Available online at www.sciencedirect.com
ScienceDirect
Speech Communication 53 (2011) 355–378
SPEECH COMMUNICATION
www.elsevier.com/locate/specom

Suppressing the influence of additive noise on the Kalman gain for low residual noise speech enhancement

Stephen So *, Kuldip K. Paliwal

Signal Processing Laboratory, Griffith School of Engineering, Griffith University, Brisbane, QLD 4111, Australia

Received 10 March 2010; received in revised form 26 October 2010; accepted 28 October 2010
Available online 17 November 2010

Abstract

In this paper, we present a detailed analysis of the Kalman filter for the application of speech enhancement and identify its shortcomings when the linear predictor model parameters are estimated from speech that has been corrupted with additive noise. We show that when only noise-corrupted speech is available, the poor performance of the Kalman filter may be attributed to the presence of large values in the Kalman gain during low speech energy regions, which cause a large degree of residual noise to be present in the output. These large Kalman gain values result from poor estimates of the LPCs due to the presence of additive noise. This paper presents the analysis and application of the Kalman gain trajectory as a useful indicator of Kalman filter performance, which can be used to motivate further methods of improvement. As an example, we analyse the previously-reported application of long and overlapped tapered windows using Kalman gain trajectories to explain the reduction and smoothing of residual noise in the enhanced output. In addition, we investigate further extensions, such as Dolph–Chebychev windowing and iterative LPC estimation. This modified Kalman filter was found to have improved on the conventional and iterative versions of the Kalman filter in both objective and subjective testing.
© 2011 Elsevier B.V. All rights reserved.

Keywords: Kalman filtering; Speech enhancement; Linear prediction; Dolph-Chebycher windows

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 such as the half-rate PDC (Personal Digital Cellular) standard (Ohya et al., 1994) and MELPe (Mixed Excitation Linear Predictive enhanced) standard (Wang et al., 2002) 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 (Boll, 1979), MMSE–STSA (minimum mean square error, short-term spectral amplitude) estimation (Ephraim and Malah, 1984), Wiener filtering (Wiener, 1949), subspace methods (Ephraim and Van Trees, 1995), and Kalman filtering (Paliwal and Basu, 1987).

The *discrete Kalman filter*[1] is an unbiased, time-domain, linear minimum mean squared error (MMSE) estimator[2] that originated from control systems theory (Kalman, 1960). Its role is to estimate the unknown states of a dynamic system, using a weighted summation of noise-corrupted observations and the predicted states obtained from a dynamic model. The Kalman filter 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

* Corresponding author.
E-mail addresses: s.so@griffith.edu.au (S. So), k.paliwal@griffith.edu.au (K.K. Paliwal).

[1] For simplicity, we drop the prefix 'discrete' from here onwards.
[2] The Kalman filter becomes a true MMSE estimator when the noise is Gaussian (Ma et al., 2006).

0167-6393/$ - see front matter © 2011 Elsevier B.V. All rights reserved.
doi:10.1016/j.specom.2010.10.006

made inherent in the Kalman recursion equations by using a linear predictor as the dynamic model. Secondly, when accurate linear prediction coefficients (LPCs) are available, the enhanced speech from the Kalman filter contains no random frequency tones (otherwise known in the literature as *musical noise*) (Ma et al., 2006). Thirdly, the Kalman filter can process non-stationary speech signals, unlike the Wiener filter, which assumes the speech and noise are stationary. Furthermore, in contrast to other enhancement techniques such as the Wiener filter, the Kalman filter can be 'turned-on' at the first sample $n = 0$, where the recursion parameters such as the state vector and error covariance matrix are initialised with their expected values (Hayes, 1996). Lastly, in the ideal non-stationary case, the Kalman filter can be viewed as a joint estimator for both the magnitude and phase spectrum of speech (Li, 2006).

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 by Paliwal and Basu (1987). 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 LPC and excitation variance estimates obtained using conventional spectral estimation methods (such as the autocorrelation method (Makhoul, 1975)) will have large estimation variance (see Fig. 1(a)) and bias (see Fig. 1(b)). The enhanced speech from this suboptimal Kalman filter has been reported to contain a large amount of wideband residual noise (So and Paliwal, 2008).

One method of overcoming this problem is to use an iterative estimation method (Gibson et al., 1991), which is similar to the iterative Wiener filter proposed by Lim and Oppenheim (1978) and is essentially an approximated EM (Expectation–Maximisation) algorithm (Gannot et al., 1998). In this method, more refined LPC estimates are calculated using the enhanced output of the Kalman filter from the previous iteration. While this iterative LPC estimation method 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 does not adequately address the problem of poor LPC estimates. A better estimation method of LPCs from the noise-corrupted speech is required, especially during the initial iteration. Gannot et al. (1998) developed an iterative method for Kalman filtering which unlike the method of Gibson et al. (1991), is a true EM algorithm and is guaranteed to monotonically increase the likelihood function of the estimated parameters. Yet another iterative approach, which is based on the work by Mehra (1970) in the control literature, estimates the Kalman gain without requiring the knowledge of the noise and process variances (Gabrea et al., 1999). However, the method achieved lower SNRs than the one by Gibson et al. (1991).

In this paper, we analyse how the LPC estimation error affects the operation of the Kalman filter in practical situations, whereby only noise-corrupted speech is available. In particular, we show the Kalman gain trajectory to be a useful indicator of the error of the LPC estimates and ultimately the enhancement performance of the Kalman filter. This detailed study will enable the motivation of new methods to correct the Kalman gain as well as explain the effectiveness of recently reported methods for reducing residual noise in the output. These methods consist of two modifications to the conventional Kalman filtering scheme (So and Paliwal, 2008): the use of long and overlapping speech frames; and the application of tapered windows during LPC estimation. The Kalman gain analysis not only explains the effectiveness of these two modifications, but it

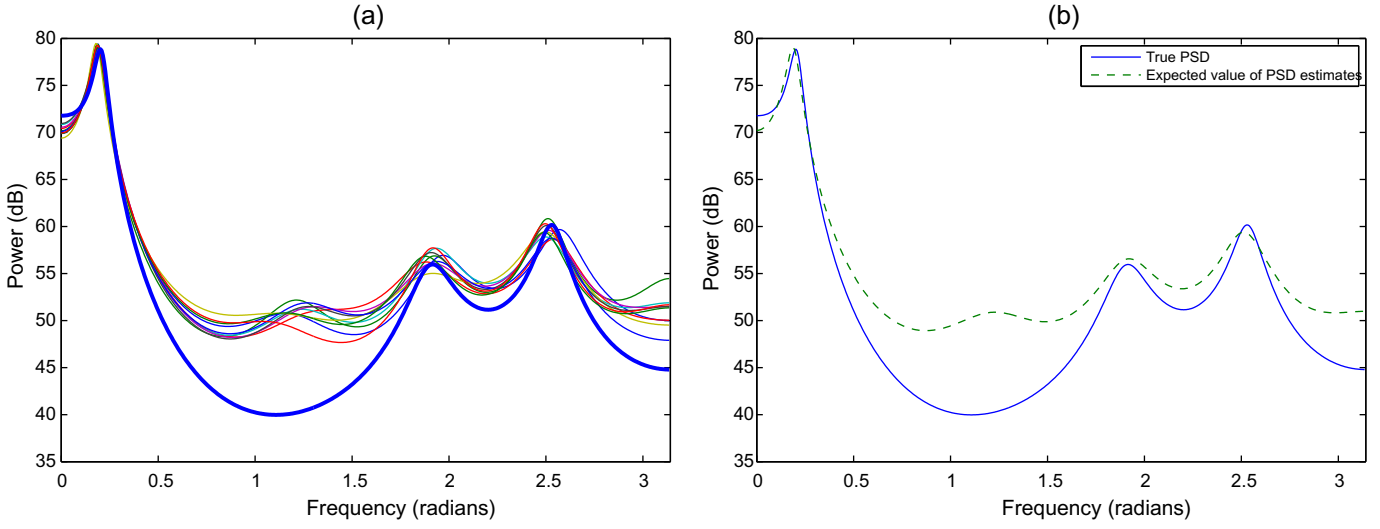

Fig. 1. The effect of white noise (10 realisations) at an SNR of 15 dB on LPC estimation showing: (a) the variance of the PSD estimates (thick line is the true PSD estimate); and (b) the expected value of the PSD estimates.

also motivates further extensions to the method, namely the use of Dolph–Chebychev windows with large sidelobe attenuation and iterative LPC estimation. The improvements provided by this method will be verified in objective and subjective speech enhancement experiments performed on the NOIZEUS speech corpus.

This paper is organised as follows. In Section 2, we give an introduction to speech enhancement using the Kalman filter as well as brief the reader on the mathematical notation used in the recursive equations. We also attempt to pinpoint the cause of suboptimality in the Kalman filter when only noise-corrupted speech is available, which will include a discussion on the importance of the Kalman gain and its time trajectories. In Section 3, we propose several modifications to the Kalman filter that effectively suppress the influence of additive noise on the Kalman gain trajectory. The effectiveness of these modifications will be shown via analysis of the Kalman gain trajectories, spectrograms, and objective measures such as SNR and segmental SNR. Section 4 describes the speech enhancement experiments that were performed on the NOIZEUS corpus (Hu and Loizou, 2006b) to verify the performance of the proposed Kalman filter and compare it against other enhancement methods, such as the iterative Kalman filter (Gibson et al., 1991) and the MMSE–STSA method (Ephraim and Malah, 1984). White Gaussian noise and coloured car noise at varying signal-to-noise ratios (SNRs) are investigated. Two sets of objective scores are presented: (1) SNR, segmental SNR, and PESQ; and (2) composite measures simulating the ITU-T P.835 methodology (Hu and Loizou, 2006a). Subjective preference scores from a blind AB listening test are also presented. Finally, we offer our concluding remarks in Section 5.

2. Speech enhancement using the Kalman filter

2.1. The Kalman recursion equations

If the clean speech is represented as $x(n)$ and the noise signal as $v(n)$ (for $n = 0, 1, \ldots, N$), 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 (Paliwal and Basu, 1987), $v(n)$ 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 $\{a_k;\ k = 1, 2, \ldots, p\}$ are the LPCs and $w(n)$ is a white Gaussian excitation with zero mean and a variance of σ_w^2. Fig. 2 shows a block diagram of the speech production model and additive noise. Rewriting Eqs. (1) and (2) using state vector representation:

$$\mathbf{x}(n) = \mathbf{A}\mathbf{x}(n-1) + \mathbf{d}w(n), \tag{3}$$

$$y(n) = \mathbf{c}^T\mathbf{x}(n) + v(n), \tag{4}$$

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

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

The Kalman filter calculates $\hat{\mathbf{x}}(n|n)$, which is an unbiased, linear MMSE estimate of the state vector $\mathbf{x}(n)$, given the samples of corrupted speech up to time n (i.e. $y(1)$, $y(2), \ldots, y(n)$), by using the following recursive equations:

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

$$\mathbf{P}(n|n-1) = \mathbf{A}\mathbf{P}(n-1|n-1)\mathbf{A}^T + \sigma_w^2\mathbf{d}\mathbf{d}^T, \tag{7}$$

$$\mathbf{K}(n) = \mathbf{P}(n|n-1)\mathbf{c}\left[\sigma_v^2 + \mathbf{c}^T\mathbf{P}(n|n-1)\mathbf{c}\right]^{-1}, \tag{8}$$

$$\hat{\mathbf{x}}(n|n) = \hat{\mathbf{x}}(n|n-1) + \mathbf{K}(n)[y(n) - \mathbf{c}^T\hat{\mathbf{x}}(n|n-1)], \tag{9}$$

$$\mathbf{P}(n|n) = [\mathbf{I} - \mathbf{K}(n)\mathbf{c}^T]\mathbf{P}(n|n-1), \tag{10}$$

We briefly explain what each variable in the above recursive equations represents:

- $\hat{\mathbf{x}}(n|n-1)$ is the *a priori* estimate of the current state vector at time n, given the observations up to $n-1$;
- σ_v^2 is the variance of the corrupting noise $v(n)$;
- $\mathbf{P}(n|n-1)$ is the error covariance matrix of the *a priori* estimate, $\hat{\mathbf{x}}(n|n-1)$, i.e. $E\{\mathbf{e}(n|n-1)\,\mathbf{e}^T(n|n-1)\}$, where $E\{\cdot\}$ is the expectation operator and $\mathbf{e}(n|n-1) = \mathbf{x}(n) - \hat{\mathbf{x}}(n|n-1)$;
- The expression $y(n) - \mathbf{c}^T\hat{\mathbf{x}}(n|n-1)$ in Eq. (9) is termed the *innovation*, as it represents the information contained in the current observation $y(n)$ that cannot be predicted by the dynamic model;

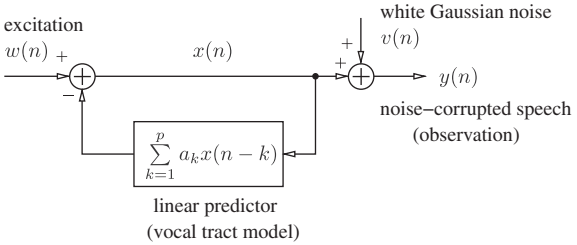

Fig. 2. Block diagram showing the speech production model (linear predictor) and white Gaussian noise that is added to the speech.

[3] For the case where speech has been corrupted by coloured noise, an additional linear predictor of order q is used to model the coloured noise, which augments the Kalman state vector to a size of $p + q$ (Gibson et al., 1991).

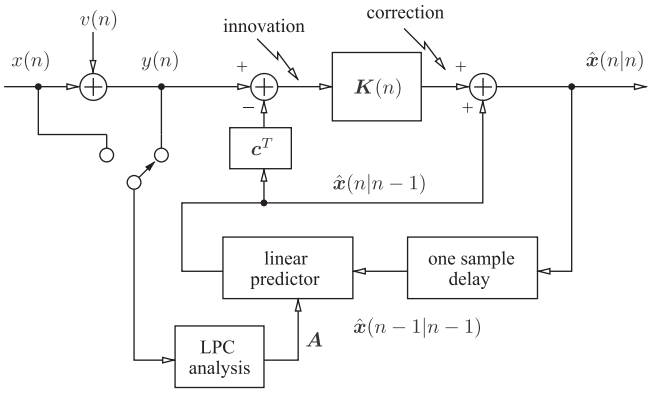

Fig. 3. Block diagram of the Kalman filter. Bolded variables are vectors.

- $\hat{\mathbf{x}}(n|n)$ is the *a posteriori* estimate of $\mathbf{x}(n)$, given the observations up to n, which consists of the *a priori* estimate $\hat{\mathbf{x}}(n|n-1)$ corrected with the innovation weighted by the Kalman gain $\mathbf{K}(n)$; and
- $\mathbf{P}(n—n)$ is the error covariance matrix of the *a posteriori* estimate.

The current estimated sample is then given by $\hat{x}(n) = \mathbf{c}^T\hat{\mathbf{x}}(n|n)$, which extracts the first component of the estimated state vector. Fig. 3 shows a block diagram of the Kalman filter.

During the operation of the Kalman filter, the noise-corrupted speech $y(n)$ is windowed into short (e.g. 20 ms) and non-overlapped frames, where 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 $\mathbf{K}(n)$ and error covariance $\mathbf{P}(n|n)$) and state vector estimate $\hat{\mathbf{x}}(n|n)$ are continually updated on a sample-by-sample basis (regardless of which frame we are in).

2.2. Zero-lag and fixed-lag Kalman filters

Two types of Kalman filter were discussed by Paliwal and Basu (1987), which we will refer to in this paper as the *zero-lag* and *fixed-lag Kalman filter*.[4] These are shown in Fig. 4. In the *zero-lag Kalman filter*, the current enhanced speech sample $\hat{x}(n)$ is formed by taking the first component of the estimated state vector $\hat{\mathbf{x}}(n|n)$. 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 *fixed-lag Kalman filter* (also referred to as the delayed Kalman filter by Paliwal and Basu (1987)), the past enhanced speech sample $\hat{x}(n-p+1)$ is formed by taking the last component of the estimated state vector $\hat{\mathbf{x}}(n|n)$. In effect, this past speech sample is *re-estimated* using information that is based on:

[4] Note that the term 'lag' refers to the delay caused by block-based processing, rather than the computational delay.

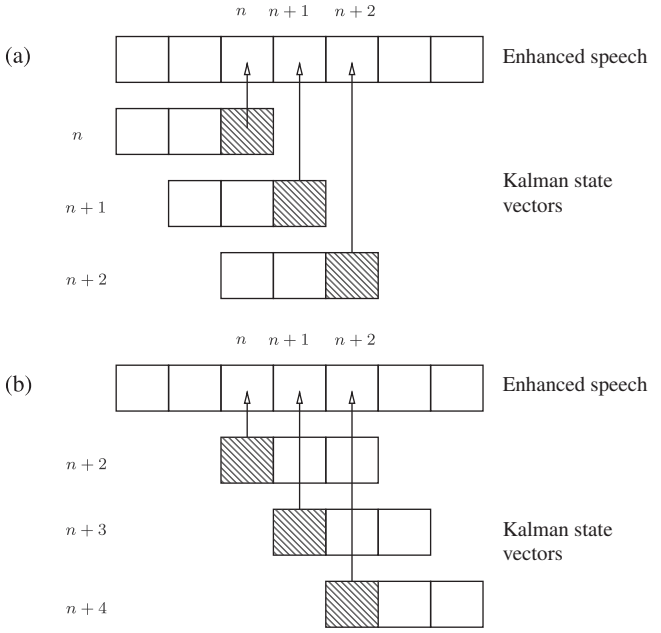

Fig. 4. Diagram showing the difference between the two types of Kalman filters (of state vector length of 3) described by Paliwal and Basu (1987): (a) In the *zero-lag Kalman filter*, the first component (shown as a shaded box) of each estimated state vector is taken as the estimated speech sample; (b) In the *fixed-lag Kalman filter*, the last component (shown as a shaded box) of each estimated state vector is taken as the estimated speech sample.

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

Therefore, the fixed-lag Kalman filter, which we will adopt in our study, can be viewed as a fixed-lag smoother, which was shown to give better enhancement than the zero-lag Kalman filter (Paliwal and Basu, 1987). Fig. 5 shows the spectrograms of the clean, noisy, and enhanced speech from the zero-lag and fixed-lag Kalman filter. Note that the LPCs were estimated from the clean speech in this example, hence this can be viewed as the best case scenario for the Kalman filter.

2.3. Using LPCs estimated from noise-corrupted speech in the Kalman filter

In practice, the clean speech is not available and so the LPCs need to be estimated from the noise-corrupted speech. Fig. 6 shows the effect of noise on the LPC estimates, where five realisations of white Gaussian noise have been added to a frame of speech at varying SNRs. The LPCs were estimated using the autocorrelation method (Makhoul, 1975). It can be seen that the spectral tilt as well as the sharpness and dynamic range of the two formants above 2 kHz have been reduced.

To understand the repercussions of inaccurate LPCs on the Kalman filter, we will firstly examine the effect that the

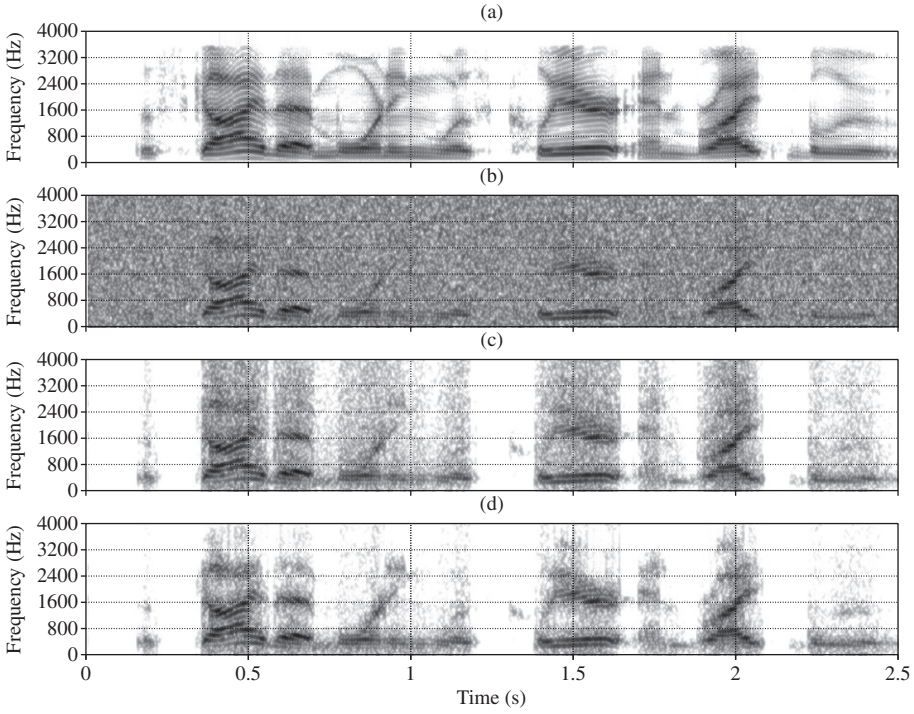

Fig. 5. Speech enhancement from the zero-lag and fixed-lag of Kalman filter that uses LPCs obtained from the clean speech: (a) spectrogram of the clean speech utterance (sp10.wav) 'The sky that morning was clear and bright blue'; (b) spectrogram of the noise-corrupted speech (SNR = 0 dB; Seg-SNR = −7.5131 dB); (c) Spectrogram of enhanced speech using the zero-lag Kalman filter (SNR = 7.39 dB, Seg-SNR = 3.9024 dB); (d) Spectrogram of enhanced speech using the fixed-lag Kalman filter (SNR = 8.8614 dB, Seg-SNR = 4.8894 dB).

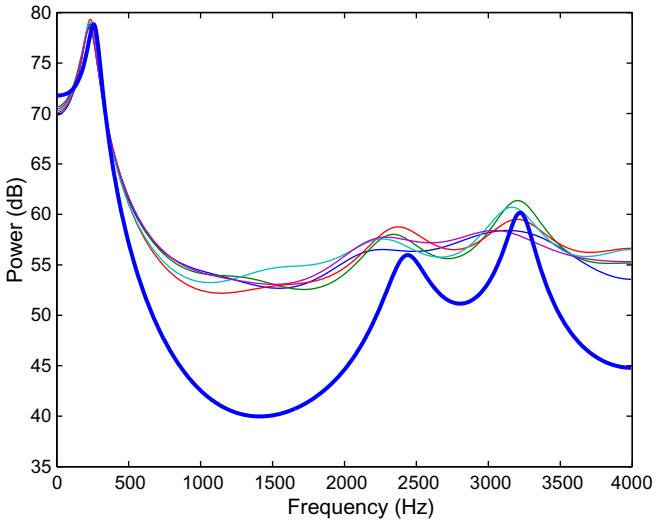

Fig. 6. Spectral envelopes from LPCs estimated from clean speech and speech corrupted with white noise (five realisations) at an SNR of 10 dB. A rectangular window has been applied before the LPC analysis. The thick line represents the true PSD envelope.

Kalman gain $\mathbf{K}(n)$ has on the enhancement performance by rearranging the Kalman recursion equations Eqs. (6)–(9). This will identify the source of Kalman-induced noise in the filter output, when noise-corrupted speech is used for LPC estimation. Following on from this, we can demonstrate the link between the second order statistics of the corrupting noise $v(n)$ and the Kalman gain $\mathbf{K}(n)$ via the *a priori* error covariance matrix $\mathbf{P}(n|n-1)$. The derivations in this section will provide the basis for using the Kalman gain trajectory as an effective indicator of the Kalman filter enhancement performance. This will motivate other methods for improving the performance of the Kalman filter, which we will examine in the second half of this paper. We should point out that other related analyses of the Kalman gain can be found in many books on control systems, estimation theory, and statistical signal processing (such as Kay, 1993; Hayes, 1996; Åström and Wittenmark, 1997; Haykin, 2002).

Let us start by examining Eq. (8), which calculates the Kalman gain vector, $\mathbf{K}(n)$. In this equation, $\mathbf{P}(n|n-1)$ is the covariance matrix of the error between the current state vector $\mathbf{x}(n)$ and *a priori* state vector estimate $\hat{\mathbf{x}}(n|n-1)$:

$$\mathbf{P}(n|n-1) = E\{\mathbf{e}(n|n-1)\mathbf{e}^T(n|n-1)\}, \tag{11}$$

where:

$$\mathbf{e}(n|n-1) = \mathbf{x}(n) - \hat{\mathbf{x}}(n|n-1), \tag{12}$$

$$= \mathbf{x}(n) - \mathbf{A}\hat{\mathbf{x}}(n-1|n-1). \tag{13}$$

If the LPCs, $\{a_1, a_2, \ldots, a_p\}$, were estimated from noise-corrupted speech $y(n)$ with a low SNR, then the prediction error $\mathbf{e}(n|n-1)$ will tend to be large and consequently, the diagonal elements of $\mathbf{P}(n|n-1)$ will be large as well, since the trace of $\mathbf{P}(n|n-1)$ is equal to the mean squared error. In Eq. (8), the term $\mathbf{c}^T\mathbf{P}(n|n-1)\mathbf{c}$ extracts the top left element of $\mathbf{P}(n|n-1)$ (which we refer to as $P_{0,0}(n|n-1)$),

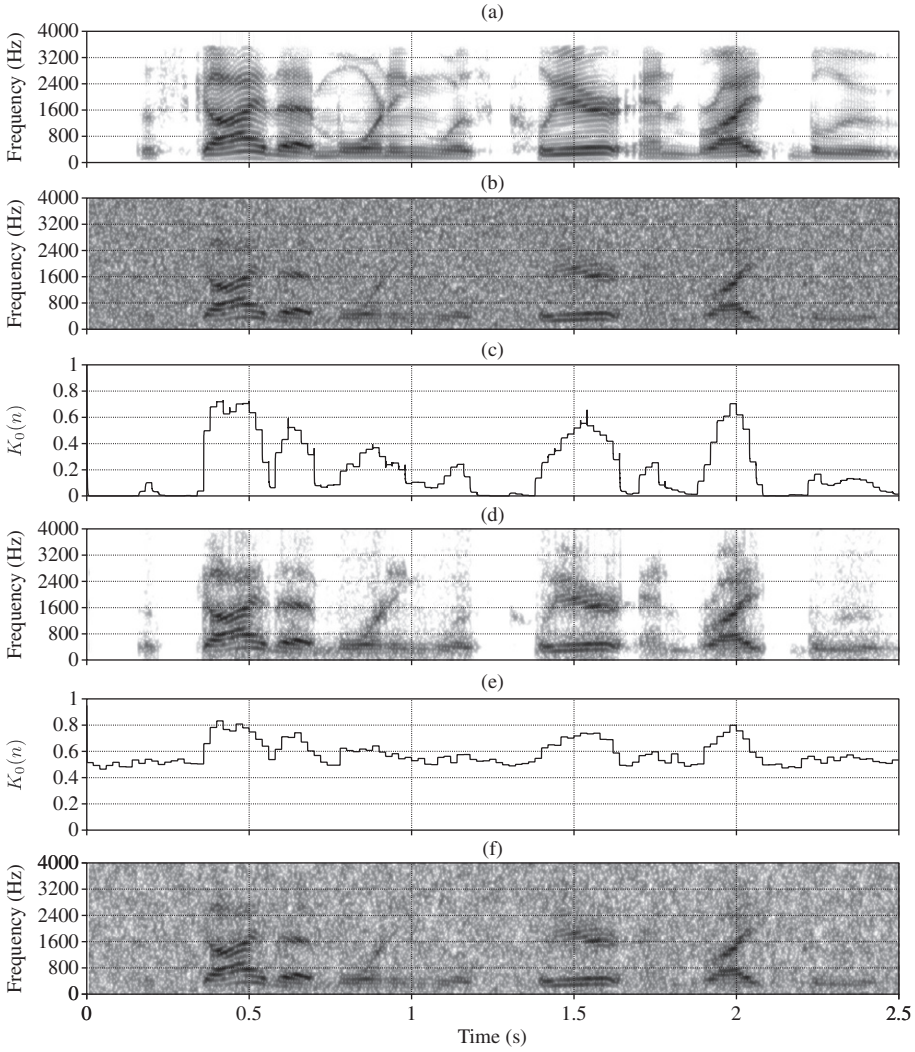

Fig. 7. The influence of LPC estimates and Kalman gain on the enhancement performance of the Kalman filter: (a) spectrogram of the clean speech utterance (sp10.wav); (b) spectrogram of speech corrupted with white noise at SNR of 0 dB; (c) first component of Kalman gain, $K_0(n)$, when *LPCs are estimated from clean speech*; (d) spectrogram of enhanced speech using Kalman filter of part (c); (e) plot of the first component of Kalman gain, $K_0(n)$, when *LPCs are estimated from noise-corrupted speech*; (f) spectrogram of enhanced speech using Kalman filter of part (e) (SNR = 4.293 dB, Seg-SNR = −3.194 dB).

which is the predicted error variance of the first component of $\hat{\mathbf{x}}(n|n-1)$, while the term $\mathbf{P}(n|n-1)\mathbf{c}$ takes the first column of $\mathbf{P}(n|n-1)$ (which we refer to as $\mathbf{P}_0(n|n-1)$).

The first component of the Kalman gain vector (which we will refer to as $K_0(n)$), can be expressed as (by rewriting Eq. (8)):

$$K_0(n) = \frac{P_{0,0}(n|n-1)}{\sigma_v^2 + P_{0,0}(n|n-1)}. \tag{14}$$

If the variance of the corrupting white noise σ_v^2 is close to zero, then $K_0(n)$ will be close to one. If σ_v^2 is much larger than the predicted error variance, then $K_0(n)$ will be close to zero. We rearrange Eq. (9) to give the following form:

$$\hat{\mathbf{x}}(n|n) = [\mathbf{I} - \mathbf{K}(n)\mathbf{c}^T]\hat{\mathbf{x}}(n|n-1) + \mathbf{K}(n)y(n). \tag{15}$$

We can see in Eq. (15) that the Kalman gain acts as a regulator, controlling the relative proportions of the noisy observation and predicted state vector that are added to give the *a posteriori* estimate $\hat{\mathbf{x}}(n|n)$. If we consider the first component[5] of all vector quantities in Eq. (15), which we denote with a subscript 0, we can further reduce it down to:

$$\hat{x}_0(n|n) = [1 - K_0(n)]\hat{x}_0(n|n-1) + K_0(n)y(n) \tag{16}$$

$$= [1 - K_0(n)]\hat{x}_0(n|n-1) + K_0(n)x(n) + \underbrace{K_0(n)v(n)}_{\text{residual noise}} \tag{17}$$

Here, we can see that if $K_0(n)$ is equal to one, then the enhanced speech sample would be comprised entirely of the observation $y(n)$. In this case, the observation is deemed by the Kalman filter to be reliable since the corrupting noise variance is zero. On the other hand, if the observations

[5] This is the case for the zero-lag Kalman filter. The analysis of the fixed-lag Kalman filter is slightly more complicated due to the recursive nature of the state vector estimates. We assume here that the fixed-lag Kalman filter can do no worse than the zero-lag Kalman filter.

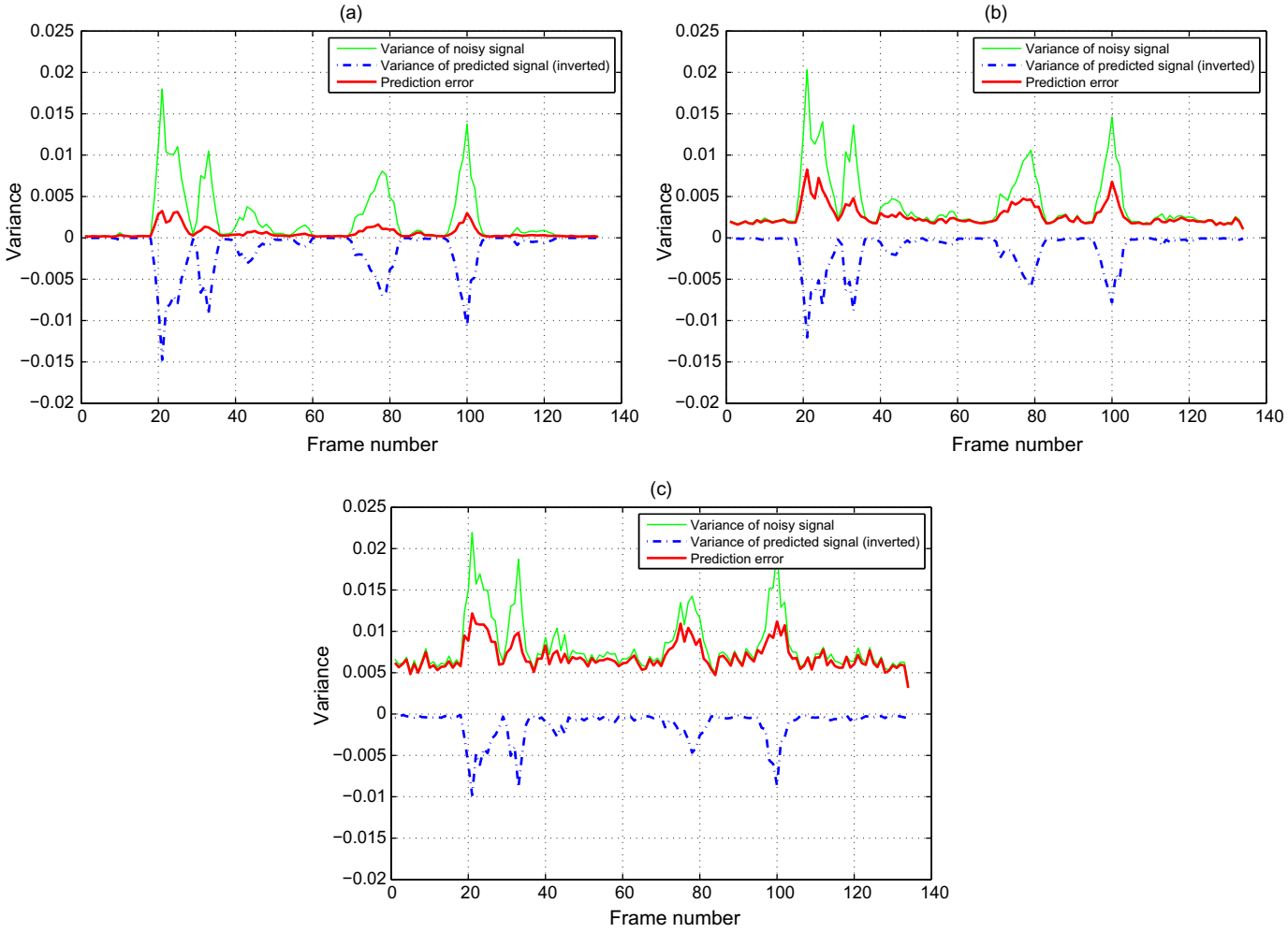

Fig. 8. Comparing variances of noise-corrupted signal $y(n)$, predicted signal (inverted), and prediction error for different SNRs: (a) 10 dB; (b) 0 dB; (c) −5 dB. Note that the thick red line is the summation of the other two.

have a low SNR, then $K_0(n)$ will approach zero, so the enhanced speech sample would be comprised entirely of the predicted component.

Fig. 7 shows the influence of the Kalman gain on the enhancement performance of the Kalman filter. We can see in Fig. 7(c) and (d), where the LPCs have been estimated from the clean speech, that when $K_0(n)$ is close to zero, the output of the Kalman filter is free from noise (during silence periods). In the section of speech starting from 0.36 s, $K_0(n)$ rises to about 0.7, which suggests that the enhanced output consists of 70% of the noisy observation. As can be observed in the enhanced output spectrogram, this observation signal provides the long-term correlation information (i.e. fine structure) that cannot be predicted by the low order linear predictor, accompanied by some noise (represented by the last term in Eq. (17)), which we will refer to as the *residual noise*, since it is a fraction of the observation noise that is passed to the output. Therefore, the amount of residual noise that is present in the enhanced speech is dependent on the value of $K_0(n)$.

If we look at Fig. 7(e) and (f), where the LPCs have been estimated from the noise-corrupted speech, we notice that on average, $K_0(n)$ is quite high (above 0.5) for the entire utterance. This means that the enhanced output is formed from 50% of the noise component $v(n)$, which explains the presence of residual noise. As the variance of the noise, σ_v^2, is relatively stationary throughout the entire utterance, then based on Eq. (14), we may pinpoint the cause to large diagonal elements in $\mathbf{P}(n|n-1)$, or more specifically $P_{0,0}(n|n-1)$. Correspondingly from Eq. (7), we attribute these large values to the variance of the model excitation, σ_w^2, which for the autocorrelation method is given by:

$$\sigma_w^2 = \underbrace{R_{yy}(0)}_{\text{signal variance}} + \underbrace{\sum_{k=1}^{p} a_k R_{yy}(k)}_{\text{prediction variance}}, \tag{18}$$

where $R_{yy}(k)$ is the kth autocorrelation coefficient of the signal $y(n)$.[6] Fig. 8 shows the variances of the noise-corrupted signal (i.e. $R_{yy}(0)$), predicted signal (i.e. $\sum_{k=1}^{p} a_k R_{yy}(k)$), and

[6] Note that Eq. (18) is effectively calculating the error variance by subtracting the predicted signal variance from the variance of the original signal. The summation operation is due to the negative sign convention that we have used for linear prediction.

mean squared prediction error (i.e. σ_w^2) for different SNRs. We can see that as the SNR is lowered, the variance of the noise-corrupted signal (thin green line) goes up due to increasing levels of noise but the variance of the predicted signal (dashed blue line) decreases. This results in a net increase in the prediction error variance.

Assuming the clean speech and noise signals ($x(n)$ and $v(n)$, respectively) are zero-mean and uncorrelated, we can write:

$$R_{yy}(k) = R_{xx}(k) + R_{vv}(k). \tag{19}$$

Hence we can rewrite Eq. (7) as:

$$\mathbf{P}(n|n-1) = \mathbf{A}\mathbf{P}(n-1|n-1)\mathbf{A}^T + \left(R_{xx}(0) + \sum_{k=1}^{p} a_k R_{xx}(k) + R_{vv}(0) + \sum_{k=1}^{p} a_k R_{vv}(k) \right) \mathbf{d}\mathbf{d}^T. \tag{20}$$

We can see from Eq. (20) the presence of extra terms related to the noise $v(n)$ causes estimates of the *a priori* error covariance $\mathbf{P}(n|n-1)$ to be large, resulting in a higher-than-usual Kalman gain and hence more residual noise in the output.

Since $P_{0,0}(n|n-1)$ has been offset by the extra terms related to the noise $v(n)$, we may rewrite Eq. (14) (with $\sigma_v^2 = R_{vv}(0)$):

$$K_0(n) = \frac{\sigma_v^2 + \sum_{k=1}^{p} a_k R_{vv}(k) + P^c_{0,0}(n|n-1)}{2\sigma_v^2 + \sum_{k=1}^{p} a_k R_{vv}(k) + P^c_{0,0}(n|n-1)}, \tag{21}$$

where $P^c_{0,0}(n|n-1)$ denotes top-left element of $\mathbf{P}^c(n|n-1)$, which we denote as the *a priori* error covariance matrix obtained from *clean speech*. That is:

$$\mathbf{P}^c(n|n-1) = \mathbf{A}\mathbf{P}(n-1|n-1)\mathbf{A}^T + \left(R_{xx}(0) + \sum_{k=1}^{p} a_k R_{xx}(k) \right) \mathbf{d}\mathbf{d}^T. \tag{22}$$

If the corrupting noise $v(n)$ is white, then the term $\sum_{k=1}^{p} a_k R_{vv}(k)$ reduces to zero. Therefore, in the non-speech regions, where $P^c_{0,0}(n|n-1) = 0$, Eq. (21) gives us $K_0(n) = 0.5$, as was observed in Fig. 7(e).

We can see that the time trajectories of the first component of the Kalman gain provide useful information about the enhancement performance of the Kalman filter, especially in relation to the amount of residual noise as well as the error of the LPC estimates. In the following sections, we will use the $K_0(n)$ trajectory information to assess the effectiveness of some LPC estimation techniques.

2.4. Iterative estimation of LPCs for Kalman filtering

One way of overcoming the problem of LPC estimates with error from the noise-corrupted speech is to use an iterative method that is similar to that of Lim and Oppenheim (1978). In the first iteration, the LPCs are estimated from the noise-corrupted speech frame and Kalman filtering is performed on this frame to produce an enhanced speech

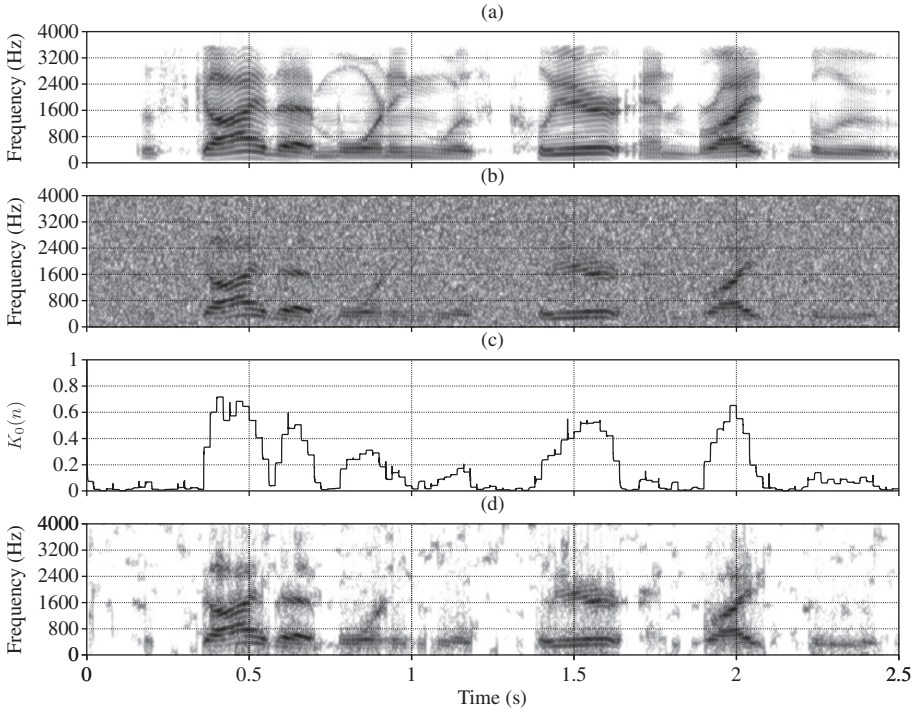

Fig. 9. Kalman filter using iterative estimation of LPCs from noise-corrupted speech (four iterations): (a) spectrogram of clean speech (sp10.wav); (b) spectrogram of speech corrupted with white Gaussian noise at SNR of 0 dB; (c) plot of first component of Kalman gain, $\mathbf{K}_0(n)$; (d) Spectrogram of enhanced speech (SNR = 8.087 dB, Seg-SNR = 3.189 dB).

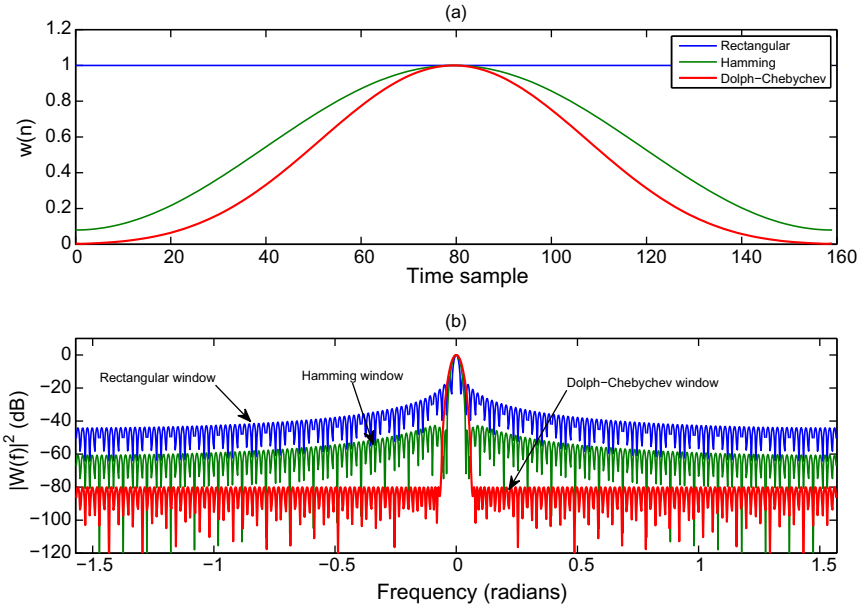

Fig. 10. Comparing the rectangular, Hamming, and Dolph–Chebychev (with −80 dB side lobe attenuation) windows in the: (a) time domain; and (b) frequency domain.

frame. Then in subsequent iterations, the LPCs are re-estimated using the enhanced speech frames from previous iterations.

Fig. 9 shows some enhancement results of the Kalman filter that uses iterative estimation of LPCs from the noise-corrupted speech. We can see that $K_0(n)$ behaves more like the one in Fig. 7(c) and there is evidently less residual noise in the spectrogram of Fig. 9(d) than Fig. 7(f), which indicates that the LPC estimates are more accurate. This improvement in the enhancement of the Kalman filter is evident in the objective measures as well and confirms the results that were reported by Gibson et al. (1991).

However, we notice in Fig. 9(d) that there are still remnants of residual noise that appear as random isolated tones. Informal listening of the enhanced speech suggests that the residual noise is 'musical' in nature and this effect appears to be rarely mentioned in the speech enhancement literature on Kalman filtering. There is also speech distortion that is present in the enhanced speech. While it has been reported that the Kalman filter does not introduce musical noise (Ma et al., 2006), this appears to be based on the assumption that reliable LPC estimates (obtained from the clean speech) are available, which is not the case in practice. The computational complexity of the Kalman filter also increases as more iterations are used.

3. Residual noise reduction with temporal smoothing of the Kalman gain

3.1. Tapered LPC analysis windows with large side lobe attenuation

In this section, we investigate the effect of applying different analysis windows during the LPC estimation stage of the Kalman filter and assess their effectiveness in improving the Kalman filter using $K_0(n)$ trajectories. It is well known in spectral estimation that the application of tapered analysis windows in the time domain de-emphasises the edges of a finite frame of data, which reduces spectral leakage via lower spectral side lobes of the window. It was found in a previous experimental study (So and Paliwal, 2008) that the application of a Hamming window during LPC analysis had improved the enhancement ability of the Kalman filter. In this study, we will examine why the Hamming window is beneficial to the Kalman filter as well as investigate another tapered analysis window, the Dolph–Chebychev window, whose frequency characteristics are shown in Fig. 10.

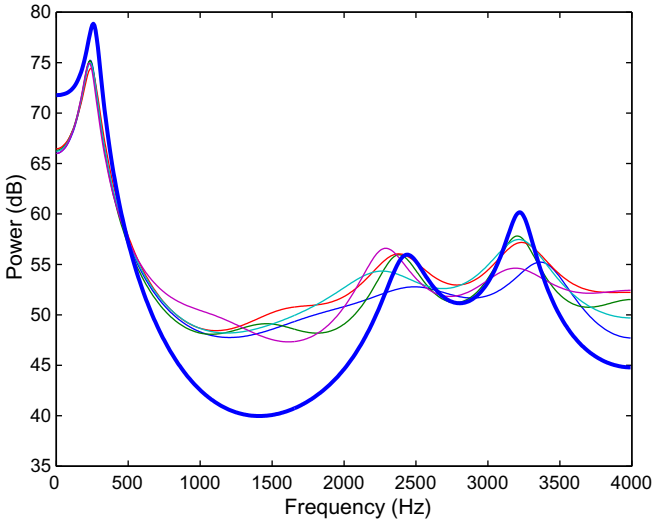

Fig. 11. Spectral envelopes from LPCs estimated from clean speech and speech corrupted with white noise (five realisations) at an SNR of 10 dB. A Hamming window has been applied before the LPC analysis. The thick line represents the true PSD envelope.

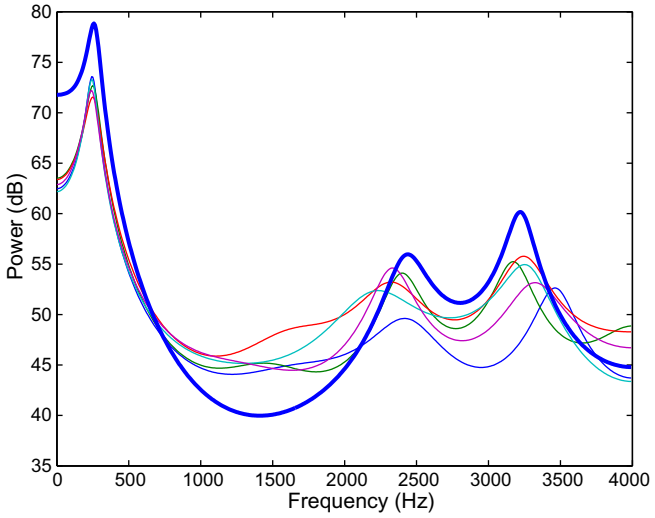

Fig. 12. Spectral envelopes from LPCs estimated from clean speech and speech corrupted with white noise (five realisations) at an SNR of 10 dB. A Dolph–Chebychev window (with −200 dB side lobe attenuation) has been applied before the LPC analysis. The thick line represents the true PSD estimate.

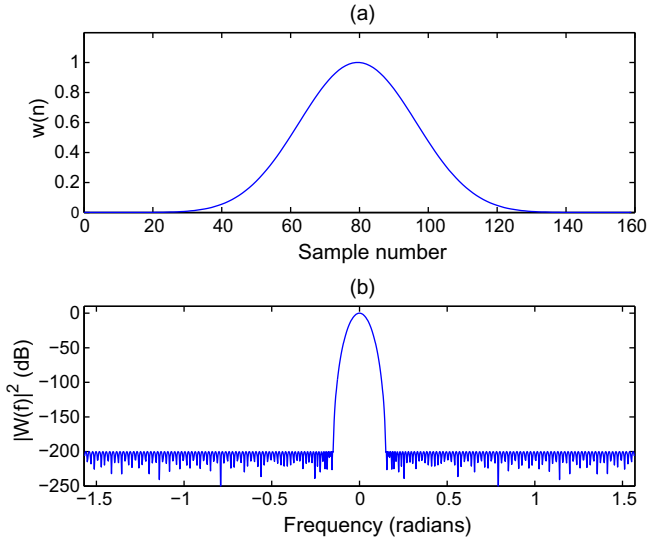

Fig. 13. The 20 ms Dolph–Chebychev window (with −200 dB side lobe attenuation) in the: (a) time domain; and (b) frequency domain.

Let us examine the effect of applying the Hamming window on the spectral envelope of speech that has been corrupted by white noise. The non-iterative fixed-lag Kalman filter was used in the experiments that are reported in this section. Fig. 11 shows the spectral envelope from LPCs that have been estimated from clean and five realisations of noisy speech, where a Hamming window is applied during the LPC analysis. Comparing this with Fig. 6, we can see that the spectral envelopes have lower power, which suggests a reduction in the excitation variance σ_w^2. Also, the dynamic range of the two formants has improved slightly with lower spectral valleys. This indicates that the tapered analysis window has an impact on LPC estimation in the presence of noise or more specifically, the mean squared prediction error, $P_{0,0}(n|n-1)$.

In order to test this theory further, we investigated the Dolph–Chebychev window, whose side lobe attenuation can be adjusted. This window, as shown in Fig. 13, was designed to have a very large side lobe attenuation (−200 dB). Fig. 12 shows the spectral envelopes from LPCs, where the Dolph–Chebychev window has been applied before LPC analysis. We can see that for each realisation, the formants appear to be sharper and more resolvable than either the Hamming or rectangular window. In particular, the spectral valleys are the lowest, which in effect also enhances the formants. We should note that the variance of the spectral envelope estimates appears to have increased with the tapered windows.

These results are consistent with those reported by Erkelens and Broersen (1997), where tapered analysis windows were shown to reduce the bias in the excitation or residual variance of the autocorrelation method. The benefit that this result brings to the Kalman filter is made obvious when we consider the effect of a lower $P_{0,0}(n|n-1)$ in Eq. (21). The reduction of excitation variance via the application of a tapered window counteracts the bias introduced by the corrupting noise (i.e. σ_v^2), which leads to lower Kalman gain trajectories.

Fig. 14 shows the time trajectories of $K_0(n)$ for the different analysis windows. Also shown in this plot is the $K_0(n)$ when LPCs are estimated from clean speech, which serves as the 'best case scenario'. Recall that $K_0(n)$ gives the relative proportion of the observation signal $y(n)$ and we have seen previously that it tends to be high in the voiced speech regions, since they contain long-term correlation information. When using the rectangular window for estimating LPCs from noise-corrupted speech, $K_0(n)$ is high in the regions where there is very little speech and this results in a large amount of residual noise. We can see in Fig. 14 that the tapered analysis windows (Hamming and Dolph–Chebychev) result in lower $K_0(n)$ values than the rectangular window and therefore we expect a reduced level of residual noise. We can see in Fig. 15(d) that there is less residual noise in the Dolph–Chebychev window case. However, we can also see that the $K_0(n)$ for the Dolph–Chebychev window is lower than the $K_0(n)$ of the clean LPC case in the voiced speech regions, hence we expect enhanced speech to have diminished fine structure. In terms of objective measures, the tapered windows give higher SNRs and segmental SNRs than the rectangular window.

3.2. Using long and overlapping speech frames for temporal smoothing

It can be seen in Fig. 14 that $K_0(n)$ appears to fluctuate over time. As $K_0(n)$ also represents the proportion of noise $v(n)$ that is added to the enhanced speech (see the last term of Eq. (17)), these fluctuations result in a non-stationary residual noise that may be annoying to listeners. Looking at Eq. (8) and assuming that the variance of the noise σ_v^2 remains relatively constant, we may attribute these fluctuations to high variance LPC estimates. The increased LPC

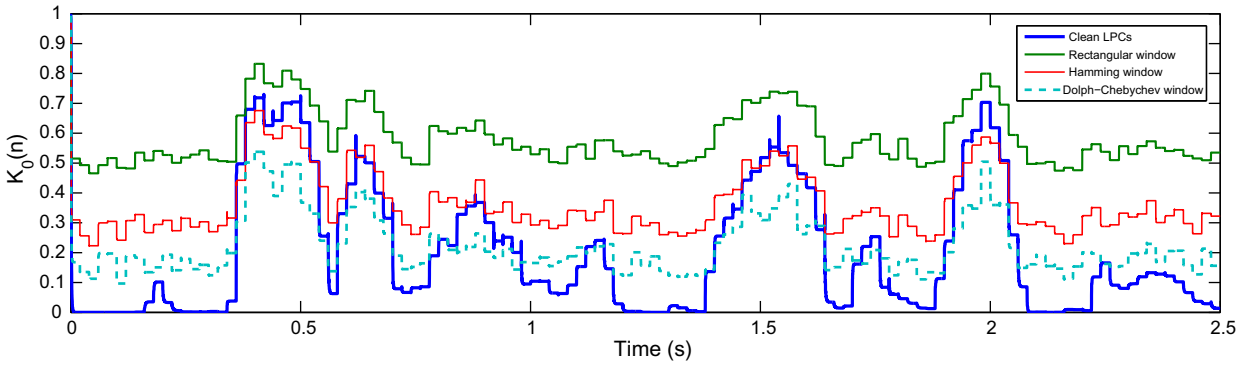

Fig. 14. Time trajectories of the first component of the Kalman gain $K_0(n)$ for clean LPCs and noisy LPCs with different analysis windows.

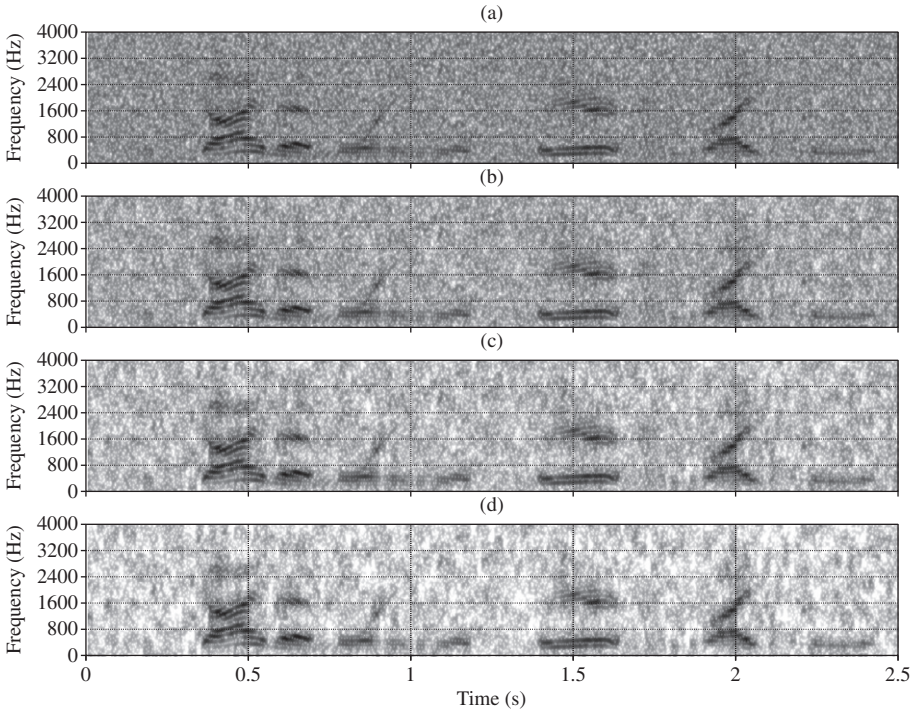

Fig. 15. Speech spectrograms showing the effect of different LPC analysis windows: (a) corrupted speech (SNR = 0 dB; Seg-SNR = − 7.5131 dB; (b) rectangular window (SNR = 4.293 dB, Seg-SNR = −3.194 dB); (c) Hamming window (SNR = 6.531 dB, Seg-SNR = −0.513 dB); (d) Dolph–Chebychev window with −200 dB side lobe attenuation (SNR = 6.522 dB, Seg-SNR = 0.601 dB).

estimation variance is due to the tapering at the two edges of the data frame by the Hamming and Dolph–Chebychev windows, which is equivalent to a decrease in the number of observations available for the estimation of the autocorrelation coefficients (Erkelens and Broersen, 1997).

In order to lower the variance of the LPC estimates, we first seek to lower the variance of autocorrelation estimates $\widehat{R}_{yy}(k)$:

$$\widehat{R}_{yy}(k) = \frac{1}{N}\sum_{n=0}^{N-k-1} y(n)y(n+k). \tag{23}$$

It is well known that when provided with only a fixed number of samples N, the higher lag autocorrelation coefficients become less reliable since the number of summation terms diminishes. We counter this problem by increasing the frame length. Increasing frame length also has the effect of narrowing the main lobe of the Dolph–Chebychev window (shown in Fig. 17), which improves the frequency resolution. In our study, we investigated frame lengths of up to 80 ms. Fig. 18 shows the trajectory of $K_0(n)$ for 20 ms and 80 ms frames, and we can see that effectively, the first component of Kalman gain remains constant for a longer period. However, because speech is generally assumed to be quasi-stationary over a 20–30 ms frame length, the LPC estimates are not being updated frequently enough to capture the changes. This leads to sharp transitions of $K_0(n)$ between successive frames.

We deal with the problem of sharp transitions by using overlapping frames, as shown in Fig. 16. In this procedure,

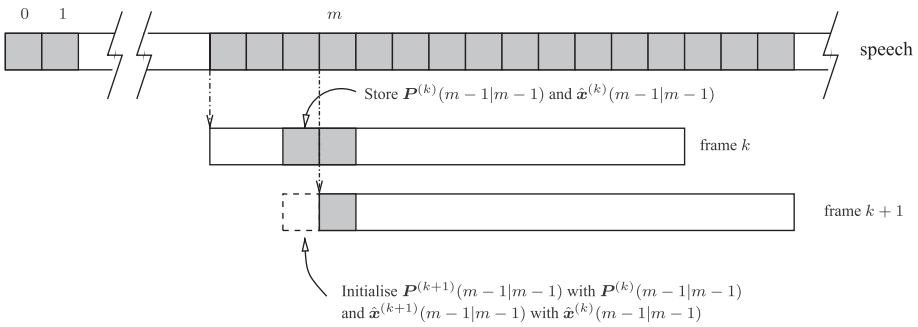

Fig. 16. Diagram showing the use of overlapping frames and initialisation of error covariance and state estimate in the Kalman filter. The previous *a posteriori* error covariance of frame $k+1$, $\mathbf{P}^{(k+1)}(m-1|m-1)$ and state vector estimate $\hat{\mathbf{x}}^{(k+1)}(m-1|m-1)$ are set to $\mathbf{P}^{(k)}(m-1|m-1)$ and $\hat{\mathbf{x}}^{(k)}(m-1|m-1)$ of the previous frame k.

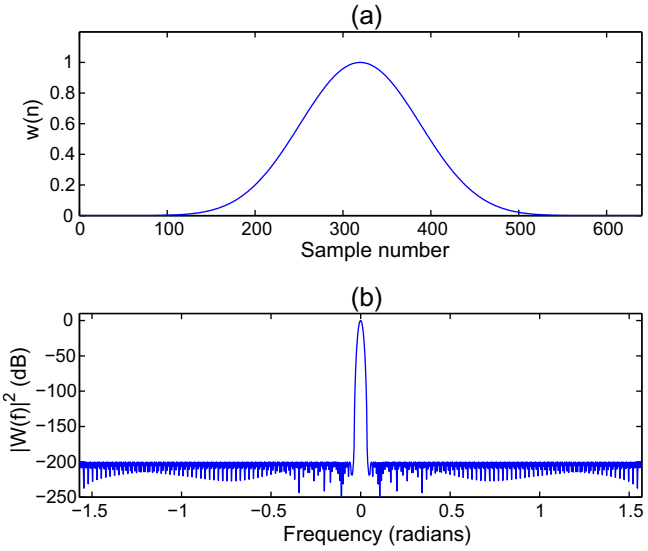

Fig. 17. The 80 ms Dolph–Chebychev window (with −200 dB side lobe attenuation) in the: (a) time domain; and (b) frequency domain.

we provide for an overlap between successive frames by using a window length that is greater than the frame shift. A tapered analysis window $w_a(n)$ (Hamming or Dolph–Chebychev window) is applied during the estimation of the LPCs and excitation variance while the Kalman filtering is performed on the non-windowed, original samples. For frame k, at the sample $(m-1)$ prior to the overlap with the next frame, we store the error covariance matrix $\mathbf{P}^{(k)}(m-1|m-1)$ and state vector estimate $\hat{\mathbf{x}}^{(k)}(m-1|m-1)$. Before we start to filter the next frame $k+1$, we initialise the *a priori* error covariance $\mathbf{P}^{(k+1)}(m-1|m-1)$ and previous state estimate $\hat{\mathbf{x}}^{(k+1)}(m-1|m-1)$ with their corresponding values from the previous frame. As is done in the windowed overlap-add (WOLA) method (Crochiere, 1980), the enhanced frames are multiplied by a synthesis window $w_s(n)$, overlapped and then added together. In this study, we applied a frame shift that was equal to 1/8th of the frame length and a modified Hanning window as the synthesis window:

$$w_s(n) = 0.5\left[1 - \cos\left(\frac{2\pi n + \pi}{N}\right)\right] \tag{24}$$

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 provides some time-based averaging which seems to smooth the transition between successive frames. When we factor the averaging of the synthesis window, we

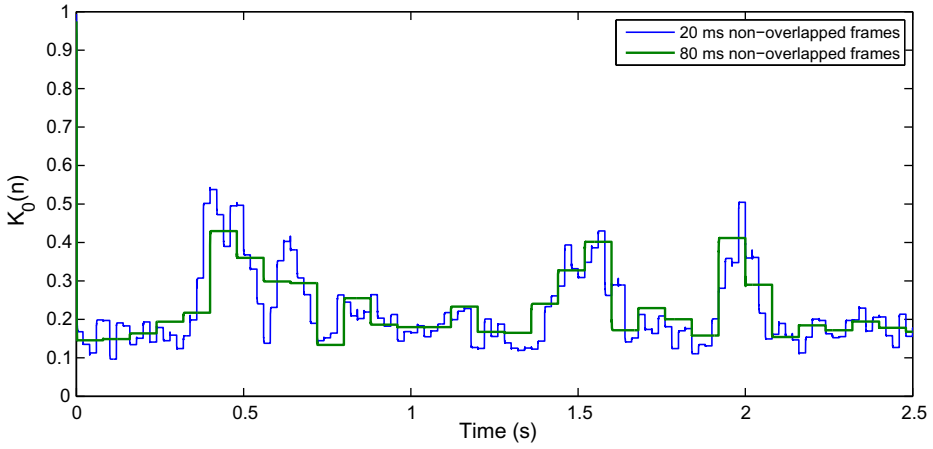

Fig. 18. Comparing time trajectories of the first component of Kalman gain $K_0(n)$ for 20 ms and 80 ms non-overlapping frames. A Dolph–Chebychev window with −200 dB side lobe attenuation was applied before LPC analysis.

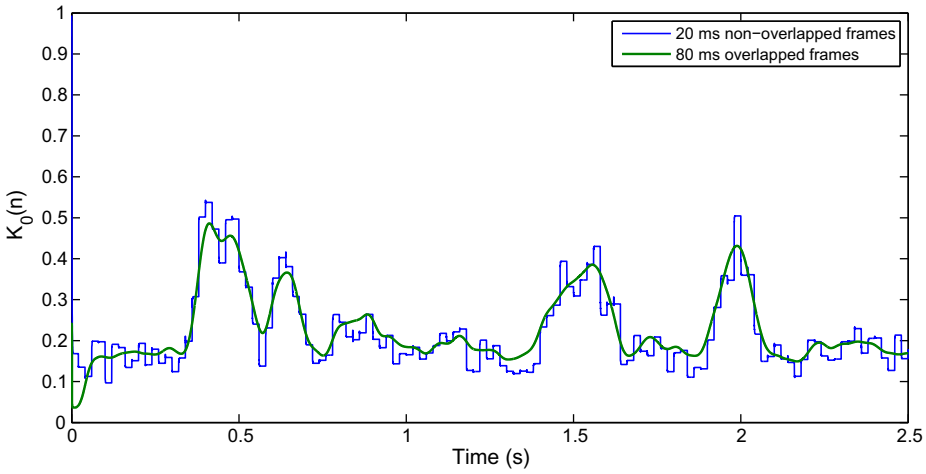

Fig. 19. Comparing time trajectories of the first component of Kalman gain $K_0(n)$ for 20 ms non-overlapping and 80 ms overlapping frames. A Dolph–Chebychev window with −200 dB side lobe attenuation was applied before LPC analysis.

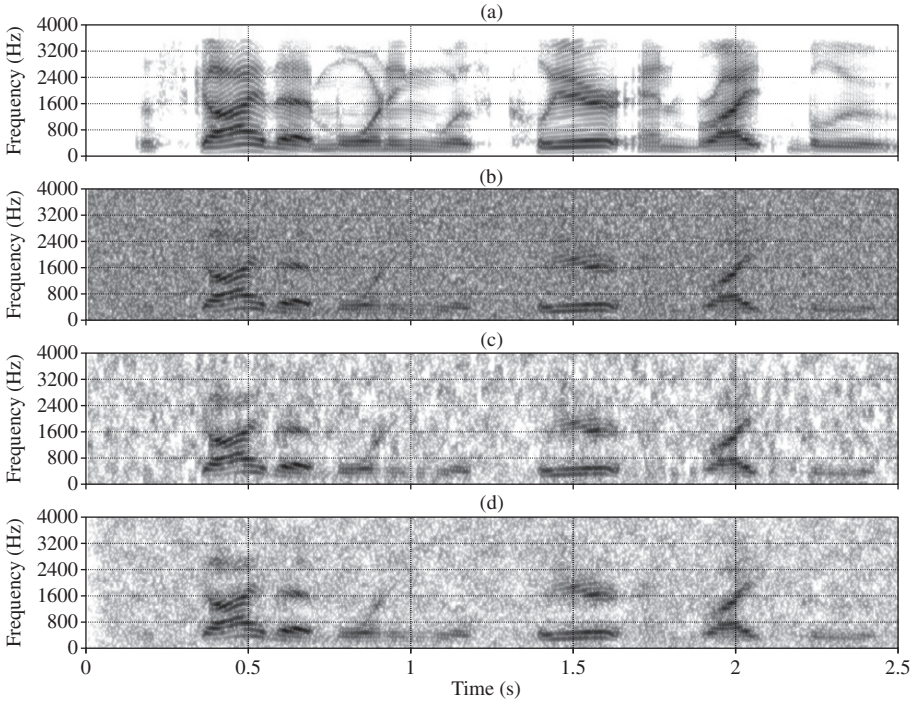

Fig. 20. Speech spectrograms comparing the Kalman filter that uses short and non-overlapping frames with one that uses long and overlapping frames: (a) clean speech; (b) corrupted speech (SNR = 0 dB; Seg-SNR = −7.5131 dB; (c) Kalman-enhanced speech using Dolph–Chebychev window and 20 ms non-overlapping frames (SNR = 6.52 dB; Seg-SNR = 0.601 dB); (d) Kalman-enhanced speech using Dolph–Chebychev analysis window and 80 ms overlapping frames (SNR = 6.85 dB; Seg-SNR = 0.941 dB).

can see in Fig. 19 that the $K_0(n)$ trajectory is not only smoother for the 80 ms overlapped frames, but when compared with the 80 ms non-overlapped frame case (in Fig. 18), it captured more temporal changes (e.g. the peak at around 0.65 s).

Fig. 20 shows the spectrograms of the Kalman-enhanced speech when using 20 ms non-overlapped frames and 80 ms overlapped frames. In both cases, the Dolph–Chebychev window with −200 dB side lobe attenuation was used during LPC analysis. We can see that the residual noise in Fig. 20(d) is more stationary than the noise in (c). Informal listening of the enhanced speech in (d) reveals the residual noise to be less harsh and more 'white' than the residual noise in (c).

Gannot et al. (1998) have reported that the use of 16 ms overlapped analysis frames did not yield improved performance for the Kalman filter. Table 1 shows average objective scores on the NOIZEUS database for different

Table 1
Average objective scores on the NOIZEUS database (white Gaussian noise with SNR of 0 dB) comparing the non-overlapped and overlapping modes (frame update of 1/8th the frame duration) of the Kalman filter. (Note that 'rect' and 'D–C' stands for rectangular and Dolph–Chebychev windows, respectively.)

Method (ms)	SNR (dB)	Seg-SNR (dB)	PESQ
20, no overlap, rect window	4.461	−3.868	1.739
20, with overlap, rect window	4.50	−3.859	1.74
80, no overlap, rect window	4.539	−3.853	1.733
80, with overlap, rect window	4.66	−3.811	1.745
20, no overlap, D–C window	6.756	−0.011	1.829
20, with overlap, D–C window	7.052	0.14	1.866
80, no overlap, D–C window	6.239	0.01	1.808
80, with overlap, D–C window	7.074	0.381	1.908

configurations of the Kalman filter. We can see that when a rectangular analysis window is applied, overlapping the frames did not produce any appreciable improvement, as was observed by Gannot et al. (1998). However, when using long frames with a tapered analysis window, we can see that overlapping the frames had yielded improved performance. This may be explained by the fact that symmetric tapered windows emphasise only the centre of a frame, so overlapping ensures all samples receive the same weighting when used for LPC estimation in successive frames. In addition to this, overlapping long frames ensures more frequent updates, which is important for non-stationary speech signals.

3.3. Iterative LPC estimation with tapered analysis windows and long overlapping frames

In this section, we investigate the use of long and overlapping tapered windows for LPC analysis in the iterative Kalman filter. The problem with the iterative Kalman filter of Gibson et al. (1991) is that it uses LPCs derived from the noise-corrupted speech to be used in the initialisation stage. We have observed that the non-iterative Kalman filter does not do such a good job at enhancing the speech (see Fig. 7), so further iterations are needed to obtain better enhancement. Our aim is to use the methods that we have discussed previously to initialise the iterative LPC estimation, so that less iterations are required. Specifically, we use 80 ms overlapping frames in all iterations. The tapered analysis window is used in the LPC analysis of the initialisation stage only, while the rectangular window is used in subsequent iterations.

Fig. 21 compares the spectrograms of the enhanced speech using the Hamming and Dolph–Chebychev windows, where only two iterations have been performed. We can see that the residual noise in Fig. 21(d) has been reduced when comparing with the spectrograms of Fig. 20(d). When comparing the spectrograms of using the Hamming window (Fig. 21(c)) with Dolph–Chebychev (Fig. 21(d)), the level of residual noise is higher in the former. This is better indicated in Fig. 22, where we can see $K_0(n)$ is always lower in the non-speech regions for the Dolph–Chebychev window. Both tapered windows can be

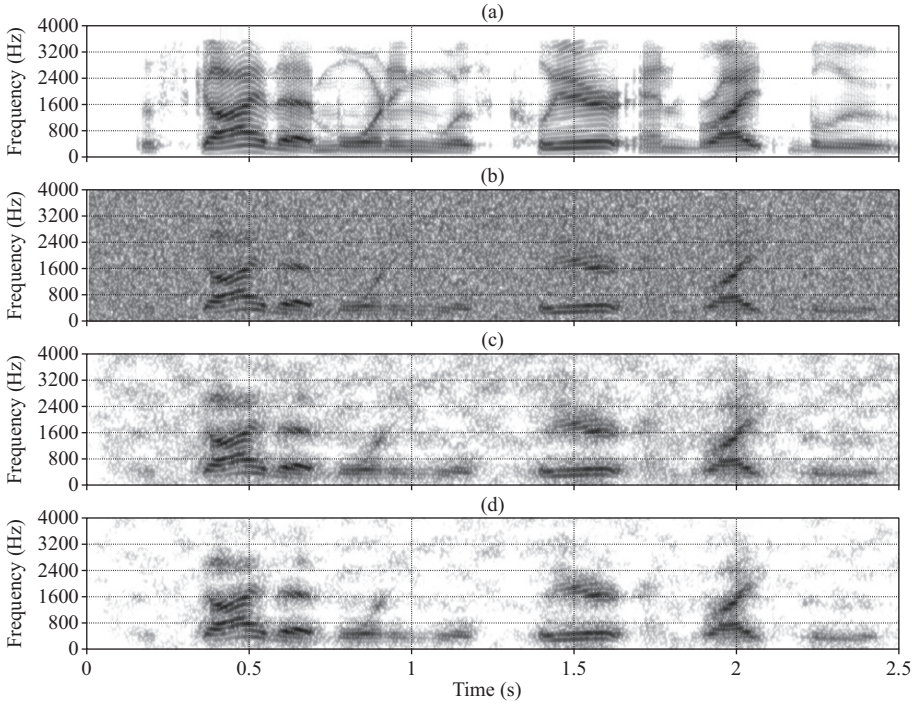

Fig. 21. Speech spectrograms of the iterative Kalman filter (two iterations) that uses 80 ms overlapping frames and tapered analysis window: (a) clean speech; (b) corrupted speech (SNR = 0 dB; Seg-SNR = −7.5131 dB; (c) enhanced speech from iterative Kalman filter with 80 ms overlapping Hamming window (SNR = 8.199 dB; Seg-SNR = 2.516 dB); (d) Kalman-enhanced speech using Dolph–Chebychev analysis window and 80 ms overlapping frames (SNR = 7.67 dB; Seg-SNR = 3.25 dB).

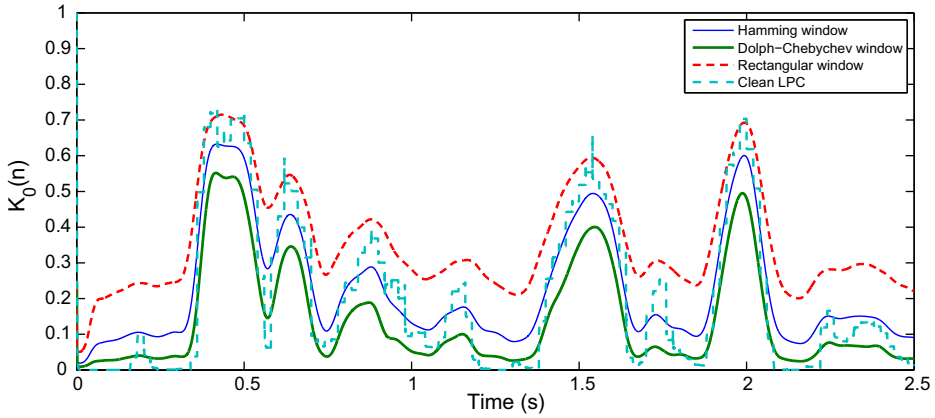

Fig. 22. Comparing time trajectories of the first component of Kalman gain $K_0(n)$ of the iterative Kalman filter (two iteration) with 80 ms overlapping frames for different windows.

seen in Fig. 22 to produce lower $K_0(n)$ values in the non-speech regions than the rectangular window, hence there is less residual noise. However, we should point out that in the voiced speech regions, because $K_0(n)$ is lower, then there is less long-term correlation information in the enhanced speech and this may result in speech distortion. Therefore, one may argue that the Hamming window may be the best compromise between residual noise reduction and speech distortion.

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 (Hu and Loizou, 2006b). 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 five SNR levels ($-5, 0, 5, 10$ and 15 dB) and coloured car noise at four SNR levels (0, 5, 10 and 15 dB). For the coloured noise case, we have based the Kalman filter implementation on the one by Gibson et al. (1991), where a noise LPC order q of 4 was used.

In this study, we conducted both objective evaluations as well as spectrogram analysis for both noise types. Two objective evaluations were carried out on the NOIZEUS corpus using:

1. three objective speech quality measures, namely SNR, segmental SNR, and PESQ (perceptual evaluation of speech quality) (Rix et al., 2001);
2. composite measures of signal distortion (SIG), background intrusiveness (BAK), and mean opinion score (OVRL) derived by Hu and Loizou (2006a) that simulate the ITU-T P.835 methodology.

Seven treatment types were used, with five corresponding to speech enhancement methods and two corresponding to clean and noisy speech. In addition, blind AB listening tests were undertaken to determine subjective method preference (Sorqvist et al., 1997). Two NOIZEUS sentences that were corrupted with white Gaussian noise case at 5 dB SNR, belonging to different speakers, were included in these tests. Stimuli pairs were played back to the listeners. The listeners were asked to make a subjective preference for each stimuli pair. Fifteen English speaking listeners participated in the listening tests. 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 with four iterations (Gibson et al., 1991), 20 ms, $p = 10$, no overlap, rectangular window (**Kalman iterative**);
6. Proposed 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

Table 2
Average composite measure scores on the NOIZEUS corpus of the proposed method for a varying number of iterations (corrupted with white Gaussian noise at an SNR of 5 dB).

Treatment type	Composite measure		
	SIG	BAK	OVRL
Noisy	2.03	1.91	1.82
Kalman proposed (iteration 1)	2.72	2.38	2.40
Kalman proposed (iteration 2)	2.82	2.58	2.55
Kalman proposed (iteration 3)	2.49	2.56	2.30
Kalman proposed (iteration 4)	2.18	2.55	2.11

Table 3
Average objective scores on the NOIZEUS corpus of the proposed method as well as the non-iterative and iterative Kalman filter after two, three, and four iterations (corrupted with white Gaussian noise at an SNR of 5 dB).

Treatment type	Objective measure		
	SNR (dB)	SegSNR (dB)	PESQ
Noisy	5.00	−3.31	1.82
Kalman noisy	9.03	0.70	2.06
Kalman iterative (iteration 2)	11.19	3.48	2.29
Kalman iterative (iteration 3)	11.82	5.19	2.40
Kalman iterative (iteration 4)	11.84	5.84	2.36
Kalman proposed (iteration 2)	11.32	5.64	2.50

7. MMSE–STSA method (Ephraim and Malah, 1984) (**MMSE**).

4.2. Results and discussion

4.2.1. Objective comparison using composite measures of the proposed method for a varying number of iterations (white Gaussian noise case)

Table 2 shows the average composite measures of the proposed Kalman filter for a varying number of iterations, where the speech has been corrupted by white Gaussian noise at an SNR of 5 dB. We can see that the speech quality (SIG) is highest for two iterations but degrades significantly when more iterations are used. This is consistent with the fact that the iterative Kalman filter of Gibson et al. (1991) is an approximate EM algorithm only (Gannot et al., 1998), where the likelihood function of the LPC estimates is not guaranteed to increase with each iteration. Informal listening tests reveal the speech became more

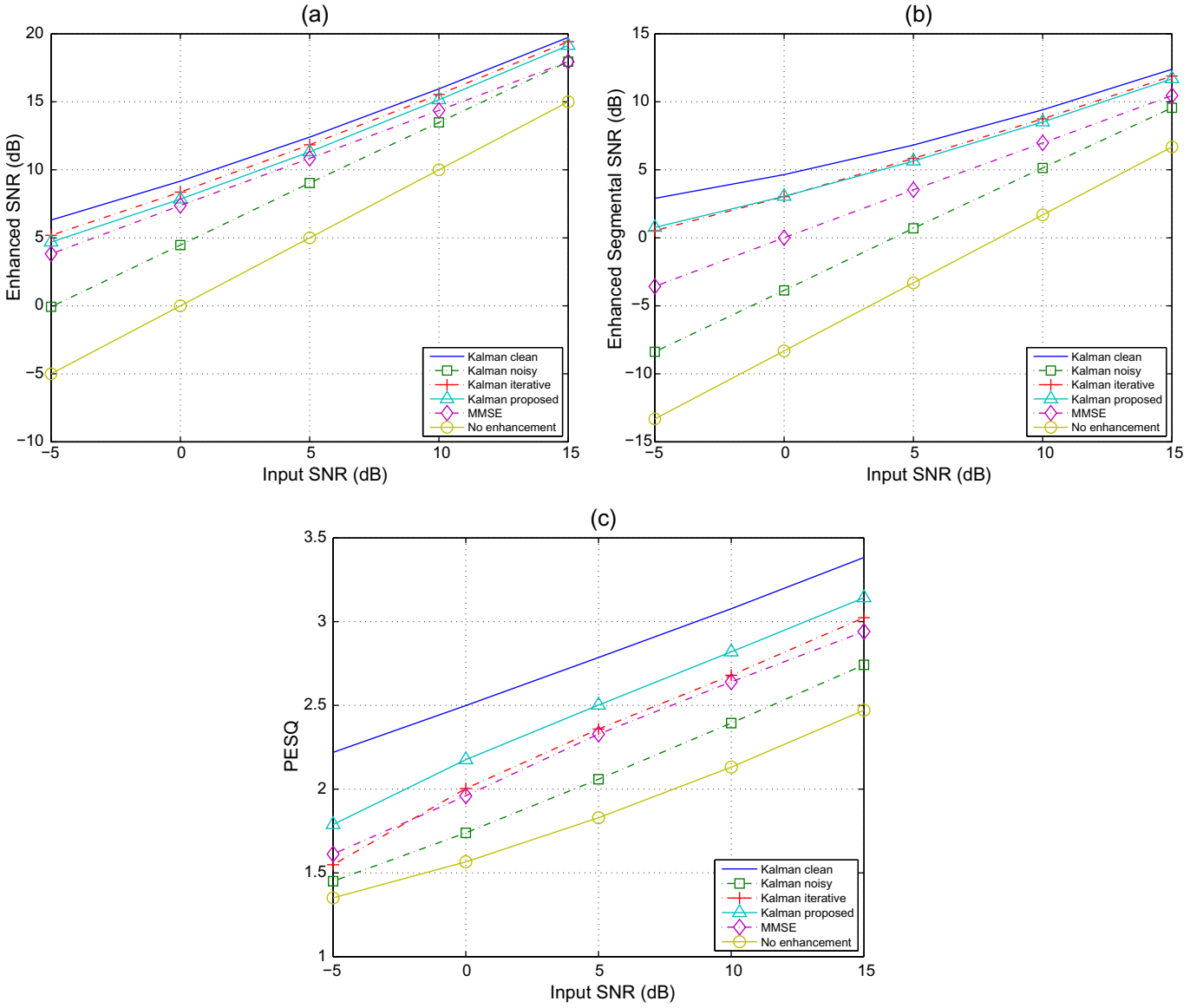

Fig. 23. Objective results for all treatment types on the NOIZEUS corpus corrupted with white Gaussian noise: (a) SNR results; (b) segmental SNR results; (c) PESQ results.

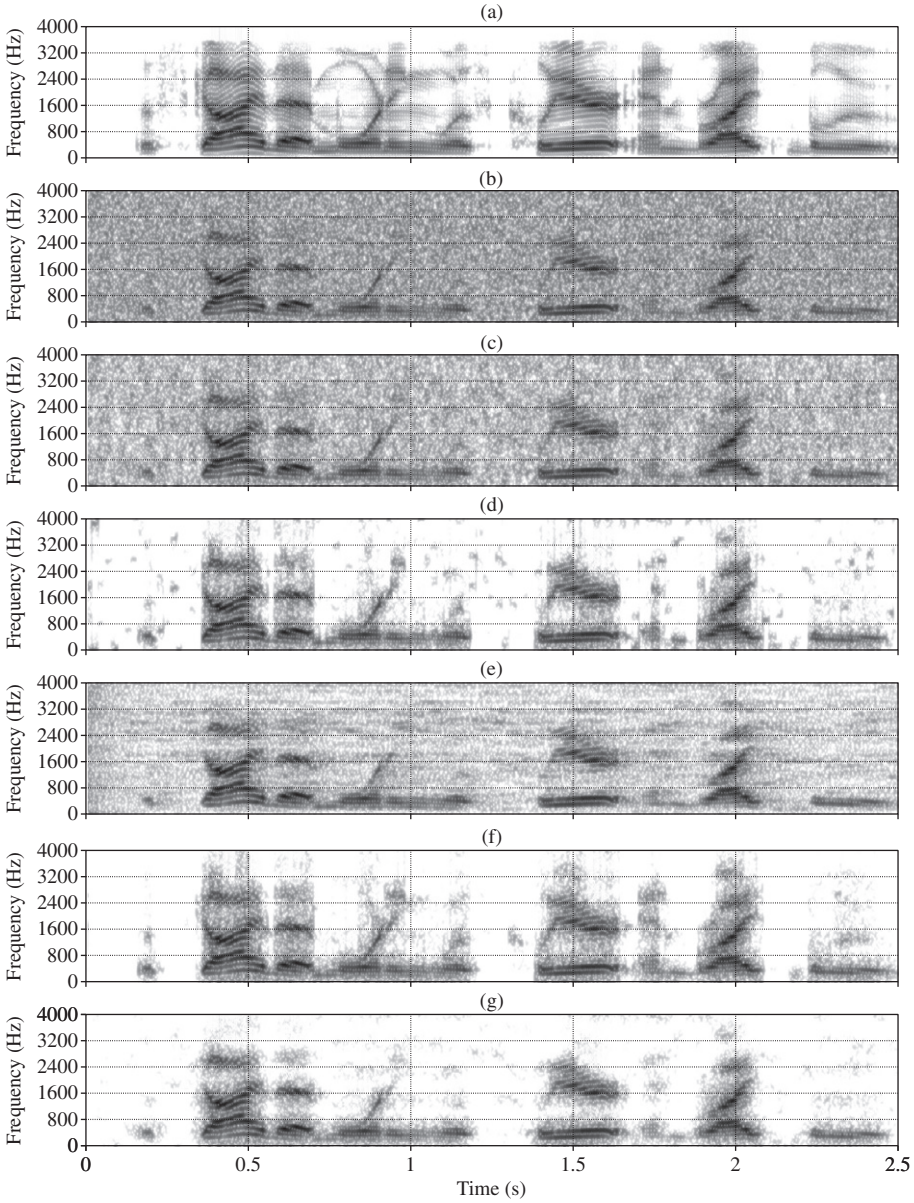

Fig. 24. Spectrograms of all treatment types on the NOIZEUS corpus corrupted with white Gaussian noise at an SNR of 5 dB: (a) clean speech (sp10.wav) 'The sky that morning was clear and bright blue'; (b) noise-corrupted speech; (c) Kalman noisy; (d) Kalman iterative; (e) MMSE–STSA; (f) Kalman clean; (g) Kalman proposed.

and more suppressed in subsequent iterations. On the other hand, the residual noise intrusiveness (BAK) score rose sharply when going to the second iteration but then remained relatively constant in subsequent iterations.

4.2.2. Objective comparison using SNR, segmental SNR and PESQ with the iterative Kalman filter (white Gaussian noise case)

Table 3 shows the objective results of the proposed Kalman filter as well as the non-iterative and iterative Kalman filter after a varying number of iterations for 5 dB of white Gaussian noise. We can see that large improvements are obtained by using iterative estimation of the LPCs, as noted in Gibson et al. (1991). The proposed Kalman filter, which uses effectively two iterations only, has achieved similar SNR and segmental SNR scores as the four-iteration Kalman filter. However, in terms of PESQ, the proposed Kalman filter outperforms the iterative Kalman filter while using less iterations and hence, less computations.

4.2.3. Objective comparison using SNR, segmental SNR and PESQ with other enhancement methods (white Gaussian noise case)

Fig. 23 shows the average objective scores for each of the different treatment types for the white Gaussian noise case. We can see that the proposed method is competitive with the iterative Kalman filter in terms of SNR and segmental SNR (Fig. 23(a) and (b)), despite the former requiring only half the number of iterations as the latter, while outperforming the other treatment types except for the

Table 4
Average composite measure scores on the NOIZEUS corpus corrupted with white Gaussian noise at an SNR of 5 dB.

Treatment type	Composite measure		
	SIG	BAK	OVRL
Noisy	2.03	1.91	1.82
Kalman clean	3.72	2.94	3.18
Kalman noisy	2.39	2.16	2.11
Kalman iterative	2.41	2.44	2.25
Kalman proposed	2.82	2.58	2.55
MMSE	2.83	2.45	2.49

Kalman filter that uses clean-speech-derived LPCs. In terms of the PESQ score (Fig. 23(c)), it can be seen that the proposed Kalman filter, which uses a Dolph–Chebychev window with large side lobe attenuation and long 80 ms frame lengths, outperforms all other enhancement methods except for the Kalman filter that uses clean-speech-derived LPCs.

Fig. 24 shows the spectrograms of all the treatment types. From these, we can see a large amount of residual noise in the Kalman noisy case. This observation is consistent with our earlier analysis of the Kalman gain, where a large value resulted in the passing through of noise from input to output. The residual noise in Kalman-iterative (Fig. 24(d)) is much less, owing to better estimates of the LPCs. However, the residual noise appears to be musical in nature and this was confirmed in informal listening tests. The speech was also quite distorted. In the output from the MMSE–STSA method (Fig. 24(e)), there is a high degree of structured residual noise that spans temporally. We can see that enhanced speech from the proposed Kalman filter (Fig. 24(g)) has low residual noise. However, when compared with the original clean speech, there appears to be

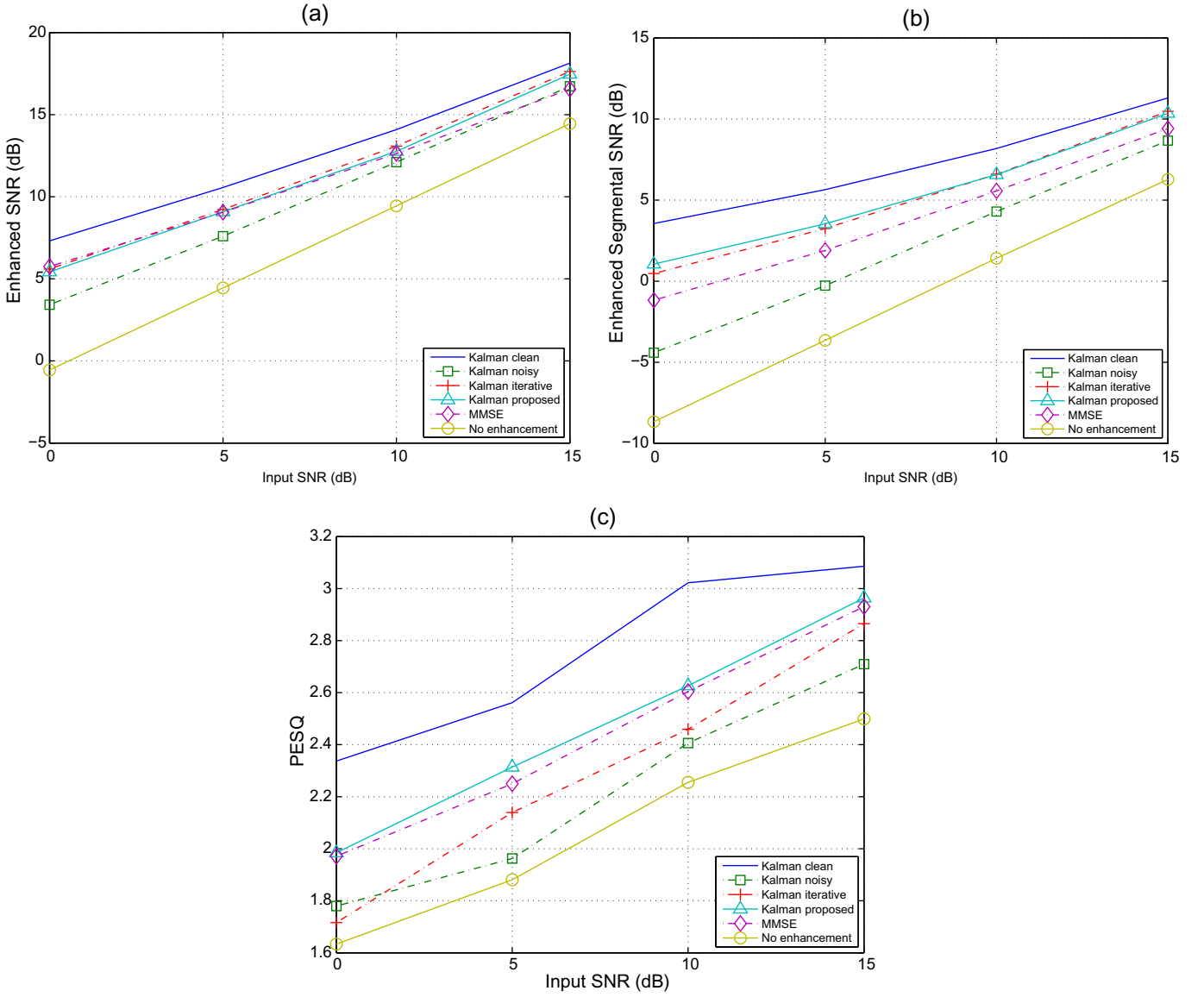

Fig. 25. Objective results for all treatment types on the NOIZEUS corpus corrupted with car noise (coloured): (a) SNR results; (b) segmental SNR results; (c) PESQ results.

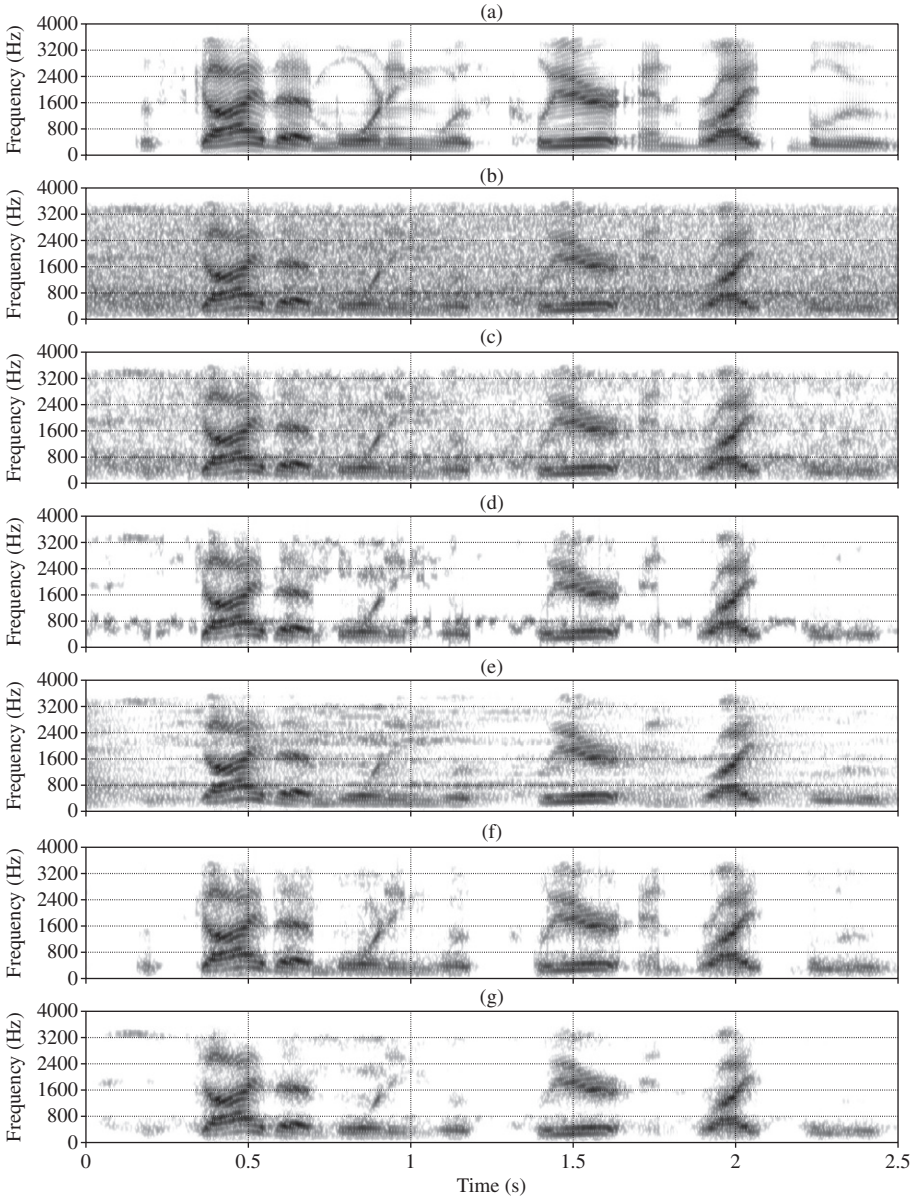

Fig. 26. Spectrograms of all treatment types on the NOIZEUS corpus corrupted with car noise (coloured) at an SNR of 5 dB: (a) clean speech (sp10.wav) 'The sky that morning was clear and bright blue'; (b) noise-corrupted speech; (c) Kalman noisy; (d) Kalman iterative; (e) MMSE–STSA; (f) Kalman clean; (g) Kalman proposed.

some loss of speech information. This was also confirmed in informal listening tests, where the residual noise was determined to be less annoying, but the speech was slightly distorted. This may be due to an 'over-suppression' of the Kalman gain in the speech regions. It was also noted that increasing the number of iterations reduced the residual noise further but also resulted in a higher level of speech distortion.

4.2.4. Objective comparison using composite measures with other enhancement methods (white Gaussian noise case)

Table 4 shows the average composite measures (SIG, BAK, OVRL) for the proposed Kalman filter compared with the other enhancement methods at 5 dB SNR of white Gaussian noise. We can see that under white noise, the proposed Kalman filter achieves the highest scores for all three composite measures in the Kalman 'family' of methods, apart from the ideal case (Kalman clean). In order words,

Table 5
Average composite measure scores on the NOIZEUS corpus corrupted with car noise at an SNR of 5 dB.

Treatment type	Composite measure		
	SIG	BAK	OVRL
Noisy	2.93	2.02	2.33
Kalman clean	3.69	2.82	3.08
Kalman noisy	2.39	2.16	2.11
Kalman iterative	2.60	2.23	2.23
Kalman proposed	2.89	2.34	2.48
MMSE	3.28	2.39	2.60

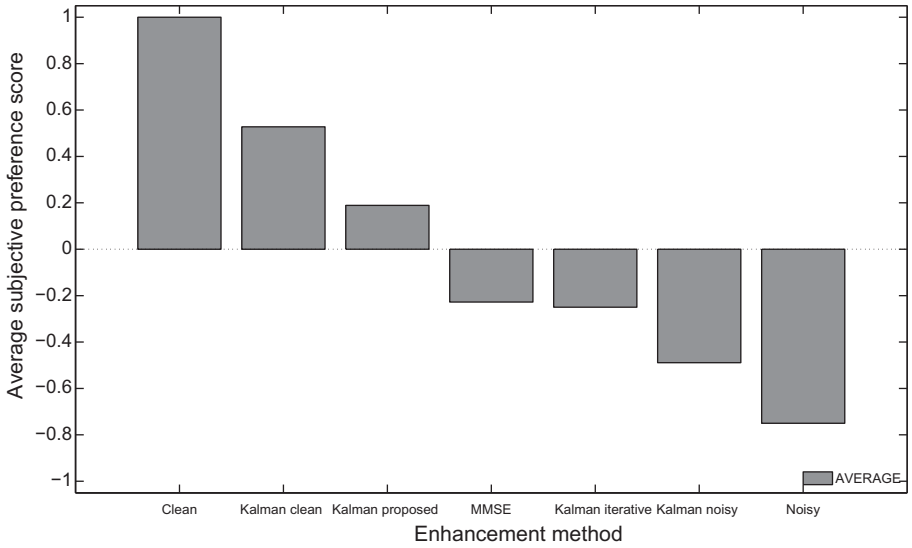

Fig. 27. Average subjective preference scores on the NOIZEUS corpus corrupted with white Gaussian noise at an SNR of 5 dB.

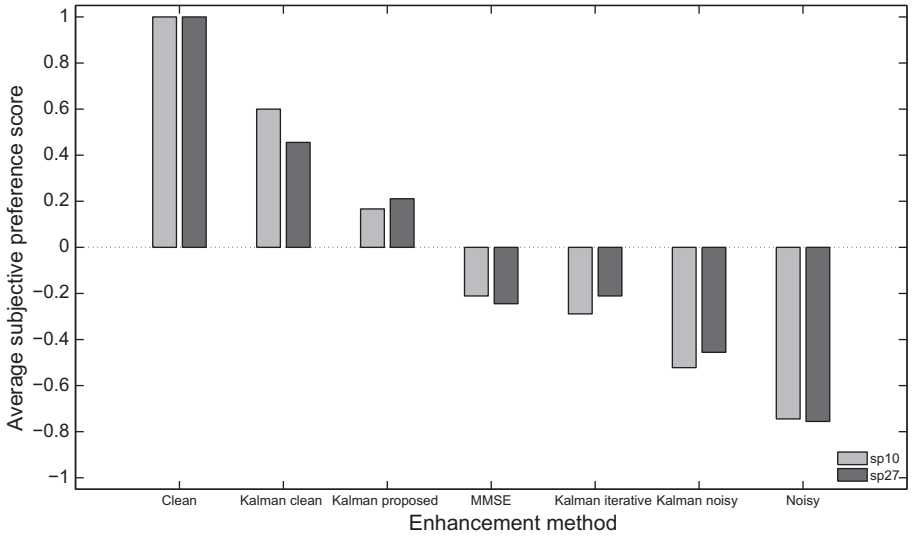

Fig. 28. Average subjective preference scores for different speakers on the NOIZEUS corpus corrupted with white Gaussian noise at an SNR of 5 dB. Sp10 was a male utterance, 'The sky that morning was clear and bright blue'. Sp27 was a female utterance, 'Bring your best compass to the third class'.

this means that the output speech from the proposed method has less speech distortion, less residual noise, and better overall quality than the iterative Kalman filter of Gibson et al. (1991). When compared with the MMSE–STSA method, the proposed method possesses similar speech quality but lower residual noise and better overall quality.

4.2.5. Objective comparison using SNR, segmental SNR and PESQ with other enhancement methods (coloured car noise case)

Fig. 25 shows the average objective results for each of the different treatment types for the coloured car noise case. It can be seen in Fig. 25(a) and (b) that the SNR scores of the proposed Kalman filter were similar to those of the MMSE–STSA method and the iterative Kalman filter. In terms of segmental SNR, the proposed method were similar to the iterative Kalman filter while outperforming MMSE–STSA. The segmental SNR results indicate that the proposed method has, on average, lower errors in the low power speech regions, which suggests less residual noise.[7] In Fig. 25(c), we can see that the proposed method achieves a slightly higher PESQ score to the MMSE–STSA method while outperforming the iterative Kalman filter.

Fig. 26 shows the spectrograms of all treatment types for car noise at an SNR of 5 dB. We can see in Fig. 26(d), that

[7] This is because the global SNR makes no distinction between errors in high and low energy regions, in spite of errors in low energy regions being perceptually more noticeable (Holmes and Holmes, 2001).

the iterative Kalman filter has reduced a lot of the noise in the background but suffers from random residual peaks, which result in popping noises, as noticed in informal listening tests. The MMSE–STSA method (Fig. 26(e)) can be seen to introduce a spread of residual background noise. In contrast, the proposed Kalman filter in Fig. 26(g) can be seen to produce little-to-no observable residual noise.

Therefore, the Dolph–Chebychev analysis window with large side lobe attenuation and long 80 ms frame lengths have improved the objective enhancement performance of the Kalman filter for coloured car noise.

4.2.6. Objective comparison using composite measures with other enhancement methods (coloured car noise case)

Table 5 shows the average composite measures (SIG, BAK, OVRL) for the proposed Kalman filter compared with the other enhancement methods at 5 dB SNR of car noise. We can see that the proposed Kalman filter achieves the highest scores for all three composite measures in the Kalman 'family' of methods, apart from the ideal case (Kalman clean). However, when compared with the MMSE–STSA method, the proposed method produces speech with more speech distortion but similar residual noise suppression. The cause of speech distortion can be pinpointed to over-suppression of the Kalman gain trajectory in the speech regions, which increases the contribution of the predicted component.

4.2.7. Subjective comparison using blind AB listening tests

Fig. 27 shows the average preference scores from the subjective listening tests. We can see that the proposed Kalman filter was preferred on average over all other methods except for the Kalman filter with clean-speech-derived LPCs and clean speech. The significance of this result can be appreciated when we consider that the proposed method involved the use of longer frames and a special tapered analysis window only. Fig. 28 shows the average preference scores for each of the two utterances. The listeners also preferred on average the proposed Kalman filter for both utterances over the other methods except for the Kalman clean method and clean speech.

5. Conclusion

In this paper, we have analysed the effect of poor linear predictive coefficient estimates on the performance of the Kalman filter for speech enhancement. Higher-than-usual prediction errors due to the presence of noise result in large Kalman gain values, even during regions where speech energy is low. Since the Kalman gain regulates the contribution of the noisy observation to the output, large Kalman gain values result in more residual noise being passed through to the output, which can be annoying to the listener, as errors in low speech energy regions are more perceptible. By using tapered windows (such as the Dolph–Chebychev window) during LPC estimation, the influence of additive noise on the Kalman gain was suppressed, which resulted in lower residual noise. We have also shown that overlapped frames are beneficial to the Kalman filter when long frames and tapered analysis windowing are used. The proposed Kalman filter, which uses two iterations for LPC estimation, Dolph–Chebychev windowing in the first iteration, and long and overlapped 80 ms frames, was found to have improved on conventional Kalman filtering schemes and in objective and subjective listening tests.

Acknowledgments

The authors would like to acknowledge and thank the anonymous reviewers for their expert opinions during the review process. Their valuable insights as well as careful guidance have greatly enhanced the clarity and breadth of this paper.

Appendix A. The relative importance of accurate LPC and excitation variance estimates in Kalman filtering

In this appendix, we investigate the relative importance of accurate LPCs and excitation variance estimates in the Kalman filter by performing two oracle Kalman filtering experiments on a speech file (sp10.wav) that has been corrupted with white Gaussian noise at an SNR of 0 dB:

- **Experiment 1**: with clean LPCs + noisy excitation variances.
- **Experiment 2**: with noisy LPCs + clean excitation variances.

The aim is to gain some insight into whether the enhancement performance of the Kalman filter can be improved by correcting either the excitation variance or LPC estimates, as we have shown earlier in this paper that tapered windows reduce the bias of the excitation variance of the autocorrelation method. It should be stressed though that the excitation variance and LPCs are dependent to each other, so the application of tapered windows will influence both parameters.

The spectrograms as well as the PESQ scores and SNRs are shown in Fig. A.29. We can see in the spectrograms that when noisy excitation variances are used with clean LPC estimates in Fig. A.29(d) (Experiment 1), the enhanced speech contains some coloured residual noise, which is consistent with our analysis of the dependence of the Kalman gain on the excitation variance of the model. On the other hand, in Fig. A.29(e) (Experiment 2), where noisy LPCs are paired with clean excitation variance estimates, we can see that the residual noise is much lower but the speech appears to be distorted (and is more noticeable in the low frequency range). Also, the remaining residual noise appears to exist only in the speech regions.

While observing the objective results, we can see that the enhanced speech from both Experiments 1 and 2 have a higher PESQ score than the noisy Kalman filter (where

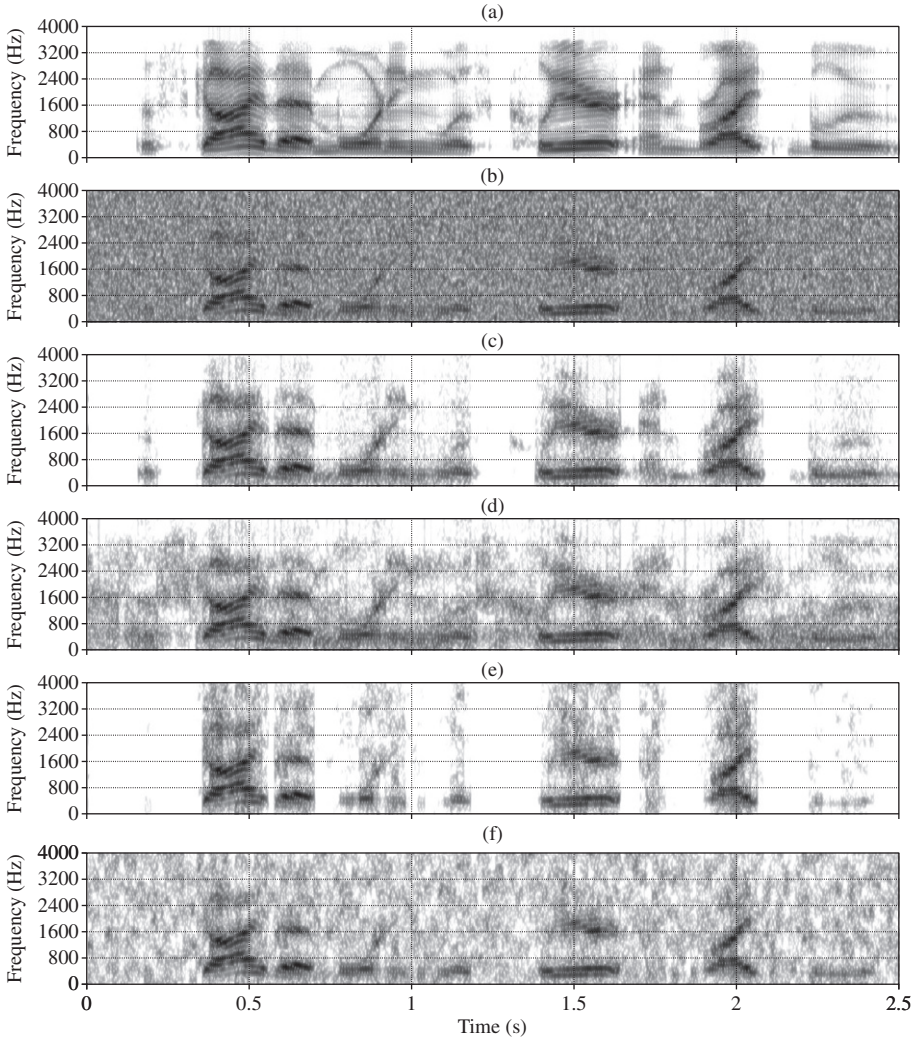

Fig. A.29. Spectrograms from oracle experiments comparing the relative importance of accurate LPC and excitation variance estimates: (a) clean speech (sp10.wav) 'The sky that morning was clear and bright blue'; (b) speech corrupted with white Gaussian noise at an SNR of 0 dB (PESQ = 1.59); (c) Kalman clean (PESQ = 2.86, SNR = 8.86 dB); (d) Experiment 1 (PESQ = 2.43, SNR = 6.80 dB); (e) Experiment 2 (PESQ = 2.05, SNR = 6.20 dB); (f) Kalman noisy (PESQ = 1.79, SNR = 6.52 dB).

both LPCs and excitation variances are estimated from noisy speech), though Experiment 2 produced a lower SNR, which is likely due to the higher speech distortion. However, the PESQ improvement appears greater for Experiment 1 than Experiment 2. Therefore, these experiments indicate that correcting the LPC estimates leads to more significant improvements in Kalman filtering performance than correcting the excitation variance.

Appendix B. Comparison with a Kalman filter using spectral-enhanced LPC estimates

In this appendix, we present some results of the Kalman filter, where the LPCs have been computed from noisy speech that has been enhanced by the MMSE–STSA algorithm (Ephraim and Malah, 1984), which we refer to as the *Kalman-MMSE method*. The purpose of this auxiliary study is to compare the proposed Kalman filter with the Kalman-MMSE method, as both procedures aim to compute better LPC estimates via some form of preprocessing of the speech.

Table B.6 shows the average objective results from the NOIZEUS database comparing the Kalman-MMSE method with the proposed Kalman filter. The results for the Kalman noisy method (i.e. Kalman filter using LPCs

Table B.6
Average objective scores on the NOIZEUS corpus of the proposed method and the Kalman-MMSE method (speech has been corrupted with white Gaussian noise at an SNR of 5 dB).

Treatment type	Objective measure		
	SNR (dB)	SegSNR (dB)	PESQ
Noisy	5.00	−3.31	1.82
Kalman noisy	9.03	0.70	2.06
Kalman-MMSE	11.58	4.90	2.43
Kalman proposed	11.32	5.64	2.50

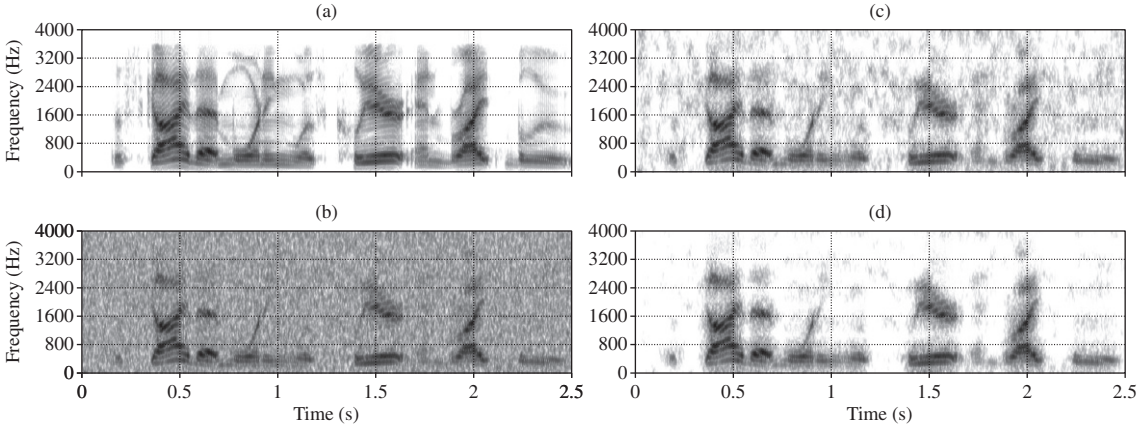

Fig. B.30. Spectrograms comparing the Kalman-MMSE method with the proposed Kalman filter: (a) clean speech (sp10.wav) 'The sky that morning was clear and bright blue'; (b) speech corrupted with white Gaussian noise at an SNR of 5 dB (PESQ = 1.59); (c) Kalman-MMSE (PESQ = 2.41); (d) Kalman proposed (PESQ = 2.52).

from unprocessed noisy speech) are also provided. We can see that both the Kalman-MMSE and the proposed Kalman filter have achieved improvements over Kalman noisy, which may be attributed to more accurate LPC estimates. It is interesting to note that the proposed Kalman filter (which applied the Dolph–Chebychev window during LPC analysis and used long and overlapping frames) has produced enhanced speech with a higher segmental SNR and PESQ score than the Kalman-MMSE (which uses the more sophisticated MMSE–STSA algorithm). These objective results are consistent with the spectrograms in Fig. B.30, where the enhanced speech from the proposed Kalman filter can be seen to contain less residual noise than the enhanced speech from the Kalman-MMSE method. Informal listening of the speech files confirmed that the proposed Kalman filter produced less residual noise than the Kalman-MMSE method, though the latter method produced less speech distortion.

References

Åström, K.J., Wittenmark, B., 1997. Computer-Controlled Systems: Theory and Design, third ed.. In: Prentice Hall Information and System Sciences Series Prentice-Hall, New Jersey.

Boll, S., 1979. Suppression of acoustic noise in speech using spectral subtraction. IEEE Trans. Acoust. Speech, Signal Process. ASSP-27 (2), 113–120.

Crochiere, R.E., 1980. A weighted overlap-add method of short-time Fourier analysis/synthesis. IEEE Trans. Acoust. Speech Signal Process. ASSP-28 (1), 99–102.

Ephraim, Y., Malah, D., 1984. Speech enhancement using a minimum-mean square error short-time spectral amplitude estimator. IEEE Trans. Acoust. Speech, Signal Process. 32, 1109–1121.

Ephraim, Y., Van Trees, H.L., 1995. A signal subspace approach for speech enhancement. IEEE Trans. Speech Audio Process. 3, 251–266.

Erkelens, J.S., Broersen, P.M.T., 1997. Bias propagation in the autocorrelation method of linear prediction. IEEE Trans. Speech Audio Process. 5 (2), 116–119.

Gabrea, M., Grivel, E., Najim, M., 1999. A single microphone Kalman filter-based noise canceller. IEEE Signal Process. Lett. 6 (3), 55–57.

Gannot, S., Burshtein, D., Weinstein, E., 1998. Iterative and sequential Kalman filter-based speech enhancement algorithms. IEEE Trans. Speech Audio Process. 6 (4), 373–385.

Gibson, J.D., Koo, B., Gray, S.D., 1991. Filtering of colored noise for speech enhancement and coding. IEEE Trans. Signal Process. 39 (8), 1732–1742.

Hayes, M.H., 1996. Statistical Digital Signal Processing and Modeling. John Wiley, New Jersey.

Haykin, S., 2002. Adaptive Filter Theory, fourth ed.. In: Prentice Hall Information and System Sciences Series Prentice-Hall, New Jersey.

Holmes, J.N., Holmes, W.J., 2001. Speech Synthesis and Recognition, second ed. CRC Press.

Hu, Y., Loizou, P., 2006a. Evaluation of objective measures for speech enhancement. In: Proc. INTERSPEECH 2006, pp. 1447–1450.

Hu, Y., Loizou, P., 2006b. Subjective comparison of speech enhancement algorithms. In: Proc. IEEE Int. Conf. Acoust. Speech, Signal Processing, Vol. 1. pp. 153–156.

Kalman, R.E., 1960. A new approach to linear filtering and prediction problems. J. Basic Eng. Trans. ASME 82, 35–45.

Kay, S.M., 1993. Fundamentals of Statistical Signal Processing. In: Prentice Hall Signal Processing Series, Vol. 1. Prentice-Hall, New Jersey.

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

Lim, J.S., Oppenheim, A.V., 1978. All-pole modeling of degraded speech. IEEE Trans. Acoust. Speech Signal Process. ASSP-26, 197–210.

Ma, N., Bouchard, M., Goubran, R.A., 2006. Speech enhancement using a masking threshold constrained Kalman filter and its heuristic implementations. IEEE Trans. Speech Audio Process. 14 (1), 19–32.

Makhoul, J., 1975. Linear prediction: a tutorial review. Proc. IEEE 63 (4), 561–580.

Mehra, M.K., 1970. On the identification of variances and adaptive Kalman filtering. IEEE Trans. Autom. Control AC-15 (2), 175–184.

Ohya, T., Suda, H., Miki, T., 1994. 5.6 kbits/s PSI-CELP of the half-rate PDC speech coding standard. In: IEEE 44th Vehicular Technology Conf., pp. 1680–1684.

Paliwal, K.K., Basu, A., 1987. A speech enhancement method based on Kalman filtering. In: Proc. IEEE Int. Conf. Acoust. Speech, Signal Processing, Vol. 12. pp. 177–180.

Rix, A., Beerends, J., Hollier, M., Hekstra, A., 2001. Perceptual evaluation of speech quality (PESQ), an objective method for end-to-end speech quality assessment of narrowband telephone networks and speech codecs. ITU-T Recommendation P.862. Tech. rep., ITU-T.

So, S., Paliwal, K.K., Sep. 2008. A long state vector Kalman filter for speech enhancement. In: Proc. Int. Conf. Spoken Language Processing. pp. 391–394.

Sorqvist, P., Handel, P., Ottersten, B., 1997. Kalman filtering for low distortion speech enhancement in mobile communication. In: Proc. IEEE Int. Conf. Acoust. Speech, Signal Processing, Vol. 2. pp. 1219–1222.

Wang, T., Koishida, K., Cuperman, V., Gersho, A., Collura, J.S., 2002. A 1200/2400 bps coding suite based on MELP. In: IEEE Workshop on Speech Coding.

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