A DUAL KALMAN FILTER-BASED SMOOTHER FOR SPEECH ENHANCEMENT

Hong Cai, Eric Grivel and Mohamed Najim

ESI, UMR LAP5131, ENSEIRB-Université Bordeaux 1
BP 99, F 33402 TALENCE Cedex, France

ABSTRACT

Kalman algorithms have been widely applied, for instance in single-channel speech enhancement. However, when carrying out Kalman smoothing, the computational cost and the data storage requirements are two specific problems. In this paper, a dual-filter-based smoother is proposed and used in the framework of speech enhancement. Our approach comprises a forward-in-time Kalman filter and a backward-in-time Kalman filter. Both filters are based on their respective forward-in-time linear prediction (LP) model and backward-in-time LP model. This method does not require a large storage space as a standard Kalman smoother does. The algorithm is evaluated by considering a speech signal embedded in a white Gaussian noise. Simulation Results show that the proposed algorithm provides a higher improvement of signal to noise ratio (SNR) than the Kalman filtering.

Keywords: speech enhancement, Kalman filter, smoothing, expectation-maximization algorithm

1. INTRODUCTION

When using a state space representation of the system, the Kalman filter (KF) is a way to recursively obtain the optimal estimation of the state, given the past and present observation data [9]. Among its various applications, Kalman filtering has been successfully used for single-channel speech enhancement (see for instance [3][4][5][6][7][8][11]). In this area, the clean speech, denoted $s(k)$ is often represented by a linear prediction (LP) model [2] ; in addition, the background noise is usually assumed stationary and its second order statistics can be estimated during the periods of silence, between utterances. In [11], Paliwal et al. propose a Kalman filter-based speech enhancement, in which the LP parameters are estimated directly from the clean speech, supposed to be available. However, this approach cannot be computed in practice. For this reason, an Expectation-Maximization (EM) [10][12] can be considered; this is an iterative likelihood maximization method used when it is difficult to obtain a direct maximum likelihood (ML) estimate, operating in two steps (E and M steps). In [4][6], the E-step consists in carrying out a Kalman filter. It should be noted that a Kalman smoother [1][5] can also be used since it improves the estimation precision and hence weakens the residual noise in the enhanced speech. However, when completing standard Kalman smoothing [1][13], data storage space must be taken into account. This comprises the Kalman gain matrix, filtered mean vector and covariance matrix, one-step-forward predicted mean vector and covariance matrix. These quantities are calculated by the Kalman filter and used for backward in time recursive calculation. So, due to the enormous storage requirements and the computational cost, applying Kalman smoothing is confined to very specific applications.

In this paper, we propose to investigate a dual-filter-based smoother that makes it possible to reduce the storage space; In addition, we apply it in the framework of speech enhancement and compare it to the Kalman filter scheme.

The remainder of the paper is organized as follows: in section 2, a dual-filter-based smoother for speech enhancement is presented; simulation results are given in section 3, and finally a brief concluding remark is given in section 4.

2. A DUAL-FILTER-BASED SMOOTHER

Let us consider the observations $z(k)$ of a speech signal $s(k)$ contaminated by an additive noise $v(k)$:

$$z(k) = s(k) + v(k) \qquad (1)$$

Here, we assume that $v(k)$ is the zero mean white noise with variance σ_v^2 . It should be noted that if the additive noise is colored, then a pre-whitening step can be considered like in [8].

Here, we propose to retrieve the speech from the noisy observations. For this purpose, we propose a frame-by-frame approach using a dual-filter-based smoother. The basic idea is to run a Kalman filter forward in time, in order to estimate the mean and the covariance of the state at the instant *k*, given the observations $z(1),\cdots,z(k)$. Meanwhile, a second Kalman filter is used backward in time to produce a one-step backward-time predicted mean and covariance, given the future data $z(N),\cdots,z(k+1)$. Both obtained estimates are then combined to provide a smoothed estimate of the speech signal. See Figure 1.

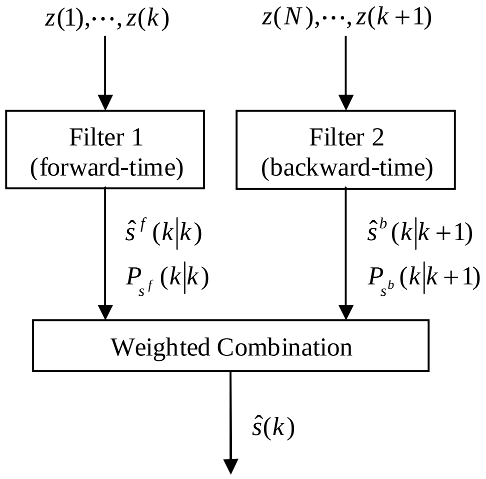

Figure 1. a smoother based on forward filter and backward prediction

Let us introduce the forward in time LP model and the backward in time LP model during an analyzed frame with time index $(1,\cdots,N)$.

The forward in time M^{th} order LP process can be expressed as follows:

$$s(k)=\sum_{i=1}^{M} a_i^p s(k-i)+u^p(k),\quad k=M+1,\cdots,N \qquad (2)$$

where $u^p(k)$ is the driving process with zero mean and variance $\sigma^2_{u^p}$. $\Theta^p=(a_1^p,\cdots,a_M^p,\sigma^2_{u^p})$ is the forward in time LP model parameter which can be estimated from $s(1),\cdots,s(N)$ by using ML estimator, Yule-Walker equation or LS estimator, etc. [9][12].

Similarly, the backward in time M^{th} LP process can be defined as follows:

$$s(k)=\sum_{i=1}^{M} a_i^b s(k+i)+u^b(k),\quad k=N-M,\cdots,1 \qquad (3)$$

where $u^b(k)$ is the driving process with zero mean and variance $\sigma^2_{u^b}$. $\Theta^b=(a_1^b,\cdots,a_M^b,\sigma^2_{u^b})$ is the backward in time LP model parameter which can be estimated from $s(N),\cdots,s(1)$.

By respectively denoting the space vector for the forward in time LP model and the backward in time LP model:

$$\underline{X}^f(k)=\left[s(k-M+1),\cdots,s(k)\right]^T,$$

$$\underline{X}^b(k)=\left[s(k+M-1),\cdots,s(k)\right]^T,$$

the state space representation can be written as follows:

$$\begin{aligned}\underline{X}^f(k)&=A^f\,\underline{X}^f(k-1)+Bu^p(k)\\ z(k)&=C\underline{X}^f(k)+v(k)\end{aligned} \qquad (4)$$

and

$$\begin{aligned}\underline{X}^b(k)&=A^b\,\underline{X}^b(k+1)+Bu^b(k)\\ z(k)&=C\underline{X}^b(k)+v(k)\end{aligned} \qquad (5)$$

where

$$A^f=\begin{bmatrix}0 & 1 & 0 & \cdots & 0\\ \vdots & \ddots & \ddots & \ddots & \vdots\\ \vdots & \ddots & \ddots & \ddots & 0\\ 0 & \cdots & \cdots & 0 & 1\\ a_M^p & a_{M-1}^p & \cdots & \cdots & a_1^p\end{bmatrix},\ C=\begin{bmatrix}0 & \cdots & 0 & 1\end{bmatrix}$$

$$A^b=\begin{bmatrix}0 & 1 & 0 & \cdots & 0\\ \vdots & \ddots & \ddots & \ddots & \vdots\\ \vdots & \ddots & \ddots & \ddots & 0\\ 0 & \cdots & \cdots & 0 & 1\\ a_M^b & a_{M-1}^b & \cdots & \cdots & a_1^b\end{bmatrix},B=\begin{bmatrix}0 & \cdots & 0 & 1\end{bmatrix}^T$$

Based on equation (4), and by taking the following initial conditions :

$$\hat{\underline{X}}^f(M|M)=\left[z(1),\cdots,z(M)\right]^T,$$

$$P^f(M|M)=\sigma_v^2 I$$

we can recursively calculate the mean and covariance of the state, denoted by $\left(\hat{\underline{X}}^f(k|k),\ P^f(k|k)\right)$, for $k=M+1,\cdots,N$.

Based on equation (5) and by taking the initial conditions:

$$\hat{\underline{X}}^b(N-M+1|N-M+2)=\left[z(N),\cdots,z(N-M+1)\right]^T,$$

$$P^b(N-M+1|N-M+2)=\sigma_v^2 I$$

we can recursively calculate the one-step backward-time predicted mean and covariance of the state, $\left(\hat{\underline{X}}^b(k|k+1),\ P^b(k|k+1)\right)$, for $k = N-M,\cdots,1$.

Accordingly, we can obtain the forward-time filter and the one-step backward-time prediction of $s(k)$ and their variances, denoted respectively as $\left(\hat{s}^f(k|k),\ P_{s^f}(k|k)\right)$ and $\left(\hat{s}^b(k|k+1),\ P_{s^b}(k|k+1)\right)$.

A smoothed estimate of $s(k)$ can then be obtained by combining $\hat{s}^f(k|k)$ and $\hat{s}^b(k|k+1)$, such as

$$P_s^{-1}(k) = P_{s^f}^{-1}(k|k) + P_{s^b}^{-1}(k|k+1) \tag{5}$$

$$\hat{s}(k) = P_s(k)\left(P_{s^f}^{-1}(k|k)\hat{s}^f(k|k) + P_{s^b}^{-1}(k|k+1)\hat{s}^b(k|k+1)\right) \tag{6}$$

When the clean speech was available, the above smoothing algorithm could be directly applied for speech enhancement, by estimating the forward in time LP parameters and the backward in time parameters, directly from the clean speech.

However, in practice, only the noisy speech is available. For this reason, an EM algorithm based approach is considered.

In this framework, $s(1),\cdots,s(N)$ define the so-called complete data whereas $z(1),\cdots,z(N)$ are the so-called incomplete data. The ordinary ML estimation consists in maximizing the likelihood of the complete data $s(1),\cdots,s(N)$. Since the complete data are not available, estimating the LP model parameters can be done through the maximization of the conditional expectation of the likelihood. This leads to the iterative and successive processing of the so-called M-step and E-step. During the M-step, the LP parameters are estimated by means of the estimated complete data $\hat{s}(1),\cdots,\hat{s}(N)$. These quantities can be obtained during the E-step by using the estimated LP parameters $\hat{\Theta}^p,\hat{\Theta}^b$ obtained during the former M-step and by applying a Kalman algorithm. Here, the above dual-filter-based smoother is used.

The procedure can be described as follows:

Step 1 Initialization: select the initial parameter estimates $\hat{\Theta}_0^p$, $\hat{\Theta}_0^b$, and for $i = 0,1,\cdots$, until convergence;

Step 2 Expectation: based on the estimated parameters $\hat{\Theta}_i^p,\hat{\Theta}_i^b$ in the i^{th} iteration, exploit the proposed dual-filter-based smoother to calculate the enhanced speech of the analysed frame in the i^{th} iteration, denoted as $\hat{s}_i(k)$;

Step 3 Maximization: based on the estimated speech $\hat{s}_i(k)$ at the i^{th} iteration, re-estimate the parameters of the forward in time LP model and the backward in time LP model for the $i+1^{th}$ iteration;

Step 4 Convergence test: if the convergence test is not satisfied, then go to Step 2.

3. RESULTS

The comparative study is carried out with speech signal sampled at 8kHz. The length of the frame N is equal to 256 and an overlap of 50 % is used.

First of all, we assume that the clean speech is available, like in [11]. Figure 2 shows the plots of the clean speech, the noisy speech with a Signal to Noise Ratio (SNR) equal to 5 and the enhanced speech using the proposed dual-filter-based smothering algorithm. Table 1 shows the comparison of the SNR improvement for the KF based enhancement algorithm [11] and the proposed DFBS based enhancement algorithm, for the input SNR ranked from –10dB to 15dB.

input SNR (dB)	-10	-5	0	5	10	15
KF[11]: output SNR (dB)	2.2	4.2	6.8	9.9	13.5	18.8
DFBS: output SNR (dB)	2.9	5.1	8.0	11.2	14.6	19.2

Table 1. comparison of DFBS based enhancement algorithm and KF based enhancement algorithm over SNR

The second part of our simulations is completed when only noisy speech is available. At that stage, we use EM algorithms for speech enhancement, in which the E-step are respectively based on KF and DFBS. The convergence test rule is taken as follows: If the difference of SNR improvement between two neighboring iterations is less than 0.1dB, the convergence of the performance is achieved. The comparison of the SNR improvement of the two algorithms is showed in Table 2. The average iteration numbers (AIN) in a frame for the two algorithms are showed in Table 3. From Tables 2 and 3, it is obvious that the proposed DFBS based EM algorithm has more SNR improvement and less iteration numbers than the KF based EM algorithm.

input SNR (dB)	-10	-5	0	5	10	15
KF-EM: output SNR (dB)	1.6	3.6	6.4	9.7	13.3	17.2
DFBS-EM: output SNR (dB)	1.8	4.1	7.1	10.6	14.3	18.1

Table 2. comparison of DFBS based EM algorithm and KF based EM algorithm

input SNR (dB)	-10	-5	0	5	10	15
KF-EM: AIN	4.5	4.1	4.0	3.9	3.6	3.0
DFBS-EM: AIN	3.4	3.1	3.0	2.9	2.9	2.8

Table 3. comparison of average iteration numbers (AIN) of DFBS based EM algorithm and KF based EM algorithm

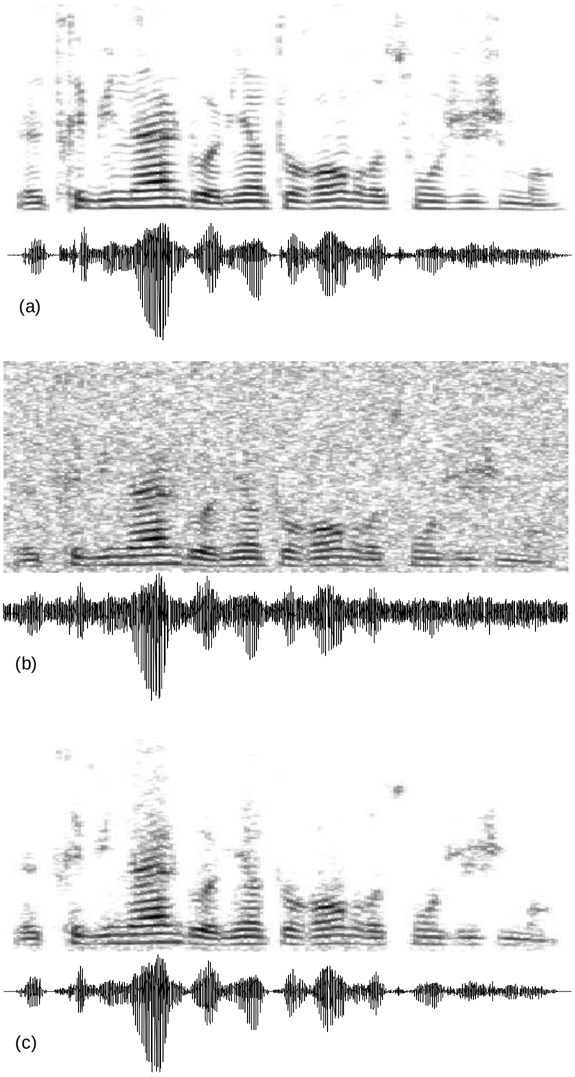

Figure 2. (a) clean speech. (b) noisy speech with SNR=5dB. (c) enhanced speech using DFBS.

4. CONCLUSION

The Kaman Filter has been widely used in many areas from tracking to speech enhancement. Since the speech signal is often assumed stationary during an analysed frame (20-30 ms), the Kalman smoother can be carried out and provides better estimates of the state since it is based on a higher number of observations. However, the enormous storage requirements of the Kalman smoother confine its applications in practice. In this paper, our purpose was to present an alternative to the Kalman smoother, named dual-filter-based smoother (DFBS), which does not require such a storage space for the vectors and matrices during the Kalman filtering process.

The evaluation of the DFBS-based speech enhancement algorithm has been performed in comparison to the KF based speech enhancement algorithm. The results show that the proposed DBFS based algorithm can provide a higher SNR improvement than the KF based algorithm.

REFFERENCES

[1] J. Casals, M. Jerez and S. Sotoca, “Exact Smoothing for Stationary and Nonstationary Time Series”, *International Journal of Forecasting*, vol.16, pp.59-69,2000.

[2] J. R. Deller, J. H. L. Hanson and J. G.Proakis, *Discrete-Time Processing of Speech Signal*, IEEE PRESS, New York, 2000.

[3] M. Gabrea, E. Grivel and M. Najim, “A Single Microphone Kalman Filter-Based Noise Canceller”, IEEE Signal Processing Letters, march 1999, pp. 53-55.

[4] M. Gabrea, “Adaptive Kalman Filtering for Speech Signal Recovery in Colored Noise”, *EUSIPCO’2002*, Toulouse, France, 2002.

[5] S. Gannot, D. Burchtein and E. Weinstein, “Iterative and Sequential Kalamn Filter-Based Speech Enhancement algorithms”, *IEEE Trans. on Speech and audio Processing*, pp.373-385, July 1998

[6] 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, pp. 1732-1742, August 1991.

[7] Z. Goh, K.-C. Tan and B. T. G. Tan, “Kalman-Filtering Speech Enhancement Method Based on a Voiced-Unvoiced Speech Model”, *IEEE Trans. On Speech and Audio Processing*, vol. 7, no. 5, pp. 510--524, 1999.

[8] E. Grivel, M. Gabrea and M. Najim, “Speech Enhancement as a realization issue”, *Signal Processing*, Dec. 2002.

[9] S. Haykin, *Adaptive Filter Theory*, 3rd Edition, Prentice-Hall, N. J., 1996.

[10] T.K.Moon, “The Expectation-Maximization Algorithm”, *IEEE Signal Processing Magazine*, pp. 47-60, Nov. 1996.

[11] K. K. Paliwal and A.Basu, “A speech enhancement method based on Kalman filtering, *Proceedings of ICASSP’87*, pp.177-180, Dallas, TX, USA, 1987.

[12] S. V. Vaseghi, *Advanced Signal Processing and Digital Noise Reduction*, Wiley Teubner, 1996.

[13] J. Vermaak, M. Niranjan and S. J. Godsill, “Comparing Solution Methodologies for Speech Enhancement within a Bayesian Framework”, *Technical Report CUED/F-INFENG/TR.329*, Cambridge University Engineering Department, August 1998.