A long state vector Kalman filter for speech enhancement

Stephen So and Kuldip K. Paliwal

Signal Processing Laboratory, Griffith School of Engineering
Griffith University, Brisbane, QLD, Australia, 4111
s.so@griffith.edu.au, k.paliwal@griffith.edu.au

Abstract

In this paper, we investigate a long state vector Kalman filter for the enhancement of speech that has been corrupted by white and coloured noise. It has been reported in previous studies that a vector Kalman filter achieves better enhancement than the scalar Kalman filter and it is expected that by increasing the state vector length, one may improve the enhancement performance even further. However, any enhancement improvement that may result from an increase in state vector length is constrained by the typical use of short, non-overlapped speech frames, as the autocorrelation coefficient estimates tend to become less reliable at higher lags. We propose to overcome this problem by incorporating an analysis-modification-synthesis framework, where long, overlapped frames are used instead. Our enhancement experiments based on the NOIZEUS corpus show that the proposed long state vector Kalman filter achieves higher mean SNR and PESQ scores than the scalar and short state vector Kalman filter, therefore fulfilling the notion that a longer state vector can lead to better enhancement.

Index Terms: Kalman filtering, speech enhancement, analysis-modification-synthesis

1. Introduction

In the problem of speech enhancement, where a speech signal corrupted by noise is given, we are primarily interested in suppressing the noise so that the speech quality and intelligibility are improved. Speech enhancement is useful in many applications where corruption by noise is undesirable and unavoidable. For example, speech enhancement techniques are used as a preprocessor in speech coding standards for cellular telephony (e.g. in [1]) to suppress the background noise prior to coding. Various speech enhancement methods have been reported in the literature and these include spectral subtraction [2], MMSE-STSA (short-term spectral amplitude) estimation [3], Wiener filtering [4], subspace methods [5], and Kalman filtering [6].

The Kalman filter is a recursive, time-domain linear MMSE estimator [7] that has been of particular interest in speech enhancement, due to several advantages it has over other spectral domain-based enhancement methods. For instance, the speech production model is inherent in the Kalman recursion equations. Also, no stationarity assumption needs to be made on the speech signal to be enhanced, unlike for the Wiener filter.

Two types of Kalman filter were introduced by Paliwal and Basu [6], which we refer to here as the *scalar* and *vector Kalman filter*[1]. In the *scalar Kalman filter*, the current enhanced speech sample $\hat{x}(n)$ is formed by taking the first component of the estimated state vector, $\hat{\boldsymbol{x}}(n|n)$[2]. This component is calculated based on information from past and current observations of the noise-corrupted speech as well as past estimated speech samples. For the *vector Kalman filter* (also referred to as the delayed Kalman filter in [6]), the past enhanced speech sample $\hat{x}(n-p+1)$ is formed by taking the last component of $\hat{\boldsymbol{x}}(n|n)$. Therefore, this component is estimated using information from future speech estimates as well as future observations of the noise-corrupted speech, which can be seen as a 'smoother on a component-by-component basis' [8].

The state vector length is equal to the order of the linear predictor model. Therefore, it is reasonable to suggest that to increase the state vector length, a higher order linear predictor model must be used. However, this method has its drawbacks when applied in practice, since the linear prediction coefficients (LPCs) need to be estimated from the noise-corrupted speech only. Increasing the linear predictor order requires the higher-lag autocorrelation coefficients to be estimated reliably. Gibson, *et al.* [8] reported that increasing the LPC order from 4 to 20 improved the SNR by less than 1 dB for the white noise case.

In this paper, we propose a long state vector Kalman filter for speech enhancement. To utilise long state vectors while avoiding the problem associated with unreliable autocorrelation estimates, the proposed scheme incorporates various elements of an analysis-modification-synthesis (AMS) framework, whereby the linear prediction analysis and Kalman filtering is performed on overlapped frames and the output is synthesised using an windowed overlap-add (WOLA) operation [9]. By using the AMS framework, longer speech frames can be used and this in turn allows us to increase the state vector length of the Kalman filter without suffering from unreliable LPC estimates. We show in this paper that a long state vector Kalman filter leads to better enhancement performance as well as residual noise that is less annoying to the listener. To demonstrate its effectiveness, we present some experimental results of the proposed scheme and compare them with that of the scalar and short state vector Kalman filter of [6]. Both white and coloured noise were considered and the LPCs were derived from the noise-corrupted speech.

2. Kalman filtering-based speech enhancement

2.1. The scalar Kalman filter

If the clean speech is represented as $x(n)$ and the noise signal as $v(n)$, then the noise-corrupted speech $y(n)$, which is the only

[1] Note that these definitions are slightly different from those in [8].

[2] The mathematical notation is explained in Section 2. Note that it is different from the notation used in [6].

observable signal in practice, is expressed as:

$$y(n) = x(n) + v(n) \tag{1}$$

In the scalar Kalman filter that is used for speech enhancement [6], $v(n)$ is a zero-mean, white Gaussian noise that is uncorrelated with $x(n)$[3]. A pth order linear predictor is used to model the speech signal:

$$x(n) = -\sum_{k=1}^{p} a_k x(n-k) + w(n) \tag{2}$$

where $\{a_k, k = 1, 2, \dots, p\}$ are the LPCs and $w(n)$ is the white Gaussian excitation with zero mean and a variance of σ_w^2. Rewriting Eqs. (1) and (2) using state vector representation:

$$\begin{aligned} \boldsymbol{x}(n) &= \boldsymbol{A}\boldsymbol{x}(n-1) + \boldsymbol{d}w(n) && (3) \\ y(n) &= \boldsymbol{c}^T\boldsymbol{x}(n) + v(n) && (4) \end{aligned}$$

where $\boldsymbol{x}(n) = [x(n), x(n-1), \dots, x(n-p+1)]^T$ is the 'hidden' state vector, $\boldsymbol{d} = [1, 0, \dots, 0]^T$ and $\boldsymbol{c} = [1, 0, \dots, 0]^T$ are the measurement vectors for the excitation noise and observation, respectively. The linear prediction state transition matrix $\boldsymbol{A}$ is given by:

$$\boldsymbol{A} = \begin{bmatrix} -a_1 & -a_2 & \dots & -a_{p-1} & -a_p \\ 1 & 0 & \dots & 0 & 0 \\ 0 & 1 & \dots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \dots & 1 & 0 \end{bmatrix} \tag{5}$$

When provided with the current sample of corrupted speech $y(n)$, the Kalman filter calculates $\hat{\boldsymbol{x}}(n|n)$, which is an unbiased, linear MMSE estimate of the state vector $\boldsymbol{x}(n)$, by using the following recursive equations:

$$\begin{aligned} \boldsymbol{P}(n|n-1) &= \boldsymbol{A}\boldsymbol{P}(n-1|n-1)\boldsymbol{A}^T + \sigma_w^2\boldsymbol{d}\boldsymbol{d}^T \\ \boldsymbol{K}(n) &= \boldsymbol{P}(n|n-1)\boldsymbol{c}\left[\sigma_v^2 + \boldsymbol{c}^T\boldsymbol{P}(n|n-1)\boldsymbol{c}\right]^{-1} \\ \hat{\boldsymbol{x}}(n|n-1) &= \boldsymbol{A}\boldsymbol{x}(n-1|n-1) \\ \boldsymbol{P}(n|n) &= [\boldsymbol{I} - \boldsymbol{K}(n)\boldsymbol{c}^T]\boldsymbol{P}(n|n-1) \\ \hat{\boldsymbol{x}}(n|n) &= \hat{\boldsymbol{x}}(n|n-1) + \boldsymbol{K}(n)[y(n) - \boldsymbol{c}^T\hat{\boldsymbol{x}}(n|n-1)] \end{aligned}$$

The current estimated sample is then given by $\hat{x}(n) = \boldsymbol{c}^T\hat{\boldsymbol{x}}(n|n)$, which extracts the first component of the estimated state vector.

During the operation of the Kalman filter, the noise-corrupted speech $y(n)$ is windowed into non-overlapped, short (e.g. 20 ms) frames and the LPCs and excitation variance σ_w^2 are estimated. These LPCs remain constant during the Kalman filtering of speech samples in the frame, while the Kalman parameters (such as Kalman gain $\boldsymbol{K}(n)$ and error covariance $\boldsymbol{P}(n|n)$) and state vector estimate $\hat{\boldsymbol{x}}(n|n)$ are continually updated on a sample-by-sample basis (regardless of whichever frame we are in).

2.2. The proposed long state vector Kalman filter

In the proposed long state vector Kalman filter, we provide for an overlap between successive frames by using a window length that is greater than the frame shift. Let us assume that the frame shift is m samples. A tapered analysis window $w_a(n)$ is applied during the estimation of the LPCs and excitation variance while the Kalman filtering is performed on the non-windowed, original samples. At the point of overlap with the next frame, we store the error covariance matrix $\boldsymbol{P}(m|m)$ and state vector estimate $\hat{\boldsymbol{x}}(m|m)$. Before we start to filter the next frame, we initialise the *a priori* error covariance $\boldsymbol{P}(n|n-1)$ and previous state estimate $\hat{\boldsymbol{x}}(n-1|n-1)$ with $\boldsymbol{P}(m|m)$ and $\hat{\boldsymbol{x}}(m|m)$ from the previous frame, respectively. As is done in the windowed overlap-add (WOLA) method [9], the enhanced frames are multiplied by a synthesis window $w_s(n)$, overlapped and then added together.

In contrast to the scalar Kalman filter, the current as well as past estimated speech samples are replaced by the *current* estimated state vector $\hat{\boldsymbol{x}}(n|n)$ in the vector Kalman filter. In effect, the *past* speech sample $\hat{x}(n-p+1)$ is *re-estimated* using information that is based on:

- future noisy observations,
  i.e. $\{y(n-p+2), y(n-p+3), \dots, y(n)\}$; and
- future speech estimates,
  i.e. $\{\hat{x}(n-p+2), \hat{x}(n-p+3), \dots, \hat{x}(n-1)\}$.

Past studies (e.g. [6] and [8]) have reported better enhancement performance with the vector Kalman filter, when compared with the scalar Kalman filter. In particular, it was noted in [6] that the current state vector $\hat{\boldsymbol{x}}(n|n)$ provided a 'fixed-lag-smoothed estimate' of $\hat{x}(n-p+1)$.

The state vector length is equal to p samples. In order to increase the state vector length and thereby enhance the amount of 'smoothing', we need to use a linear predictor with a higher order. LPC estimation requires $p+1$ autocorrelation coefficient estimates:

$$\hat{R}(k) = \frac{1}{N}\sum_{n=0}^{N-1-k} y(n)y(n+k) \tag{6}$$

for $k = 0, 1, \dots, p$, where N is the number of samples in a frame. For a fixed N, these autocorrelation coefficient estimates become less reliable at high lags because the number of terms in the summation of Eq. (6) diminishes. Therefore, in order to improve the reliability of these autocorrelation estimates, N should be increased.

3. Experimental setup

The NOIZEUS corpus [10], which consists of 30 phonetically-balanced sentences belonging to six speakers (three males and three females), sampled at 8 kHz, was used in the speech enhancement experiments. In the scalar and short state vector Kalman filter of [6], the speech was windowed into non-overlapped frames of 20 ms. Linear prediction analysis was performed on the noise-corrupted speech $y(n)$ using the autocorrelation method.

In the proposed long state vector Kalman filter using longer frames, the noise-corrupted speech was windowed into overlapped frames (where the shift was 1/8th of the frame length) of 100 ms and the Hamming window was applied to the speech prior to the linear prediction analysis. The Kalman filter parameters that were modified in the experiments are listed below:

- p = order of the linear predictor model
- L = length of the frame (in seconds)
- R = ratio of overlap ($R = 0$ means there is no overlap)

[3] Coloured noise can be modelled by another linear predictor model, which can be augmented into the Kalman state vector equations [8].

Table 1: Mean and standard deviation (σ) of SNR and PESQ for each Kalman filter case for speech corrupted at 0 dB with white noise. Bolded values show the best results.

Kalman filter case	**SNR (dB)**		**PESQ**	
	Mean	σ	**Mean**	σ
No enhancement	0.000	0.00	1.566	0.19
(I)	4.223	0.16	1.694	0.17
(II)	4.461	0.15	1.739	0.16
(III)	4.083	0.15	1.749	0.15
(IV)	**7.404**	0.33	**1.922**	0.14

Table 2: Mean and standard deviation (σ) of SNR and PESQ for each Kalman filter case for speech corrupted at 0 dB with car noise. Bolded values show the best results.

Kalman filter case	**SNR (dB)**		**PESQ**	
	Mean	σ	**Mean**	σ
No enhancement	-0.555	0.20	1.634	0.22
(I)	3.398	0.51	1.714	0.21
(II)	3.475	0.55	1.747	0.19
(III)	3.082	0.54	1.776	0.18
(IV)	**5.658**	0.75	**1.932**	0.15

The modified Hanning window was used in the WOLA synthesis stage:

$$w_s(n) = 0.5 \left[1 - \cos\left(\frac{2\pi n + \pi}{N}\right)\right], \tag{7}$$

for $n = 0, 1, \ldots, N-1$, where N is the number of samples in each frame. We have tested the following cases at an initial SNR of 0 dB:

(I) Scalar Kalman filter with shorter frames, ($p = 10$, $L = 20$ ms, $R = 0$)

(II) Short state vector Kalman filter with shorter frames, ($p = 10$, $L = 20$ ms, $R = 0$)

(III) Long state vector Kalman filter with shorter frames, ($p = 40$, $L = 20$ ms, $R = 0$)

(IV) Long state vector Kalman filter with longer frames, ($p = 40$, $L = 100$ ms, $R = 1/8$)

Both white noise and coloured noise (car noise) were considered, where the first 1000 samples were assumed to be noise only. To handle the coloured noise, we have used the approach proposed by Gibson, *et al.* [8], where the noise is modelled as a 10th order linear predictor and included in augmented Kalman state recursion equations.

4. Results and discussion

Tables 1 and 2 show the average SNR and PESQ results for the Kalman filter cases on the NOIZEUS corpus, where speech has been corrupted by white noise and car noise at 0 dB, respectively. We can see that the Kalman filter case (II), which is the short state vector Kalman filter of [6], achieves slightly better enhancement than the scalar Kalman filter (case (I)). The difference is not as large as that reported in [6], mainly because we are estimating the LPCs from the noise-corrupted speech, rather than the clean speech.

In case (III), the order of the linear predictor was increased to achieve longer state vectors for the Kalman filter, with the frame length kept at 20 ms. As we can see in both Tables, the enhancement performance has not improved by much (SNR decreased by about 0.3 dB). This may be attributed to the use of unreliable estimates of the high-lag autocorrelation coefficients. Case (IV) is the long state vector Kalman filter that is proposed in this paper. We can see that for both white and car noise, the proposed Kalman case (IV) achieves the highest SNR and PESQ scores. We can attribute these improvements to the use of long state vectors as well as high-lag autocorrelation estimates being more reliable.

Figures 1 and 2 show the spectrograms of the original, corrupted, and enhanced speech signal for white and car noise, respectively. We observe that the amount of residual noise in both Figs. 1 (e) and 2(e) is noticeably less. The residual noise in case (IV), in comparison with cases (I) and (II), also appears to be less localised in time. Preliminary listening tests have suggested that the proposed long vector Kalman filter produces residual noise that tends to be less annoying to the listener than the musical noise that accompanies other spectral-based enhancement methods such as the Wiener filter.

5. Conclusion

In this paper, we presented a long state vector Kalman filter using long frames (i.e. frames formed from using longer window lengths). One of the drawbacks of the vector Kalman filter over short frames is that improving the enhancement performance via an increase in the state vector length is inherently constrained, due to unreliable estimates of the autocorrelation coefficients. This was confirmed in our experiments, where the enhancement performance did not improve by much when the linear predictor order was increased while keeping the frames to be short. We overcame this constraint by incorporating an AMS framework into the vector Kalman filter, whereby speech was divided into long, overlapped frames. Our experimental results and spectrograms showed that the proposed long state vector Kalman filter using long frames achieved better SNR and PESQ scores than the scalar and short state vector Kalman filters.

6. References

[1] T. Ohya, H. Suda, and T. Miki, “5.6 kbits/s PSI-CELP of the half-rate PDC speech coding standard,” in *IEEE 44th Vehicular Technology Conference*, 1994, pp. 1680–1684.

[2] S. Boll, “Suppression of acoustic noise in speech using spectral subtraction,” *IEEE Trans. Acoust., Speech, Signal Process.*, vol. ASSP-27, no. 2, pp. 113–120, 1979.

[3] Y. Ephraim and D. Malah, “Speech enhancement using a minimum mean square error log-spectral amplitude estimator,” *IEEE Trans. Acoust., Speech, Signal Process.*, vol. ASSP-33, pp. 443–445, 1985.

[4] N. Wiener, *The Extrapolation, Interpolation, and Smoothing of Stationary Time Series with Engineering Applications*. New York: Wiley, 1949.

[5] Y. Ephraim and H. L. Van Trees, “A signal subspace approach for speech enhancement,” *IEEE Trans. Speech Audio Process.*, vol. 3, pp. 251–266, Jul. 1995.

[6] K. K. Paliwal and A. Basu, “A speech enhancement method based on Kalman filtering,” in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, vol. 12, Apr. 1987, pp. 177–180.

[7] R. E. Kalman, “A new approach to linear filtering and prediction problems,” *J. Basic Eng., Trans. ASME*, vol. 82, pp. 35–45, Mar. 1960.

[8] J. D. Gibson, B. Koo, and S. D. Gray, “Filtering of colored noise for speech enhancement and coding,” *IEEE Trans. Signal Process.*, vol. 39, no. 8, pp. 1732–1742, Aug. 1991.

[9] R. E. Crochiere, “A weighted overlap-add method of short-time Fourier analysis/synthesis,” *IEEE Trans. Acoust., Speech, Signal Process.*, vol. ASSP-28, no. 1, pp. 99–102, Feb. 1980.

[10] Y. Hu and P. Loizou, “Subjective comparison of speech enhancement algorithms,” in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, vol. 1, 2006, pp. 153–156.

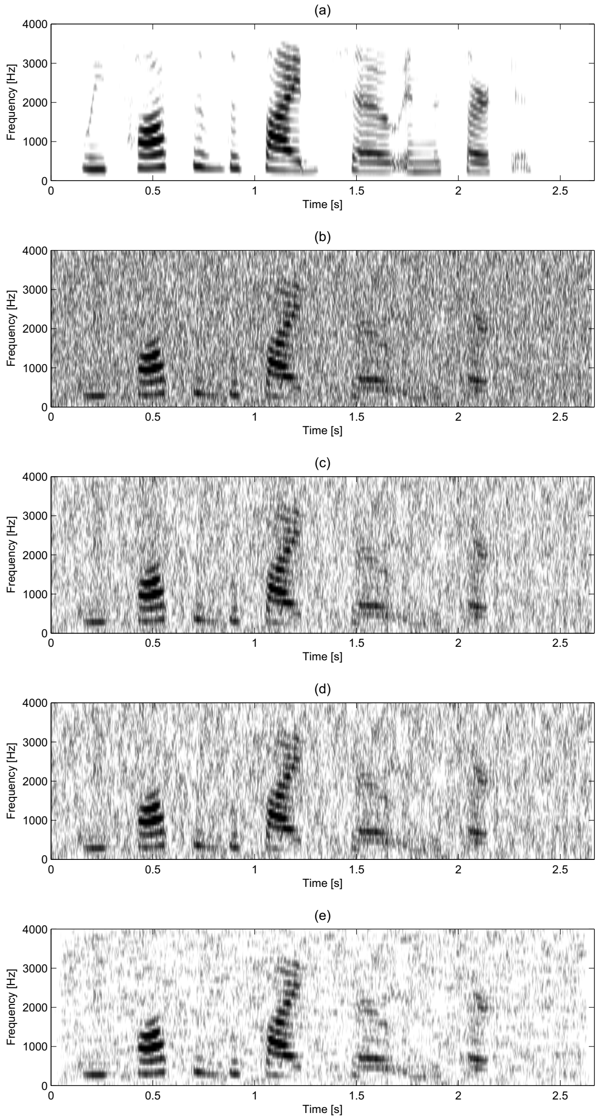

Figure 1: Spectrograms of the utterance sp19.wav ("We talked of the sideshow in the circus") corrupted with white noise at 0 dB: (a) Clean speech; (b) Noise-corrupted speech (SNR = 0.000 dB, PESQ = 1.570) (c) Kalman filter case (I) (SNR = 4.441 dB, PESQ = 1.729); (d) Kalman filter case (II) (SNR = 4.591 dB, PESQ = 1.770); (e) Kalman filter case (IV) (SNR = 8.084 dB, PESQ = 2.001)

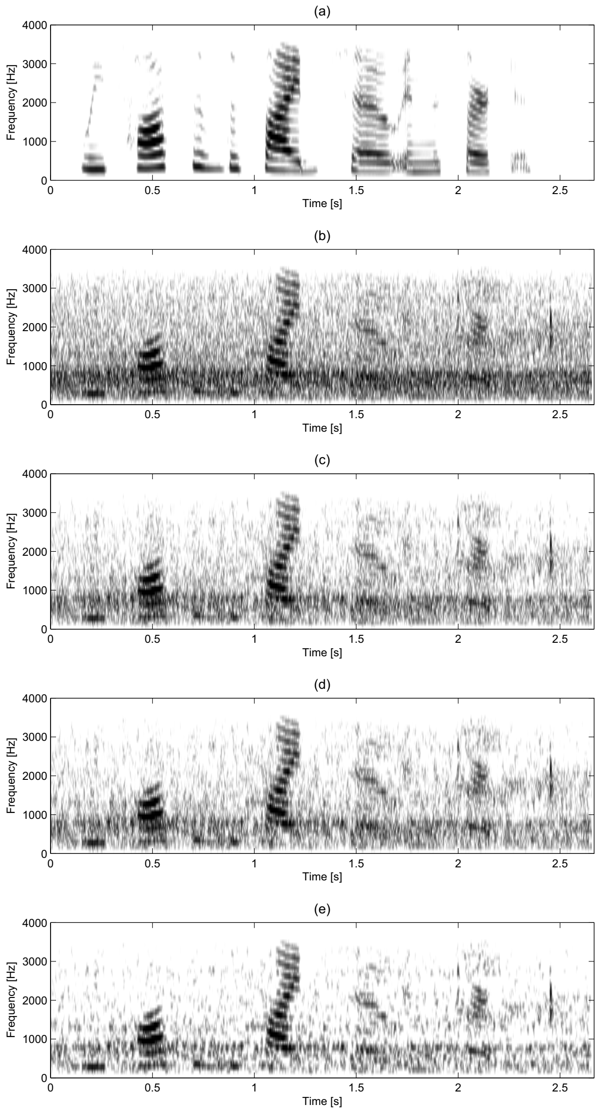

Figure 2: Spectrograms of the utterance sp19.wav ("We talked of the sideshow in the circus") corrupted with car noise at 0 dB: (a) Clean speech; (b) Noise-corrupted speech (SNR = -0.797 dB, PESQ = 1.477) (c) Kalman filter case (I) (SNR = 3.447 dB, PESQ = 1.609); (d) Kalman filter case (II) (SNR = 3.497 dB, PESQ = 1.679); (e) Kalman filter case (IV) (SNR = 6.222 dB, PESQ = 1.951)