SUBSPACE STATE SPACE MODEL IDENTIFICATION FOR SPEECH ENHANCEMENT

Eric Grivel, Marcel Gabrea and Mohamel Najim

Equipe Signal et Image, ENSERB and GDR-ISIS, CNRS
B.P. 99
F- 33402 Talence Cedex, France
e-mail : najim@goelette.tsi.u-bordeaux.fr

ABSTRACT

This paper deals with Kalman filter-based enhancement of a speech signal contaminated by a white noise, using a single microphone system. Such a problem can be stated as a realization issue in the framework of identification. For such a purpose we propose to identify the state space model by using subspace non-iterative algorithms based on orthogonal projections. Unlike Estimate-Maximize (EM)-based algorithms, this approach provides, in a single iteration from noisy observations, the matrices related to state space model and the covariance matrices that are necessary to perform Kalman filtering. In addition no voice activity detector is required unlike existing methods. Both methods proposed here are compared with classical approaches.

1. INTRODUCTION

So far, classical adaptive noise cancellation devices have used two microphones. But speech enhancement using a single microphone has become an active research issue.

Given a sequence of speech signal corrupted by an additive white noise, our purpose is to retrieve the speech signal. This problem occurs especially in free hand mobile phones and teleconferences.

Various approaches based on Kalman filtering have been referenced in the literature. They usually operate in two steps:

1. first, noise and driving process variances and speech model parameters are estimated;
2. second, the speech signal is estimated by using Kalman filtering.

In fact these approaches essentially differ in the way of choosing speech model and estimating speech model parameters and noise variances. In [1] the estimated speech parameters are obtained from the clean speech, before being contaminated by white noise. Then the authors use a delayed version of Kalman filter in order to estimate speech signal. In [2] the proposed method provides a sub-optimal solution which is a simplified version of the Estimate-Maximize (EM) algorithm based on the maximum likelihood argument. However noise variance is estimated during silent period, which implies the use of voice activity detector. Furthermore the estimation of the driving process covariance requires the estimation of the observation correlation function. In [3] the authors propose a solution to these various problems, that especially occur for EM-based algorithms, by employing the Kalman EM Iterative (KEMI) algorithm. In [4], [5] an alternative approach is proposed. Indeed Kalman gain calculation is computed without an explicit estimation of noise and driving process variances. But the AR parameters are estimated by solving the modified Yule-Walker equations, which needs the estimate of the observation autocorrelation.

In this paper a new approach is presented. Indeed speech enhancement process can be stated as a realization problem in the framework of state space representation. Thus we propose to enhance speech signal by using Kalman filtering and state space system identification methods introduced by Van Overschee [6], [7]. These methods are based on the concept of orthogonal projections and use non-iterative subspace algorithms. This approach has the advantage of directly providing, from noisy observations, the matrices related to state space model and the covariance matrices that are necessary to perform Kalman filtering. Furthermore, unlike existing methods, no voice activity detector is required to estimate noise variance. Besides no observation covariance estimation is involved in the determination of the driving process covariance matrix.

This paper is organized as follows: in section 2 we present the state space model of the noisy speech signal and the Kalman filter. In section 3 the principles of the methods are introduced. In sections 4 and 5 a description of the algorithms is presented. In the last section we give experimental results and compare both methods proposed here with classical approaches.

2. NOISY SPEECH MODEL AND KALMAN FILTERING

2.1 Model equations

Let us consider the speech signal $s(k)$ modeled as a n order AR process:

$$s(k) = \sum_{j=1}^{n} a_j s(k-j) + w(k) \quad (1)$$

$$y(k) = s(k) + v(k) \quad (2)$$

where $s(k)$ is the k^{th} sample of the speech signal;
$y(k)$ the noisy observation;
$v(k)$ the measurement noise;
$w(k)$ the driving process;
a_j the j^{th} AR parameter.

This system can be represented by the following state space model:

$$\underline{s}(k+1) = A\underline{s}(k) + \underline{w}(k) \quad (3)$$

$$y(k) = C\underline{s}(k) + v(k) \quad (4)$$

where:

1. $\underline{s}(k)$ is the n×1 state vector:
   $\underline{s}(k) = [s(k-n+1) \quad \cdots \quad s(k)]^T$.
2. $\underline{w}(k)$, a n×1 vector, is defined as follows:
   $\underline{w}(k) = [0 \quad \cdots \quad 0 \quad w(k)]^T$.
3. A is the n×n transition matrix:
   $$A = \begin{bmatrix} 0 & 1 & 0 & \cdots & 0 \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ \vdots & & \ddots & \ddots & 0 \\ 0 & \cdots & \cdots & 0 & 1 \\ a_n & a_{n-1} & \cdots & \cdots & a_1 \end{bmatrix}.$$
4. $\underline{w}(k)$ and $\underline{v}(k)$ are zero mean gaussian white noise sequences with respective covariance matrices Q and R and cross variance matrix S.
5. C , the transition row vector, is defined as follows:
   $C = [0 \quad \cdots \quad 0 \quad 1]$.

The standard Kalman filter provides the updating state vector estimator [8]. However, the transition matrix A, the transition row vector C and the variance matrices Q and R must be estimated. Instead of directly determining the quadruplet [A, C, Q, R], we can evaluate the following similar one [9]:

$$\left[A_T = T^{-1}AT, C_T = CT, Q_T = T^{-1}QT^{-T}, R_T = R \right]$$

where T is a non singular transformation.

Indeed, the relations (3) and (4) are equivalent to:

$$\underline{s}(k) = T\underline{s}^*(k) \quad (5)$$

$$\underline{s}^*(k+1) = T^{-1}AT\underline{s}^*(k) + T^{-1}w(k) \quad (6)$$

$$y(k) = CT\underline{s}^*(k) + v(k) \quad (7)$$

Given $[A_T, C_T, Q_T, R_T]$, we can then use the Kalman filtering.

2.2 Kalman filtering

We use standard Kalman filtering equations with the quadruplet $[A_T, C_T, Q_T, R_T]$. Even if the state space matrices are not calculated in their canonical forms we can estimate speech signal $\hat{s}(k)$ as follows:

$\hat{s}(k) = C\hat{\underline{s}}(k) = C_T\hat{\underline{s}}^*(k)$ where $\hat{\underline{s}}^*(k)$ is the $\underline{s}^*(k)$ estimate.

In the next paragraphs we will present two non-iterative methods determining the quadruplet $[A_T, C_T, Q_T, R_T]$ in order to perform Kalman-based speech enhancement. Both methods are based on Van Overschee's state space identification approach [6] [7], the fundamental concepts of which will be introduced in the next paragraph.

3. PRELIMINARIES: PRINCIPLES OF $[A_T, C_T, Q_T, R_T]$ ESTIMATION

First of all let us consider the $i \times (N - 2i - 1)$ noisy observation Hankel matrices:

$$Y_{0/i-1} = \begin{bmatrix} y_0 & \cdots & y_{N-2i} \\ \vdots & & \vdots \\ y_{i-1} & \cdots & y_{N-i-1} \end{bmatrix} \quad Y_{i/2i-1} = \begin{bmatrix} y_i & \cdots & y_{N-i} \\ \vdots & & \vdots \\ y_{2i-1} & \cdots & y_{N-1} \end{bmatrix}$$

And let us define subspace orthogonal projection operator as follows:

$L / M = LM^T \left(MM^T\right)^{-1} M$ with L and M two subspaces.

These definitions being given, let us divide $[A_T, C_T, Q_T, R_T]$ quadruplet estimation into $[A_T, C_T]$ and $[Q_T, R_T]$ pair estimations.

3.1 Principles of $[A_T, C_T]$ estimation

The estimation of $[A_T, C_T]$ can be seen as a realization problem. Classical realization methods need the knowledge of the observation covariance matrices and are based on the factorization of the correlation matrix between $Y_{0/i-1}$ and $Y_{i/2i-1}$ into the observability and controllability matrices [10].

Alternative methods that avoid the knowledge of the observation covariance matrix were proposed by Van Overschee [6], [7]. The idea is to project $Y_{i/2i-1}$ subspace onto $Y_{0/i-1}$ subspace in order to approximate the observability matrix of the system Γ_i :

$$\Gamma_i = \left[C_T{}^T \quad (C_T A_T)^T \quad \cdots \quad \left(C_T A_T{}^{n-1}\right)^T \right]^T.$$

The first step of this approach is to determine two sequences, Z_i and Z_{i+1}, respectively from the projections $Y_{i/2i-1} / Y_{0,i-1}$ and $Y_{i+1/2i-1} / Y_{0,i}$.

Z_i and Z_{i+1} can be considered as the outputs of a bank of a non-steady state Kalman filters after i and i+1 time steps and verify:

$$Y_{i/2i-1} / Y_{0,i-1} = \Gamma_i Z_i \quad (8)$$

$$Y_{i+1/2i-1} / Y_{0,i} = \underline{\Gamma}_i Z_{i+1} \quad (9)$$

with $\underline{\Gamma}_i$ equal to Γ_i without its last row.

The second step consists in deriving the pair $[A_T, C_T]$ from the sequences Z_i and Z_{i+1}, by solving the following least squares problem [6], [7]:

$$\min_{A_T C_T} \left\| \begin{pmatrix} Z_{i+1} \\ Y_{i,i} \end{pmatrix} - \left[\begin{pmatrix} A_T \\ C_T \end{pmatrix} Z_i \right] \right\|_F^2 \quad (10)$$

where the $\|.\|_F$ is the Froebenius norm.

The matrix division of $\begin{pmatrix} Z_{i+1} \\ Y_{i,i} \end{pmatrix}$ into Z_i allows to determine $[A_T, C_T]$.

3.2 Principles of $[Q_T, R_T]$ estimation

For both methods, we obtain the pair $[Q_T, R_T]$ from the residuals ρ of the least square solution of (10) [6]:

$$\rho = \begin{pmatrix} Z_{i+1} \\ Y_{i,i} \end{pmatrix} - \left[\begin{pmatrix} A_T \\ C_T \end{pmatrix} Z_i \right] \text{ and } \frac{1}{N} \rho \rho^T = \begin{bmatrix} Q_T & S_T \\ S_T^T & R_T \end{bmatrix}$$

4. ESTIMATION OF $[A_T, C_T]$: METHOD 1

If the matrices Z_i and Z_{i+1} are known, we can determine $[A_T, C_T]$ as follows [7]:

$$A_T = E_N \left[Z_{i+1} Z_i^T \right] \left(E_N \left[Z_i Z_i^T \right] \right)^{-1}$$

$$C_T = E_N \left[Y_{i/i} Z_i^T \right] \left(E_N \left[Z_i Z_i^T \right] \right)^{-1}$$

with $E_N[.] = \lim_{N \to \infty} \frac{1}{N} [.]$

However such an approach cannot be implemented directly. This is the reason why we introduce principal angles between subspaces and principal vectors [7], [10].

Indeed, we can determine Z_i from the principal angles between the two subspaces $Y_{0/i-1}$ and $Y_{i/2i-1}$ and their principal vectors.

Let P_i denote the principal vectors of $Y_{0/i-1}$, Q_i those associated to $Y_{i/2i-1}$ and S_i the principal angles between $Y_{0/i-1}$ and $Y_{i/2i-1}$.

Denote with the square matrix S_i^n the matrix S_i with its n first rows and columns, where n is the AR process order.

$$Z_i = \left(S_i^n \right)^{1/2} P_i^n$$

In the same way, we can evaluate Z_{i+1} from $Y_{0/i}$ and $Y_{i+1/2i-1}$.

Algorithm 1, that determines Z_i and Z_{i+1}, is based on the RQ factorization of the 2i×(N-2i-1) noisy observation Hankel matrix, $Y_{0/2i-1}$, and quotient singular value decompositions [7].

Step 1: RQ Factorization:

$$\frac{Y_{0/2i-1}}{\sqrt{N-2i+1}} = \begin{pmatrix} R_{11} & 0 & 0 & 0 \\ R_{21} & R_{22} & 0 & 0 \\ R_{31} & R_{32} & R_{33} & 0 \\ R_{41} & R_{42} & R_{43} & R_{44} \end{pmatrix} \begin{pmatrix} Q_1^T \\ Q_2^T \\ Q_3^T \\ Q_4^T \end{pmatrix}$$

with $R_{11} \in R^{(i-1)\times(i-1)}, R_{22} \in R, R_{33} \in R, R_{44} \in R^{(i-1)\times(i-1)}$

Step 2: quotient singular decompositions:

$$\begin{pmatrix} R_{31} & R_{32} \\ R_{41} & R_{42} \end{pmatrix} = U_i S_i X_i^T \text{ and } \begin{pmatrix} R_{33} & 0 \\ R_{43} & R_{44} \end{pmatrix} = V_i T_i X_i^T$$

with X_i is a non singular square matrix.

Denote with U_i^1 the matrix U_i with its n first columns.

$$Z_i = \left(S_i^n \right)^{1/2} U_i^{1\,T} Q_{1:2}^{1\,T}$$

$$\begin{pmatrix} R_{41} \\ R_{42} \\ R_{43} \end{pmatrix} = U_{i-1} S_{i-1} X_{i-1}^T \text{ and } R_{44}^T = V_{i-1} T_{i-1} X_{i-1}^T$$

$$Z_{i+1} = \left(S_i^n \right)^{-1/2} \underline{X}_i^{1\#} X_{i-1}^1 \left(S_{i-1}^n \right)^{1/2} \left(U_{i-1}^1 \right)^T Q_{1:3}^{1\,T}$$

where $\underline{X}_i^1$ is equal to X_i^1 without its last row and $\underline{X}_i^{1\#}$ denotes the pseudo inverse of $\underline{X}_i^1$.

Step 3: Determination of Γ_i

$$\Gamma_i = X_i^1 \left(S_i^1 \right)^{1/2}$$

Step 4: Determination of $[A_T, C_T]$

$$A_T = \left(S_i^1 \right)^{-1/2} \left(\underline{X}_i^1 \right)^{\#} X_{i-1}^1 S_{i-1}^1 \left(\underline{U}_{i-1}^1 \right)^T U_i^1 \left(S_i^1 \right)^{-1/2}$$

C_T is defined from the first row of Γ_i

5. ESTIMATION OF $[A_T, C_T]$: METHOD 2

In this method the evaluation of Z_i and Z_{i+1} needs the estimation of the projections $Y_{i/2i-1} / Y_{0,i-1}$, $Y_{i+1/2i-1} / Y_{0,i}$ and Γ_i [6].

Indeed (8) and (9) are equivalent to:

$$Z_i = \Gamma_i^{\#} \left(Y_{i/2i-1} / Y_{0,i-1} \right) \quad (11)$$

$$Z_{i+1} = \underline{\Gamma}_i^{\#} \left(Y_{i+1/2i-1} / Y_{0,i} \right) \quad (12)$$

where $\Gamma_i^{\#}$ denotes the pseudo inverse of Γ_i

Algorithm 2 is a reformulation of N4SID algorithm (which stands for Numerical Algorithm for Subspace State Space System Identification) for state space model representing noisy speech signal. It is based on the RQ factorization of the 2i×(N-2i-1) noisy observation Hankel matrix, $Y_{0/2i-1}$, and singular value decompositions (SVD) [6].

Step 1: RQ Factorization:

$$\frac{Y_{0/2i-1}}{\sqrt{N-2i+1}} = \begin{pmatrix} R_{11} & 0 & 0 \\ R_{21} & R_{22} & 0 \\ R_{31} & R_{32} & R_{33} \end{pmatrix} \begin{pmatrix} Q_1^T \\ Q_2^T \\ Q_3^T \end{pmatrix}$$

with $R_{11} \in R^{i\times i}, R_{22} \in R, R_{33} \in R^{(i-1)\times(i-1)}$

Step 2: Determination of the projections $Y_{i/2i-1} / Y_{0,i-1}$ *and* $Y_{i+1/2i-1} / Y_{0,i}$

$$Y_{i/2i-1} / Y_{0,i-1} = \begin{pmatrix} R_{21} \\ R_{31} \end{pmatrix} R_{11}^{-1} Y_{0,i-1}$$

$$Y_{i+1/2i-1} / Y_{0,i} = \begin{pmatrix} R_{31} & R_{32} \end{pmatrix} \begin{pmatrix} R_{21} & R_{22} \\ R_{31} & R_{32} \end{pmatrix}^{-1} Y_{0,i}$$

Step 3: Determination of Γ_i

We can approximate Γ_i by computing the SVD of $Y_{i/2i-1} / Y_{0,i-1}$.

$$Y_{i/2i-1} / Y_{0,i-1} = U\Sigma V^T \text{ and } \Gamma_i = U^1 \left(\Sigma^n\right)^{1/2}$$

Step 4: Determination of $[A_T, C_T]$.

6. SIMULATIONS AND CONCLUSION

For a noisy speech signal sampled at 8 kHz, we first estimate the quadruplet $[A_T, C_T, Q_T, R_T]$ according to the methods developed in paragraphs 3, 4 and 5. Second we compare our approaches to Paliwal's method developed in [1], Gibson's iterative approach, [2] and the approach proposed by Gabrea in [4] and [5].

Table 1 illustrates the performance of the various approaches, from 200 tests.

Input SNR	-10	-5	0	5	10
Output SNR Paliwal's Method	12,37	9,27	6,76	4,8	3,34
Output SNR Method n°2	11,30	8,51	6,34	4,52	3,06
Output SNR Gibson's Method	11,30	8,45	6,22	4,51	3,14
Output SNR Method n°1	11,10	8,32	6,21	4,45	3,03
Output SNR Method [4], [5]	10.48	8.13	5.78	3.50	1.57

Table 1: SNR GAIN for VARIOUS INPUT SNR (dB)

Paliwal's method stands as a reference since the AR parameters are estimated from the noise free signal. Furthermore noise model variance is directly obtained from the white noise sequence as it is separately available [1].

Gibson's approach needs three to four iterations to get the highest SNR gain. But the iteration number that provides best results is a priori unknown.

In this context we can point out the fact that both methods presented here provide, in one iteration, significant SNR gain. Unlike EM-based algorithms, no other iteration is necessary to improve the estimate of the enhanced speech. In addition no voice activity detector is required. Last no covariance information is involved in the estimation of the driving process covariance matrix.

Figure 1 shows an example of enhancement of a speech signal corrupted by white noise. The noise free speech, the noisy speech and the enhanced speech are given in figure 1.

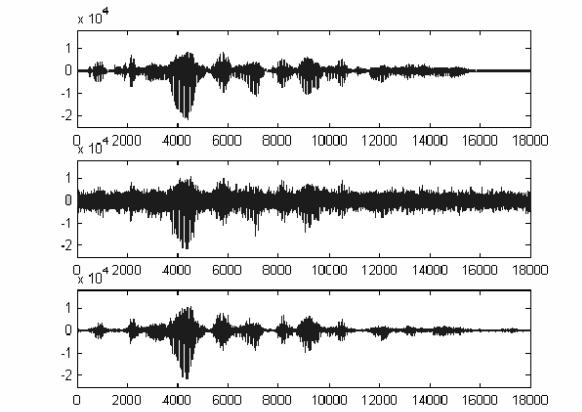

Figure 1: Example of enhancement speech corrupted by white noise (Input SNR=0dB)

7. ACKNOWLEDGEMENT

We would like to acknowledge Matra Communication Company for gently providing us with various recorded speech signals.

8. REFERENCES

[1] K. K. Paliwal and A. Basu, "A Speech Enhancement Method Based on Kalman Filtering", ICASSP 87, pp. 177-180.

[2] J. D. Gibson, B. Koo and S. D. Gray, "Filtering of Colored Noise for Speech Enhancement and Coding", IEEE Trans. on Signal Processing, Vol. 39, n°8, pp. 1732-1742, August 1991.

[3] S. Gannot, D. Burchtein and E. Weinstein "Iterative and Sequential Kalman Filter-Based Speech Enhancement Algorithms", IEEE Trans. On Speech and Audio Processing, July 1998.

[4] M. Gabrea, E. Mandridake and M. Najim "A Single Microphone Noise Canceller Based on Adaptive Kalman Filter", EUSIPCO 96, Vol n°2, pp. 979-982, 1996.

[5] M. Gabrea, E. Grivel and M. Najim " A Single Microphone Kalman Filter-Based Noise Canceller", submitted to IEEE Signal Processing Letters, 1998.

[6] P. Van Overschee and B. de Moor, "N4SID: Susbspace Algorithm for the Identification of Combined Deterministic and Stochastic Systems", Automatica, Vol. 30, n°1, pp. 75-93, 1994.

[7] P. Van Overschee and B. de Moor, "Subspace Algorithms for the Stochastic Identification Problem", Automatica, Vol 29, n°3, pp. 649-660, 1993.

[8] M. Najim. "Modelization and Identification in Signal Processing", in french, Masson Publisher, Paris 1988.

[9] T. Kailath. "Linear Systems", Prentice Hall Information and System Sciences Series, 1980.

[10] S. K. Kung, "A New low-Order Approximation Algorithm via Singular Value Decomposition", Proc. 12th Asilomar Conference on Circuits, Systems and Computers, pp. 705-714, 1978.

[11] G. H. Golub and C. F. Van Loan "Matrix Computation", North Oxford Academic, John's Hopkins University Press, 1983.