Available online at www.sciencedirect.com
ScienceDirect
Speech Communication 53 (2011) 818–829
SPEECH COMMUNICATION
www.elsevier.com/locate/specom

Modulation-domain Kalman filtering for single-channel speech enhancement

Stephen So *, Kuldip K. Paliwal

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

Received 30 August 2010; received in revised form 16 December 2010; accepted 1 February 2011
Available online 16 February 2011

Abstract

In this paper, we investigate the modulation-domain Kalman filter (MDKF) and compare its performance with other time-domain and acoustic-domain speech enhancement methods. In contrast to previously reported modulation domain-enhancement methods based on fixed bandpass filtering, the MDKF is an adaptive and linear MMSE estimator that uses models of the temporal changes of the magnitude spectrum for both speech and noise. Also, because the Kalman filter is a joint magnitude and phase spectrum estimator, under non-stationarity assumptions, it is highly suited for modulation-domain processing, as phase information has been shown to play an important role in the modulation domain. We have found that the Kalman filter is better suited for processing in the modulation-domain, rather than in the time-domain, since the low order linear predictor is sufficient at modelling the dynamics of slow changes in the modulation domain, while being insufficient at modelling the long-term correlation speech information in the time domain. As a result, the MDKF method produces enhanced speech that has very minimal distortion and residual noise, in the ideal case. The results from objective experiments and blind subjective listening tests using the NOIZEUS corpus show that the MDKF (with clean speech parameters) outperforms all the acoustic and time-domain enhancement methods that were evaluated, including the time-domain Kalman filter with clean speech parameters. A practical MDKF that uses the MMSE-STSA method to enhance noisy speech in the acoustic domain prior to LPC analysis was also evaluated and showed promising results.

Keywords: Modulation domain; Kalman filtering; Speech enhancement

1. Introduction

In the problem of speech enhancement, where a speech signal is corrupted by noise, 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. The Kalman filter (Kalman, 1960) is an unbiased, time-domain, linear minimum mean squared error (MMSE) estimator, where the enhanced speech is recursively estimated on a sample-by-sample basis. The Kalman filter can be viewed as a joint estimator for both the magnitude and phase spectrum of speech, under non-stationarity assumptions (Li, 2006). This is in contrast to the short-time Fourier transform (STFT)-based enhancement methods, such as spectral subtraction (Boll, 1979), Wiener filtering (Wiener, 1949; Chen et al., 2006), and MMSE estimation (Ephraim and Malah, 1984, 1985), where only the clean magnitude spectrum is estimated. No processing is performed on the *noisy* phase spectrum before it is combined with the estimated clean magnitude spectrum to produce the enhanced speech frame.

The Kalman filter was first introduced for speech enhancement by Paliwal and Basu (1987), where significant noise reduction was reported when linear prediction coefficients (LPCs) estimated from clean speech were provided. In practice though, poor parameter estimates from noisy speech result in degraded enhancement performance.

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

0167-6393/$ - see front matter © 2011 Elsevier B.V. All rights reserved.
doi:10.1016/j.specom.2011.02.001

Iterative Kalman filters (Gibson et al., 1991) have been shown to alleviate the effects of poor parameter estimates in the Kalman filter, resulting in an improvement in SNR and reduction in background noise level. However, the enhanced quality was not guaranteed to improve after further iterations since the iterative LPC estimation was essentially an approximated Expectation-Maximisation (EM) algorithm, where the likelihood function of the LPC estimates was not guaranteed to monotonically increase (Gannot et al., 1998). The subband Kalman filter was proposed by Wu and Chen (1998), whereby the speech signal was first decomposed into subbands and then each temporal subband signal was enhanced using a low-order Kalman filter. As well as possessing lower computational complexity, the subband Kalman filter was found to perform better than the full-band Kalman filter.

There has been recent interest in using the modulation domain as an alternative to the acoustic domain for speech enhancement, where we define the *acoustic spectrum* as the STFT of a signal and the *modulation domain* as the temporal trajectories of the magnitude spectrum at all acoustic frequencies (Atlas et al., 2003). There is growing psychoacoustic and physiological evidence to support the significance of the modulation domain for speech analysis and processing. For example, neurones in the auditory cortex are thought to decompose the acoustic spectrum into spectro-temporal modulation content (Mesgarani and Shamma, 2005). Low frequency modulations of sound have been shown to be the fundamental carriers of information in speech (Atlas et al., 2003). Drullman et al. (1994a,b) investigated the importance of modulation frequencies for intelligibility by applying low-pass and high-pass filters to the temporal envelopes of acoustic frequency subbands. They showed frequencies between 4 and 16 Hz to be important for intelligibility, with the region around 4–5 Hz being the most significant. In a similar study, (Arai et al., 1999) showed that applying passband filters between 1 and 16 Hz did not impair speech intelligibility. While the envelope of the acoustic magnitude spectrum represents the shape of the vocal tract, the modulation domain represents how the vocal tract changes as a function of time. It is these temporal changes that convey most of the linguistic information (or intelligibility) of speech. For a detailed review of studies on the importance of the modulation domain, the reader can refer to (Paliwal et al., 2010).

Hermansky et al. (1995) proposed to bandpass filter the time trajectories of cubic-root compressed short-time power spectrum for enhancement of speech corrupted by additive noise. Similar bandpass filtering was applied to the time trajectories of the short-time power spectrum for speech enhancement in Falk et al. (2007) and Lyons and Paliwal (2008). These bandpass filtering methods have several limitations: (1) the filters are fixed in nature and therefore assume the speech and noise signals are stationary in time; (2) the properties of the noise are not exploited in the design of the filters; and (3) noise contained in the filter passband (the speech modulation regions) is preserved. These limitations were addressed recently in Paliwal et al. (2010), whereby the spectral subtraction algorithm was used to process the modulation spectrum on a frame-by-frame basis. This meant that the speech and noise signals were assumed to be quasi-stationary in short-time frames, which is in contrast to the earlier bandpass filtering methods that assumed stationarity for all time.

In this paper, we investigate the use of Kalman filtering for estimating the modulating signals of speech, which are the temporal trajectories of the magnitude spectrum along each acoustic frequency. We believe the ability of the Kalman filter to process non-stationary signals as well as estimate both the magnitude and phase spectrum makes it preferable over STFT-based enhancement methods, because phase information has been shown to play an important role in the modulation domain (Kanedera et al., 1998; Greenberg et al., 1998; Greenberg and Arai, 2001). Furthermore, we make the observation that the Kalman filter with low order linear predictor is more suitable for enhancing slow changing modulating signals than for enhancing the speech signal in the time domain, as the latter contains long-term correlation information that the low order linear predictor cannot capture. Using objective and blind subjective tests on the NOIZEUS speech corpus (Loizou, 2007), we show that in the ideal case where accurate model parameters are available, the modulation domain Kalman filter (MDKF) outperforms all acoustic and time-domain speech enhancement methods that were evaluated (including the time-domain Kalman filter (TDKF)) for both white and coloured noise. We also present some results of a practical MDKF that uses the MMSE-STSA algorithm in the acoustic domain as a pre-processor for LPC estimation.

The rest of this paper is structured as follows. In Section 2.1, we describe the analysis-modification-synthesis (AMS) framework that is used to obtain the modulation domain. Following this, the modulation-domain Kalman filter and its operation are detailed in Section 2.2, where we also discuss the validity of some Kalman filtering assumptions in the modulation domain. In Section 2.3, we present a comparative analysis of the MDKF and the TDKF in the ideal case, where LPCs from clean speech are available. This analysis will highlight the advantages of performing Kalman filtering in the modulation domain, rather than in the time domain. The objective and blind subjective listening experiments that were performed in this study are described in Section 3.1 and the results and discussion follow in Section 3.2. Finally, we conclude in Section 4.

2. Modulation domain Kalman filtering for speech enhancement

2.1. Acoustic analysis-modification-synthesis framework

The analysis-modification-synthesis (AMS) framework consists of three stages: (1) the analysis stage, where the input speech is processed using STFT analysis; (2) the

modification stage, where the noisy spectrum undergoes some kind of modification; and (3) the synthesis stage, where the inverse STFT is followed by the overlap-add synthesis to reconstruct the output signal.

Let us consider an additive noise model:

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

where $y(n)$, $x(n)$ and $v(n)$ denote zero-mean signals of noisy speech, clean speech and noise, respectively. Since speech can be assumed to be quasi-stationary, it is analysed framewise using short-time Fourier analysis. The STFT of the corrupted speech signal $y(n)$ is given by:

$$Y(n,k) = \sum_{l=-\infty}^{\infty} y(l)w(n-l)e^{-j\frac{2\pi kl}{N}} \tag{2}$$

where k refers to the index of the discrete acoustic frequency, N is the acoustic frame duration (in samples) and $w(n)$ is an acoustic analysis window function. In speech processing, the Hamming window with 20–40 ms duration is typically employed. Using STFT analysis, we can represent Eq. (2) as:

$$Y(n,k) = X(n,k) + V(n,k) \tag{3}$$

where $Y(n,k)$, $X(n,k)$ and $V(n,k)$ are the STFTs of noisy speech, clean speech, and noise, respectively. Each of these can be expressed in terms of acoustic magnitude and acoustic phase spectrum. For instance, the STFT of the noisy speech signal can be written in polar form as:

$$Y(n,k) = |Y(n,k)|e^{j\angle Y(n,k)} \tag{4}$$

where $|Y(n,k)|$ denotes the acoustic magnitude spectrum and $\angle Y(n,k)$ denotes the acoustic phase spectrum.

Traditional AMS-based speech enhancement methods modify, or enhance, only the noisy acoustic magnitude spectrum while keeping the noisy acoustic phase spectrum unchanged. Let us denote the enhanced magnitude spectrum as $|\widehat{X}(n,k)|$, then the modified acoustic spectrum is constructed by combining $|\widehat{X}(n,k)|$ with the noisy phase spectrum, as follows:

$$\widehat{X}(n,k) = |\widehat{X}(n,k)|e^{j\angle Y(n,k)} \tag{5}$$

The enhanced speech $\hat{x}(n)$ is reconstructed by taking the inverse STFT of the modified acoustic spectrum followed by synthesis windowing and overlap-add reconstruction (Quatieri, 2002).

2.2. Kalman filtering in the modulation domain

The modulation domain views the acoustic magnitude spectrum as a series of N *modulating signals* that span across time. Each modulating signal represents the temporal evolution of each acoustic magnitude spectral component, as shown in Fig. 1. In the proposed modulation-domain Kalman filter (MDKF), each modulating signal, $|Y(n,k)|$ (where $k = 1,2,\ldots,N$) is processed using a Kalman filter (see Fig. 2).

In the modulation-domain Kalman filter, we assume an additive noise model for each modulating signal:

$$|Y(n,k)| = |X(n,k)| + |V(n,k)| \tag{6}$$

where $|V(n,k)|$ is the kth modulating signal of white Gaussian noise. A pth order linear predictor can be used to model the temporal evolution of the kth modulating signal of speech:

$$|X(n,k)| = -\sum_{j=1}^{p} a_{j,k}|X(n-j,k)| + W(n,k) \tag{7}$$

where $\{a_{j,k};\ j = 1,2,\ldots,p\}$ are the linear prediction coefficients (LPCs) and $W(n,k)$ is a white random excitation with a variance of $\sigma^2_{W(k)}$. Together with the corrupting noise, we can write the following state space representation for $|Y(n,k)|$:

$$\boldsymbol{X}(n,k) = \boldsymbol{A}(k)\boldsymbol{X}(n-1,k) + \boldsymbol{d}W(n,k) \tag{8}$$

$$|Y(n,k)| = \boldsymbol{c}^T\boldsymbol{X}(n,k) + |V(n,k)| \tag{9}$$

where $\boldsymbol{X}(n,k) = [|X(n,k)|,\ |X(n-1,k)|,\ldots,|X(n-p+1,k)|]^T$ is the clean modulation state vector, $\boldsymbol{d} = [1,0,\ldots,0]^T$ and $\boldsymbol{c} = [1,0,\ldots,0]^T$ are the measurement vectors for the excitation noise $W(n,k)$ and observation, respectively, and $\boldsymbol{A}(k)$ is the state transition matrix:

$$\boldsymbol{A}(k) = \begin{bmatrix} -a_{1,k} & -a_{2,k} & \ldots & -a_{p-1,k} & -a_{p,k} \\ 1 & 0 & \ldots & 0 & 0 \\ 0 & 1 & \ldots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \ldots & 1 & 0 \end{bmatrix} \tag{10}$$

The Kalman filter recursively computes an unbiased and linear MMSE estimate $\widehat{\boldsymbol{X}}(n|n,k)$ of the kth modulation state vector at time n, given the noisy modulating signal up to time n (i.e. $|Y(1,k)|$, $|Y(2,k)|$, …, $|Y(n,k)|$), by using the following equations:

$$\boldsymbol{P}(n|n-1,k) = \boldsymbol{A}(k)\boldsymbol{P}(n-1|n-1,k)\boldsymbol{A}(k)^T + \sigma^2_{W(k)}\boldsymbol{d}\boldsymbol{d}^T \tag{11}$$

$$\boldsymbol{K}(n,k) = \boldsymbol{P}(n|n-1,k)\boldsymbol{c}\left[\sigma^2_{V(k)} + \boldsymbol{c}^T\boldsymbol{P}(n|n-1,k)\boldsymbol{c}\right]^{-1} \tag{12}$$

$$\widehat{\boldsymbol{X}}(n|n-1,k) = \boldsymbol{A}(k)\widehat{\boldsymbol{X}}(n-1|n-1,k) \tag{13}$$

$$\boldsymbol{P}(n|n,k) = [\boldsymbol{I} - \boldsymbol{K}(n,k)\boldsymbol{c}^T]\boldsymbol{P}(n|n-1,k) \tag{14}$$

$$\widehat{\boldsymbol{X}}(n|n,k) = \widehat{\boldsymbol{X}}(n|n-1,k) + \boldsymbol{K}(n,k)[|Y(n,k)| - \boldsymbol{c}^T\hat{\boldsymbol{X}}(n|n-1,k)] \tag{15}$$

During the operation of the Kalman filter, the noisy modulating signal $|Y(n,k)|$ is windowed into short modulation frames and the LPCs and excitation variance $\sigma^2_{W(k)}$ are estimated. In this study, we investigated short modulation frame durations of 10–20 ms, which has been reported to maintain good intelligibility (Paliwal et al., 2011). These LPCs remain constant during the Kalman filtering of the modulating signal in the frame, while the Kalman

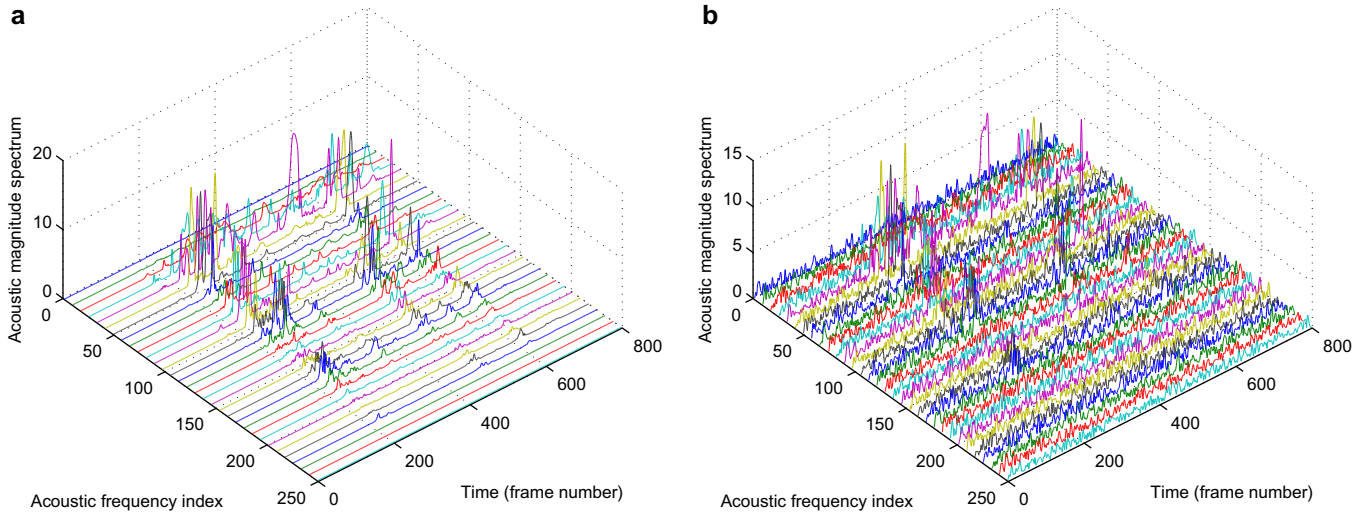

Fig. 1. The modulation domain representation of speech ('The sky that morning was clear and bright blue'), showing the temporal evolution of the modulating signals: (a) clean speech; (b) speech corrupted with white Gaussian noise at an SNR of 0 dB.

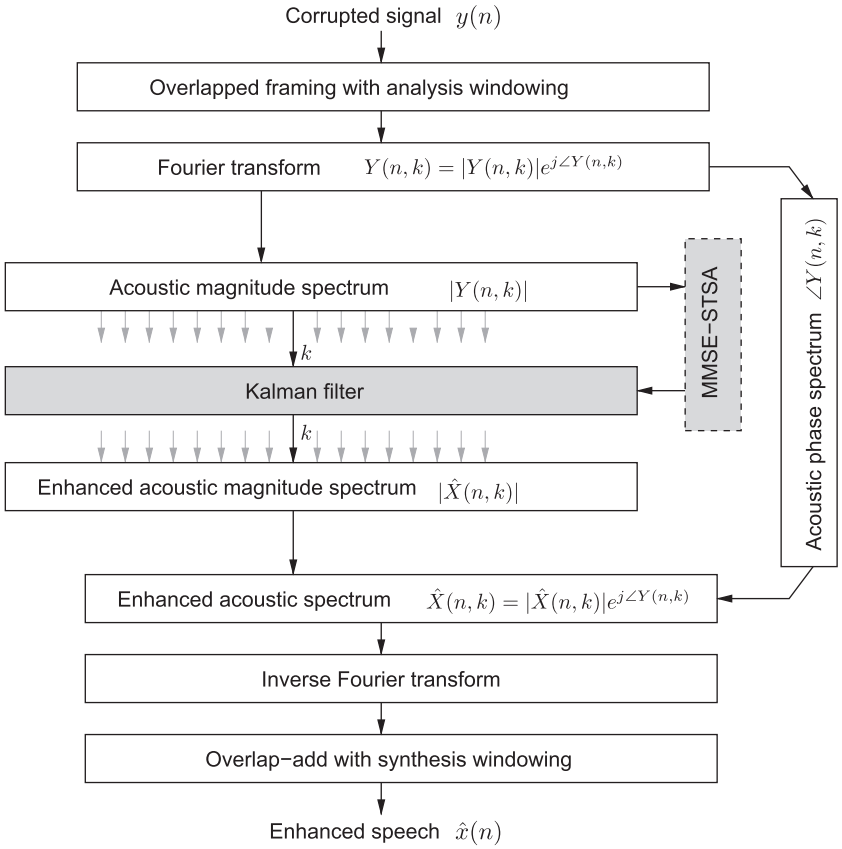

Fig. 2. Schematic diagram of the proposed AMS-based modulation-domain Kalman filtering framework (the MMSE-STSA block with dashed outline is an additional component for the MDKF-MMSE method).

parameters (such as Kalman gain $\boldsymbol{K}(n,k)$ and error covariance $\boldsymbol{P}(n|n,k)$) and state vector estimate $\widehat{\boldsymbol{X}}(n \mid n,k)$ are continually updated on a sample-by-sample basis (regardless of whichever frame we are in).

When applying the Kalman filter in the modulation domain, there are some time domain-based assumptions that may not necessarily be satisfied in the modulation domain:

- additive noise in the time domain may not be additive in the modulation domain (Eq.(6));
- white noise in the time domain may not be spectrally white in the modulation domain; and
- the linear predictor may not be the best dynamic model of modulating signals.

In regards to the additive noise assumption in the modulation domain, let us consider Eq. (3) in polar form:

$$|Y(n,k)|e^{j\angle Y(n,k)} = |X(n,k)|e^{j\angle X(n,k)} + |V(n,k)|e^{j\angle V(n,k)} \quad (16)$$

Using a geometric approach (Loizou, 2007), it is easy to see that the additive noise assumption of Eq. (6) is approximately satisfied if either $\angle X(n,k) \approx \angle V(n,k)$ or $|X(n,k)| \gg |V(n,k)|$. The first condition is more difficult to show since it is assumed that clean speech and noise signals are not correlated. However, the second condition is related to the instantaneous spectral SNR at acoustic frequency index k, i.e. $|X(n,k)|^2/|V(n,k)|^2$. Hence it can be inferred that the additive noise assumption in the modulation domain is roughly satisfied in high spectral SNR regions.

Fig. 3 shows the autocorrelation function of the modulating signal at eight acoustic frequencies for 32 ms of white Gaussian noise. We can see that the modulating signals of white noise do contain some correlation at higher lags and hence their modulation spectrum is not white. Therefore, in order to accommodate this fact, the coloured-noise Kalman filter (Gibson et al., 1991) was chosen for use in the proposed MDKF-MMSE, where an extra qth linear predictor is used to model the noise and the state vectors and transition matrices are augmented to sizes of $p+q$. The Kalman recursive equations for the coloured-noise case are provided in the Appendix.

In order to handle non-stationary noise, we require the q linear predictor coefficients to be updated for each Kalman filter whenever speech is absent in the modulating signal. The noise estimate is obtained in a similar fashion to Paliwal et al. (2010), where it is based on a decision from a simple voice activity detector (VAD) (Loizou, 2007) applied in the modulation domain. As mentioned before, the modulating signals $|Y(n,k)|$ are windowed into short acoustic frames prior to the LPC analysis. The modulation spectrum is computed using STFT analysis (Paliwal et al., 2010):

$$\mathcal{Y}(\eta,k,m) = \sum_{l=-\infty}^{\infty} |Y(l,k)|t(\eta-l)e^{-j\frac{2\pi ml}{M}} \quad (17)$$

where η is the acoustic frame number, m refers to the index of the discrete modulation frequency, M is the modulation frame duration (in terms of acoustic frames) and $t(\eta)$ is a modulation analysis window function. The VAD classifies each modulation domain frame as either 1 (speech present) or 0 (speech absent), using the following binary rule:

$$\Phi(\eta,k) = \begin{cases} 1, & \text{if } \phi(\eta,k) \geqslant \theta \\ 0, & \text{otherwise} \end{cases} \quad (18)$$

where η is the modulation frame number, θ is an empirically determined speech presence threshold, and $\phi(\eta,k)$ denotes a modulation frame SNR computed as follows:

$$\phi(\eta,k) = 10\log_{10}\left(\frac{\sum_m |\mathcal{Y}(\eta,k,m)|^2}{\sum_m |\widehat{\mathcal{V}}(\eta-1,k,m)|^2}\right) \quad (19)$$

where $|\widehat{\mathcal{V}}(\eta-1,k,m)|$ is the estimated modulation magnitude spectrum of the noise in the previous modulation frame. The noise estimate is updated during speech absence using the following averaging rule (Virag, 1999):

$$|\widehat{\mathcal{V}}(\eta,k,m)|^2 = \lambda|\widehat{\mathcal{V}}(\eta-1,k,m)|^2 + (1-\lambda)|\widehat{\mathcal{Y}}(\eta,k,m)|^2 \quad (20)$$

where $|\widehat{\mathcal{V}}(\eta,k,m)|^2$ is the modulation power spectrum of the noise and λ is a forgetting factor chosen depending on the stationarity of the noise. Once the modulation power spectrum of the noise has been updated, an inverse discrete Fourier transform is applied to obtain $q+1$ autocorrelation coefficients and these are used in the Levinson–Durbin algorithm to compute the updated q linear predictor coefficients of the noise.

Finally, in regards to the dynamic model, we have observed in our experiments that for the MDKF in the ideal case (where clean speech parameters are available), the linear predictor is sufficient at modelling the modulating signals of clean speech. Since temporal changes in the vocal tract tend to be relatively slow due to physiological constraints, we have found that low LPC orders ($p=2$) are sufficient for modelling the modulating signals. However, the presence of noise will introduce bias in the LPC estimates, which will degrade the performance of the Kalman filter. In this study, we evaluate the MDKF-MMSE method, which pre-processes the noisy speech using the MMSE-STSA method (as shown in Fig. 2) prior to LPC

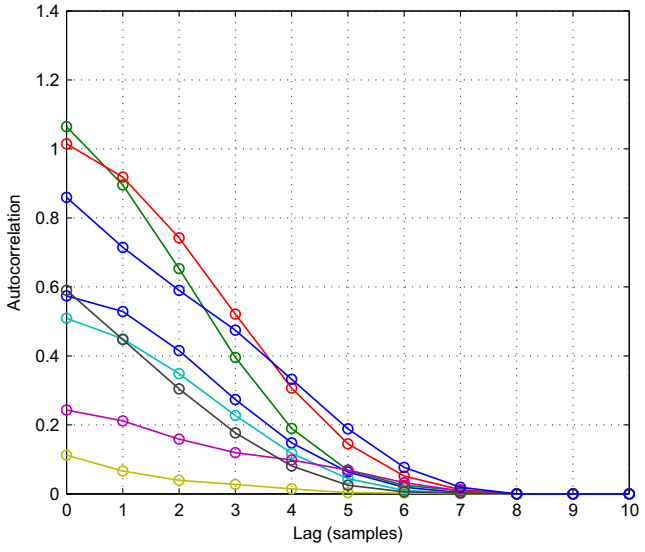

Fig. 3. Plot of autocorrelation function of the modulating signals at eight acoustic frequencies for 32 ms of white Gaussian noise.

estimation in the modulation domain in order to reduce the effect of noise, in a similar manner to the Kalman-PSC filter proposed by So et al. (2009).

2.3. Performance analysis and comparison between modulation-domain and time-domain Kalman filtering with LPCs from clean speech

For the purposes of explaining the limitations of the time-domain Kalman filter, we include the Kalman recursive equations for reference (So et al., 2009):

$$\hat{\boldsymbol{x}}(n|n-1) = \boldsymbol{A}\hat{\boldsymbol{x}}(n-1|n-1) \tag{21}$$

$$\boldsymbol{P}(n|n-1) = \boldsymbol{A}\boldsymbol{P}(n-1|n-1)\boldsymbol{A}^T + \sigma_w^2\boldsymbol{d}\boldsymbol{d}^T \tag{22}$$

$$\boldsymbol{K}(n) = \boldsymbol{P}(n|n-1)\boldsymbol{c}\left[\sigma_v^2 + \boldsymbol{c}^T\boldsymbol{P}(n|n-1)\boldsymbol{c}\right]^{-1} \tag{23}$$

$$\hat{\boldsymbol{x}}(n|n) = \hat{\boldsymbol{x}}(n|n-1) + \boldsymbol{K}(n)[y(n) - \boldsymbol{c}^T\hat{\boldsymbol{x}}(n|n-1)] \tag{24}$$

$$\boldsymbol{P}(n|n) = [\boldsymbol{I} - \boldsymbol{K}(n)\boldsymbol{c}^T]\boldsymbol{P}(n|n-1) \tag{25}$$

where $\hat{\boldsymbol{x}}(n|n-1)$ and $\hat{\boldsymbol{x}}(n|n)$ are the *a priori* and *a posteriori* state vectors, respectively; $\boldsymbol{P}(n|n-1)$ and $\boldsymbol{P}(n|n)$ are the *a priori* and *a posteriori* error covariance matrices, respectively; $\boldsymbol{K}(n)$ is the Kalman gain; and σ_v^2 and σ_w^2 is the variance of the noise and excitation, respectively.

Fig. 4 shows spectrograms of white noise-corrupted speech that has been enhanced by the TDKF [Fig. 4(c)] and MDKF [Fig. 4(d)], where LPCs from the clean speech are available. While these are not available in practice, the aim of this section is to compare the empirical upper-bound performance of the two enhancement methods. We can see in the spectrograms that both methods do a good job at suppressing the noise, particularly in the regions where there is no speech. However, it can be seen in the TDKF output that some noise is present in the speech, where it is particularly noticeable in between the pitch harmonics. Also, the harmonics above 1600 Hz that are seen in the original clean speech appear to have been replaced by noise. This was confirmed by informal listening of the TDKF output, where we noticed the speech to sound breathy and partially voiceless.

This characteristic is a limitation of the Kalman filter for speech enhancement, since the enhanced output is formed by a linear combination of the observed speech and predicted speech (by rearranging Eq. (24)):

$$\hat{\boldsymbol{x}}(n|n) = [\boldsymbol{I} - \boldsymbol{K}(n)\boldsymbol{c}^T]\underbrace{\hat{\boldsymbol{x}}(n|n-1)}_{\text{predicted}} + \boldsymbol{K}(n)\underbrace{y(n)}_{\text{observed}} \tag{26}$$

We can see that the relative weighting of the two components is controlled by the Kalman gain, which itself is dependent on the power of the prediction error versus that of the noise (see Eq. (23)). When there is no speech present, $\boldsymbol{P}(n|n-1) = \boldsymbol{0}$, which means that $\boldsymbol{K}(n) = \boldsymbol{0}$, hence the estimated state vector contains no (noisy) observed component. However, the limitation arises during regions where speech is present and both components are combined to form the estimated state vector. Since the low-order linear predictor model uses only short-term correlation information, which does not capture the harmonic structure of voiced speech, the predicted component will contribute only to the formant structure, while introducing unvoiced and noise-like characteristics. In relation to the observed component, it is essentially the noisy speech, from which we can observe in Fig. 4(b) that its harmonic structure above 1600 Hz has been overcome with noise due to the inherent spectral tilt of the speech power spectrum. Therefore, the observed speech component only preserves the strong harmonic structure below 1600 Hz. As a result, the enhanced speech from the Kalman filter suffers from breathy voice characteristics, especially at low SNRs where the predicted component would be more favoured over the observed one due to Eq. (23).

On observing the spectrogram of the MDKF enhanced speech in Fig. 4(d), we can see that the MDKF has overcome the limitations of the TDKF and a large part of the harmonic structure above 1600 Hz has been preserved. There is also noticeably less residual noise in regions where

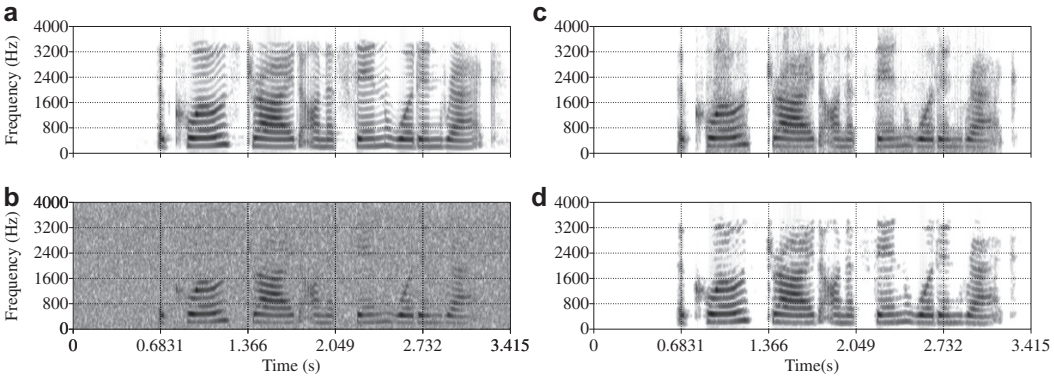

Fig. 4. Spectrograms of the sp15 utterance 'The clothes dried on a thin wooden rack' (female speaker) corrupted with white Gaussian noise, showing the enhancement provided by the modulation domain Kalman filter compared with the time domain Kalman filter in the ideal case: (a) clean speech; (b) speech corrupted with white Gaussian noise at 5 dB SNR (PESQ = 1.62); (c) time-domain Kalman filter with $p = 10$, $q = 4$ (PESQ = 2.43); (d) modulation domain Kalman filter with $p = 2$ (PESQ = 3.50).

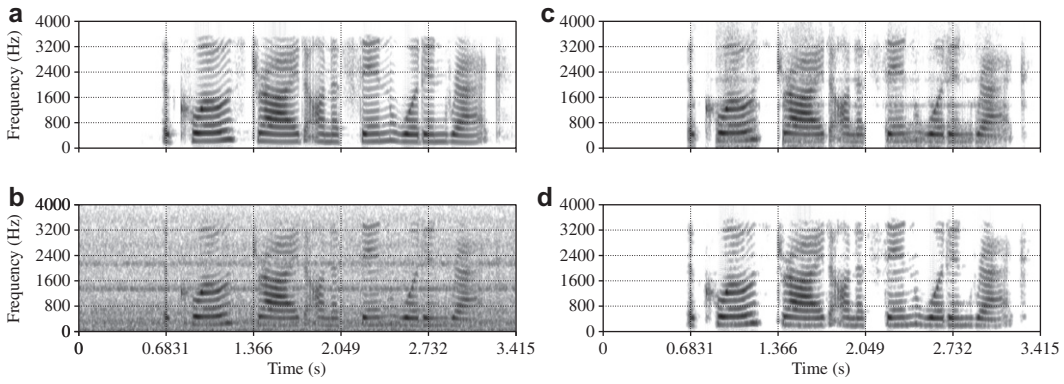

Fig. 5. Spectrograms of the sp15 utterance 'The clothes dried on a thin wooden rack' (female speaker) corrupted with coloured noise, showing the enhancement provided by the modulation domain Kalman filter compared with the time domain Kalman filter in the ideal case: (a) clean speech; (b) speech corrupted with coloured noise (F-16 noise) at 5 dB SNR (PESQ = 1.92); (c) time-domain Kalman filter with $p = 10$, $q = 4$ (PESQ = 2.41); (d) modulation domain Kalman filter (PESQ = 3.59) with $p = 2$.

speech is present when compared with the TDKF output in Fig. 4(c). As a result, the PESQ (perceptual evaluation of speech quality) score of the MDKF is much higher than that of the TDKF. The advantage of the MDKF over the TDKF lies in the linear predictor model used in the Kalman filter. In the TDKF, the linear predictor is used to model speech using short-term autocorrelation coefficients and as we have noted, this dynamic model is not sufficient at reproducing the harmonic structure of speech, which require autocorrelation lags in the order of the number of samples in a pitch period. On the other hand, the linear predictor in the MDKF is modelling the time trajectories of the acoustic magnitude spectrum of speech, which represents the changes of the vocal tract as a function of time. Therefore, the residual noise that accompanies the MDKF is mostly manifested in the modulation frequency spectrum, rather than the acoustic frequency spectrum (as is the case with the TDKF). Another advantage of Kalman filtering in the modulation domain is that low-order linear predictors are sufficient at modelling the modulating signal dynamics, due to the physiological limitation of how fast the vocal tract is able to change with time (Paliwal et al., 2010).

Fig. 5 compares the performance of the TDKF and MDKF for coloured noise (F-16 noise) at 5 dB SNR. In a similar way to the white noise case, both methods suppress the noise very well in the regions where speech is absent. The harmonic structure above 1600 Hz appears better reconstructed in the TDKF in Fig. 5(c) than in the white noise case (in Fig. 4(c)) because of the lower level of noise at those frequencies. However, there is still the problem of noise 'leaking' into the enhanced output via the observed component and this is noticeable in Fig. 5(c), especially the remnants of the two dominating noise tones at approximately 1400 Hz and 2000 Hz, respectively. We can see that the MDKF output in Fig. 5(d) does not suffer the problems of the TDKF output and therefore, the former has a higher PESQ score. These trends between the ideal MDKF and TDKF are also validated in the average objective and subjective scores in Section 3.2.

3. Speech enhancement experiments

3.1. Experimental setup

In our experiments, we use the NOIZEUS speech corpus, which is composed of 30 phonetically balanced sentences belonging to six speakers (Loizou, 2007). 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 and coloured F-16 noise[1] at four SNR levels (0, 5, 10 and 15 dB). The noise-only sections of all the stimuli have been extended to approximately 500 ms to allow for reliable noise estimation for acoustic and modulation-domain enhancement methods. The FFT size (N) was 512. The objective evaluation was carried out on the NOIZEUS corpus using the PESQ measure (Rix et al., 2001) and the log likelihood ratio (LLR) distortion (Sambur and Jayant, 1976).

In addition, two sets of blind AB listening tests were undertaken to determine subjective method preference (Sorqvist et al., 1997). In the first set of listening tests, the NOIZEUS sentence, 'The clothes dried on a thin wooden rack', was corrupted with white Gaussian noise at 5 dB SNR. In the second set, the sentence was corrupted with coloured F16 noise at 5 dB SNR. Stimuli pairs were played back to several English-speaking listeners, who were asked to make a subjective preference for each stimuli pair. The total number of stimuli pair comparisons for seven treatment types (listed below) in each test was 42. This method was preferred over conventional MOS (mean opinion score)-based listening tests, which we have found to be prone to producing scores with a large variance.

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

1. original clean speech (**Clean**);

[1] The F-16 noise was obtained from the *Signal Processing Information Base* (SPIB) at <http://spib.rice.edu/spib/data/signals/noise/f16.html>.

2. speech corrupted with white Gaussian noise or coloured F16 noise (**Noisy**);
3. time-domain Kalman filter with LPCs estimated from clean speech, $p = 10$, $q = 4$, 20 ms frame duration with no overlap (**TDKF-clean**);
4. modulation-domain Kalman filter with LPCs estimated from clean speech, $p = 2$, 10 ms frame duration with 2.5 ms update in modulation domain, (**MDKF-clean**);
5. modulation-domain Kalman filter with LPCs estimated from noisy speech using three iterations (Gibson et al., 1991), $p = 2$, $q = 4$, 20 ms frame duration with no overlap in modulation domain, (**MDKF-iter**);
6. modulation-domain Kalman filter with LPCs estimated from MMSE-STSA enhanced speech, $p = 2$, $q = 4$, 20 ms frame duration with no overlap in modulation domain (**MDKF-MMSE**);
7. MMSE-STSA method (Ephraim and Malah, 1984) (**MMSE-STSA**);

For the methods that use an AMS framework, we have used 32 ms frames with 4 ms update.

3.2. Results and discussion

3.2.1. Objective results

Tables 1 and 2 show the average PESQ scores comparing the different speech enhancement methods for white Gaussian noise and F16 noise, respectively. PESQ scores for the acoustic and time-domain enhancement methods are given in the top half of the tables while the bottom half contain the PESQ scores for modulation-domain Kalman filtering methods. From these results, we can see that in almost all cases and for both noise types, the MDKF methods give higher PESQ scores than the acoustic and time-domain methods. In particular, the MDKF-ideal method, which represents the upper bound performance of Kalman filtering in the modulation domain, has achieved the highest PESQ scores, even outperforming the TDKF-clean, which also had the benefit of using clean LPC estimates. This reaffirms our observation in Section 2.3 that the Kalman filter appears better suited for enhancement in the modulation domain than in the time domain. We can also see that the proposed MDKF-MMSE method makes up for some of the performance loss when only noisy speech is available for LPC estimation. Finally, these objective scores suggest that the combination of MMSE-STSA preprocessing prior to LPC estimation is superior to iterative LPC estimation, when used within the MDKF.

Tables 3 and 4 present the average LLR distortions for each of the speech enhancement methods that were evaluated for white and coloured F16 noise, respectively. From these results, we can see that the enhanced speech from the MDKF-ideal method consistently had the lowest LLR distortion, even when compared with the TDKF-clean. In the case of F16 noise, the LLR distortion was less than half of the distortion from the TDKF-clean method. Together

Table 1
Average PESQ scores comparing the different speech enhancement methods for speech from the NOIZEUS corpus that have been corrupted by white Gaussian noise. Bold numbers show the best score.

Method	Input SNR (dB)			
	0	5	10	15
No enhancement	1.57	1.83	2.14	2.48
Acoustic and time-domain methods:				
TDKF-clean	2.49	2.77	3.09	3.41
MMSE-STSA	1.96	2.33	2.66	2.94
Modulation-domain Kalman filtering:				
MDKF-ideal	**3.34**	**3.54**	**3.72**	**3.89**
MDKF-iter	1.94	2.33	2.70	3.07
MDKF-MMSE	2.19	2.51	2.81	3.06

Table 2
Average PESQ scores comparing the different speech enhancement methods for speech from the NOIZEUS corpus that have been corrupted by F16 noise. Bold numbers show the best score.

Method	Input SNR (dB)			
	0	5	10	15
No enhancement	1.84	2.17	2.49	2.82
Acoustic and time-domain methods:				
TDKF-clean	2.54	2.80	3.13	3.47
MMSE-STSA	2.32	2.63	2.91	3.15
Modulation-domain Kalman filtering:				
MDKF-ideal	**3.47**	**3.65**	**3.82**	**3.97**
MDKF-iter	2.25	2.62	3.00	3.38
MDKF-MMSE	2.42	2.74	3.05	3.31

Table 3
Average LLR distortions comparing the different speech enhancement methods for speech from the NOIZEUS corpus that have been corrupted by white noise. Bold numbers show the best score.

Method	Input SNR (dB)			
	0	5	10	15
No enhancement	1.52	1.40	1.25	1.08
Acoustic and time-domain methods:				
TDKF-clean	0.62	0.54	0.47	0.38
MMSE-STSA	1.34	1.18	1.03	0.89
Modulation-domain Kalman filtering:				
MDKF-ideal	**0.49**	**0.40**	**0.30**	**0.22**
MDKF-iter	1.47	1.32	1.16	1.00
MDKF-MMSE	1.34	1.22	1.09	0.96

Table 4
Average LLR distortions comparing the different speech enhancement methods for speech from the NOIZEUS corpus that have been corrupted by F16 noise. Bold numbers show the best score.

Method	Input SNR (dB)			
	0	5	10	15
No enhancement	1.13	0.97	0.81	0.66
Acoustic and time-domain methods:				
TDKF-clean	0.55	0.48	0.40	0.31
MMSE-STSA	0.95	0.81	0.69	0.60
Modulation-domain Kalman filtering:				
MDKF-ideal	**0.27**	**0.21**	**0.15**	**0.10**
MDKF-iter	1.10	0.92	0.75	0.62
MDKF-MMSE	1.00	0.86	0.75	0.65

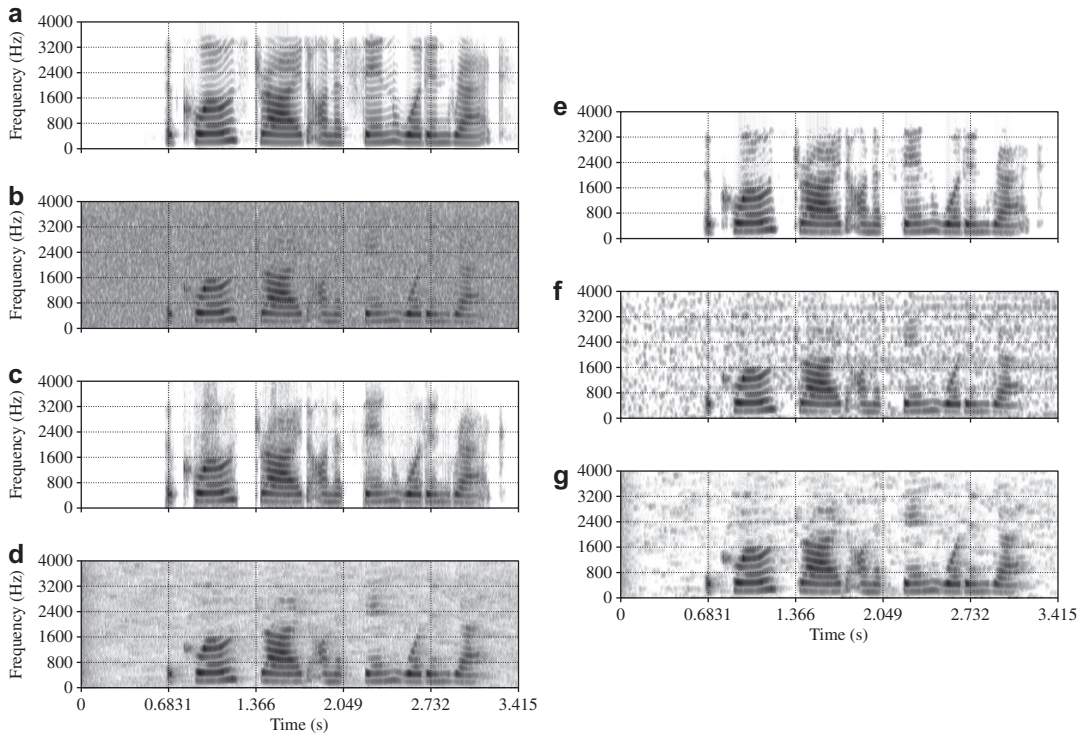

Fig. 6. Spectrograms from the treatment types for the sp15 utterance 'The clothes dried on a thin wooden rack': (a) clean speech; (b) speech corrupted with white Gaussian noise at 5 dB SNR (PESQ = 1.62); (c) TDKF-clean (PESQ = 2.46); (d) MMSE-STSA (PESQ = 2.30); (e) MDKF-clean (PESQ = 3.50); (f) MDKF-iter (PESQ = 2.30); (g) MDKF-MMSE (PESQ = 2.54).

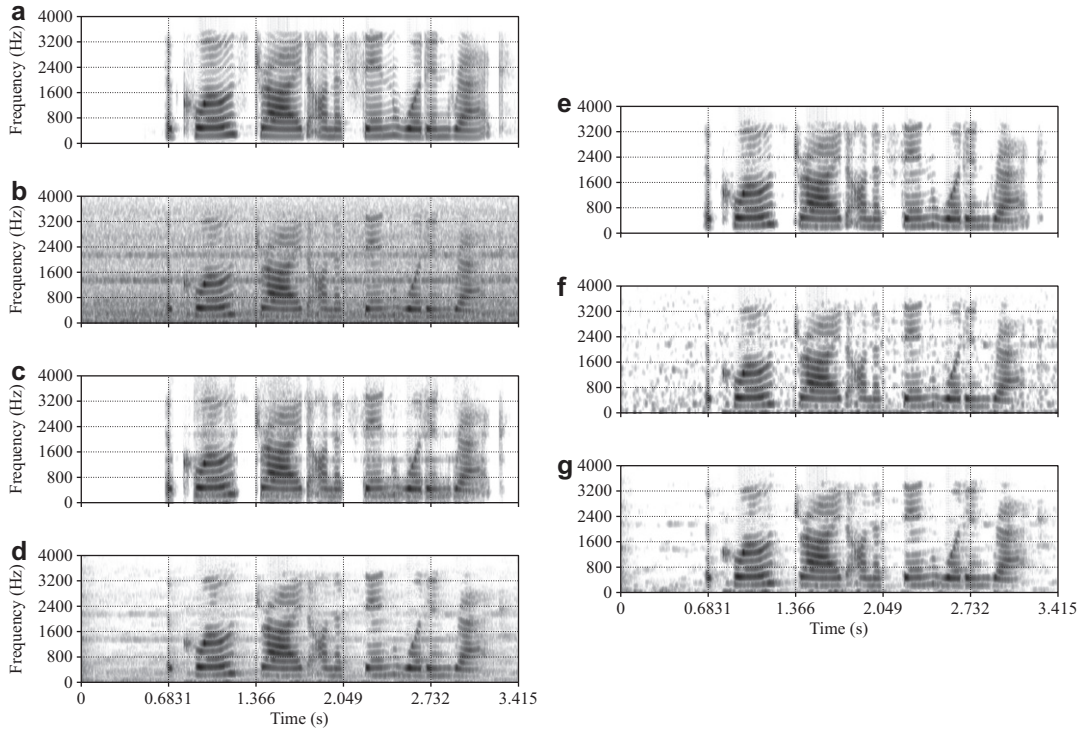

Fig. 7. Spectrograms from the treatment types for the sp15 utterance 'The clothes dried on a thin wooden rack' : (a) clean speech; (b) speech corrupted with coloured F16 noise at 5 dB SNR (PESQ = 1.92); (c) TDKF-clean (PESQ = 2.41); (d) MMSE-STSA (PESQ = 2.60); (e) MDKF-clean (PESQ = 3.59); (f) MDKF-iter (PESQ = 2.68); (g) MDKF-MMSE (PESQ = 2.75).

with the PESQ scores, these objective results suggest that the Kalman filter performs enhances speech more effectively when processing in the modulation domain than it does in the time domain.

3.2.2. Spectrogram analysis

Figs. 6 and 7 show spectrogram comparisons between the various enhancement methods for white Gaussian and F16 noises at an SNR of 5 dB. We can see that the output speech from the MDKF-iter method in Figs. 6(f) and 7(f) suffer from musical noise, which was also observed previously for the iterative TDKF in So and Paliwal (2011). In comparison, the spectrograms of the speech from the MDKF-MMSE method in Figs. 6(g) and 7(g) do not show signs of strong and localised musical-like tones. The residual noise level of the MDKF-MMSE also appears lower than that of the MMSE-STSA method.

A further observation can be made when we compare the spectrograms from the TDKF-clean and MDKF-MMSE in Figs. 7(c) and 7(g), respectively. We can see that in the regions where speech is present, the MDKF-MMSE method does not introduce the noise that we observe in the TDKF-clean output at frequencies above 1600 Hz.

3.2.3. Subjective listening tests

Figs. 8 and 9 show the mean preference scores for the subjective listening tests for white Gaussian noise and coloured F16 noise. We can see that for both noise types, the MDKF-clean method was consistently preferred over the other enhancement methods (second only to clean speech) by the listeners, who noted that the speech enhanced by MDKF-clean sounded very similar to the clean speech with no residual noise detected. Because the LPCs were estimated from the clean speech, this result is considered the upper performance bound of the MDKF. When the LPCs were iteratively estimated from the noise-corrupted speech using the method proposed by Gibson et al. (1991) in the MDKF-iter method, we note that the mean subjective preference score decreased dramatically to below that of the MMSE-STSA method. This correlates with our spectrogram analysis, where a large amount of musical noise was observed for the MDKF-iter method. On the other hand, the proposed MDKF-MMSE method had the third highest mean preference score, outperforming MDKF-iter as well as the other time and acoustic-domain enhancement methods.

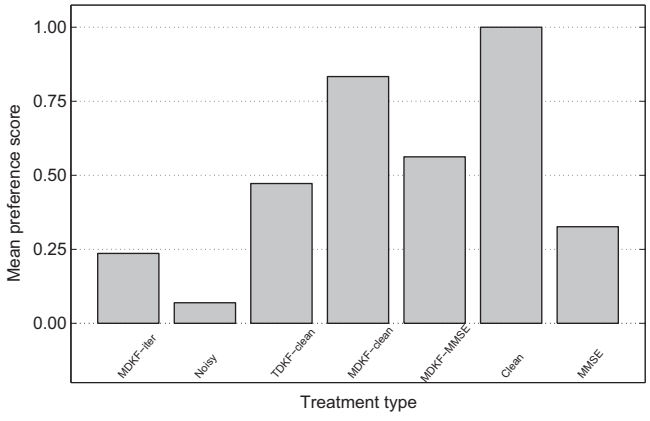

Fig. 8. Mean preference scores from subjective listening tests of sp15 utterance 'The clothes dried on a thin wooden rack' corrupted with white Gaussian noise at 5 dB.

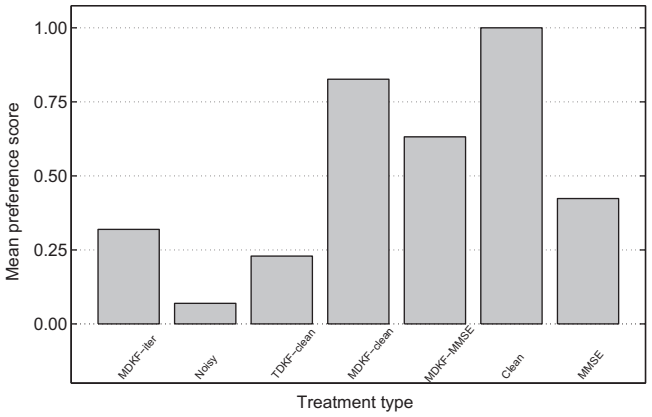

Fig. 9. Mean preference scores from subjective listening tests of sp15 utterance 'The clothes dried on a thin wooden rack' corrupted with coloured F16 noise at 5 dB.

It is interesting to point out that the MDKF-MMSE subjectively scored higher than the TDKF-clean, which had the advantage of using LPC estimates from the clean speech. Comments from the listeners suggested that they did not like the residual noise that 'leaked' into the TDKF-clean output during the regions where speech was present, even though the silent regions were mostly noise-free. In other words, the listeners preferred residual noise levels that were uniformly spread out in time, rather than in short bursts during the speech, which was the case with TDKF-clean. On the other hand, the MDKF-clean does not suffer from residual noise problems. Therefore, we can infer that in a speech enhancement scenario where accurate LPC estimates are available, the Kalman filter performs best when applied in the modulation domain, rather than the time domain.

4. Conclusions

In this paper, we have investigated the use of Kalman filtering in the modulation domain and compared its performance with other time-domain and acoustic-domain speech enhancement methods. In contrast to previously reported modulation domain-enhancement methods which consisted of fixed bandpass filtering, the modulation-domain Kalman filter (MDKF) is an adaptive MMSE estimator that uses the statistics of temporal changes in the magnitude spectrum for both speech and noise. Furthermore, since the modulation phase plays a more important role than acoustic phase, the Kalman filter is highly suited since it is a joint magnitude and phase spectrum estimator, under non-stationarity assumptions. We have shown empirically that the upper bound performance of the MDKF exceeds that of the conventional time-domain

Kalman filter (TDKF). This was attributed to the inability of the TDKF and its low order dynamic model to predict long-term correlation information (such as pitch harmonics), which resulted in breathy unvoiced speech that contained short bursts of residual noise. Due to the physiological limitations of the temporal dynamics of the vocal tract, the MDKF with a low order dynamic model was found to be more effective at enhancing the modulating signals, producing speech that had very minimal distortion and no trace of residual noise. Experimental results from objective tests and blind subjective listening tests from the NOIZEUS corpus showed the MDKF (with clean speech parameters) to outperform all the acoustic and time-domain enhancement methods evaluated.

Acknowledgements

The authors would like to thank the anonymous reviewer for their valuable and constructive feedback during the review process. In addition, the authors would like to acknowledge Kamil Wójcicki for providing the AMS and modulation domain processing framework code as well as his preliminary work on the MDKF.

Appendix A. Kalman recursion equations for the coloured noise case

In this appendix, we provide the recursion equations for the Kalman filter for the coloured noise case (Gibson et al., 1991), which we have used in the MDKF-MMSE method. The kth modulating signal of the coloured noise $|V(n,k)|$ is modelled using a qth order linear predictor:

$$|V(n,k)| = -\sum_{j=1}^{q} b_{j,k}|V(n-j,k)| + U(n,k) \tag{A.1}$$

where $U(n,k)$ is a white random signal with a variance of $\sigma^2_{U(k)}$. We define the following state vector:

$$\boldsymbol{V}(n,k) = \begin{bmatrix} |V(n,k)| \\ |V(n-1,k)| \\ \vdots \\ |V(n-q+1,k)| \end{bmatrix} \tag{A.2}$$

Therefore, the state-space representation for the coloured noise can be written as:

$$\boldsymbol{V}(n,k) = \boldsymbol{B}(k)\boldsymbol{V}(n-1,k) + \boldsymbol{d}_v U(n,k) \tag{A.3}$$

$$|V(n,k)| = \boldsymbol{c}_v^T \boldsymbol{V}(n,k) \tag{A.4}$$

where $\boldsymbol{c}_v = [1,0,\ldots,0]^T$, $\boldsymbol{d}_v = [1,0,\ldots,0]^T$, and:

$$\boldsymbol{B}(k) = \begin{bmatrix} -b_{1,k} & -b_{2,k} & \cdots & -b_{q-1,k} & -b_{q,k} \\ 1 & 0 & \cdots & 0 & 0 \\ 0 & 1 & \cdots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \cdots & 1 & 0 \end{bmatrix} \tag{A.5}$$

We can combine the modulating signal of the speech $|X(n,k)|$ and coloured noise $|V(n,k)|$ into one set of state-space equations:

$$\begin{bmatrix} \boldsymbol{X}(n,k) \\ \boldsymbol{V}(n,k) \end{bmatrix} = \begin{bmatrix} \boldsymbol{A}(k) & 0 \\ 0 & \boldsymbol{B}(k) \end{bmatrix} \begin{bmatrix} \boldsymbol{X}(n-1,k) \\ \boldsymbol{V}(n-1,k) \end{bmatrix} + \begin{bmatrix} \boldsymbol{d} & 0 \\ 0 & \boldsymbol{d}_v \end{bmatrix} \begin{bmatrix} W(n,k) \\ U(n,k) \end{bmatrix} \tag{A.6}$$

$$|Y(n,k)| = \begin{bmatrix} \boldsymbol{c}^T & \boldsymbol{c}_v^T \end{bmatrix} \begin{bmatrix} \boldsymbol{X}(n-1,k) \\ \boldsymbol{V}(n-1,k) \end{bmatrix} \tag{A.7}$$

These can be rewritten in augmented matrix form:

$$\widetilde{\boldsymbol{X}}(n,k) = \widetilde{\boldsymbol{A}}(k)\widetilde{\boldsymbol{X}}(n-1,k) + \boldsymbol{D}\widetilde{\boldsymbol{W}}(n,k) \tag{A.8}$$

$$|Y(n,k)| = \tilde{\boldsymbol{c}}^T \widetilde{\boldsymbol{X}}(n,k) \tag{A.9}$$

Using this augmented matrix notation, we can therefore write the Kalman recursive equations as:

$$\boldsymbol{P}(n|n-1,k) = \widetilde{\boldsymbol{A}}(k)\boldsymbol{P}(n-1|n-1,k)\widetilde{\boldsymbol{A}}(k)^T + \widetilde{\boldsymbol{D}}\boldsymbol{Q}\widetilde{\boldsymbol{D}}^T \tag{A.10}$$

$$\boldsymbol{K}(n,k) = \boldsymbol{P}(n|n-1,k)\tilde{\boldsymbol{c}}[\tilde{\boldsymbol{c}}^T\boldsymbol{P}(n|n-1,k)\tilde{\boldsymbol{c}}]^{-1} \tag{A.11}$$

$$\widetilde{\boldsymbol{X}}(n|n-1,k) = \widetilde{\boldsymbol{A}}(k)\widetilde{\boldsymbol{X}}(n-1|n-1,k) \tag{A.12}$$

$$\boldsymbol{P}(n|n,k) = [\boldsymbol{I} - \boldsymbol{K}(n,k)\tilde{\boldsymbol{c}}^T]\boldsymbol{P}(n|n-1,k) \tag{A.13}$$

$$\widetilde{\boldsymbol{X}}(n|n,k) = \widetilde{\boldsymbol{X}}(n|n-1,k) + \boldsymbol{K}(n,k)[|Y(n,k)| - \tilde{\boldsymbol{c}}^T\widetilde{\boldsymbol{X}}(n|n-1,k)] \tag{A.14}$$

Since $W(n,k)$ and $U(n,k)$ are assumed to be uncorrelated, then:

$$\boldsymbol{Q} = \begin{bmatrix} \sigma^2_{W(k)} & 0 \\ 0 & \sigma^2_{U(k)} \end{bmatrix} \tag{A.15}$$

References

Arai, T., Pavel, M., Hermansky, H., Avendano, C., 1999. Syllable intelligibility for temporally filtered LPC cepstral trajectories. J. Acoust. Soc. Amer. 105 (5), 2783–2791.

Atlas, L., Shamma, S.A., 2003. Joint acoustic and modulation frequency. EURASIP J. Appl. Signal Process. 2003, 668–675.

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

Chen, J., Benesty, J., Huang, Y., Doclo, S., 2006. New insights into the noise reduction Wiener filter. IEEE Trans. Audio Speech Lang. Process. 14 (4), 1218–1234.

Drullman, R., Festen, J.M., Plomp, R., 1994a. Effect of reducing slow temporal modulations on speech reception. J. Acoust. Soc. Amer. 95 (2), 2670–2680.

Drullman, R., Festen, J.M., Plomp, R., 1994b. Effect of temporal envelope smearing on speech reception. J. Acoust. Soc. Amer. 95 (2), 1053–1064.

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., Malah, D., 1985. Speech enhancement using a minimum mean square error log-spectral amplitude estimator. IEEE Trans. Acoust. Speech Signal Process. ASSP-33, 443–445.

Falk, T., Stadler, S., Kleijn, W.B., Chan, W.Y., 2007. Noise suppression based on extending a speech-dominated modulation band. In: Proc. European Signal Processing Conference. pp. 970–973.

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.

Greenberg, S., Arai, T., 2001. The relation between speech intelligibility and the complex modulation spectrum. In: Proc. European Conference on Speech Communication and Technology. pp. 473–476.

Greenberg, S., Arai, T., Silipo, R., 1998. Speech intelligibility derived from exceedingly sparse spectral information. In: Proc. Int. Conf. Spoken Language Processing. pp. 2803–2806.

Hermansky, H., Wan, E., Avendano, C., 1995. Speech enhancement based on temporal processing. In: Proc. IEEE Int. Conf. Acoustics, Speech, Signal Processing, vol. 1, pp. 405–408.

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

Kanedera, N., Hermansky, H., Arai, T., 1998. On properties of modulation spectrum for robust automatic speech recognition. In: Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing, vol. 2, pp. 613–616.

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.

Loizou, P., 2007. Speech Enhancement: Theory and Practice, first ed. CRC Press LLC.

Lyons, J.G., Paliwal, K.K., 2008. Effect of compressing the dynamic range of the power spectrum in modulation filtering based speech enhancement. In: Proc. INTERSPEECH 2008, pp. 387–390.

Mesgarani, N., Shamma, S., 2005. Speech enhancement based on filtering the spectrotemporal modulations. In: Proc. IEEE Int. Conf. Acoustics, Speech, Signal Processing, pp. 1105–1108.

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

Paliwal, K.K., Wojcicki, K.K., Schwerin, B., 2010. Single-channel speech enhancement using spectral subtraction in the short-time modulation domain. Speech Comm. 52 (5), 450–475.

Paliwal, K.K., Schwerin, B., Wojcicki, K.K., 2011. Role of modulation magnitude and phase spectrum towards speech intelligibility. Speech Commun. 53 (3), 327–339.

Quatieri, T., 2002. Discrete-Time Speech Signal Processing: Principles and Practice. Prentice-Hall, Upper Saddle River, NJ.

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. Technical Report, ITU-T.

Sambur, M.R., Jayant, N.S., 1976. LPC analysis/synthesis from speech inputs containing quantizing noise or additive white noise. IEEE Trans. Acoust. Speech Signal Process. ASSP-24, 488–494.

So, S., Paliwal, K.K., 2011. Suppressing the influence of additive noise on the Kalman filter gain for low residual noise speech enhancement. Speech Commun. 53 (3), 355–378.

So, S., Wojcicki, K.K., Lyons, J.G., Stark, A.P., Paliwal, K.K., 2009. Kalman filter with phase spectrum compensation algorithm for speech enhancement. In: Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing, pp. 4405–4408.

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.

Virag, N., 1999. Single channel speech enhancement based on masking properties of the human auditory system. IEEE Trans. Speech Audio Process. 7 (2), 126–137.

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

Wu, W.R., Chen, P.C., 1998. Subband Kalman filtering for speech enhancement. IEEE Trans. Circuits Syst. II 45 (8), 1072–1083.