singular values of H_s, by setting

$$g_{SVA}(i) = \frac{\sqrt{\sigma_i^2 - \sigma_w^2}}{\sigma_i}$$

The mapping operator of the SVA method is defined by [11]

$$\min_{rk(\hat{H}_{\mathrm{SVA}})=p} ||H_s - \hat{H}_{\mathrm{SVA}}||_F^2$$

Time Domain Constrained (TDC) The TDC estimate is found by minimising the signal distortion while setting a user-defined upper bound on the residual noise level via a control parameter $\mu \geq 0$. In the modified SVD of H_x, $g_{\mathrm{TDC}}(i)$ is equal to

$$\frac{1 - \frac{\sigma_w^2}{\sigma_i^2}}{1 - \frac{\sigma_w^2}{\sigma_i^2}(1-\mu)}$$

A detailed description can be found in [2].

Spectral Domain Constrained (SDC) A simple form of residual noise shaping is provided by the SDC estimator. Here, the estimate is found by minimising the signal distortion subject to constraints on the energy of the projections of the residual noise onto the signal subspace. However, it is not possible to exploit the

information obtained from a masking model. More than one solution for the gain factors in the modified SVD exists. One possible expression for $g_{SDC}(i)$ is [2]

$$g_{SDC-1}(i) = \sqrt{\exp\left(\frac{-\beta\,\sigma_w^2}{\sigma_i^2 - \sigma_w^2}\right)}$$

with $\beta \geq 1$. We will further refer to this estimator as SDC-1. An alternative solution [2] is to choose

$$g_{SDC-2}(i) = \left(1 - \frac{\sigma_w^2}{\sigma_i^2}\right)^{\gamma/2}$$

with $\gamma \geq 1$, further denoted as SDC-2. The amount of noise reduction can be controlled by the parameters β and γ.

2.3. Implementation issues

The use of SVD based filtering implies the careful choice of certain parameters. In this section we discuss the impact of the most important ones, namely the frame length N, the dimensions of H_x, and the dimension p of the signal subspace.

Signal Subspace dimension In theory the dimension of the signal subspace is defined by the order of the linear signal model. However, in practice the speech contents will strongly vary (e.g. voiced versus unvoiced segments) and the entire signal will never exactly obey one model. Several techniques, such as Minimum Description Length (MDL) were developed to estimate the model order. Sometimes, the order p is chosen on a frame-by-frame basis, and e.g. chosen as the number of positive eigenvalues of R_s. For 16 kHz data the value of p is usually around 12.

Frame length The frame length N must be larger than the order of the assumed signal model, such that the correlation that is embedded in the speech signal can be fully exploited to split the latter signal from the noise. On the other side, the frame length is limited by the time over which the speech and noise can be assumed stationary (usually 20 to 30 ms). Besides, N must not be too large to avoid prohibitively large computations in the SVD of H_x.

Matrix dimension Observe that the dimensions $m \times q$ of H_x cannot be chosen independently due to the relation $m+q = N+1$. The smaller dimension q of H_x should be larger than the order of the assumed signal model, such that the correlation between the clean signal samples can be exploited to separate the speech from the noise. A sufficiently high value for q is beneficial for effective noise removal, since the (pre)white(ned) noise will be equally distributed over all dimensions.

3. EXTENSION TO COLOURED NOISE

If the additive noise is not white, the assumptions made above are no longer valid and a different procedure should be applied. However, the modified SVD noise reduction scheme can easily be extended to the general coloured noise case.

If the noise covariance matrix R_n can be estimated (from noise-only input segments), a prewhitening operation can be applied based on the QR factorisation of R_n. Indeed, if

$$H_x R^{-1} = (H_s + H_n) R^{-1}$$

then

$$\left(H_n R^{-1}\right)^T \left(H_n R^{-1}\right) = Q^T Q = I$$

A corresponding dewhitening operation should be included after the SVD modification.

Because subsequent pre- and dewhitening can cause a loss of accuracy due to numerical instability, usually an implicit pre- and dewhitening is performed by working with the quotient SVD (QSVD) of the matrix pair (H_s, H_n) [7].

A major drawback of pre- and dewhitening is that not only the additive noise but also the original signal is effected by the transformation matrices. The optimisation criteria (e.g. minimal signal distortion) will hence be applied to a transformed (= distorted) version of the speech and not to the original speech. It can be theoretically shown that in this case only an upper-bound of the signal distortion is minimised. Alternatively, the pre- and dewhitening can be avoided by projecting the coloured noise onto the *clean* signal subspace,

$$\Sigma_{c,\text{proj}} = \sqrt{U^T R_n U}$$

with U obtained from $R_s = UEU^T$ [12].

4. RECOGNITION EXPERIMENTS

In this section we describe the results of LV-CSR experiments, in which the SVD based speech enhancement procedure is used as a preprocessing step, prior to the recognisers' feature extraction module. Experiments are carried out with all five above mentioned estimators.

Evaluation Database As test material we took the Resource Management (RM) database. These data are considered as clean data, to which distortions were artificially added. The SNR-ratio is determined in the same way as in the Aurora 4 benchmark database [13]. The ratio of signal to noise energy is defined after filtering both signals with the G.712 characteristic. To determine the speech energy the ITU recommendation P.56 is applied by using the corresponding ITU software. The noise energy is calculated as RMS value with the same software. Two noise types were added to the clean speech, namely (1) white noise $w(k)$, and (2) coloured noise $c(k)$ (obtained as low-pass filtered white noise, $c(z) = w(z) + w(z^{-1})$), both at various SNR ratios (5, 10, 15, 20, 25 and 30 dB).

Speech Recogniser For the assessment of the different subspace approaches we use the speaker-independent LV-CSR system that has been developed by the ESAT-PSI speech group of the K.U.Leuven. The system is beneficial for this purpose because of its fast experiment turn-around time and good baseline accuracy. In the preprocessing, the common MEL cepstral coefficients are combined with their first and second order derivatives (25 features in total). To remove convolutional noise distortions, a CMS step is included. The acoustic modeling is based on a set of 46 phones. Each of the 139 HMM states is modelled by a mixture of 128 tied gaussian distributions, which are selected from a total set of 4526 gaussians. Training is performed with the original clean RM data; no retraining with SVD enhanced speech material is conducted. A word-pair grammar (WPG) language model for the 1k-word vocabulary is used, while decoding is done with a time-synchronous beam search algorithm. The training material consists of the SI-109 train set, while testing is done with the feb89 testset.

Results The estimation criteria mentioned above, are compared

SNR (dB)	White Noise						Coloured Noise					
	5	10	15	20	25	30	5	10	15	20	25	30
Ref	2.30	4.57	25.07	52.13	73.45	85.63	1.91	12.10	41.62	67.51	83.16	90.82
LS	2.73	14.17	41.62	67.67	82.43	89.34	2.42	19.29	51.19	71.81	84.97	91.14
MV	14.14	42.68	71.22	86.26	91.21	93.05	17.53	50.06	75.79	88.95	91.64	92.97
SVA	6.60	31.12	64.86	82.35	90.12	92.31	9.14	37.13	69.97	84.07	89.50	91.84
TDC	18.00	46.00	73.72	87.15	91.57	93.17	24.95	53.30	77.39	88.99	91.80	92.89
SDC-1	7.77	38.34	67.24	83.52	88.64	90.63	15.50	42.33	72.20	86.22	89.54	89.81
SDC-2	16.75	47.56	74.81	86.84	91.37	93.06	22.18	51.27	75.95	88.99	91.68	92.98
COPY-SPEC	34.24	61.81	80.94	89.22	91.57	92.82	33.78	58.38	77.35	87.39	91.80	92.82

Table 1. *Recognition rates (%) with SVD based speech enhancment — RM feb89 test set.*

in a series of recognition experiments. For each estimator we varied the values of the main algorithm parameters (frame length, subspace order, ...) to find the "optimal" performance.
Table 1 presents the reference recognition rates (i.e. without noise reduction) together with the best recognition rates for each of the estimation criteria. The analysed frames (no windowing) have a 30 ms length with 50% overlap; the enhanced frames are hamming windowed for resynthesis of the enhanced signal. The clear differences in reference recognition rates between the white and coloured noise case are mainly due to the definition of SNR that "underestimates" the SNR for low-frequency noise sources. For completeness, we mention that the recognition accuracy for the original clean data is 95.12%. For the TDC and SDC estimators, the best results are obtained with $\mu=3$, $\beta=1$ and $\gamma=4$.
The results of the COPY-SPEC estimator are obtained by a cheating experiment, in which the noisy singular values are replaced by the singular values of the clean Hankel matrix H_s: if H_s is known with SVD given by $U_s \Sigma_s V_s^T$, and if $H_x = U \Sigma V^T$ then $\hat{H} = U \Sigma_s V^T$. Intuitively, these results give an indication of an upper-bound on the recognition accuracy that could be obtained with SVD based filtering.
From our experiments we further learn that (1) the MV, TDC and SDC-2 estimators are most effective, (2) except for the LS case, the order p of the signal subspace should be almost equal to q (no nulling of the noise subspace), (3) the optimal order q of the Hankel matrix is between 8 and 20, (4) the frame length is best between 10 and 30 ms, and (5) the QSVD and the noise projection method yield comparable results for the coloured noise case.

5. CONCLUSIONS

In this paper we compared several subspace filtering estimators and showed that these techniques can significantly increase the noise robustness of LV-CSR. The MV estimator and its generalisation, the SDC estimator, proved to give the best recognition accuracy. Interestingly, we found that their performance remains rather constant under mild violations of the optimal parameter values in the algorithm. Our current focus is on the assessment of these techniques in non-stationary noise conditions and on the incorporation of a psycho-acoustic model to achieve a perceptual shaping of the residual noise.

6. REFERENCES

[1] B. De Moor, "The singular value decomposition and long and short spaces of noisy matrices," *IEEE Trans. on SP*, vol. 41, no. 9, pp. 2826–2838, Sept. 1993.

[2] Y. Ephraim and H.L. Van Trees, "A signal subspace approach for speech enhancement," *IEEE Trans. on SAP*, vol. 3, no. 4, pp. 251–266, July 1995.

[3] F. Jabloun and B. Champagne, "Incorporating the human hearing properties in the signal subspace approach for speech enhancement," *IEEE Trans. on SAP*, vol. 11, no. 6, pp. 700–708, Nov. 2003.

[4] Y. Hu and P. Loizou, "Perceptual weighting motivated subspace based speech enhancement approach," in *Proc. ICSLP*, Denver, Colorado, U.S.A., Sept. 2002, pp. 1797–1800.

[5] M. Dendrinos, S. Bakamidis, and G. Carayannis, "Speech enhancement from noise: a regenarative approach," *Speech Comm.*, vol. 10, pp. 45–57, Feb. 1991.

[6] P.S.K. Hansen, Hansen P.C., Hansen S.D., and Sørensen J.A., "Experimental comparison of signal subspace based noise reduction methods," in *Proc. ICASSP*, Phoenix, U.S.A., Mar. 1999, vol. I, pp. 101–104.

[7] S.H. Jensen, P.C. Hansen, S.D. Hansen, and J.A. Sørensen, "Reduction of broad-band noise in speech by truncated QSVD," *IEEE Trans. on SAP*, vol. 3, pp. 439–448, Nov. 1995.

[8] J. Huang and Y. Zhao, "Energy-constrained signal subspace method for speech enhancement and recognition," *IEEE SP Letters*, vol. 4, no. 10, pp. 283–285, Oct. 1997.

[9] K. Hermus, W. Verhelst, and P. Wambacq, "Optimized subspace weighting for robust speech recognition in additive noise environments," in *Proc. ICSLP*, Beijing, China, Oct. 2000, vol. III, pp. 542–545.

[10] P.C. Hansen and S.H. Jensen, "FIR filter representations of reduced-rank noise reduction," *IEEE Trans. on SP*, vol. 46, no. 6, pp. 1737–1741, June 1998.

[11] S. Van Huffel, "Enhanced resolution based on minimum variance estimation and exponential data modeling," *Signal Processing*, vol. 33, no. 3, pp. 333–355, Sept. 1993.

[12] U. Mittal and N. Phamdo, "Signal/noise KLT based approach for enhancing speech degraded by colored noise," *IEEE Trans. on SAP*, vol. 8, no. 2, pp. 159–167, Mar. 2000.

[13] H.G. Hirsch and D. Pearce, "The AURORA experimental framework for the performance evaluation of speech recognition systems under noisy conditions," in *Proc. ISCA ITRW ASR*, Paris, France, Sept. 2000, pp. 181–188.