Speech Communication 125 (2020) 142–151

Contents lists available at ScienceDirect

Speech Communication

journal homepage: www.elsevier.com/locate/specom

Speech enhancement using a DNN-augmented colored-noise Kalman filter

Hongjiang Yu [*,a], Wei-Ping Zhu [a], Benoit Champagne [b]

[a] Department of Electrical and Computer Engineering, Concordia University, Montreal, Quebec, Canada
[b] Department of Electrical and Computer Engineering, McGill University, Montreal, Quebec, Canada

ARTICLE INFO

Keywords:
Speech enhancement
Deep neural network
Corlored-noise Kalman filter
Spectral subtraction

ABSTRACT

In this paper, we propose a new speech enhancement system using a deep neural network (DNN)-augmented colored-noise Kalman filter. In our system, both clean speech and noise are modelled as autoregressive (AR) processes, whose parameters comprise the linear prediction coefficients (LPCs) and the driving noise variances. The LPCs are obtained through training a multi-objective DNN that learns the mapping from the noisy acoustic features to the line spectrum frequencies (LSFs), while the driving noise variances are obtained by solving an optimization problem aiming to minimize the difference between the modelled and observed AR spectra of the noisy speech. The colored-noise Kalman filter with DNN estimated parameters is then applied to the noisy speech for denoising. Finally, a post-subtraction technique is adopted to further remove the residual noise in the Kalman-filtered speech. Extensive computer simulations show that the proposed speech enhancement system achieves significant performance gains when compared to conventional Kalman filter based algorithms as well as recent DNN-based methods under both seen and unseen noise conditions.

1. Introduction

Speech enhancement, which aims to suppress the background noise and improve the quality and intelligibility of a speech signal, has been widely adopted as a pre-processing means in a variety of speech-related applications to provide better user experience. Numerous speech enhancement techniques have been proposed in the literature over the past decades, but due to their limited performance, the problem continues to be intensively studied.

Spectral subtraction (Boll, 1979), one of the earliest techniques for speech enhancement, modifies the noisy speech power spectrum by subtracting the estimated noise power spectrum. Although spectral subtraction is easy to employ, the difficulty in accurately estimating the noise spectrum hinders the enhancement performance. Extra distortion, such as the musical noise, can degrade the perceptual quality of the enhanced speech if the noise spectrum is not accurately estimated. More flexible spectral subtraction algorithms with better performance were proposed in Berouti et al. (1979), Kushner et al. (1989), where two techniques, i.e., the use of oversubtraction factor and spectral flooring parameter, were introduced along with the standard spectral subtraction. These techniques are used to adjust the estimated noise spectrum, and thereby control the ratio of the remaining residual noise and perceived musical noise in the enhanced speech. In Singh and Sridharan (1998), Kamath and Loizou (2002), a multiband spectral subtraction was proposed based on the fact that the noise affects the speech at different levels depending on frequency bands. In the multiband approach, the speech spectrum is divided into several non-overlapping frequency bands, and then spectral subtraction is performed independently in each band.

The statistical filter based speech enhancement methods have also received considerable attention. Wiener filtering, one of the most famous algorithms in this class, aims to find the minimum mean square error (MMSE) estimate of the clean speech's discrete Fourier transform (DFT) coefficients (Lim and Oppenheim, 1979; Chen et al., 2006; Srinivasan et al., 2005). Compared with spectral subtraction, Wiener filtering introduces less distortion in the enhanced speech. However, Wiener filters are derived under the assumption that the processed signals are stationary, which is rarely satisfied in real-world applications. Kalman filters, which can handle non-stationary signals, have therefore attracted the interests of speech enhancement researchers (Paliwal and Basu, 1987). In this context, the Kalman filter can be viewed as a time-domain, sequential linear MMSE estimator of the noise corrupted speech, in which the clean speech is characterized by a dynamical or state-space model, such as the autoregressive (AR) model. As such, the enhancement performance is largely dependent on the estimation accuracy of the AR parameters, which include the linear prediction

* Corresponding author.
E-mail address: ho_yu@encs.concordia.ca (H. Yu).

https://doi.org/10.1016/j.specom.2020.10.007
Received 21 April 2020; Received in revised form 16 September 2020; Accepted 29 October 2020
Available online 4 November 2020

coefficients (LPCs) and the variances of the driving and observation noises.

Ideally, the AR parameters of the clean speech can lead to excellent performance of the Kalman filter (Paliwal and Basu, 1987), but they are not accessible in practice. Therefore, various estimation algorithms have been proposed to obtain the above parameters from the noisy speech, which can be divided into two categories: online estimation (Gibson et al., 1991; Gannot et al., 1998; Mellahi and Hamdi, 2015; Xia and Wang, 2015) and offline estimation (Nower et al., 2015; Kavalekalam et al., 2016). The former algorithms usually estimate and update the denoised speech and the model parameters in an iterative manner, while the latters require a training stage on a clean speech database to predict the parameters beforehand. To further improve the speech enhancement performance, several advanced versions of Kalman filters have been proposed. For example, the subband Kalman filtering technique (Wu and Chen, 1998; Roy et al., 2016; Yu et al., 2020a; 2020b) divides the noisy speech into several contiguous frequency bands, and performs Kalman filtering separately as the noise level dynamic varies in each band. The improved Kalman filter in Popescu and Zeljkovic (May, 1998), Grancharov et al. (2005) models both clean speech and noise as AR processes, and achieves a better performance in color noise environments. The perceptual Kalman filter (Ma et al., 2004; 2005) incorporates an additional post-filter to further remove the residual noise by scaling the estimation error of the Kalman filter below the masking threshold.

In recent years, deep learning, and especially deep neural network (DNN), has been successfully applied in many areas. Compared with the unsupervised statistical filter based methods, the use of DNN for speech enhancement offers several advantages: (1) powerful learning capability to model various non-linear mapping relationships; (2) no reliance on assumptions about the statistical properties of the speech and noise, and; (3) no specific need for the noise spectrum estimation. Early works in this area, e.g., Xu et al. (2013, 2015) employed DNN to directly estimate the clean speech magnitude spertrum, where the DNN acts as a regression model to implement a mapping function between the log-power spectra (LPS) of the noisy and clean speech signals. Subsequent works seeked to estimate ratio masks via DNN-based approaches, and then remove the background noise in the spectral domain by means of the estimated masks (Narayanan and Wang, 2013; Erdogan et al., 2015; Han et al., 2016; Williamson et al., 2016; Tu and Zhang, 2017; Yu et al., 2020c). For instance, in Narayanan and Wang (2013), the ideal ratio mask (IRM) is predicted by a DNN and then applied to the noisy magnitude spectrum to recover desired speech signal.

Nonetheless, deep learning algorithms require large training databases to improve their generalization capability (Xu et al., 2015). Since the statistical filter based methods can reduce different kinds and levels of noises to a sensible extent in a variety of situations, researchers have recently turned their attention to the combination of DNN and statistical filter based approaches. Indeed, the learning capability of the former makes it possible to boost speech enhancement performance under various conditions, while the latter helps better exploits the generalization capability of the enhancement system by providing an appropriate structural framework. Li and Kang (2016), Nie et al. (2018), Ouyang et al. (2018), Yu et al. (2019), Yu et al. (2020). Recently in Yu et al. (2019), we have proposed a DNN-augmented Kalman filter for speech enhancement, where the DNN is trained to predict the AR parameters needed for Kalman filtering. Experiments have shown that the AR parameters estimated in this way are less sensitive to various types of noise, leading to a better enhancement performance than the subband iterative Kalman filter algorithm (Roy et al., 2016). However, the enhanced speech still suffers from distortion at higher frequencies, partly due to the inaccurate estimation of additive noise and its harmful effects on the conventional Kalman filter.

In this paper, we propose a novel speech enhancement system consisting of a *colored-noise* Kalman filter augmented with DNN-based parameter estimation, where both clean speech and noise are modelled as AR processes. In our system, a multi-objective DNN is first employed to estimate the line spectrum frequencies (LSFs), which are used for the representation of the LPCs parameters in these models. Two kinds of DNN are used in this work, i.e. the fully-connected feed-forward DNN (denoted as FNN) (Li and Kang, 2016) and the long short-term memory (LSTM) (Chen and Wang, 2017). The driving noise variances for the clean speech process and the noise process are obtained by solving an optimization problem as in Srinivasan et al. (2005). The multi-objective DNN training is beneficial as it can simultaneously estimate the AR parameters of the clean speech and noise with a lower computational complexity, while providing more accurate estimates under noisy conditions. Subsequently, the colored-noise Kalman filter with the DNN estimated AR parameters is applied to the noisy speech for denoising. Finally, a post subtraction technique is employed to further remove the residual noise in the Kalman-filtered speech, which is caused by the parameter estimation error. Through exhaustive computer simulations, it is shown that the proposed system can not only significantly improve the performance of Kalman filtering in speech enhancement, but also offer a good generalization capability in both seen and unseen noise conditions.

The rest of the paper is organized as follows. Section 2 summarizes our previous work on DNN-based Kalman filtering. Section 3 presents the newly proposed speech enhancement system with DNN-augmented colored-noise Kalman filter, including a detailed description of its main components. Section 4 presents a series of experiments to assess the system performance. Section 5 concludes the paper.

2. Related work

Herein, we briefly review our previous work on speech enhancement using DNN and Kalman filtering (Yu et al., 2019), where the DNN is employed to estimate the AR parameters in the conventional Kalman filter.

2.1. Conventional Kalman filter

Consider the noisy speech $y(n)$ as an additive mixture of the clean speech $s(n)$ and the background noise $w(n)$,

$$y(n) = s(n) + w(n) \tag{1}$$

where $n \in \mathbb{N}$ is the discrete time index. As usual, $w(n)$ is regarded as a zero-mean white noise with variance σ_w^2, uncorrelated with $s(n)$. The clean speech $s(n)$ is usually represented by a linear model as a dynamic process of speech production. For the widely-adopted AR model, we have

$$s(n) = \sum_{i=1}^{p} a_{s,i} s(n-i) + v(n) \tag{2}$$

where $a_{s,i}$ are the LPCs of the clean speech, p the order of the model, and $v(n)$ the driving noise, i.e., a zero-mean white noise with variance σ_v^2.

To facilitate the Kalman filter presentation for speech enhancement, the above model equations for $s(n)$ and $y(n)$ can be rewritten in matrix form as,

$$\begin{cases} \mathbf{s}(\mathbf{n}) = \mathbf{F_s}\mathbf{s}(\mathbf{n}-\mathbf{1}) + \mathbf{G_s}\mathbf{v}(\mathbf{n}) \\ y(n) = \mathbf{H}_s^T \mathbf{s}(n) + w(n) \end{cases} \tag{3}$$

where $\mathbf{s}(n) = [s(n-p+1), \ldots, s(n-1), s(n)]^T$ denotes the speech state vector. Moreover, the transition matrix $\mathbf{F}_s$ is given by

$$\mathbf{F}_s = \begin{bmatrix} 0 & 1 & \cdots & 0 & 0 \\ 0 & 0 & \cdots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \cdots & 0 & 1 \\ a_{s,p} & a_{s,p-1} & \cdots & a_{s,2} & a_{s,1} \end{bmatrix} \tag{4}$$

and $\mathbf{H}_s = \mathbf{G}_s = [0, \ldots, 0, 1]^T \in \mathbb{R}^p$.

The denoising process with a Kalman filter amounts to recursively calculate an unbiased, linear MMSE estimate of the state vector $\mathbf{s}(n)$, given the corrupted speech $y(n)$. This process can be summarized by the following equations:

$$\begin{cases} e(n) = y(n) - \mathbf{H}_s^T \widehat{\mathbf{s}}(n|n-1) \\ \mathbf{K}(n) = \mathbf{P}(n|n-1)\mathbf{H}_s \left(\sigma_w^2 + \mathbf{H}_s^T \mathbf{P}(n|n-1)\mathbf{H}_s\right)^{-1} \\ \widehat{\mathbf{s}}(n|n) = \widehat{\mathbf{s}}(n|n-1) + \mathbf{K}(n)e(n) \\ \mathbf{P}(n|n) = \left(\mathbf{I} - \mathbf{K}(n)\mathbf{H}_s^T\right)\mathbf{P}(n|n-1) \\ \widehat{\mathbf{s}}(n+1|n) = \mathbf{F}_s \widehat{\mathbf{s}}(n|n) \\ \mathbf{P}(n+1|n) = \mathbf{F}_s \mathbf{P}(n|n)\mathbf{F}_s^T + \sigma_v^2 \mathbf{G}_s \mathbf{G}_s^T \end{cases} \tag{5}$$

where $\widehat{\mathbf{s}}(n|n-1)$ is the *a priori* estimate of the current state vector $\mathbf{s}(n)$, given observations up to a time index $n-1$, i.e., $y(1), \ldots, y(n-1)$, $\mathbf{P}(n|n-1)$ the predicted state error correlation matrix of $\widehat{\mathbf{s}}(n|n-1)$, $e(n)$ the innovation, $\mathbf{K}(n)$ the Kalman gain matrix, $\widehat{\mathbf{s}}(n|n)$ the filtered estimate of state vector $\mathbf{s}(n)$, and $\mathbf{P}(n|n)$ the filtered state error covariance matrix of $\widehat{\mathbf{s}}(n|n)$. The denoised speech $\widehat{s}(n)$ is finally given by

$$\widehat{s}(n) = \mathbf{G}_s^T \widehat{\mathbf{s}}(n|n). \tag{6}$$

2.2. Parameter estimation

We note that several parameters appearing in the above equations should be estimated or calculated from the noisy observations in order to perform Kalman filtering. Those parameters include the driving noise variance σ_v^2, the additive noise variance σ_w^2, and the transition matrix $\mathbf{F}_s$ which contains the LPCs of the clean speech model.

In our previous work (Yu et al., 2019), an FNN is adopted for the LPCs prediction. More specifically, the LPCs of the noisy speech and of the clean speech are first calculated and then converted into their representative LSFs, which are used as input features and output targets of the DNN, respectively. Using LSFs instead of LPCs offers a more stable DNN training process (Nower et al., 2015), due to the relatively well-behaved dynamic range of LSFs. The well-trained FNN can learn the non-linear relationship between the noisy LSFs and the clean ones. Finally, the estimated LSFs are transformed back to LPCs, as required in the transition matrix $\mathbf{F}_s$ needed to perform Kalman filtering.

The variance σ_w^2 of the additive noise $w(n)$ is usually estimated and updated during the unvoiced frames. The calculation involves a voice activity detection (VAD) procedure (Moattar and Homayounpour, 2009) to detect whether a given speech frame is voiced or unvoiced. The variance of the driving noise $v(n)$ can be then estimated as:

$$\begin{aligned} \sigma_v^2 &= \sigma_y^2 - \sigma_w^2 \\ &= \mathrm{E}[y^2(n)] - \mathbf{r}_y^T \, \mathbf{a}_y - \sigma_w^2 \end{aligned} \tag{7}$$

where $\mathbf{a}_\mathbf{y} = [a_{y,1}, \ldots, a_{y,p}]^T$ is the LPC vector of the noisy speech, and $\mathbf{r}_\mathbf{y} = \mathrm{E}[\mathbf{y}(n)y(n)]$ the autocorrelation vector of the noisy speech $y(n)$ with its past p samples, represented by the vector $\mathbf{y}(n) = [y(n-1), \ldots, y(n-p)]^T$.

Although the performance of the conventional Kalman filter method for speech enhancement has been improved notably by using the FNN for parameter estimation, several limitations have been identified. Firstly, the additional VAD procedure needed for the estimation of the additive noise variance increases the computational and structural complexity of the system. In addition, accurately detecting the unvoiced frames remains a difficult task, and the detection errors lead to inaccurate variance estimation of the additive noise, which brings further distortion to the enhanced speech.

3. Proposed system

To counter the difficulties posed by the VAD procedure and improve the accuracy of the variance estimation, we propose a hybrid speech enhancement system that combines DNN-based parameter estimation with colored-noise Kalman filter. The overall block diagram of our new system is depicted in Fig. 1, which is composed of two stages, namely: the training stage and the enhancement stage. In the training stage, the input feature set to the DNN consists of the combination of the noisy speech LSFs along with four acoustic features from (Wang et al., 2013). The output targets are the LSFs of both the clean speech and the noise. Then, a multi-objective DNN is trained to learn the mapping from the noisy input feature set to the targets. In the enhancement stage, given a noisy speech signal, we obtain first the input feature set, and then process it by the trained DNN to predict the clean speech LSFs and noise LSFs. The estimated LPCs are then obtained from the LSFs, and applied to both variance estimation and Kalman filtering. Subsequently, the noisy speech is enhanced by the colored-noise Kalman filter. This operation is followed by a post subtraction to further remove the residual noise in the filtered speech. The key components and steps involved in the proposed system are described in further details below.

3.1. Colored-noise Kalman filter

As mentioned before, in a conventional Kalman filter the clean speech is modelled as an AR process, while the additive noise is assumed to be white, which is not suitable for the complex noises encountered in real-world environment. To overcome this limitation, we herein adopt the colored-noise Kalman filter. In this method, the additive noise $w(n)$ in (1) is now modelled as an AR process, expressed as,

$$w(n) = \sum_{i=1}^{q} a_{w,i} w(n-i) + z(n) \tag{8}$$

where $a_{w,i}$ are the LPCs of the colored noise, q the order of the AR model, and $z(n)$ the zero-mean white driving noise with variance σ_z^2.

The underlying AR signal model in the colored-noise Kalman filter can be conveniently incorporated into the following state-space matrix form,

$$\begin{aligned} \mathbf{x}(n) &= \mathbf{F}\mathbf{x}(n-1) + \mathbf{G}\mathbf{u}(n) \\ y(n) &= \mathbf{H}^T \mathbf{x}(n) \end{aligned} \tag{9}$$

where $\mathbf{x}(n) = [\mathbf{s}(\mathbf{n}), \mathbf{w}(\mathbf{n})]^T$ is the $p+q$ dimensional concatenated state vector constituted by the clean speech vector $\mathbf{s}(n) = [s(n-p+1), \ldots, s(n-1), s(n)]$ together with the noise vector $\mathbf{w}(n) = [w(n-q+1), \ldots, w(n-1), w(n)]$, and $\mathbf{u}(n) = [v(n), z(n)]^T$ is the concatenated driving noise vector. Moreover, the augmented matrices $\mathbf{G}$, $\mathbf{H}$, and the overall transition matrix $\mathbf{F}$ are given as follows:

$$\mathbf{F} = \begin{bmatrix} \mathbf{F}_s & \mathbf{0} \\ \mathbf{0} & \mathbf{F}_w \end{bmatrix}, \quad \mathbf{G} = \begin{bmatrix} \mathbf{G}_s & \mathbf{0} \\ \mathbf{0} & \mathbf{G}_w \end{bmatrix}, \quad \mathbf{H} = \begin{bmatrix} \mathbf{H}_s \\ \mathbf{H}_w \end{bmatrix} \tag{10}$$

with

$$\mathbf{F}_w = \begin{bmatrix} 0 & 1 & \cdots & 0 & 0 \\ 0 & 0 & \cdots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \cdots & 0 & 1 \\ a_{w,q} & a_{w,q-1} & \cdots & a_{w,2} & a_{w,1} \end{bmatrix} \tag{11}$$

and $\mathbf{H}_w = \mathbf{G}_w = [0, \ldots, 0, 1]^T \in \mathbb{R}^q$.

Given a noisy observation $y(n)$, the estimate of the state vector $\mathbf{x}(n)$ can be obtained by the following Kalman filtering recursive equations:

$$\begin{cases} e(n) = y(n) - \mathbf{H}^T \widehat{\mathbf{x}}(n|n-1) \\ \mathbf{K}(n) = \mathbf{P}(n|n-1)\mathbf{H}\left(\mathbf{H}^T \mathbf{P}(n|n-1)\mathbf{H}\right)^{-1} \\ \widehat{\mathbf{x}}(n|n) = \widehat{\mathbf{x}}(n|n-1) + \mathbf{K}(n)e(n) \\ \mathbf{P}(n|n) = \left(\mathbf{I} - \mathbf{K}(n)\mathbf{H}^T\right)\mathbf{P}(n|n-1) \\ \widehat{\mathbf{x}}(n+1|n) = \mathbf{F}\widehat{\mathbf{x}}(n|n) \\ \mathbf{P}(n+1|n) = \mathbf{F}\mathbf{P}(n|n)\mathbf{F}^T + \mathbf{G}\mathbf{Q}_u\mathbf{G}^T \end{cases} \tag{12}$$

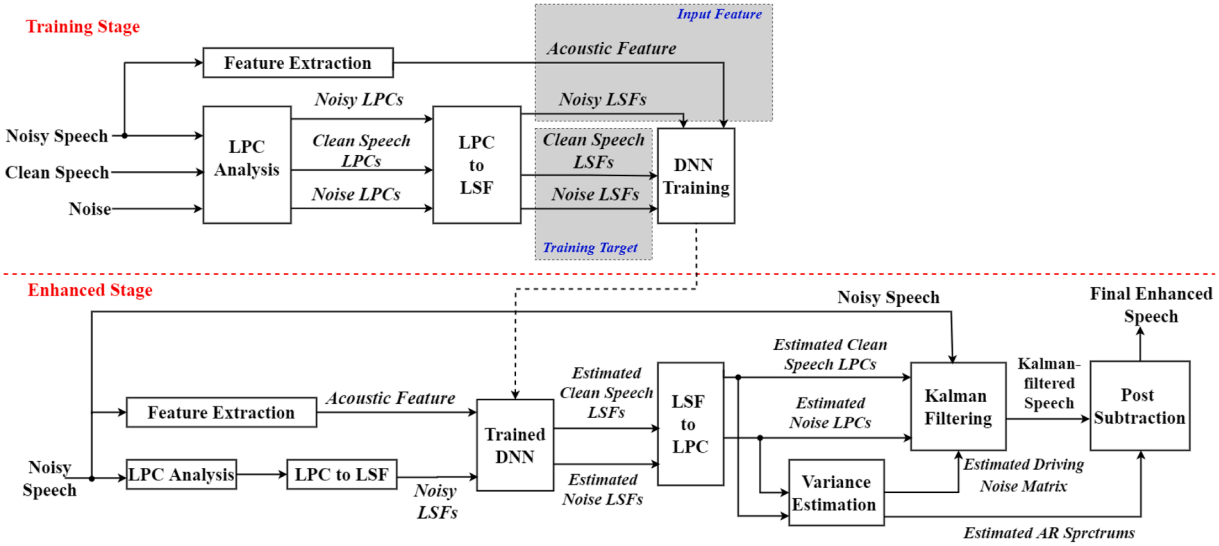


Fig. 1. Block diagram of proposed speech enhancement system using DNN-augmented colored noise Kalman filter.

where $e(n)$ is the innovation, $\mathbf{K}(n)$ the Kalman gain matrix, $\hat{\mathbf{x}}(n|n)$ the filtered estimate of state vector $\mathbf{x}(n)$, $\hat{\mathbf{x}}(n|n-1)$ *a priori* estimate of the state vector $\mathbf{x}(n)$. $\mathbf{P}(n|n)$ is the filtered state error covariance matrix, and $\mathbf{P}(n|n-1)$ the predicted state error correlation matrix. $\mathbf{Q}_u$ is the covariance matrix of the driving noise vector $\mathbf{u}(\mathbf{n})$, which is given by

$$\mathbf{Q}_u = E\left[\mathbf{u}(n)\mathbf{u}(n)^T\right] = \begin{bmatrix} \sigma_v^2 & 0 \\ 0 & \sigma_z^2 \end{bmatrix}. \tag{13}$$

The denoised speech is the output of the colored-noise Kalman filter, i.e.,

$$\hat{s}(n) = \left[\mathbf{G}_s^T, \mathbf{0}^T\right]\hat{\mathbf{x}}(n|n) \tag{14}$$

Note that two parameterized matrices that appear in the process Eq. (12) should be estimated from the noisy speech to carry out Kalman filtering, namely, the overall transition matrix $\mathbf{F}$ and the covariance matrix $\mathbf{Q}_u$ of the concatenated driving noise vector $\mathbf{u}(n)$. The first depends on the clean speech and noise LPCs, which can be converted to the LSFs and predicted through a DNN, while the second one is obtained by solving an optimization problem. The details of this parameter estimation are provided in the following subsections.

3.2. DNN-based LSFs estimation

Recently, we have demonstrated that FNN offers a convenient means for LSFs estimation in speech processing applications (Yu et al., 2019). Here, we propose to employ two different networks, i.e., FNN and LSTM, to predict both the clean speech LSFs and noise LSFs. The specific configuration of each network is described in Section 4.2.

For the input features, we extract 12-dimensional LSFs along with several complementary features from the noisy speech, in order to collect more information about the speech characteristics. Specifically, the following additional acoustic features are utilized: the 15-dimensional amplitude modulation spectrum (AMS); the 31-dimensional relative spectral transform and perceptual linear prediction (RASTA-PLP); the 13-dimensional Mel-frequency cepstral coefficients (MFCC) and their deltas; and the 64-dimensional Gammatone filterbank energies (GF), and their deltas (Wang et al., 2013). The total dimension of the input feature set is 258, i.e., (12+2 × (15+31+13+64))

The input features are computed for each frame of the noisy speech, and represented as a row vector $\mathbf{f}(m)$ with m denoting the frame index. To make full use of the temporal information of the speech, it is common to incorporate the features of adjacent frames into a single extended feature vector. Hence, the extended feature vector centered at the mth frame is constructed as $\tilde{\mathbf{f}}(m) = [\mathbf{f}(m-m_0), \ldots, \mathbf{f}(m), \ldots, \mathbf{f}(m+m_0)]$, where m_0 is the number of adjacent frames to be included on each side. The value of m_0 is set to 2 in our experiment. Note that all the different features are normalized to the range [0,1) in order to balance the training errors.

For the training targets, we adopt a multi-objective learning architecture to estimate both the clean speech LSFs and noise LSFs. Compared to a standard DNN, the output layer in the proposed architecture is divided into two parts: one for the clean speech LSFs and the other for the noise LSFs. The advantages of multi-objective learning are twofold. On one hand, it has lower computational complexity compared to training two separate DNNs (i.e., one for clean speech and one for noise). On the other hand, estimating the two sets of LSFs simultaneously can help better exploit the relationship between the clean speech and noise.

In the training stage, back propagation is used to adjust the weights and biases so as to minimize the cost function, which is defined as the mean square error (MSE) between the reference LSFs and the estimated ones for each training utterance. Note that the cost function is composed of two parts: one for the clean speech LSFs and the other for the noise LSFs, as given by,

$$\begin{aligned} MSE_{LSF} = & \frac{1}{M}\sum_{m=1}^{M}\left\{\frac{1}{p}\sum_{i=1}^{p}\left[\hat{L}_{s,i}(m) - L_{s,i}(m)\right]^2\right. \\ & \left. + \frac{1}{q}\sum_{j=1}^{q}\left[\hat{L}_{w,j}(m) - L_{w,j}(m)\right]^2\right\} \end{aligned} \tag{15}$$

where m is the frame index of the input noisy speech and M the total number of the frames. The quantities $L_{s,i}(m)$ and $\hat{L}_{s,i}(m)$ are the reference clean speech LSFs and the estimated ones at frame m, where $i \in \{1, \ldots, p\}$ is the order index of the clean speech AR model. Similarly, $L_{w,j}$ are the reference noise LSFs and $\hat{L}_{w,j}$ the estimated ones at frame m, where $i \in \{1, \ldots, q\}$ is the order index of the noise AR model.

In the enhancement stage, the clean speech LSFs and noise LSFs are first obtained by the well-trained DNN, and then converted to their respective LPCs. The estimated LPCs are used along with the estimated variances in the Kalman filter Eq. (12) in order to estimate the desired speech signal.

3.3. Variance estimation

The covariance matrix $\mathbf{Q}_u$ in (13) is another key parameter that needs to be estimated prior to the application of the Kalman filtering

equations. Proceeding as in Srinivasan et al. (2005), we now formulate an optimization problem to estimate σ_v^2 and σ_z^2. Our goal is to minimize the difference between the noisy spectrum and the sum of the estimated clean speech spectrum and noise spectrum.

From Eqs. (1), (2) and (8), the spectrum of the AR-modelled noisy speech can be expressed as:

$$\begin{aligned}\widehat{P}_y(k) &= \widehat{P}_s(k) + \widehat{P}_w(k) \\ &= \frac{\sigma_v^2}{|A_s(k)|^2} + \frac{\sigma_z^2}{|A_w(k)|^2}\end{aligned} \tag{16}$$

with

$$\begin{aligned}A_s(k) &= 1 - \sum_{i=1}^{p} a_{s,i} e^{-j2\pi ik/K} \\ A_w(k) &= 1 - \sum_{i=1}^{q} a_{w,i} e^{-j2\pi ik/K}\end{aligned} \tag{17}$$

where K is the frame length. Note that the clean speech LPCs $a_{s,i}$ and the noise LPCs $a_{w,i}$ can be obtained from the LSFs at the output of the trained DNN.

The AR spectrum of the observed noisy speech $P_y(k)$ can be written as,

$$P_y(k) = \frac{\sigma_y^2}{|A_y(k)|^2} \tag{18}$$

with

$$A_y(k) = 1 - \sum_{i=1}^{p} a_{y,i} e^{-j2\pi ik/K} \tag{19}$$

$$\sigma_y^2 = E\left[y(n)^2\right] - \mathbf{r}_y^T \, \mathbf{a}_y. \tag{20}$$

We can obtain the variance estimates by minimizing the difference between the AR spectrum of the modelled noisy speech $\widehat{P}_y(k)$ and that of the observed one $P_y(k)$, that is,

$$\sigma_v^{*2}, \sigma_z^{*2} = \underset{\sigma_v^2, \sigma_z^2}{\operatorname{argmin}} d\left(\widehat{P}_y(k), P_y(k)\right) \tag{21}$$

where the difference is measured in the log-spectral domain as given by,

$$\begin{aligned} d\left(\widehat{P}_y(k), P_y(k)\right) &= \frac{1}{K}\sum_{k=1}^{K}\left|\ln\widehat{P}_y(k) - \ln P_y(k)\right|^2 \\ &\approx \frac{1}{K}\sum_{k=1}^{K}\left|\frac{\sigma_v^2 \big/ |A_s(k)|^2 + \sigma_z^2 \big/ |A_w(k)|^2 - P_y(k)}{P_y(k)}\right|^2. \end{aligned} \tag{22}$$

To obtain the approximate equation in (22), we have used Eq. (16) and the approximation of $\ln(x+1) \approx x$. Then by applying partial differentiation to the difference $d(\widehat{P}_y(k), P_y(k))$ with respect to σ_v^2 and σ_z^2, we obtain the following linear system of equations:

$$\begin{bmatrix} E_{ss} & E_{sw} \\ E_{sw} & E_{ww} \end{bmatrix} \begin{bmatrix} \sigma_v^2 \\ \sigma_z^2 \end{bmatrix} = \begin{bmatrix} E_{ys} \\ E_{yw} \end{bmatrix} \tag{23}$$

with

$$\begin{aligned} E_{ss} &= \| \frac{1}{P_y^2(k)|A_s(k)|^4} \|, E_{ww} = \| \frac{1}{P_y^2(k)|A_w(k)|^4} \| \\ E_{sw} &= \| \frac{1}{P_y^2(k)|A_s(k)|^2|A_w(k)|^2} \| \\ E_{ys} &= \| \frac{1}{P_y(k)|A_s(k)|^2} \|, E_{yw} = \| \frac{1}{P_y(k)|A_w(k)|^2} \| \end{aligned} \tag{24}$$

The norms involved in Eq. (24) are defined as $\| f(k) \| \triangleq \sum_{k=1}^{K} |f(k)|$.

When the AR spectrum of the observed noisy speech $P_y(k)$ is calculated and $A_s(k)$ and $A_w(k)$ are obtained with the estimated LPCs from the trained DNN, we can finally obtain the optimal variances σ_v^2 and σ_z^2 using Eq. (23).

3.4. Post subtraction

To further remove the residual noise in the Kalman-filtered speech, a post subtraction algorithm is applied right after Kalman filtering. We adopt multiband spectral subtraction because of its good performance in reducing speech distortion (Kamath and Loizou, 2002). The main idea of this method is described as follows.

The fast Fourier transform (FFT) is first applied to the windowed Kalman-filtered speech to obtain the magnitude spectrum. Next, the noise spectrum is estimated and updated during the unvoiced frames. The detection of unvoiced frames is accomplished by comparing the total power of the estimated clean speech, say $\widehat{P}_s^2$ and that of the estimated noise, $\widehat{P}_w^2$, which can easily be obtained from the estimated spectra in Section 3.3. Specifically, a frame is labelled as a voiced frame if $\widehat{P}_s^2 > \widehat{P}_w^2$, and as an unvoiced frame otherwise.

Then, the magnitude spectra of the filtered speech and noise are divided into L subbands. In each subband, the Kalman-filtered magnitude spectrum is enhanced by subtracting a noise power spectrum term,

$$|\widehat{C}_l(k)|^2 = |\widehat{S}_l(k)|^2 - \alpha\,\delta_l \left|\widehat{D}_l(k)\right|^2 \tag{25}$$

where $|\widehat{C}_l(k)|^2$ denotes the modified subband speech power spectrum, $|\widehat{S}_l(k)|^2$ the Kalman-filtered speech power spectrum and $|\widehat{D}_l(k)|^2$ the estimated noise power spectrum (obtained and updated during unvoiced frames), with k being the discrete frequency and l the subband index. Moreover, α is the oversubtraction factor and δ_l the additional subtraction factor that can be individually set for each subband to customize the noise removal process.

The factors α and δ_l are used to control the noise subtraction level within each band. The value of α is defined as a function of the segmental signal-to-noise ratio (SNR) (in dB), i.e.,

$$\alpha = \begin{cases} 4.75 & , \quad \text{SNR} < -5 \\ 4 - \frac{3}{20}\text{SNR} & , \quad -5 \le \text{SNR} \le 20 \\ 1 & , \quad \text{SNR} > 20 \end{cases} \tag{26}$$

with

$$\text{SNR} = 10\log_{10}\left(\frac{|\widehat{S}_l(k)|^2}{|\widehat{D}_l(k)|^2}\right) \tag{27}$$

The value of δ_l is determined as,

$$\delta_l = \begin{cases} 1 & , \quad f_l < 1\text{kHz} \\ 2.5 & , \quad 1\text{kHz} \le f_l \le \frac{F_s}{2} - 2\text{kHz} \\ 1.5 & , \quad f_l > \frac{F_s}{2} - 2\text{kHz} \end{cases} \tag{28}$$

where f_l is the upper frequency of the lth band and F_s the sampling frequency. The above values of the factors α and β are taken from Kamath and Loizou (2002) where they have been determined empirically based large experiments.

Finally, we synthesize the modified subband spectrum from the modified magnitude (25) and the phase of the Kalman-filtered speech. The final enhanced speech is obtained by computing the inverse FFT of

the modified subband spectrums.

4. Experimental results

4.1. Experimental setup

Databases: The clean speech is selected from the IEEE sentence database (IEEE Subcommittee, 1969), where we choose 670 utterances for training and 50 utterances for enhancement. The noise is from the NOISEX-92 database (Varga and Steeneken, 1993), where four types of noises (babble, white, street, factory) are employed as seen noise, and another four (pink, buccaneer2, destroyerengine, hfchannel) as unseen noise. In the training stage, the noisy speech is obtained by mixing the clean speech with seen noise at four levels SNRs, i.e., -3dB, 0dB, 3dB and 6dB, which results in 10,720 utterances. In the enhancement stage, both seen noise and unseen noise are mixed with the clean speech at the above mentioned SNR levels. The number of noisy utterances used in the enhancement stage is 800 for both seen noise and unseen noise. The sampling frequency is set to 16 kHz for both clean speech and noise.

Objective metrics: To evaluate the enhancement performance, two objective metrics are selected: the perceptual evaluation of speech quality (PESQ) measure (ITU-R, 2001) and the short-time objective intelligibility (STOI) measure (Taal et al., 2011). PESQ and STOI evaluate the processed speech from two different perspectives, i.e., speech quality and intelligibility, and are widely adopted in speech-related applications. PESQ measures the perceptual distortion by comparing the original and processed signals. A score given by PESQ evaluation ranges from -0.5 to 4.5. Although PESQ is an objective metric for evaluating the speech quality, it also reflects faithfully the subjective score of the processed speech. STOI has been put forward in recent years for objective assessment of the speech intelligibility. The score of STOI ranges from 0 to 1, and shows a good correlation with the subjective score in listening test for the speech intelligibility. For both metrics, a higher score means a better speech quality or intelligibility.

4.2. Reference methods

To evaluate the proposed speech enhancement system, we adopt several existing methods for performance comparison. Our reference methods include both Kalman filter based algorithms and DNN-based approaches. The following provides a brief conceptual summary of each one of the reference methods.

- IKF (Iterative Kalman filtering) (Gannot et al., 1998): This algorithm iteratively performs conventional Kalman filtering, in which the LPCs are updated in each iteration
- P-IKF (Perceptual IKF) (Ma et al., 2005): This algorithm calculates a perceptual mask according to human hearing system and applies it to the Kalman-filtered speech in order to further remove the residual noises.
- S-IKF (Subband IKF) (Roy et al., 2016): In this method, the noisy speech is first divided into subband signals. Iterative Kalman filtering is then applied separately for each subband noisy speech. The final enhanced speech is obtained by synthesising the subband enhanced speech signals.
- FNN-MAG (Xu et al., 2013): A FNN is employed to directly explore the mapping from the noisy speech magnitude spectrum to the clean one. The enhanced speech is synthesised with the estimated clean magnitude and noisy phase.
- FNN-WF (Li and Kang, 2016): A FNN is trained for the estimation of AR parameters of the clean speech. Then, a Wiener filter is estimated by calculating the ratio of the estimated clean speech power spectrum to that of the noisy speech. The enhanced speech is then obtained by applying the estimated Wiener filter to the noisy speech.
- FNN-KF (Yu et al., 2019): A FNN is used to predict the LPCs needed for conventional Kalman filtering. The DNN learns the mapping from the acoustic features of the noisy speech to the LSFs of the clean speech. The estimated LSFs are then converted to the desired LPCs.

Besides these benchmarks, we consider three versions of our proposed DNN-augmented colored-noise Kalman filter method, namely,

- FNN-CKF: FNN for LSFs estimation and without post subtraction.
- FNN-CKFS: FNN for LSFs estimation and with post subtraction.
- LSTM-CKFS: LSTM for LSFs estimation and with post subtraction.

In order to make fair comparisons, we use the same configuration for the FNN in the related methods, i.e., one input layer, one output layer and three hidden layers with 1024 units in each layer. The LSTM network is obtained by stacking by one input layer, two LSTM layers with 512 units in each layer, one feed-forward layer with 512 units and one output layer.

For FNN-MAG, a Hamming window is selected to divide each utterance into 20 ms time frames with an 10 ms frame shift (50% overlap). A 320-point DFT is then computed for each frame. For the other reference methods and the proposed system, a rectangular window is used to divide the audio signals into 20 ms frames with no overlap.

For the conventional Kalman filter, we set $\mathbf{s}(0|0) = \mathbf{0}$, $\mathbf{P}(0|0) = \mathbf{I}$, and the AR model order of the clean speech as $p = 12$. For the colored-noise Kalman filter, we set $\mathbf{x}(0|0) = \mathbf{0}$, $\mathbf{P}(0|0) = \mathbf{I}$, and the orders of AR models for clean speech and additive noise as $p = q = 12$. For the post subtraction in FNN-CKFS and LSTM-CKFS, the spectrum is evenly divided into 4 bands.

4.3. Evaluation of input feature sets

In the training stage, we use the following feature sets as the input of our proposed system: LPS-only set, LSF-only set, multi-feature set consisting of AMS+RASTAPLP+MFCC+GF, and joint set formed by combining the LSF-only set with the multi-feature set. In this experiment, we investigate the performance of the proposed system with these different feature sets when using FNN for LSFs estimation. The objective results of the enhanced speech are shown in Table 1.

The final enhanced speech for the LPS-only and LSF-only feature sets exhibit similar PESQ and STOI scores, while the objective scores could be improved notably for the multi-feature and joint sets, using more indicates that more acoustic features provides useful additional information about the speech. Finally, the enhanced speech from the joint set achieves the highest PESQ and STOI scores. As a result, the joint set is considered as the optimal input feature set for the proposed system.

4.4. Evaluation of LPCs estimation accuracy

In this subsection, the LPCs estimation error is evaluated to verify the learning capability of the proposed multi-objective DNN training. We first define the LPCs estimation error of the speech as the mean square error (MSE) between the estimated LPCs and the ideal LPCs calculated from the clean speech for each utterance as given below,

Table 1
Objective results with different feature sets.

		-3 dB	0 dB	3 dB	6 dB
PESQ	Noisy	1.41	1.52	1.68	1.86
	LPS-only	1.67	1.90	2.10	2.29
	LSF-only	1.69	1.92	2.12	2.33
	Multi Set	1.80	2.06	2.27	2.46
	Joint Set	**1.88**	**2.12**	**2.32**	**2.51**
STOI	Noisy	0.66	0.72	0.78	0.83
	LPS-only	0.69	0.75	0.80	0.83
	LSF-only	0.68	0.74	0.80	0.84
	Multi Set	0.72	0.78	0.83	0.86
	Joint Set	**0.73**	**0.79**	**0.85**	**0.88**

$$MSE_{LPC} = \frac{1}{M}\sum_{m=1}^{M}\left\{\frac{1}{p}\sum_{i=1}^{p}\left[\hat{a}_{s,i}(m) - a_{s,i}(m)\right]^2\right\} \quad (29)$$

where M denotes the number of the speech frames in the utterance, $a_{s,i}(m)$ the ideal LPCs of the clean speech and $\hat{a}_{s,i}(m)$ the estimated ones. The estimated LPCs are obtained by three methods for comparison. The first one applies the Levinson-Durbin (LD) algorithm to obtain the LPCs of the noisy speech directly (Shimamura et al., 1998). The second and third ones adopt the proposed DNN based LSFs estimation algorithm, where FNN and LSTM are used to estimate the LSFs, which are then converted to LPCs. Similarly, we compute the LPCs estimation error of the additive noise for each noise type by using (29), where the estimated and ideal LPCs of the speech are replaced by those of the additive noise, and the order of the speech model, p is replaced by that of the noise, q.

Fig. 2 shows the LPCs estimation error comparison for the speech. The average MSE is computed over all the testing utterances for both seen and unseen noise. In general, the FNN and LSTM based approaches give a slightly smaller error than the LD method does for the eight types of noises and different SNRs. In addition, the error from LSTM is smaller than that from FNN in most cases. Another important finding is that the error from the DNN methods decrease with an increase of the SNR, which means that DNN achieves a better performance at higher SNR. The LPCs estimation performance also varies for different noise types. In particular, the best estimation accuracy is achieved for street noise, and the worst for white noise. Interestingly, we note that the estimation error of FNN and LSTM based algorithm under unseen noise does not increase considerably compared with that under seen noise, which indicates that using DNN in LPCs estimation offers robustness and has a good generalization capability.

The LPCs estimation error comparison for the additive noise is shown in Fig. 3, where we notice important differences with the case of clean speech. Firstly, as SNR increases, the strong speech component more strongly affect the noise LPCs estimation, and hence the LPCs estimation error of additive noise gets larger. Secondly, compared with speech, noise exhibits less structure and correlation, and the mapping from the noisy speech feature to the noise LSFs is thus more difficult to learn. Therefore, we can find that the DNN estimation result is not always better than the traditional LD method, especially at low SNR.

4.5. Speech enhancement performance under seen noise

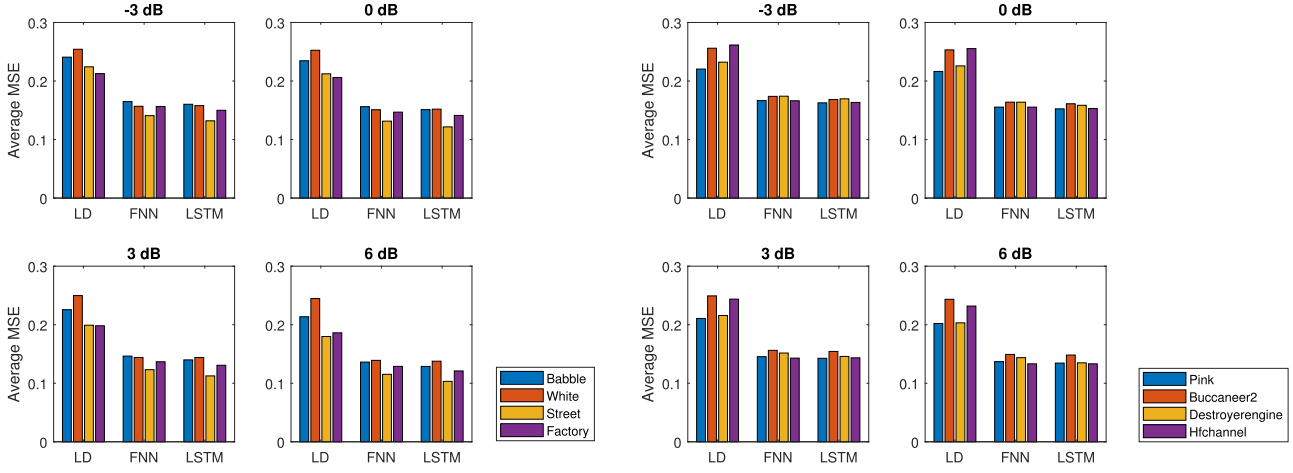
Here, we campare the different speech enhancement methods under seen noise. Table 2 gives the average objective scores of different speech enhancement methods on seen noise. We first note that the performances of the unsupervised Kalman filtering algorithms are worse than those of the DNN-based methods. The P-IKF, which incorporates a perceptual mask to further suppress the residual noise, is the best among the three unsupervised Kalman filtering algorithms. However, P-IKF still can not achieve as good performance as FNN-KF, not to mention our FNN-CKF, FNN-CKFS and LSTM-CKFS. These results demonstrate the benefit from employing DNN in parameter estimation. The DNN can predict more accurate LPCs from the noisy speech, thus improving the performance of Kalman filtering algorithms.

Moreover, FNN-KF has lower objective scores compared with the proposed methods. This is because FNN-KF requires a VAD procedure to detect the unvoiced frame for estimating and updating the additive noise variance σ_w^2. However, VAD in noisy condition is a difficult task, which causes variance estimation error and introduces extra distortion to the enhanced speech. In our proposed system, an AR model is adopted to represent the background noise. As such, the Kalman filtering equations in (12) no longer involve σ_w^2, and we can therefore overcome the speech distortion problem due to the inaccurate estimation of σ_w^2. The performance can be further improved by employing post subtraction to remove the residual noise due to the inaccurate parameters of the noise AR model. Indeed, FNN-CKFS achieves a better performance than FNN-CKF, which approaches closely that of FNN-MAG. Finally, although FNN-MAG has the best performance among all tested FNN based approaches, by employing LSTM for LSFs estimation in our proposed system, LSTM-CKFS can achieve the best PESQ scores, which demonstrates the LSTM's advantage in modelling long temporal dependencies.

4.6. Speech enhancement performance under unseen noise

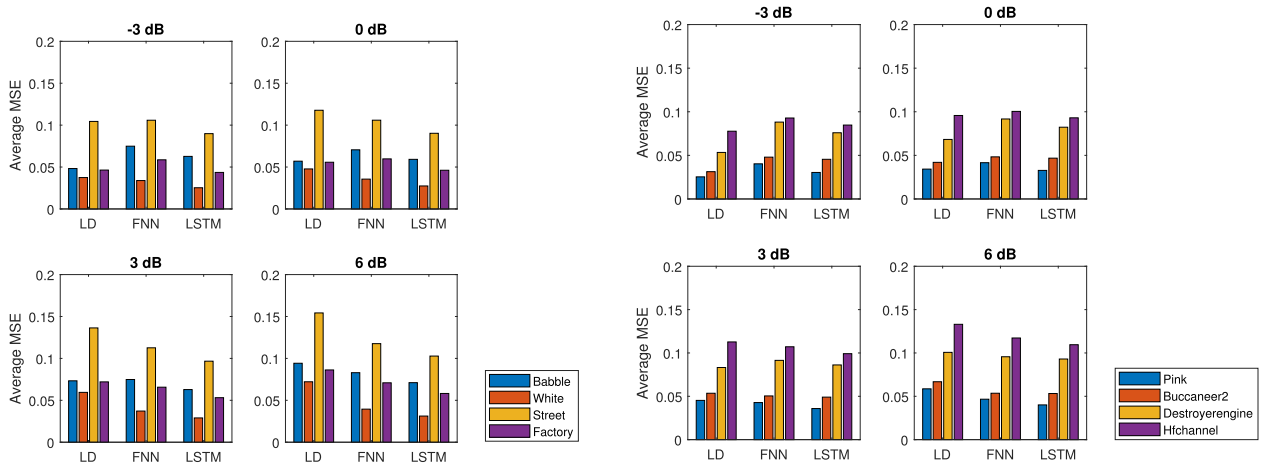
Table 3 gives the average objective scores of the different speech enhancement methods in the case of unseen noise. In this case, the performances of the unsupervised Kalman filtering algorithms are still worse than those of the FNN-based methods. Comparing FNN-KF with our proposed system, we can find again that FNN-CKF, FNN-CKFS and LSTM-CKFS outperform FNN-KF because of the adoption of colored-noise Kalman filter. However, the STOI scores of FNN-CKF are slightly lower than those of FNN-KF at low SNR. This degradation is possibly caused by the inaccuracy in estimating the noise LPCs, as shown in Fig. 3 where the FNN estimation error is higher than the LD estimation error under low input SNR conditions.

In the case of unseen noise, we find that LSTM-CKFS achieves the best objective scores due to its advanced network structure. More interestingly, FNN-MAG no longer holds the best performance among FNN based methods. In fact, the objective scores of FNN-MAG decease



(a) Seen noise (b) Unseen noise

Fig. 2. LPC estimation error comparison for speech among Levinson-Durbin (LD), FNN and LSTM methods.



(a) Seen noise (b) Unseen noise

Fig. 3. LPC estimation error comparison for additive noise among LD, FNN and LSTM methods.

Table 2
Objective scores of different speech enhancement methods on seen noise.

Method	PESQ				STOI			
	-3 dB	0 dB	3 dB	6 dB	-3 dB	0 dB	3 dB	6 dB
Noisy	1.41	1.52	1.68	1.86	0.66	0.72	0.78	0.83
IKF	1.55	1.79	2.01	2.25	0.67	0.74	0.80	0.85
P-IKF	1.57	1.83	2.08	2.31	0.68	0.75	0.81	0.85
S-IKF	1.56	1.81	2.04	2.29	0.67	0.75	0.81	0.84
FNN-MAG	1.89	2.13	2.34	2.55	**0.75**	**0.82**	**0.86**	**0.88**
FNN-WF	1.65	1.83	2.15	2.36	0.71	0.78	0.82	0.86
FNN-KF	1.70	1.93	2.13	2.30	0.71	0.77	0.81	0.85
FNN-CKF	1.73	2.01	2.26	2.49	0.72	0.78	0.84	0.87
FNN-CKFS	1.88	2.12	2.32	2.51	0.73	0.79	0.85	**0.88**
LSTM-CKFS	**1.93**	**2.16**	**2.38**	**2.58**	0.74	0.80	0.85	**0.88**

Table 3
Objective scores of different speech enhancement methods on unseen noise.

Method	PESQ				STOI			
	-3 dB	0 dB	3 dB	6 dB	-3 dB	0 dB	3 dB	6 dB
Noisy	1.37	1.51	1.65	1.82	0.65	0.72	0.78	0.83
IKF	1.64	1.84	2.04	2.26	0.68	0.75	0.81	0.85
P-IKF	1.67	1.88	2.09	2.32	0.69	0.76	0.81	0.85
S-IKF	1.66	1.87	2.08	2.31	0.68	0.75	0.81	0.84
FNN-MAG	1.73	1.92	2.13	2.32	0.70	0.76	0.82	0.87
FNN-WF	1.68	1.92	2.15	2.33	0.67	0.74	0.81	0.85
FNN-KF	1.73	1.95	2.21	2.38	0.71	0.77	0.82	0.85
FNN-CKF	1.76	2.02	2.26	2.48	0.70	0.76	0.82	0.86
FNN-CKFS	1.89	2.11	2.32	2.50	0.71	0.78	0.83	0.87
LSTM-CKFS	**1.91**	**2.15**	**2.36**	**2.55**	**0.73**	**0.79**	**0.84**	**0.88**

largely, indicating that mapping the noisy magnitude spectrum to the clean one is prone to errors when the noise is unmatched with those in the training stage. In contrast, FNN-WF, FNN-KF and our proposed system suffer less performance degradation. Indeed, the denoising process in these methods is accomplished by Wiener and Kalman filtering. Therefore, as the DNN can provide more accurate parameters, their performances would not fluctuate as much whether on seen noise or unseen noise. Based on these results, and considering the robustness of the DNN-based LPCs estimation, we can conclude that our FNN-CKF, FNN-CKFS and LSTM-CKFS, have a better generalization capability than FNN-MAG.

Finally, we make comparison in terms of each objective metric. Although the enhanced speech from LSTM-CKFS has the best speech quality according to the PESQ scores, the improvement of speech intelligibility is not obvious as seen from the STOI scores. In fact, the LSTM-CKFS gives similar STOI scores to FNN-MAG. Actually, there is a trade-off between residual noise and speech distortion for speech enhancement algorithms, leading to decreased speech intelligibility. For our LSTM-CKFS, the enhanced speech achieves similar speech intelligibility as that of FNN-MAG but far better speech quality, indicating that LSTM-CKFS could preserve the information content of clean speech well, while significantly removing the additive noise.

4.7. Spectrograms of enhanced speeches

To better understand the characteristics of the enhanced speech, Fig. 4 shows the spectrograms of the enhanced speech signals from several selected methods, demonstrating the effects of the residual noises and the distortions in the harmonic structures in the time-frequency domain. The noisy speech is obtained by mixing a selected clean speech utterance with buccaneer noise at 3dB SNR. For the best unsupervised Kalman filtering in our experiment, i.e., P-IKF, we can find the musical noise structure in the spectrogram in the region between 4kHz and 8kHz. The spectrogram of FNN-MAG also exhibits some musical noise structures in the high-frequency component as well as residual noise in the low-frequency component. For FNN-WF, the high-frequency components look better than the previous two spectrograms, but still have undesired structures, which are likely caused by the difficulty of Wiener filter in removing non-stationary noise. Finally, for the four Kalman filtering related methods, it is observed that FNN-KF, FNN-CKFS and LSTM-CKFS can remove the background noise quite well. However, the high-frequency components of FNN-KF still suffer from various degradations. While this situation is improved in the cases of FNN-CKFS and LSTM-CKFS, the LSTM-CKFS can preserve the harmonic structures best among all the tested methods, thus achieving the best objective scores.

5. Conclusion

In this paper, we have proposed a hybrid speech enhancement system with DNN-aided parameter estimation and colored-noise Kalman filtering. Our system first employs a multi-objective FNN or LSTM to estimate the AR model parameters of both clean speech and noise. Then a colored-noise Kalman filter with the estimated parameters is applied to the noisy speech for denoising. By doing so, the proposed system can more efficiently cope with color noises encountered in real-world environments. To further improve the enhancement performance, a post subtraction algorithm is adopted to better remove the residual noise.

Experiments have shown the superiority of the proposed system from

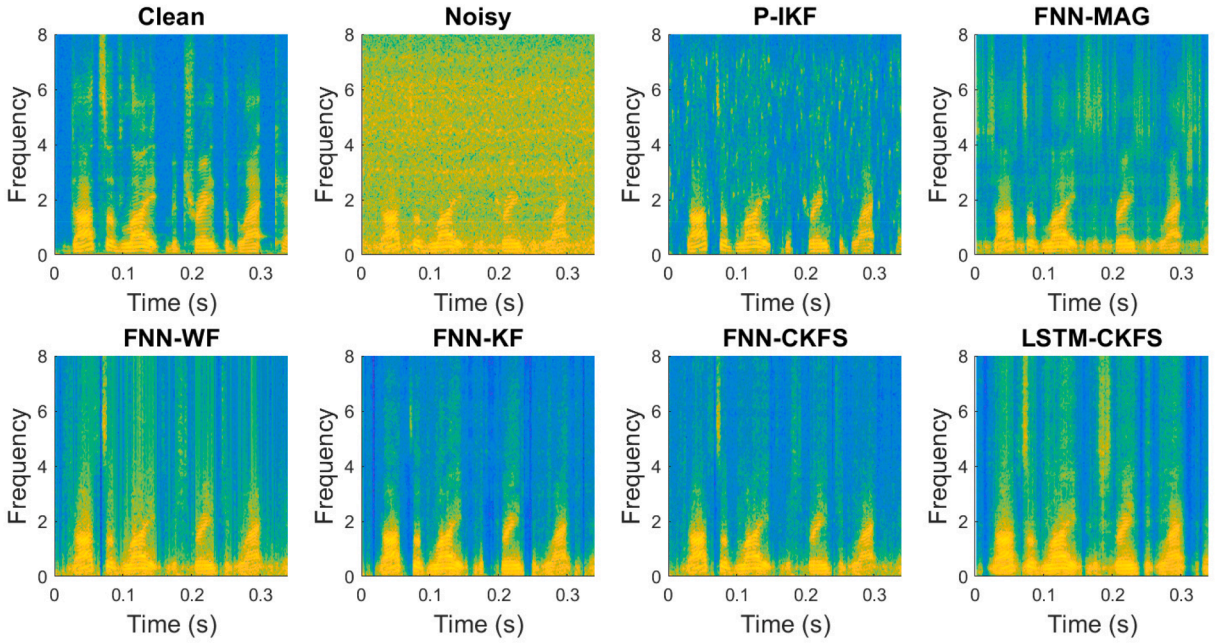


Fig. 4. Spectrograms of the clean, noisy and enhanced speech signals for different methods.

two aspects. First, the employment of DNN for parameter estimation and post subtraction for residual noise suppression largely improves the enhancement performance of colored-noise Kalman filtering. Second, our proposed system takes advantages of both unsupervised and supervised methods, and thus exhibits a better generalization capability. Indeed, while it achieves comparable performance as recent DNN-based approaches on seen noise, it offers notably better results on unseen noise.

CRediT authorship contribution statement

Hongjiang Yu: Conceptualization, Methodology, Software, Formal analysis, Writing - original draft. **Wei-Ping Zhu:** Validation, Formal analysis, Writing - review & editing, Supervision, Resources. **Benoit Champagne:** Validation, Formal analysis, Writing - review & editing.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

The authors acknowledge the support from China Scholarships Council (CSC No.201606270200) and NSERC of Canada under a CRD project sponsored by Microchip in Ottawa, Canada.

References

Berouti, M., Schwartz, R., Makhoul, J., April, 1979. Enhancement of speech corrupted by acoustic noise. IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP), pp. 208–211.
Boll, S., 1979. Suppression of acoustic noise in speech using spectral subtraction. IEEE Trans. Acoust. Speech Signal Process. 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.
Chen, J., Wang, D., 2017. Long short-term memory for speaker generalization in supervised speech separation. J. Acoust. Soc. Am. 141 (6), 4705–4714.
Erdogan, H., Hershey, J.R., Watanabe, S., Le Roux, J., April, 2015. Phase-sensitive and recognition-boosted speech separation using deep recurrent neural networks. IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP), pp. 708–712.
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.
Grancharov, V., Samuelsson, J., Kleijn, W.B., March, 2005. Improved Kalman filtering for speech enhancement. IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP), vol. 1, pp. 1109–1112.
Han, W., Zhang, X., Min, G., Sun, M., Yang, J., July, 2016. Joint optimization of audible noise suppression and deep neural networks for single-channel speech enhancement. IEEE Int. Conf. on Multimedia and Expo (ICME), pp. 1–6.
IEEE Subcommittee, 1969. IEEE Recommended practice for speech quality measurements. IEEE Trans. Audio Electroacoust. 17 (3), 225–246.
ITU-R, 2001. Perceptual evaluation of speech quality (PESQ) an objective method for end-to-end speech quality assessment of narrowband telephone networks and speech codecs. Recommendation P. 862.
Kamath, S., Loizou, P., May, 2002. A multi-band spectral subtraction method for enhancing speech corrupted by colored noise. IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP), pp. 4160–4164.
Kavalekalam, M.S., Christensen, M.G., Gran, F., Boldt, J.B., March, 2016. Kalman filter for speech enhancement in cocktail party scenarios using a codebook-based approach. IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP), pp. 191–195.
Kushner, W.M., Goncharoff, V., Wu, C., Nguyen, V., Damoulakis, J.N., May, 1989. The effects of subtractive-type speech enhancement/noise reduction algorithms on parameter estimation for improved recognition and coding in high noise environments. IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP), pp. 211–214.
Li, Y., Kang, S., 2016. Deep neural network-based linear predictive parameter estimations for speech enhancement. IET Signal Proc. 11 (4), 469–476.
Lim, J.S., Oppenheim, A.V., December, 1979. Enhancement and bandwidth compression of noisy speech. Proc. IEEE 67 (12), 1586–1604.
Ma, N., Bouchard, M., Goubran, R.A., 2005. Speech enhancement using a masking threshold constrained Kalman filter and its heuristic implementations. IEEE Trans. Audio Speech Lang.Process. 14 (1), 19–32.
Ma, N., Bouchard, M., Goubran, R.A., May, 2004. Perceptual Kalman filtering for speech enhancement in colored noise. IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP), vol. 1, pp. 717–720.
Mellahi, T., Hamdi, R., 2015. LPC-based formant enhancement method in Kalman filtering for speech enhancement. AEU-Int. J. Electron.Commun. 69 (2), 545–554.
Moattar, M.H., Homayounpour, M.M., August, 2009. A simple but efficient real-time voice activity detection algorithm. European Signal Processing Conference, pp. 2549–2553.
Narayanan, A., Wang, D., May, 2013. Ideal ratio mask estimation using deep neural networks for robust speech recognition. IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP), pp. 7092–7096.May
Nie, S., Liang, S., Liu, B., Zhang, Y., Liu, W., Tao, J., Sept., 2018. Deep noise tracking network: a hybrid signal processing/deep learning approach to speech enhancement. Proc. of Interspeech, pp. 3219–3223.
Nower, N., Liu, Y., Unoki, M., June, 2015. Restoration scheme of instantaneous amplitude and phase using Kalman filter with efficient linear prediction for speech enhancement. Speech Commun 70, 13–27.

Ouyang, Z., Yu, H., Zhu, W.-P., Champagne, B., Sept., 2018. A deep neural network based harmonic noise model for speech enhancement. Proc. of Interspeech, pp. 3224–3228.
Paliwal, K., Basu, A., April, 1987. A speech enhancement method based on Kalman filtering. IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP), vol. 12, pp. 177–180.
Popescu, D.C., Zeljkovic, I., May, 1998. Kalman filtering of colored noise for speech enhancement. IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP), vol. 2, pp. 997–1000.
Roy, S.K., Zhu, W.-P., Champagne, B., May, 2016. Single channel speech enhancement using subband iterative Kalman filter. IEEE International Symposium on Circuits and Systems (ISCAS), pp. 762–765.
Shimamura, T., Kunieda, N., Suzuki, J., May, 1998. A robust linear prediction method for noisy speech. IEEE Int. Symposium on Circuits and Systems (ISCAS), vol. 4, pp. 257–260.
Singh, L., Sridharan, S., Nov., 1998. Speech enhancement using critical band spectral subtraction. Int. Conf. on Spoken Language Processing, pp. 2827–2830.
Srinivasan, S., Samuelsson, J., Kleijn, W.B., 2005. Codebook driven short-term predictor parameter estimation for speech enhancement. IEEE Trans. Audio Speech Lang. Process. 14 (1), 163–176.
Taal, C.H., Hendriks, R.C., Heusdens, R., Jensen, J., 2011. An algorithm for intelligibility prediction of time-frequency weighted noisy speech. IEEE/ACM Trans. Audio Speech Lang.Process. (TASLP) 19 (7), 2125–2136.
Tu, M., Zhang, X., March, 2017. Speech enhancement based on deep neural networks with skip connections. IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP), pp. 5565–5569.
Varga, A., Steeneken, H.J., 1993. Assessment for automatic speech recognition: II. NOISEX-92: a database and an experiment to study the effect of additive noise on speech recognition systems. Speech Commun. 12 (3), 247–251.
Wang, Y., Han, K., Wang, D., 2013. Exploring monaural features for classification-based speech segregation. IEEE/ACM Trans. Audio Speech Lang.Process. (TASLP) 21 (2), 270–279.
Williamson, D.S., Wang, Y., Wang, D., 2016. Complex ratio masking for monaural speech separation. IEEE/ACM Trans. Audio Speech Lang.Process. (TASLP) 24 (3), 483–492.
Wu, W.-R., Chen, P.-C., 1998. Subband Kalman filtering for speech enhancement. IEEE Trans. Circuits Syst. II 45 (8), 1072–1083.
Xia, Y., Wang, J., 2015. Low-dimensional recurrent neural network-based Kalman filter for speech enhancement. Neural Netw. 67, 131–139.
Xu, Y., Du, J., Dai, L.-R., Lee, C.-H., 2013. An experimental study on speech enhancement based on deep neural networks. IEEE Signal Process. Lett. 21 (1), 65–68.
Xu, Y., Du, J., Dai, L.-R., Lee, C.-H., 2015. A regression approach to speech enhancement based on deep neural networks. IEEE/ACM Trans. Audio Speech Lang.Process. (TASLP) 23 (1), 7–19.
Yu, H., Ouyang, Z., Zhu, W.-P., Champagne, B., May, 2019. A deep neural network based Kalman filter for time domain speech enhancement. IEEE International Symposium on Circuits and Systems (ISCAS), pp. 1–5.
Yu, H., Zhu, W.-P., Champagne, B., 2020a. High-frequency component restoration for Kalman filter based speech enhancement. International Symposium on Circuits and Systems (ISCAS). IEEE, pp. 1–5.
Yu, H., Zhu, W.-P., Champagne, B., 2020b. Subband Kalman filtering with DNN estimated parameters for speech enhancement. Proc. of Interspeech, pp. 2697–2701.
Yu, H., Zhu, W.-P., Ouyang, Z., Champagne, B., 2020. A hybrid speech enhancement system with DNN based speech reconstruction and Kalman filtering. Multimed. Tools Appl. 79, 32643–32663.
Yu, H., Zhu, W.-P., Yang, Y., 2020c. Constrained ratio mask for speech enhancement using DNN. Proc. of Interspeech, pp. 2427–2431.