Fast converging iterative Kalman filtering for speech enhancement using long and overlapped tapered windows with large side lobe attenuation

Fast converging iterative Kalman filtering for speech enhancement using long and overlapped tapered windows with large side lobe attenuation

Stephen So, 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 propose an iterative Kalman filtering scheme that has faster convergence and introduces less residual noise, when compared with the iterative scheme of Gibson, et al. This is achieved via the use of long and overlapped frames as well as using a tapered window with a large side lobe attenuation for linear prediction analysis. We show that the Dolph-Chebychev window with a -200 dB side lobe attenuation tends to enhance the dynamic range of the formant structure of speech corrupted with white noise, reduce prediction error variance bias, as well as provide for some spectral smoothing, while the long overlapped frames provide for reliable autocorrelation estimates and temporal smoothing. Speech enhancement experiments on the NOIZEUS corpus show that the proposed method outperformed conventional iterative and non-iterative Kalman filters as well as other enhancement methods such as MMSE-STSA and PSC.
Index Terms: speech enhancement, Kalman filtering

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 quality and intelligibility of speech 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 in order to suppress the background noise prior to coding. Various speech enhancement methods have been reported in the literature and these include spectral subtraction, MMSE estimation methods, Wiener filtering, subspace methods, and Kalman filtering [1].

The Kalman filter is an unbiased, time-domain, linear minimum mean squared error (MMSE) estimator, where the unknown states of a dynamic system are estimated using a linear combination of noise-corrupted observations and predicted states. The Kalman filter has been of particular interest in speech enhancement because of several advantages it has over other spectral domain-based enhancement methods: (1) the speech production model is inherent in the Kalman recursion equations by using a linear predictor as the dynamic model; (2) the enhanced speech from the ideal Kalman filter contains no random frequency tones (otherwise known in the literature as *musical noise*); (3) the Kalman filter makes no stationarity assumptions; (4) the Kalman filter can be 'turned-on' at the first sample $n = 0$, where the recursion parameters are initialised with their expected values; and (5) the non-stationary Kalman filter can be viewed as a joint estimator for both the magnitude and phase spectrum of speech [2].

The enhancement performance of the Kalman filter is somewhat dependent on the accuracy of the LPC and excitation variance estimates. Ideally, these coefficients should be obtained from the clean speech, as was done in [3]. However, in practice, the LPCs and variances are generally not known *a priori*, so they must be estimated from the noise-corrupted speech. Depending on the noise characteristics and signal-to-noise ratio (SNR), the LPCs and excitation variance obtained using conventional spectral estimation methods will be poor. The enhanced speech from this suboptimal Kalman filter has been reported previously to suffer from wideband residual noise [4]. Several iterative methods have been proposed that address the issue of unreliable LPC or noise estimates [5, 6, 7]. While the iterative LPC estimation method in [5] generally results in improved SNRs after three or four iterations, 'musical' residual noise accompanies the enhanced speech. The enhanced speech also suffers from distortion, which can degrade the intelligibility. Therefore the iterative LPC estimation method of [5] does not adequately address the problem of poor LPC estimates, especially during the first iteration.

In this paper, we propose the use of long and overlapped tapered windows with large side lobe attenuation in the linear prediction analysis to reduce the presence of background residual noise in iterative Kalman filter-enhanced speech. The proposed method aims to provide a better initial estimate, so that the subsequent iteration results in improved performance. Using objective tests on the NOIZEUS speech corpus [1], we show that the proposed enhancement method (using only two iterations) performs better than conventional iterative and non-iterative Kalman filtering schemes. We also compare the proposed Kalman filter with the phase spectrum compensation method [8] and MMSE-STSA (short-time spectral amplitude) methods [9].

2. Conventional Kalman filtering for speech enhancement

2.1. Non-iterative (conventional) Kalman filtering

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 observable signal in practice, is expressed as:

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

In the Kalman filter that is used for speech enhancement [3], $v(n)$ is a zero-mean, white Gaussian noise that has a variance of σ_v^2 and is uncorrelated with $x(n)$. A pth order linear predictor is used to model the speech signal and together with the corrupting

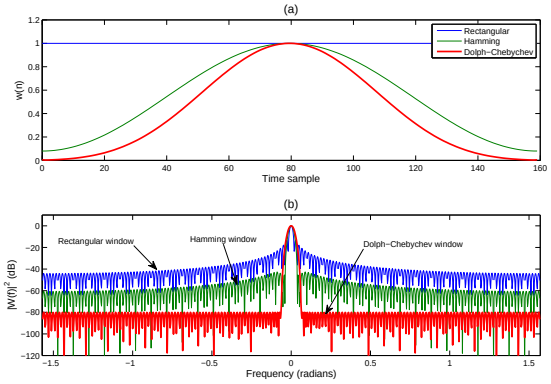

Figure 1: Comparing the rectangular, Hamming, and Dolph-Chebychev (with −80 dB side lobe attenuation) windows in the: (a) time domain; and (b) frequency domain.

noise, we can represent in state vector representation:

$$\boldsymbol{x}(n) = \boldsymbol{A}\boldsymbol{x}(n-1) + \boldsymbol{d}w(n) \tag{2}$$

$$y(n) = \boldsymbol{c}^T\boldsymbol{x}(n) + v(n) \tag{3}$$

where $\boldsymbol{A}$ is the state transition matrix (containing the model parameters), $\boldsymbol{x}(n) = [x(n) \quad x(n-1) \quad \ldots \quad x(n-p+1)]^T$ is the 'hidden' state vector, $\boldsymbol{d} = [1 \quad 0 \quad \ldots \quad 0]^T$ and $\boldsymbol{c} = [1 \quad 0 \quad \ldots \quad 0]^T$ are the measurement vectors for the excitation noise and observation, respectively.

The Kalman filter recursively computes an unbiased and linear MMSE estimate $\hat{\boldsymbol{x}}(n|n)$ of the hidden state vector at time n, given the noisy observation $y(n)$, by using the following equations:

$$\boldsymbol{P}(n|n-1) = \boldsymbol{A}\boldsymbol{P}(n-1|n-1)\boldsymbol{A}^T + \sigma_w^2\boldsymbol{d}\boldsymbol{d}^T \tag{4}$$

$$\boldsymbol{K}(n) = \boldsymbol{P}(n|n-1)\boldsymbol{c}\left[\sigma_v^2 + \boldsymbol{c}^T\boldsymbol{P}(n|n-1)\boldsymbol{c}\right]^{-1} \tag{5}$$

$$\hat{\boldsymbol{x}}(n|n-1) = \boldsymbol{A}\hat{\boldsymbol{x}}(n-1|n-1) \tag{6}$$

$$\boldsymbol{P}(n|n) = [\boldsymbol{I} - \boldsymbol{K}(n)\boldsymbol{c}^T]\boldsymbol{P}(n|n-1) \tag{7}$$

$$\hat{\boldsymbol{x}}(n|n) = \hat{\boldsymbol{x}}(n|n-1) + \boldsymbol{K}(n)[y(n) - \boldsymbol{c}^T\hat{\boldsymbol{x}}(n|n-1)] \tag{8}$$

During the operation of the Kalman filter, the noise-corrupted speech $y(n)$ is windowed into non-overlapped and 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).

When the LPC parameters from clean speech are available, the Kalman filter performs remarkably well [3]. However, when applied in practice, where LPC parameters are estimated from noise-corrupted speech, the performance of the Kalman filter degrades rapidly at low SNRs [4].

2.2. Iterative Kalman filtering

Several iterative Kalman filtering methods have been reported in the literature [5, 6, 7]. In this study, we have focused on the implementation of Gibson et al. [5], where in the first iteration, the LPC parameters are estimated using the noise-corrupted speech

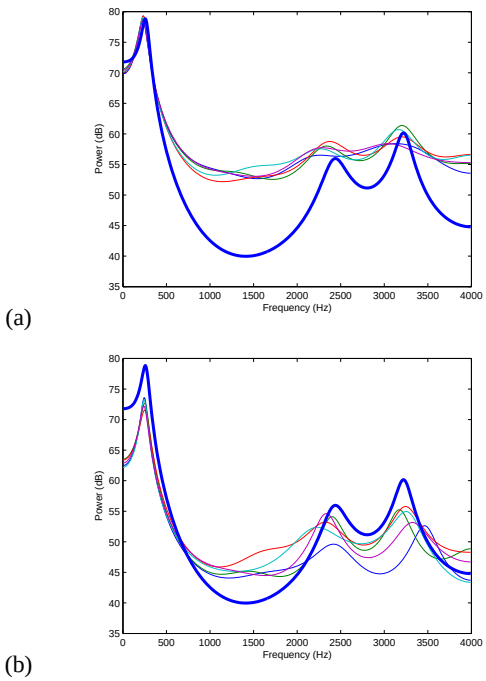

Figure 2: Power spectral estimates from linear predictive model (five realisations) of windowed speech corrupted with white noise at 10 dB SNR (thick line represents clean speech): (a) rectangular window; (b) Dolph-Chebychev window.

frame and the Kalman filter enhances the speech frame using the recursive equations given in Section 2.1. Then in subsequent iterations, the Kalman filter-enhanced speech frame is used to re-estimate the LPCs and excitation variances and the frame is filtered again.

This method of iterating between Kalman filtering and parameter re-estimation from the filtered speech was shown to be an approximated EM (Expectation Maximisation) algorithm that does not guarantee improved performance in subsequent iterations [6]. While objective measures such as SNR have been shown to improve in the first few iterations [5], we have found experimentally the enhanced output to contain musical noise and speech distortion at low SNRs.

3. Proposed iterative Kalman filter for improved speech enhancement

3.1. LPC analysis window with large side lobe attenuation

Tapered windows are often used in spectral estimation to reduce the effect of spectral leakage caused by abrupt frame boundaries. This is possible due to the lower spectral side lobes in the magnitude spectrum of tapered windows (such as the Hamming window) when compared with that of the rectangular window, as can be seen in Figure 1. On the other hand, the main spectral lobe of tapered windows is wider, which tends to smooth the magnitude spectrum of the signal. In addition to this, the variance of the windowed signal is also reduced when compared that of the original signal.

We can exploit these properties of tapered windows to obtain a better estimate of the linear predictive model in the initial

iteration. To better appreciate the benefits that tapered windows bring to the linear predictive model, Figures 2(a) and (b) each show five realisations of PSD estimates from a frame of white noise-corrupted speech at 10 dB SNR that has been windowed with rectangular and Dolph-Chebychev windows, respectively. We can see that applying the tapered window has enhanced the dynamic range of the formants and reduced some of the bias in the prediction error variance introduced by the white noise.

In the proposed iterative Kalman filter, we apply a Dolph-Chebychev window with a -200 dB side lobe attenuation during the LPC analysis in the first iteration only. In the subsequent iteration, a rectangular window is used during the LPC analysis.

3.2. Long and overlapped frames

As was done in [4], we operate the iterative Kalman filter on long and overlapped frames of 80 ms, which is shown in Figure 3. Using long frames ensures that autocorrelation estimates (used in the linear prediction analysis) are more statistically reliable while overlapping frames ensures that the model estimates are updated frequently enough. In this study, we applied a frame update of 10 ms and a modified Hanning window as the synthesis window:

$$w_s(n) = 0.5\left[1 - \cos\left(\frac{2\pi n + \pi}{N}\right)\right] \quad (9)$$

for $n = 0, 1, \ldots, N-1$, where N is the number of samples in each frame. Together with the synthesis window, the overlap-add method also provides some temporal averaging, which smooths the transition between successive frames.

4. Speech enhancement experiments

4.1. Experimental setup

In our experiments, we use the NOIZEUS speech corpus, which is composed of 30 phonetically balanced sentences belonging to six speakers [1]. The corpus is sampled at 8 kHz. For our objective experiments, we generate a stimuli set that has been corrupted by additive white Gaussian noise at four SNR levels (0, 5, 10 and 15 dB). The objective evaluation was carried out on the NOIZEUS corpus using the PESQ (perceptual evaluation of speech quality) measure.

The treatment types used in the evaluations are listed below (p is the order of the LPC analysis):

1. Original clean speech (**Clean**);
2. Speech corrupted with white Gaussian noise (**Noisy**);
3. Non-iterative Kalman filter with LPCs estimated from clean speech, 20 ms, $p = 10$, no overlap, rectangular window (**Kalman clean**);
4. Non-iterative Kalman filter with LPCs estimated from noise-corrupted speech, 20 ms, $p = 10$, no overlap, rectangular window (**Kalman noisy**);
5. Iterative Kalman filter [5] with **three iterations** , 20 ms, $p = 10$, no overlap, rectangular window (**Kalman iterative**);
6. Proposed iterative Kalman filter using the Dolph-Chebychev analysis window (-200 dB side lobe attenuation), long 80 ms frames, and **two iterations**, $p = 10$ (**Kalman proposed**); and
7. MMSE-STSA method [9] (**MMSE**).
8. Phase spectrum compensation [8] (**PSC**).

Table 1: Average PESQ results comparing the proposed method with the iterative Kalman filter of [5] for speech corrupted by white noise.

Method	Input SNR (dB)			
	0	5	10	15
No enhancement	1.566	1.829	2.131	2.471
Kalman iterative (1 iter)	1.739	2.059	2.394	2.742
Kalman iterative (2 iter)	1.921	2.288	2.628	2.978
Kalman iterative (3 iter)	2.019	2.396	2.736	3.086
Kalman iterative (4 iter)	2.004	2.359	2.680	3.023
Kalman proposed (2 iter)	**2.176**	**2.502**	**2.819**	**3.142**

Table 2: Average PESQ results comparing the different speech enhancement methods with the proposed method for speech corrupted by white noise. (Iterative Kalman filter results are shown in the bottom half of the table)

Method	Input SNR (dB)			
	0	5	10	15
No enhancement	1.566	1.829	2.131	2.471
Kalman clean	2.499	2.786	3.077	3.383
Kalman noisy	1.739	2.059	2.394	2.742
MMSE-STSA	1.960	2.328	2.640	2.941
PSC	1.965	2.335	2.702	3.065
Kalman iterative	2.019	2.396	2.736	3.086
Kalman proposed	**2.176**	**2.502**	**2.819**	**3.142**

4.2. Results and discussion

Table 1 shows the average PESQ of the proposed method when compared with the iterative Kalman filter of [5], for a varying number of iterations. We can see that the performance of iterative Kalman filter improves up till three iterations and then tapers off. In addition, the proposed method has converged to better PESQ scores using less iterations, which confirms the effectiveness of using long tapered windows in the initial iteration.

Table 2 presents PESQ results from the objective evaluation of the proposed iterative Kalman filter as well as other enhancement methods. We can see that the iterative Kalman filters outperform all enhancement methods, except for the ideal clean case (i.e. where LPCs are estimated from clean speech). This correlates with the spectrograms, where we can see relatively little wideband residual noise in Figures 4 (d) and (h) when compared with the other methods. However, it can be noted that the proposed method does cause an oversuppression of speech, which may affect intelligibility.

From informal listening tests, the conventional iterative Kalman filter was found to suffer from annoying musical tones, which can be noticed in the spectrogram as isolated dots in the non-speech areas (Figure 4(d)). The proposed iterative Kalman filter did not introduce musical tones, but rather, a smooth and slow-varying 'washy' residual artifact was noted in informal listening. We believe this artifact to be less annoying than musical noise.

5. Conclusion

In this paper, we have proposed an iterative Kalman filtering scheme that improves on the iterative Kalman filter of Gibson, et al. [5] by introducing lower and less annoying residual

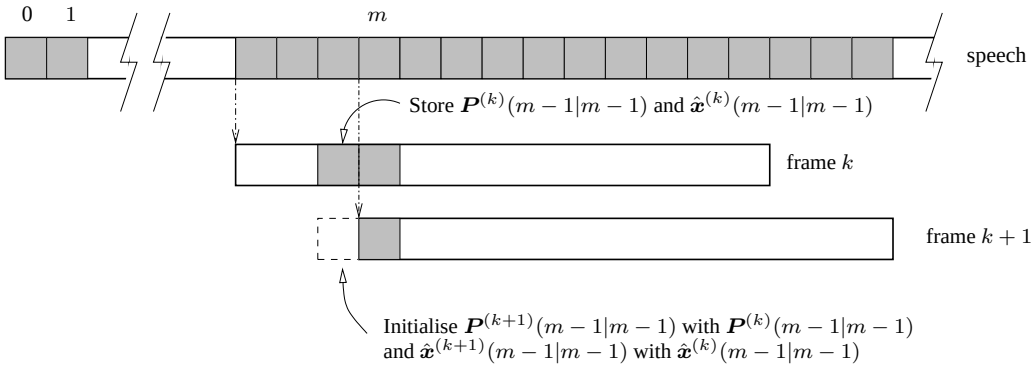

Figure 3: Diagram showing the use of overlapping frames and initialisation of error covariance and state estimate in the Kalman filter.

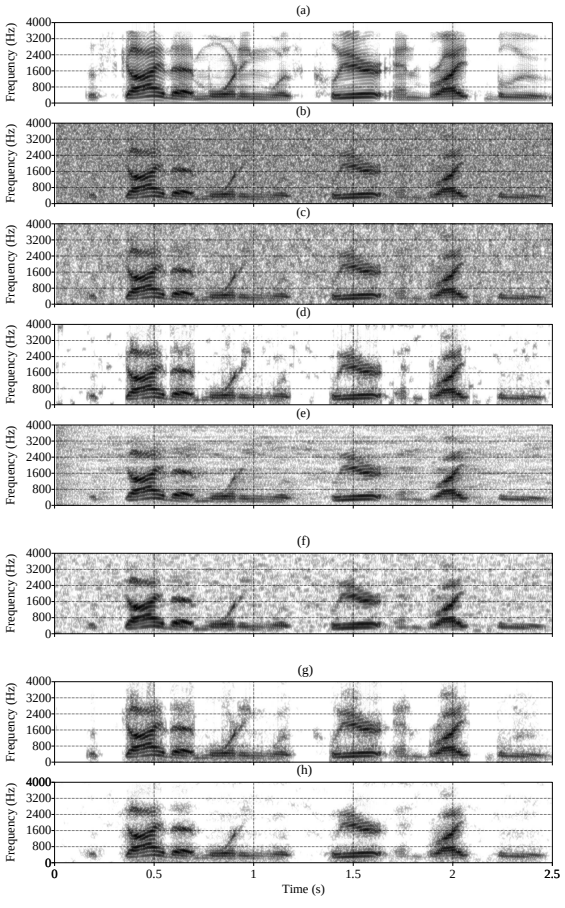

Figure 4: Spectrograms of all treatment types on the NOIZEUS corpus corrupted with white Gaussian noise at an SNR of 5 dB: (a) Clean speech (sp10.wav) 'The sky that morning was clear and bright blue'; (b) noise-corrupted speech; (c) Kalman noisy; (d) Kalman iterative; (e) MMSE-STSA; (f) PSC; (g) Kalman clean; (h) Kalman proposed.

noise as well as faster convergence. We have shown that the Dolph-Chebychev window tends to enhance the dynamic range of the formant structure of speech corrupted with white noise, reduce prediction error bias, as well as provide for some spectral smoothing due to the wider main lobe, while the long overlapped frames introduce temporal smoothing. Speech enhancement experiments that were performed show that the proposed method outperformed conventional iterative and non-iterative Kalman filters as well as other enhancement methods such as MMSE-STSA and PSC.

6. References

[1] P. Loizou, *Speech Enhancement: Theory and Practice*, 1st ed. CRC Press LLC, 2007.

[2] C. J. Li, "Non-Gaussian, non-stationary, and nonlinear signal processing methods – with applications to speech processing and channel estimation," Ph.D. dissertation, Aarlborg University, Denmark, Feb. 2006.

[3] 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.

[4] S. So and K. K. Paliwal, "A long state vector Kalman filter for speech enhancement," in *Proc. Int. Conf. Spoken Language Processing*, Sep. 2008, pp. 391–394.

[5] 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.

[6] S. Gannot, D. Burshtein, and E. Weinstein, "Iterative and sequential Kalman filter-based speech enhancement algorithms," *IEEE Trans. Speech Audio Process.*, vol. 6, no. 4, pp. 373–385, Jul. 1998.

[7] M. Gabrea, E. Grivel, and M. Najim, "A single microphone Kalman filter-based noise canceller," *IEEE Signal Process. Lett.*, vol. 6, no. 3, pp. 55–57, Mar. 1999.

[8] K. K. Wojcicki, M. Milacic, A. P. Stark, J. G. Lyons, and K. K. Paliwal, "Exploiting conjugate symmetry of the short-time Fourier spectrum for speech enhancement," *IEEE Signal Process. Lett.*, vol. 15, pp. 461–464, 2008.

[9] Y. Ephraim and D. Malah, "Speech enhancement using a minimum-mean square error short-time spectral amplitude estimator," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 32, pp. 1109–1121, Dec. 1984.