A SOFT MODEL-ORDER SUBSPACE BASED SPEECH ENHANCEMENT ALGORITHM

Jesper Jensen and Richard Heusdens

{J.Jensen, R.Heusdens}@ewi.tudelft.nl
Dept. of Mediamatics
Delft University of Technology
Delft, The Netherlands.

ABSTRACT

Subspace based schemes for noise suppression rely on decomposition of the noisy signal space into a signal (+noise) subspace and a noise subspace. Existing schemes are derived under the assumption that the dimensions of these subspaces are known with certainty, although in practice they must be inferred from the noisy data. We derive in this paper a subspace based enhancement scheme which takes into account the uncertainty of the subspace dimensions. The structure of the resulting estimator turns out to be similar to existing schemes, but the gain functions applied to the components in the signal subspace are now dependent on the likelihood that the signal subspace has a certain dimension. Simulation experiments with speech signals degraded by synthetic and natural noise sources show performance improvements of 0.3–1.6 dB in segmental SNR as compared to traditional schemes.

1. INTRODUCTION

The last decades have seen an explosion in the use of mobile digital voice communications systems, and consequently, the need for such systems to work in acoustically noisy environments. Since most such systems have been designed to work well with relatively noise-free input speech signals, their performance deteriorate when the input signals are noisy. A possible way to overcome this problem is to reduce the noise content in the noisy speech signal using a speech enhancement pre-processing step, and then apply the noise-reduced speech signal as input to the communication chain.

Traditional approaches for single-channel noise reduction include short-time Fourier transform (STFT) based methods, e.g. [1, 2], model based methods which to a larger extent try to exploit apriori speech production knowledge, e.g. [3, 4], and subspace based approaches [5, 6] which rely on vector space decomposition techniques.

The subspace based approaches exploit the idea of decomposing the covariance matrix of the noisy speech signal into two mutually orthogonal vector spaces, a signal (+noise) subspace and a noise-only subspace. Noise reduction is obtained by discarding the noise subspace completely, while modifying the noisy speech components in the signal subspace. Although essentially formulated for the white noise case, it is straightforward to apply the subspace based techniques in the case of coloured noise; this requires a pre-whitening transformation of the noisy input signal prior to enhancement, and a de-whitening step after enhancement [5, 6, 7]. Further, extensions have been presented which take into account the perceptual effects of the human auditory system, e.g. [8].

A main assumption in subspace based enhancement techniques is the existance of a lower-dimensional signal subspace. Clearly, this assumption is signal dependent. For example, for signal segments which can be approximated well by a sum of, say K, (exponentially damped) constant-frequency sinusoids such as some voiced speech segments, the corresponding covariance matrix is essentially of rank $2K$, and the signal subspace may be approximated as $2K$-dimensional. Further, since in voiced speech regions the signal-to-noise ratio (SNR) is often high, the signal subspace dimension may be readily established from a ‘gap’ in the eigenvalue spectrum of the covariance matrix of the noisy signal. However, for other speech sounds, e.g. unvoiced, transients, and transitional regions, the situation is more difficult. First, the definition of a signal subspace dimension is less clear because the eigenvalues of the corresponding covariance matrix may all be significantly larger than zero. Secondly, estimation of the signal subspace dimension is harder because the SNR for these speech sounds is often significantly lower than in the voiced case discussed above.

Existing subspace based enhancement schemes [5, 6, 9] derive estimators under the assumption that the signal subspace dimension is surely known. However, motivated by the discussion above we note that this assumption is not always valid. Therefore, we present in this paper an estimator which takes into account the uncertainty of the signal subspace dimension. The derived estimator is simply a linear combination of estimators for different assumed

This research was partly supported by Philips Research and the Technology Foundation STW, applied science division of NWO and the technology programme of the ministry of Economics Affairs.

subspace dimensions, weighted by their probability of occurence. Thus, rather than applying a hard decision on the signal subspace dimension, the proposed algorithm uses a 'soft decision' criterion.

2. SUBSPACE BASED ENHANCEMENT USING A SOFT MODEL ORDER

As in [5] we consider a signal model of the form

$$x = s + w, \text{ where } s = Va, \tag{1}$$

and $x \in \mathbb{R}^N$ denotes an observed noisy speech signal vector, $s \in \mathbb{R}^N$ denotes the clean speech signal and $w \in \mathbb{R}^N$ is an additive noise vector. The matrix $V \in \mathbb{C}^{N \times K}, K \leq N$, contains basis vectors as columns, and $a \in \mathbb{C}^K$ is a vector of zero mean random variables. We assume that the clean speech and noise processes are uncorrelated and that s and w are Gaussian distributed with probability density functions (pdfs) $f_S(s) = \mathcal{N}(0, R_S)$ and $f(w) = \mathcal{N}(0, R_w)$, where R_s and R_w denote the covariance matrices of s and w, respectively.

We consider the conditional mean estimator given by

$$\hat{s} \triangleq E(s|x) = \int_s s f(s|x) ds, \tag{2}$$

where $E(\cdot)$ denotes the statistical expectation operator, and $f(s|x)$ is the pdf of the clean signal vector conditioned on the noisy observation. Since we are going to estimate s in a subspace framework, we represent the signal subspace dimension by the integer-valued random variable m and rewrite Eq. (2) as

$$\hat{s} = \sum_m \int_s s f(s|m,x) f(m|x) ds = \sum_m f(m|x) \hat{s}_m, \tag{3}$$

where $\hat{s}_m \triangleq E(s|x,m)$ and the density $f(m|x)$ is in fact a probability mass function (pmf) because the signal subspace dimension m is an integer-valued random variable. We see from Eq. (3) that $\hat{s}$ is a linear combination of estimators $\hat{s}_m$ for different values of m, weighted by the likelihood that the signal subspace has dimension m. In the following sections we derive expressions for $\hat{s}_m$ and $f(m|x)$ used in Eq. (3).

2.1. Estimation of $\hat{s}_m$

Under our Gaussian assumptions, the conditional mean estimator is identical to the linear minimum mean-squared error estimator [10], i.e. $\hat{s}_m$ can also be found from

$$\hat{s}_m = H_m x, \tag{4}$$

where $H_m \in \mathbb{R}^{N \times N}$ is given by

$$H_m = \arg\min_H E\|s - Hx\|_2^2, \text{ s.t. } rank(H) = m.$$

The rank constraint on H ensures that the covariance matrix of $\hat{s}_m$ has the prescribed rank of m.

The procedure for estimating H_m is well-known and can be derived from e.g. [5, 6, 9]. To facilitate our further discussion, we simply state the solution here. Let us, without loss of generality, assume that the noise is white, i.e., $R_w \triangleq E(ww^T) = \sigma_w^2 I$, where $(\cdot)^T$ denotes vector transposition, σ_w^2 is the noise variance, and I is the identity operator in $\mathbb{R}^N$. Further, let $R_x = U\Lambda_x U^T$ denote the eigenvalue decomposition (EVD) of the covariance matrix R_x of the noisy signal; the unitary matrix $U \in \mathbb{R}^{N \times N}$ contains the eigenvectors as columns, while the diagonal matrix Λ_x contains the eigenvalues $\lambda_{x_i}, i = 1, \ldots, N$, in descending order. We partition U as $U = [U_1 \; U_2]$, where $U_1 \in \mathbb{R}^{N \times m}$ is an m-dimensional orthonormal basis for the (assumed) signal subspace, while $U_2 \in \mathbb{R}^{N \times (N-m)}$ constitutes a basis for the corresponding noise subspace. It can then be shown that

$$H_m = U_1 G_m U_1^T, \tag{5}$$

and

$$G_m = I - diag(\sigma_w^2/\lambda_{x_1}, \ldots \sigma_w^2/\lambda_{x_m}).$$

2.2. Estimation of $f(m|x)$

In order to estimate the posterior density $f(m|x)$ in Eq. (3) we rewrite it using Bayes rule

$$f(m|x) = \frac{f(x|m)f(m)}{f(x)}. \tag{6}$$

Here, $f(x)$ is independent of m and can be seen as a constant which ensures that the posterior integrates to one. The pmf $f(m)$ reflects a priori knowledge of observed signal subspace dimensions; in lack of any such knowledge we choose a uniform prior, $f(m) = 1/N, m = 1, \ldots, N$. In order to derive an expression for the likelihood $f(x|m)$, we first note that since the columns of the matrix U_1 form a basis for the signal subspace, we have

$$s = U_1 a',$$

where $a' \in \mathbb{R}^m$. We assume that U_1 is known (although it is estimated from the data) and express $f(x|m)$[1] as

$$f(x|m) = \int_{a'} f(x|a', m) f(a'|m) da'. \tag{7}$$

Under the Gaussian noise assumption, the density $f(x|a', m)$ is given by

$$\begin{aligned} f(x|a', m) =& (2\pi\sigma_w^2)^{-N/2} \times \\ & \exp(-\frac{1}{2\sigma_w^2}(x - U_1 a')^T (x - U_1 a')). \end{aligned} \tag{8}$$

[1] Having assumed U_1 to be given, the pdf $f(x|m)$ is, strictly speaking, conditioned on U_1

Using that $f(s) = \mathcal{N}(0, R_s)$, it follows that $f(a'|m) = \mathcal{N}(0, \Lambda_{s,m})$, i.e.

$$\begin{aligned} f(a'|m) =& (2\pi)^{-m/2} |\Lambda_{s,m}|^{-1/2} \times \\ & \exp(-\frac{1}{2} a'^T \Lambda_{s,m}^{-1} a'), \end{aligned} \tag{9}$$

where $|\cdot|$ denotes the matrix determinant, and $\Lambda_{s,m} \in \mathbb{R}^{m \times m}$ is a diagonal matrix containing the m largest eigenvalues of R_s. Inserting Eqs. (8) and (9) in Eq. (7) and using the fact that [11]

$$\begin{aligned} &\int_y \exp(-\frac{1}{2}(d + b^T y + y^T C y)) dy \\ &= (2\pi)^{m/2} |C|^{-1/2} \exp(-\frac{1}{2}(d - \frac{b^T C^{-1} b}{4})), \end{aligned}$$

where $d \in \mathbb{R}, y, b \in \mathbb{R}^l$, and $C \in \mathbb{R}^{l \times l}$, we obtain the following closed-form expression

$$\begin{aligned} f(x|m) =& (2\pi\sigma_w^2)^{-(N-m)/2} (2\pi)^{-m/2} |\Lambda_{x,m}|^{-1/2} \times \\ & \exp(-\frac{1}{2\sigma_w^2}(x^T x - x^T U_1 D U_1^T x)), \end{aligned} \tag{10}$$

where D is a diagonal matrix given by $D = I - \sigma_w^2 \Lambda_{x,m}^{-1}$, and $\Lambda_{x,m}$ contains the m largest eigenvalues of R_x on the main diagonal.

2.3. Algorithm summary

For a given noisy signal frame x, we estimate the corresponding covariance matrix R_x and compute its eigenvalue decomposition $R_x = U \Lambda_x U^T$. Assuming knowledge of the noise variance σ_w^2, we insert Eq. (10) in Eq. (6) and evaluate the expression for the possible signal subspace dimensions $m = 1, \ldots, N$. Also, using the EVD of R_x we compute the estimates of $\hat{s}_m$ in Eq. (4). Finally, we use Eq. (3) to find the clean signal estimate $\hat{s}$.

It is interesting to use Eqs. (4) and (5) to rewrite Eq. (3):

$$\hat{s}_m = \sum_{i=1}^{m} u_i g_i u_i^T x, \tag{11}$$

where u_i is the i'th column of U and g_i is the i'th diagonal element of G_m. Inserting Eq. (11) in Eq. (3) gives

$$\begin{aligned} \hat{s} &= \sum_{m=1}^{N} \sum_{i=1}^{m} u_i g_i f(m|x) u_i^T x \\ &= \sum_{i=1}^{N} \sum_{m=i}^{N} u_i g_i f(m|x) u_i^T x \\ &= U G' U^T x, \end{aligned}$$

where $G' = diag(g'_1, \ldots, g'_N)$, and $g'_i = g_i \sum_{m=i}^{N} f(m|x)$.

We conclude that $\hat{s}_m, m = 1, \ldots, N$ need not be computed explicitly, because the derived estimator has the same structure as the original estimator in Eq. (4), but the gain factors are modified in accordance with the probabilities of observing a given signal subspace dimension.

3. SIMULATION RESULTS

We evaluate the presented algorithm in simulation experiments with 48 speech signal excerpts (8 different speakers, 4 female and 4 male) with a sampling frequency of 8 kHz and a duration of 4-5 seconds each. We construct noisy speech signals by adding synthetic and natural noise sources to the clean speech signals at SNR levels of 20, 10, and 0 dB. The noise sources are taken from the Noisex data base [12] and encompass white noise, pink noise, car interior noise, and F16 cockpit noise. In all experiments, the noise power spectral density, which is roughly constant across time, is estimated from a noise-only signal region of approximately 350 ms preceding speech activity. Signal frames x of length $N = 60$ samples are taken from the noisy signal with an overlap of 50%. We estimate noisy covariance matrices R_x from segments of 300 samples, centered around the frame to be enhanced. Then the noisy signal frames are pre-whitened using the procedure outlined in [5], enhanced using different versions of the subspace based enhancement scheme described below, and de-whitened. Subsequently, the enhanced signal frames are overlap-added using a Hanning window.

We compare the proposed soft-decision based estimator with two schemes which rely on a hard signal subspace dimension. Method 1 is simply the scheme outlined in [5], where the dimension of the noise subspace of R_x is increased in steps of 1, until the largest eigenvalue in the noise subspace becomes significantly larger than the smallest eigenvalue, in which case the process is terminated. We also considered another often used scheme, where the dimension of the signal subspace is determined as the number of eigenvalues of R_x larger than σ_w^2, i.e., $\hat{s} = \hat{s}_{m^*}$ where m^* is the cardinality of the set $\{\lambda_{x_i} : \lambda_{x_i} > \sigma_w^2\}$. However, this scheme gave results which were essentially identical to those of Method 1, and has therefore not been included here. Method 2 is a slightly simplified version of the proposed algorithm where the signal subspace dimension with the largest probability $f(m|x)$ is chosen, i.e., $\hat{s} = \hat{s}_{m^*}$ with $m^* = \arg\max_m f(m|x)$. Finally, Method 3 is the proposed soft-decision scheme.

To evaluate and compare the performance of the methods, we apply the segmental SNR (Seg-SNR) defined as the average SNR, computed across each enhanced frame in the entire set of input signals. Tables 1–4 show results for different noise sources. We see that the proposed

White Noise	Input SNR [dB]		
	0	10	20
Noisy	-13.64	-3.64	6.36
Meth.1	-5.14	2.76	10.50
Meth.2	-4.65	3.12	10.76
Meth.3	-4.52	3.22	10.85

Table 1: Segmental SNR [dB], white Gaussian noise.

Pink Noise	Input SNR [dB]		
	0	10	20
Noisy	-12.30	-2.30	7.70
Meth.1	-4.32	3.38	11.39
Meth.2	-3.92	3.66	11.62
Meth.3	-3.75	3.80	11.73

Table 2: Segmental SNR [dB], pink noise.

method improves performance in all cases. More specifically, Method 3 increases the Seg-SNR with 0.3–1.6 dB as compared to Method 1. Informal listening tests confirm that Method 3 suppresses the noise better, an observation which is particularly clear in low SNR regions.

Also Method 2 performs better than Method 1, in fact, it achieves performance close to that of Method 3. Comparing Methods 2 and 3 shows that Method 2 rejects more noise in noise-only regions, while Method 3 is better in speech regions. However, signals enhanced with Method 2 contain disturbing 'switching effects' in transitions between speech presence and absence. Observing further that Method 2 is of similar computational complexity as Method 3, makes Method 3 the prefered scheme.

4. REFERENCES

[1] Y. Ephraim and D. Malah, "Speech enhancement using a minimum mean-square error short-time spectral amplitude estimator," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-32, no. 6, pp. 1109–1121, December 1984.

[2] R. Martin and C. Breithaupt, "Speech enhancement in the dft domain using laplacian speech priors," in *Int. Workshop, Acoustic Echo and Noise Control*, Kyoto, Japan, September 2003, pp. 87–90.

[3] Y. Ephraim, "Statistical-model-based speech enhancement systems," *Proc. IEEE*, vol. 80, no. 10, pp. 1526 – 1555, Oct 1993.

[4] J. H. L. Hansen and M. A. Clements, "Constrained iterative speech enhancement with application to speech recognition," *IEEE Trans. Speech, Audio Processing*, vol. 39, no. 4, pp. 795–805, April 1991.

Car Noise	Input SNR [dB]		
	0	10	20
Noisy	-5.30	4.70	14.70
Meth.1	4.19	12.64	20.95
Meth.2	5.22	13.13	21.00
Meth.3	5.78	13.72	21.61

Table 3: Segmental SNR [dB], car noise.

F16 Noise	Input SNR [dB]		
	0	10	20
Noisy	-13.03	-3.03	6.97
Meth.1	-5.14	2.70	10.73
Meth.2	-4.66	3.09	11.02
Meth.3	-4.54	3.18	11.11

Table 4: Segmental SNR [dB], Cockpit noise.

[5] Y. Ephraim and H. L Van Trees, "A signal subspace approach for speech enhancement," *IEEE Trans. Speech, Audio Processing*, vol. 3, no. 4, pp. 251–265, July 1995.

[6] S. H. Jensen et al., "Reduction of broad-band noise in speech by truncated qsvd," *IEEE Trans. Speech, Audio Processing*, vol. 3, no. 6, pp. 439–448, November 1995.

[7] H. Lev-Ari and Y. Ephraim, "Extension of the signal subspace speech enhancement approach to colored noise," *IEEE Signal Processing Lett.*, vol. 10, no. 4, pp. 104–106, April 2003.

[8] F. Jabloun and B. Champagne, "Incorporating the human hearing properties in the signal subspace approach for speech enhancement," *IEEE Trans. Speech, Audio Processing*, vol. 11, no. 6, pp. 700–708, November 2003.

[9] B. de Moor, "The singular value decomposition and long and short spaces of noisy matrices," *IEEE Trans. Signal Processing*, vol. 41, no. 9, pp. 2826–2838, September 1993.

[10] C. W. Therrien, *Discrete Random Signals and Statistical Signal Processing*, Prentice-Hall International, Inc., 1992.

[11] I. S. Gradshteyn and I. M. Ryzhik, *Table of Integrals, Series, and Products*, Academic Press, Inc., 1965.

[12] A. Varga and H. J. M. Steeneken, "Noisex-92: A database and an experiment to study the effect of additive noise on speech recognition systems," *Speech Commun.*, vol. 12, no. 3, pp. 247–253, 1993.