OPTIMIZED SUBSPACE WEIGHTING FOR ROBUST SPEECH RECOGNITION IN ADDITIVE NOISE ENVIRONMENTS

Kris Hermans, Werner Verhelst and Patrick Wambacq
Katholieke Universiteit Leuven - ESAT/PSI
Kardinaal Mercierlaan 94, B-3001 Leuven, Belgium

e-mail: Kris.Hermans@esat.kuleuven.ac.be

ABSTRACT
Signal Subspace (SS) based speech enhancement techniques obtain significant additive-noise reduction by altering the singular value spectrum of the speech observation matrix. Among the class of different possible SS weighting strategies, the Minimum Variance (MV) estimation method substantially increases the speech recognition accuracy in additive noise environments, outperforming the widely used Spectral Subtraction methods. However, these SS approaches are developed as pure speech enhancement techniques, and it is still unknown how effective they are for noise robust speech recognition. In this respect, we present the idea of ‘optimal SS weighting’ for speech recognition systems, and we illustrate in detail that the MV estimation closely approximates this optimum. We applied the SS weighting methods to a LV-CSR task with noisy data (10 dB SNR), and obtained relative reductions in Word Error Rate of more than 60%.

1. INTRODUCTION
With the increase in computing power and the advances in speech processing, current automatic speech recognizers (ASR) have attained a high level of accuracy. Large vocabulary continuous speech can fluently be transcribed with WERs of less than 10% as long as high quality speech can be assured.
As soon as the input speech signal is corrupted by disturbing sources, severe performance degradation can occur, due to the mismatch between training (laboratory) and operating (test) conditions. If no noise compensation technique is applied, many speech recognizers become completely useless in realistic operating conditions.
Many noise compensation techniques exist, and are mostly based on one of the following strategies [3]:
- match the training and test conditions by enhancing the input signal (speech enhancement techniques)
- make the recognizer a noise-independent system by deriving noise resistant features in the preprocessing [in this case the noise is not necessarily removed]
- adapt the acoustic models to the noise characteristics from the input signal (model compensation techniques)

However, many of the existing techniques for enhanced robustness still suffer from the need to accurately model the noise conditions, from erroneous noise estimation, from signal distortion (e.g. musical noise), from time consuming training,...

Recently, robust speech recognition in additive noise environments has been achieved with Signal Subspace (SS) based approaches. These techniques - previously successfully applied to signal enhancement problems [7] - aim at removing additive noise by altering the singular value spectrum of the noisy-signal observation matrix. Low energetic, noise related SVD components receive a low weight whereas high energetic components are supposed to contain almost only pure signal and are given the highest weights.

Hansen et al. [4] discussed several estimation criteria (Least Squares, Minimum Variance, Time Domain Constrained, Spectral Domain Constrained) with their corresponding weighting schemes that yield effective noise reduction for speech enhancement purposes. Huang and Zhao [6] reported remarkable speech recognition improvements with a modified Minimum Variance based SS estimation approach. In [5] the SS based methods were reported to outperform the well-known (Nonlinear) Spectral Subtraction method [1].

In this paper, we give the basics of SS weighting theory, with their most important estimation criteria. We then introduce the concept of 'optimal' SS weighting in terms of maximal recognition accuracy. This concept is used to study the performance of the MV SS estimation method. It is found that this method can be considered a near optimal weighting procedure if applied to improve the accuracy of LV-CSR in additive noise environments.

2. THEORY OF SS WEIGHTING
Signal Subspace based speech enhancement algorithms are based on the assumption that the vector space of the noisy input signal can be split in mutually orthogonal noise and signal + noise subspaces.
The spectral components in the noise subspace are suppressed (or removed completely). From the remaining signal + noise subspace, one can estimate the clean signal by changing the weight of the spectral components such that the noise is filtered out.

2.1. Algorithm
The noisy input signal $y$ consists of the desired signal $x$ and the noise $n$; both components are considered additive:

$$y[k] = x[k] + n[k]$$

The SS based noise reduction operates on a frame-by-frame basis. Consider a frame with $T$ samples: $y[k]$, $k = 0 \ldots (T-1)$. Noise removal is then obtained as follows: Constructing the Hankel matrix. First construct a Hankel matrix $Y$:

$$Y = \begin{bmatrix}
    y[0] & y[1] & \cdots & y[M-1] \\
    \vdots & \vdots & \ddots & \vdots \\
y[K-1] & y[K] & \cdots & y[T-1]
\end{bmatrix}$$

Support is acknowledged from the Flemish Community - IWT and from the Research Fund K.U. Leuven.
with dimensions $K \times M$ ($K \geq M$), and $M + K = T + 1$. According to the assumption of additive noise, we can write $Y$ as:

$$Y = X + N$$

(3)

**SVD calculation** Calculate the SVD of $Y$:

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

(4)

**Altering the singular value spectrum** The largest singular components capture almost all signal information whereas the smallest ones contain almost only noise. By adapting the weights of the different singular components, noise reduction is achieved:

$$\hat{X} = U(W_\Sigma)V^T$$

(5)

with $W$ a diagonal matrix containing the weights.

**Restoring the Hankel structure** Matrix $X$ is not Hankel anymore; an easy extraction of the improved signal $\hat{x} = \hat{x}(0), \hat{x}(1), \ldots, \hat{x}(T - 1)$ is impossible. However, the Hankel structure can be restored by constructing a new matrix $\hat{X}$ where every element from an anti-diagonal of $X$ is replaced by the average value along that anti-diagonal.

### 2.2. FIR-filter Representation

The overall procedure of altering the singular value spectrum is equivalent to a FIR-filtering operation on the noisy signal $y(k)$. In fact, the spectrum of $y(k)$ is decomposed into $M$ components, from which the weight is changed according to some estimation criterion.

### 3. SS ESTIMATION METHODS

Different SS estimation criteria can be used, each with a corresponding weighting matrix $W$.

It is assumed that the clean signal and the noise are uncorrelated:

$$X^T N = 0$$

(6)

and that the noise is white:

$$N^T N = \sigma_{\text{noise}}^2 I$$

(7)

### 3.1. Least Squares (LS)

Assume that $x$ consists of $p$ complex exponentials (this is a model that is often attributed to speech), such that $X$ is of rank $p$. The Least Squares estimate of $X$ is then the matrix of rank $p$ that is closest — in LS sense — to $Y$:

$$\|\hat{X}_{LS} - Y\|_{\text{min}}$$

(8)

This LS estimate is obtained by setting the $M - p$ smallest eigenvalues of $Y$ to zero,

$$\hat{X}_{LS} = \left[ U_p \ U_{M-p} \right] \left[ \Sigma_p \ 0 \ \right] \left[ \begin{array}{c} V_p^T \\ V_{M-p}^T \end{array} \right] = U_p \Sigma_p V_p^T$$

(9)

with $\Sigma_p$ containing the $p$ largest singular values, and $\hat{X}_{LS,p}$ the best rank-$p$ approximation of $Y$.

**Discussion** Among the different weighting methods, the LS estimate contains the highest possible residual noise level, namely $(p/M)\sigma_{\text{noise}}$ (only the noise from the noise subspace is filtered out), but has the lowest signal distortion (the signal + noise subspace is untouched).

The Least Squares estimation has a major drawback. It results in selecting some singular components, and suppressing the other components (binary approach). This often causes gaps to occur in the frequency spectrum of $x$ and results in disappointing recognition results.

### 3.2. Minimum Variance (MV)

The Minimum Variance method gives the optimal linear estimate of $x$. More precisely, it finds the $M \times M$ matrix $H$ that minimizes

$$\left\| YH - X \right\|_{X_{MV}}$$

(10)

Matrix theory states that the MV estimate $\hat{X}_{MV}$ of $X$ is the orthogonal projection of $X$ onto the column space of $Y$.

It can be proven [7] that $\hat{X}_{MV}$ is found by multiplying the singular values $\sigma_i$ of $Y$ by

$$w_i = 1 - \frac{\sigma_i^2}{\sigma_{\text{noise}}^2}$$

(11)

where $\sigma_{\text{noise}}^2$ is estimated based on speechless frames.

The weighting matrix $W$ becomes

$$W = \text{diag} \left( 1 - \frac{\sigma_1^2}{\sigma_{\text{noise}}^2}, \ldots, 1 - \frac{\sigma_p^2}{\sigma_{\text{noise}}^2} \right)$$

(12)

**Discussion** A detailed analysis of this interesting estimation method will be given in section 4.

### 3.3. Time/Spectral Domain Constrained

The Time Domain Constrained (TDC) and Spectral Domain Constrained (SDC) estimators keep the residual noise power below an adjustable threshold with minimized signal distortion. Both methods are rather robust against erroneous estimation of the noise threshold $\sigma_{\text{noise}}$. Speech recognition results were rather poor; this could be due to the substantial level of residual noise in the enhanced speech.

The weighting functions for TDC and SDC estimation can be found in [4].

### 4. ANALYSIS OF THE MV METHOD

#### 4.1. Motivation

Signal Subspace based estimation techniques, such as the MV approach, are designed in the frame work of speech enhancement. Hence, these techniques are mostly evaluated according to the obtained increase in SNR and/or the improvement in perceptual speech quality. However, in this paper we are concerned with optimal speech recognition accuracy. In this case, a particular SS weighting method can be considered optimal (within the class of all possible weighting matrices $W$) provided that it enhances the speech signal in such a way that the highest possible increase in recognition accuracy is obtained.

Of course, signal quality and recognition accuracy are related, in that a high SNR will normally result in good recognition and vice-versa. But, ASR systems are very sensitive to a mismatch between the training and test conditions. Speech enhancement techniques must not destroy information that is extracted into
the acoustic parameters and essentially needed in the decoding process for optimal discriminative power between the different sounds in the phoneme set. This is for instance why the LS estimate, that often destroys the format structure of speech (by removing some spectral components) leads to very poor results.

Before we analyze the MV approach, we first describe the main aspects of our experimental scheme.

4.2. Experimental Scheme

Experiments were carried out on the Resource Management task (for 39 test set: continuous speech, 1K words, Word Pair grammar, Context-Independent models) with the ESAT speech recognizer. For a description of the ESAT speech recognizer, the reader is referred to [2].

The SS weighting scheme precedes the normal MEL-cepstrum preprocessing. The frame-length and the order of the Hankel matrix are two important parameters that have to be optimized.

Frame Length The SS noise removal can be seen as an adaptive filtering \( \{ \text{with constant filter characteristics during one frame} \} \).

Hence, the shorter the frame length, the faster the adaptation to changing noise characteristics. Via the relation between \( T \) and the matrix dimensions \( K \) and \( M = T + M - 1 \), a minimal value for \( T \) can be derived [see below].

A frame-length of 30 msec (480 samples at 16 k Hz sampling rate) is a good compromise.

Order \( M \) The order \( M \) of the Hankel matrix is not very critical in the case of MV [7]. It must not be smaller than the order \( \{=p \} \) of the speech observation matrix and should not be too large to avoid excessive computation times.

During experiments, an order \( M = 8 \) was found to be optimal.

Recognition results A noisy database was created by adding additive white noise at 10 dB global SNR to the clean RM-data. The Word Error Rate (WER) for the clean dataset is 4.88\%, and 57.17\% for the noisy dataset.

The noise reduction with the Minimum Variance estimation method reduces the WER to 17.96\%.

4.3. The Copy-Spectrum Method

The reduction in WER obtained by the MV estimator is significant, but the optimum result is still far away, 17.96\% + 4.88\%. It is interesting to have a kind of upper bound that can be reached by the SS weighting methods.

Consider the Hankel matrices \( X \) and \( Y \), constructed from the clean signal \( x \) and the noisy signal \( y = x + n \), respectively. We calculated the SVD of \( X \) and \( Y \)

\[
X = U \Sigma V^T
\]

(13)

\[
Y = \hat{U} \Sigma \hat{V}^T
\]

(14)

The singular value spectrum of \( Y \) is strongly related to that of \( X \). In fact, \( \Sigma \) is - under the assumptions stated in eq. (6) and (7) - a noise- perturbed version of \( \Sigma \).

Each SS method will calculate an estimate \( \hat{X} \) of \( X \) by altering the singular value spectrum of \( Y \) with a (estimation-method dependent) weight matrix \( W \):

\[
\hat{X} = \hat{U} (W \Sigma) \hat{V}^T
\]

(15)

As an experiment, we consider the case that we know the singular value spectrum \( \Sigma \) of \( X \), and we use it to make an 'optimal' estimate of \( \hat{X} \) by copying this known spectrum into the SVD of \( Y \):

\[
\hat{X}_{opt} = \hat{U} \Sigma V^T
\]

(16)

We applied this copy-spectrum strategy on the testset (remember that we have both noisy and clean data), and obtained a WER of 13.32\%, which is indeed better than the 17.96\% WER for the MV method. The difference (17.96 - 13.32) is however smaller than one may expect. This is in favor of the MV method that we already found to be very powerful [5].

4.4. Visualization of the Weighting Factors

From the difference in recognition accuracy between the optimal weighting (copy-spectrum) and the MV weighting method, we learn that there must be a difference in the corresponding weighting factors also. But, how and how much do they differ?

From the test database, we select a clean sentence and its noise corrupted version (10 dB global SNR).

From all frames of these sentences, we calculate the clean singular values \( \sigma_i \) and the noisy ones \( \hat{\sigma}_i \). From the noisy sentence, we extract the noise threshold \( \sigma^2_{\text{noise}} \) which has the same value in every frame (stationary noise).

We now have a pool of matched pairs \( \{ \sigma_i, \hat{\sigma}_i \} \), for which the optimal weights \( w_{opt} \):

\[
w_{opt} = \sigma_i / \hat{\sigma}_i
\]

(17)

and the MV weights \( w_{MV} \):

\[
w_{MV} = 1 - \frac{\sigma^2_{\text{noise}}}{\hat{\sigma}_i}
\]

(18)

for \( \hat{\sigma}_i \) can be calculated.

These weights are plotted in figure 1 as a function of \( \frac{\sigma^2_{\text{noise}}}{\hat{\sigma}_i} \).

![Figure 1: Optimal weighting factors (dots), MV weighting (solid) and fitted MV weighting (dashed) as function of \( \frac{\sigma^2_{\text{noise}}}{\hat{\sigma}_i} \).](image)

We notice the following:

**Negative weights** If \( \chi^T N = 0 \) (eq. 6) and \( N^T N = \sigma^2_{\text{noise}} I \) (eq. 7), it can be proven that \( \hat{\sigma}_i > \sigma^2_{\text{noise}} \) for all \( \sigma_i \). In practice these conditions are not always fulfilled, such that a number of \( \hat{\sigma}_i < \sigma^2_{\text{noise}} \) which leads to negative MV weights. This could be avoided by working with

\[
w_{MV} = \max \left( 0, 1 - \frac{\sigma^2_{\text{noise}}}{\hat{\sigma}_i} \right)
\]

(19)

Recognition experiments pointed out that this is not a good idea (higher WER). We think that the zero weights introduce
too much signal distortion (e.g. format destruction as we reported for the LS estimation). A better idea is to replace the 0 in eq. 19 by a threshold \( \delta \).

**Underfitting** There is no overlap between the MV weights and the optimal weights. The MV weights are rather located under the cloud of optimal weights, such that the singular components are more suppressed by the MV method than seems to be necessary. In fact, we notice that almost no weights are overestimated. This may indicate that it is a good strategy to slightly overestimate the noise level in order to obtain sufficient noise reduction.

**Distribution** For the majority of the noisy singular values holds that \( \tilde{\sigma}_i \approx 1 \). Consequently, a proper determination of the weights for the \( \tilde{\sigma}_i \)’s in this region is critical to obtain a good speech enhancement.

### 4.5. Fitting of the MV Weights

It is possible to verify the importance of underfitting by generalizing the MV weighting into

\[
\omega_{MV} = 1 - \epsilon \left( \frac{\sigma_{\text{noise}}}{\tilde{\sigma}_i} \right) ^ \gamma
\]

(20)

The parameters \( \epsilon \) and \( \gamma \) are found by fitting the \( \omega_{opt} \)’s with the curve \( \omega_{MV} \) in LS sense\(^3\). For the data in figure 1, we find \( \epsilon = 0.74 \) and \( \gamma = 2.21 \). The resulting \( \omega_{MV} \) weighting curve is plotted in figure 1 as a dashed line\(^4\).

We performed the same MV fitting procedure on each of the 300 sentences in the RM dataset. The values of \( \epsilon \) and \( \gamma \) are given in figure 2. The mean values are indicated by the dashed lines. We notice that \( \epsilon \) and \( \gamma \) vary from sentence to sentence but that their mean values clearly differ from their standard MV counterparts, namely \( \epsilon = 1 \) and \( \gamma = 2 \).

Informal listening tests pointed out that the residual noise is clearly audible. This is due to the presence of weights \( \omega_i \) that were overestimated, especially in the region \( \tilde{\sigma}_i \approx \sigma_{\text{noise}} \).

A recognition experiment with this fitted MV weighting method yields a WER of 28.15% which is - as could be expected - significantly worse than the WER of 17.96% obtained with the standard MV method.

\(^3\)This problem transforms into a single linear LS fit in the log-log domain.

\(^4\)Note that the fit is best for \( \tilde{\sigma}_i \approx \sigma_{\text{noise}} \) since the majority of the points are located there. Since noise reduction is mainly achieved in this region, we did not want to equalize the importance of the range of \( \tilde{\sigma}_i \).

### 4.6. Experimental Fitting of \( \epsilon \) and \( \gamma \)

The recognition experiments with the generalized MV method seem to confirm our expectation that the standard MV estimation can be considered to be a near optimal SS based estimation algorithm for robust speech recognition.

We did an additional check by experimentally adjusting \( \epsilon \) and \( \gamma \) to reach a maximal recognition accuracy. After a series of recognition experiments we found \( \{ \epsilon, \gamma \} = \{0.98, 1.40\} \) to give the lowest WER, namely 16.32%.

This result is not significantly better than the reference WER of 17.96%, such that we can conclude that the standard MV method is indeed near optimal.

### 5. CONCLUSION

Signal Subspace based speech enhancement techniques can be applied as a powerful tool to enhance the robustness of speech recognizers in additive noise conditions.

All existing noise estimation criteria essentially obtain noise reduction by altering the singular value spectrum of the signal observation matrix. We studied in detail the most effective variant, namely the Minimum Variance (MV) estimator.

We compared this MV approach with our 'optimal' weighting scheme, that was trained with known matched pairs of clean and noisy singular spectra. We conclude that, for robust noise recognition, the MV based weighting can be considered as near-optimal in the class of SS weighting methods.

When combined with an accurate speech-noise classifier and a prewhitening step, the basic SS based algorithms are able to remove non-stationary colored additive noise.

### 6. REFERENCES