Multimedia Tools and Applications (2020) 79:32643–32663
https://doi.org/10.1007/s11042-020-09563-5

A hybrid speech enhancement system with DNN based speech reconstruction and Kalman filtering

Hongjiang Yu[1] · Wei-Ping Zhu[1] · Zhiheng Ouyang[1] · Benoit Champagne[2]

Received: 10 November 2019 / Revised: 22 June 2020 / Accepted: 6 August 2020 /
Published online: 29 August 2020

Abstract
In this paper, we propose a hybrid speech enhancement system that exploits deep neural network (DNN) for speech reconstruction and Kalman filtering for further denoising, with the aim to improve performance under unseen noise conditions. Firstly, two separate DNNs are trained to learn the mapping from noisy acoustic features to the clean speech magnitudes and line spectrum frequencies (LSFs), respectively. Then the estimated clean magnitudes are combined with the phase of the noisy speech to reconstruct the estimated clean speech, while the LSFs are converted to linear prediction coefficients (LPCs) to implement Kalman filtering. Finally, the reconstructed speech is Kalman-filtered for further removing the residual noises. The proposed hybrid system takes advantage of both the DNN based reconstruction and traditional Kalman filtering, and can work reliably in either matched or unmatched acoustic environments. Computer based experiments are conducted to evaluate the proposed hybrid system with comparison to traditional iterative Kalman filtering and several state-of-the-art DNN based methods under both seen and unseen noises. It is shown that compared to the DNN based methods, the hybrid system achieves similar performance under seen noise, but notably better performance under unseen noise, in terms of both speech quality and intelligibility.

Keywords Speech enhancement · Deep neural network · Kalman filter · Unmatched acoustic environment

1 Introduction

In real world environments, speech signals are often corrupted by a wide range of background noises. These disturbances cause problems in applications including voice communication, automatic speech recognition and speaker identification. As a result, speech enhancement, which aims to improve speech quality and intelligibility, has been intensively

✉ Hongjiang Yu
ho_yu@encs.concordia.ca

[1] Department of Electrical and Computer Engineering, Concordia University, Montreal, Canada

[2] Department of Electrical and Computer Engineering, McGill University, Montreal, Canada

studied over the past several decades, and will likely continue to be an active research topic in speech processing, recognition and communication.

Various denoising methods have been proposed in the literature, among which statistical filtering received the earliest attention. Wiener filtering is one of the well-known methods in this category, with its goal to find the optimal minimum mean square error (MMSE) estimate of the clean speech's discrete Fourier transform (DFT) coefficients [11]. Wiener filtering introduces broadband residual noise instead of musical noise in the enhanced speech, which is undesirable even though often acceptable. Kalman filter is a time-domain, linear MMSE estimator that was first applied into speech enhancement in [22], and remains of particular interest due to its several advantages: (1) ability to handle and process non-stationary signals; (2) absence of musical noise in the denoised speech given ideal parameters; (3) possibility of enhancing both the speech magnitude and phase.

In Kalman filtering, the clean speech is usually represented by a linear prediction model. As such, the enhancement performance is largely dependent on the accuracy of the model parameters, i.e.: the linear prediction coefficients(LPCs), the driving noise variance, and the additive noise variance. Various estimation algorithms have been proposed to obtain the above parameters from noisy speech, which can be divided into two categories: online estimation [3, 4, 14, 25, 40] and off-line estimation [9, 19]. The online estimation usually iterates between Kalman filtering of noisy observations and estimation of the speech parameters. In each iteration the Kalman filter enhances the speech to obtain better parameter estimation, and generally improves the final results after a few iterations. The off-line algorithms require a training stage to predict the parameters beforehand based on a clean speech database.

Recently, there has been a great deal of interest in data-driven supervised methods for speech enhancement. Among them, the most prominent ones employ deep neural networks for magnitude spectrum estimation [41, 42], where the DNN acts as a regression model to find a mapping function between the log-power spectra (LPS) of the noisy speech and that of the clean speech. DNN has also been used as a primary tool to predict key parameters in speech enhancement methods [5, 16, 17, 20, 31, 35]. For example, the authors in [16] employ DNN to estimate the ideal ratio mask (IRM) for masking based algorithms, while a long short-term memory (LSTM) network is utilised in [17] to accurately estimate the *a priori* signal-to-noise ratio (SNR) for traditional MMSE based short-time spectral amplitude (STSA) estimators. Another breakthrough includes the deep learning based generative modeling for speech enhancement, wherein the generative adversarial network (GAN) has been successfully employed to generate either clean speech waveforms [24] or spectrograms [1, 28] given the corresponding noisy counterparts.

We note that in most deep learning based methods, the noisy phase is directly used in the reconstruction of the enhanced speech, on the basis that our ears are insensitive to small phase distortions [33, 34]. In addition, the estimation of the phase remains challenging due to its unstructured characteristic and phase wrapping [12]. However, some researchers have pointed out the importance of estimating clean phase in recent works, especially at low SNRs [23, 26]. Based on this finding, phase estimation [10, 44] and complex spectrogram estimation [2, 21, 38, 39] have been proposed to enhance the magnitude and phase spectra simultaneously.

Compared with the unsupervised statistical model based methods, the use of DNN to predict clean speech magnitudes or other parameters offers many advantages. The nonlinear structures of DNN confer them with powerful learning capability, suitable to model the complex mapping relationship between the noisy and clean speech. Furthermore, deep learning based methods usually do not require the estimation of the noise power spectrum,

nor do they rely on particular assumptions about the statistical properties of the speech and noise, which allow them to handle non-stationary noises in real-world scenarios under unexpected acoustic conditions. However, deep learning based algorithms require large databases for training in order to improve their generalization capability. To achieve better performance in unseen noise condition, it is common to train a DNN with a large speech database comprising different speakers and noise types [42].

Although the conventional unsupervised statistical model based methods fail to achieve satisfactory results in real-world environments, the fact that they can reduce different kinds and levels of noises to some extent, is an attractive feature to researchers. In other words, the statistical model based methods do not employ a training stage, and thus treat all noises as unseen noise so that their denoising capability, albeit limited, remains available in all situations. Based on such considerations, hybrid approaches have been proposed and investigated by taking advantage of both unsupervised methods and deep learning methods [18]. These hybrid approaches have shown show a better generalization capability than the deep learning only method in unmatched noise conditions [42].

Recently in [43], we have proposed a DNN assisted Kalman filter for speech enhancement, where the DNN is trained to predict a variant of the LPSs, namely the line spectrum frequencies (LSFs) of the clean speech, from those of the noisy speech. Experiments have shown that with a large database for off-line training, one can reduce the sensitivity of LPCs prediction in the presence of noise, leading to a better enhancement than possible with the subband iterative Kalman filter algorithm [25]. However, the method in [43] suffers from large distortion in the high frequency components of the enhanced speech, partly due to the imperfect estimation of additive noise and driving noise variances used in the Kalman filtering of the noisy speech. In addition, its performance relative to other DNN based denoising methods is not yet known.

In this paper, we propose a new two-stage hybrid denoising system that exploits DNN based speech reconstruction in conjunction with Kalman filtering in order to achieve improved performance. In the first stage, a DNN is trained for the estimation of the speech magnitude spectrum, which is then used to reconstruct the clean speech. In the second stage, another DNN is trained for predicting the LSFs of the clean speech, which will be transformed to LPCs. Meanwhile, the additive noise and driving noise variances are extracted from the reconstructed speech. Finally, a Kalman filter with the estimated parameters is applied to the reconstructed speech to obtain further enhancements. The main features and contributions of our proposed approach in this paper are summarized as follows.

- As well-known, the above cited deep learning based methods often suffer from performance degradation due to the data mismatch between the training and testing stages. Consequently, the reconstructed speech from DNN based method inevitably contains residual noise in unmatched acoustic environment. By incorporating and combining Kalman filtering with a DNN-based speech reconstruction method, our approach makes it possible to further reduce the residual noise in unmatched conditions.
- Further advantages of employing DNN include the following: First, DNN is used to estimate clean speech amplitude in order to perform preliminary speech enhancement. The additive noise and driving noise variances required for Kalman filtering are then more accurately estimated from the DNN pre-enhanced speech. Second, DNN is used to obtain accurate LPCs estimates which is critical for improved Kalman filtering.
- The speech reconstruction is performed in the frequency domain, that is, the reconstructed speech is obtained by synthesising the estimated magnitude and the noisy phase spectra, while the denoising process of Kalman filtering is realized in the time domain.

With such a combination, our hybrid system can be viewed as a joint estimator for both magnitude and phase of the spectra of the clean speech.

The rest of the paper is organized as follows. Section 2 elaborates on the proposed hybrid speech enhancement system where the key components and processing are detailed. Section 3 presents a series of experiments to assess the system performance. Section 4 concludes the paper.

2 Proposed system

The overall block diagram of our hybrid system is depicted in Fig. 1. It consists of two parts, namely: training part and enhancement part. In the training part, we first extract noisy speech acoustic features, and then input them to two DNNs which are trained separately to learn the mapping from the noisy features to different targets: clean speech magnitudes and LSFs. In the enhancement part, the noisy speech features are extracted and processed by the well-trained DNNs to predict the clean magnitudes and LSFs. The estimated spectral magnitudes together with the noisy phase spectrum are then synthesised to obtain the reconstructed speech. Finally, a Kalman filter with the DNN-based estimated parameters is applied to the reconstructed speech to obtain the enhanced speech. The key components and processing steps involved in the proposed system are described in further details below.

2.1 Noisy speech

In our system, as several other works on speech enhancement, we consider additive noise which is the most common factor that degrades the speech quality in real-world scenarios.

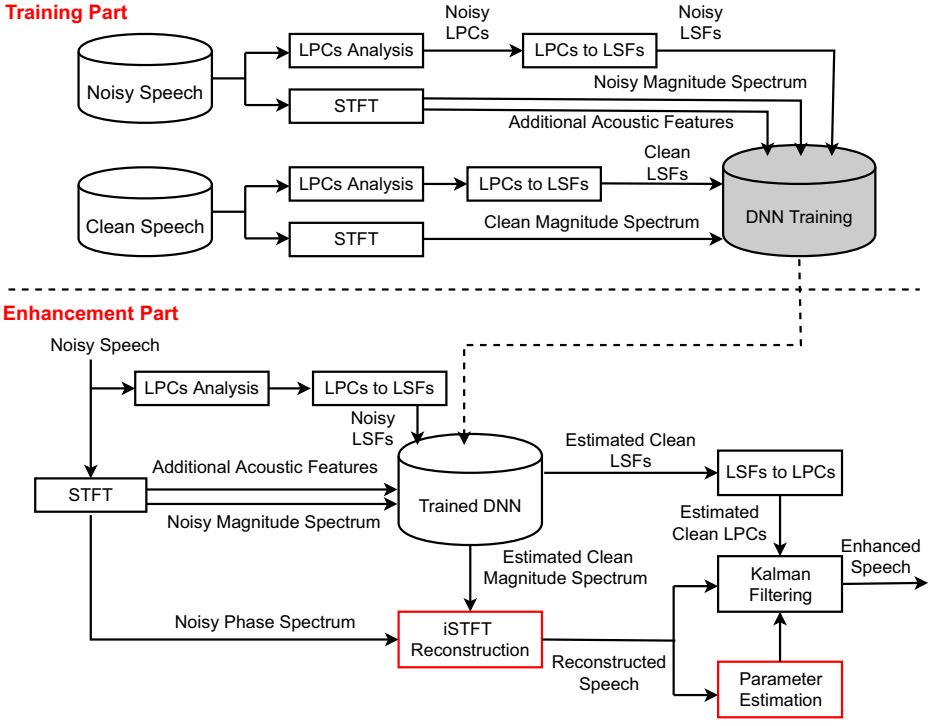

Fig. 1 A block diagram of proposed hybrid speech enhancement system

The time-domain noisy speech can be modeled as

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

where $s(n)$ is the clean speech, $w(n)$ the additive noise, $y(n)$ the noisy speech. While it is common to assume that the speech signal and noise are statistically independent in the derivation of the Kalman filtering approach, this assumption is not an essential condition for the application of the proposed method.

The time domain noisy speech is then transformed to the time-frequency domain using short-time Fourier transform (STFT). The STFT-based spectrogram of the clean speech signal $s(n)$ is denoted as $S(k, f)$, with k and f indicating the frame and frequency bin indices, respectively. $S(k, f)$ can be expressed in polar coordinates in terms of its magnitude and phase spectra as,

$$S(k, f) = |S(k, f)| e^{j\phi_s(k,f)} \tag{2}$$

The corresponding spectrograms of the additive noise and the noisy speech are donated as $W(k, f)$ and $Y(k, f)$. For simplicity, we shall refer the phase (magnitude) of the clean speech and that of the noisy speech as clean phase (or magnitude) and noisy phase (magnitude) respectively.

2.2 Targets and features

As mentioned before, two different training targets are set as the output of the DNNs, i.e.: the spectral magnitudes and LSFs. The magnitudes are employed in the speech reconstruction, while the LSFs are converted to LPCs as key parameters for Kalman filtering. In [41], the authors point out that transforming magnitudes to log-power spectral features is more relevant to human perception. However, directly using magnitudes as training target can also yield good performance and furthermore requires lower computational complexity [20]. Besides, we use LSFs instead of LPCs since the former have a well-contained dynamic range of values, while the latter require a larger dynamic range. Therefore, we can maintain the stability of the training part more easily in the LSFs domain [8].

The choice of appropriate input features is important to the performance. In [36], several monoral feature sets are introduced and discussed for DNN-based speech applications, including: the amplitude modulation spectrum (AMS); the relative spectral transform and perceptual linear prediction (RASTA-PLP); the Mel-frequency cepstral coefficients (MFCC) and their deltas; the gammatone filterbank energies (GF) and their deltas. These features are known to represent speech characteristics well and have been successfully used in many speech processing tasks. Hence, we adopt them as additional input features in our work. Note that we include the LSFs into the input feature set when the training targets are LSFs, and included the speech spectral magnitudes of the speech spectrum when the targets are magnitudes. With these two specific feature sets, we are able to better learn the mapping from the noisy features to the training targets.

2.3 LPCs-to-LSFs conversion

In the training part, LPCs are calculated using both noisy and clean speech databases, and then converted into LSFs for the DNN training. In the enhancement part, the estimated LSFs are converted to LPCs for Kalman filtering. The conversion process [13] is briefly summarized below.

A short segment of speech under the linear prediction analysis model is assumed to be generated as the output of finite impulse response filter $A(z)$. In order to define LSFs, the

p-th order linear predictor $A(z)$ is decomposed into symmetrical and anti-symmetrical parts, represented by the polynomials $P(z)$ and $Q(z)$, respectively,

$$\begin{aligned} P(z) &= A(z) + z^{-(p+1)} A\left(z^{-1}\right) \\ Q(z) &= A(z) - z^{-(p+1)} A\left(z^{-1}\right). \end{aligned} \tag{3}$$

The LSFs ω_i are expressed as the zeroes (or roots) of $P(z)$ and $Q(z)$ in terms of the angular frequency.

The conversion from LSFs back to LPCs requires to obtain $A(z)$. Since $A(z)$ is expressed as the linear combination of $P(z)$ and $Q(z)$, i.e., $A(z) = 0.5[P(z) + Q(z)]$, we can easily construct $A(z)$ by using the ordered LSFs ω_i of $P(z)$ and $Q(z)$, i.e.:

$$\begin{aligned} P(z) &= (1 - z^{-1}) \prod_{i=2,4,\cdots,p} \left(1 - 2z^{-1}\cos\omega_i + z^{-2}\right) \\ Q(z) &= (1 + z^{-1}) \prod_{i=1,3,\cdots,p-1} \left(1 - 2z^{-1}\cos\omega_i + z^{-2}\right). \end{aligned} \tag{4}$$

2.4 DNN Structure

As shown in Fig. 2, we employ two fully-connected DNNs to estimate the spectral magnitudes and LSFs separately in our work. Employing two separate DNNs can provide better

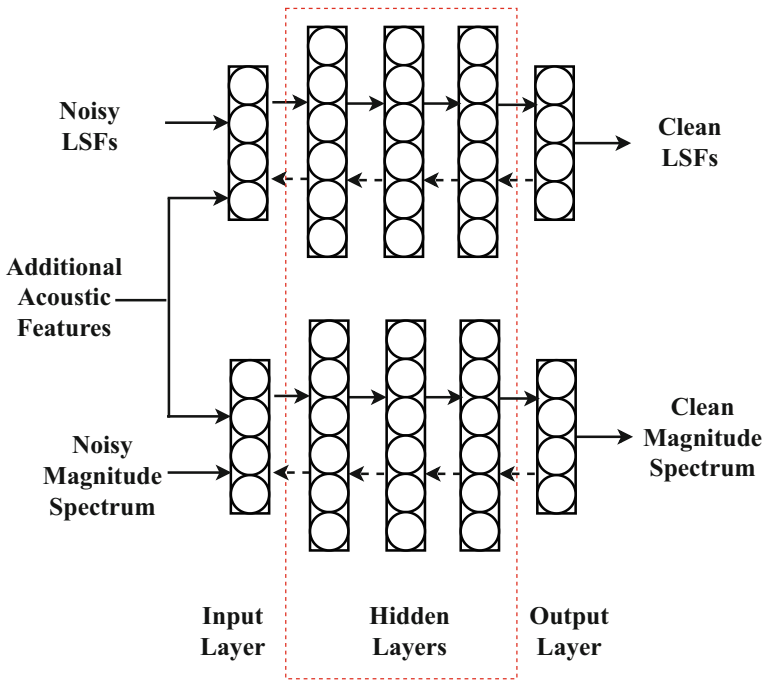

Fig. 2 DNN structure in proposed hybrid speech enhancement system

results than training only one DNN to learn the mapping to these two targets by multi-objective learning, because the LSFs and the magnitudes share little similarity in their structure.

Although the targets are different, the settings for each DNN are the same in our work. Each network consists of an input layer, an output layer and three hidden layers, each comprising 1024 units. The linear activation functions are used in the output layer, whereas the rectified linear functions are used in the hidden layers.

The input features of both DNNs are computed for each frame of the signal. To make full use of the temporal information of speech, it is common to incorporate features of adjacent time frames into a single feature vector. Moreover, we normalize different features into the range [0, 1) in order to balance the training errors.

Back propagation is used to adjust the weights and biases in the training part. The cost function for each training utterance is defined as the mean square error (MSE) of the magnitudes (or LSFs). The respective MSE is computed between the clean and estimated targets, i.e.,

$$\mathrm{MSE}_{\mathrm{MAG}} = \frac{1}{KF}\sum_{k=1}^{K}\sum_{f=1}^{F}\left(|\hat{S}(k,f)| - |S(k,f)|\right)^2 \tag{5}$$

or

$$\mathrm{MSE}_{\mathrm{LSF}} = \frac{1}{Kp}\sum_{k=1}^{K}\sum_{i=1}^{p}\left(\hat{\omega}(k,i) - \omega(k,i)\right)^2 \tag{6}$$

where $|S(k, f)|$ and $|\hat{S}(k, f)|$ are the clean and estimated magnitudes, respectively, with K indicating the number of frames and F the number of frequency bins, while $\omega(k, i)$ and $\hat{\omega}(k, i)$ are the clean and estimated LSFs, respectively, with i indicating the order index and p the AR speech model order.

2.5 Speech reconstruction

The STFT of the reconstructed speech, $R(k, f)$ is obtained by combining the estimated magnitudes from the well-trained DNN together with the noisy spectral phase values ϕ_y , i.e.,

$$R(k,f) = \left|\hat{S}(k,f)\right| e^{j\phi_y(k,f)} \tag{7}$$

The reconstructed speech $r(n)$ is then obtained by computing the inverse STFT of $R(k, f)$. Although we have used the noisy phases for synthesis, the reconstructed speech will be Kalman filtered in the time-domain, which can be regarded as a joint form of enhancement of the magnitude and phase spectra.

2.6 Kalman filtering

In Kalman filter based speech enhancement, the auto-regressive (AR) model is widely adopted to represent the clean speech and derive the Kalman recursion equations. The p-th order AR signal model is given by

$$s(n) = \sum_{i=1}^{p} a_i s(n-i) + v(n) \tag{8}$$

where $\{a_i\}_{i=1}^{p}$ are the LPCs of the speech signal and $v(n)$ is the driving white noise with variance σ_v^2.

It has been pointed out in [27] that when the AR parameters are calculated from clean speech, the enhanced speech from the ideal Kalman filter is of high quality and contains no musical noise. In practical applications, the noisy speech (1) is used to estimate the AR parameters, as the clean speech is not accessible. However, in our system, the reconstructed speech $r(n)$ is used instead of the noisy speech as input to the Kalman filter, in order to provide more accurate parameter estimates. The Kalman filtering procedure is implemented as follows.

Firstly, the clean and reconstructed speech models are expressed in terms of matrix and vector notations to facilitate the presentation the Kalman filtering equations.

$$\begin{cases} \mathbf{u}(n) = \mathbf{F}\mathbf{u}(n-1) + \mathbf{G}\mathbf{v}(n) \\ \mathbf{r}(n) = \mathbf{H}\mathbf{u}(n) + \mathbf{w}(n) \end{cases} \tag{9}$$

where the transition matrix $\mathbf{F}$ is defined by

$$\mathbf{F} = \begin{bmatrix} 0 & 1 & 0 & \cdots & 0 & 0 \\ 0 & 0 & 1 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & 0 & 1 \\ a_p & a_{p-1} & a_{p-2} & \cdots & a_2 & a_1 \end{bmatrix} \in \mathbb{R}^{p\times p}, \tag{10}$$

$\mathbf{H}$ is a p-th order identity matrix and $\mathbf{G} = [0, \cdots, 0, 1]^T \in \mathbb{R}^p$. Moreover, $\mathbf{u}(n)$ denotes the speech state vector, $\mathbf{r}(n)$ the reconstructed speech vector, $\mathbf{w}(n)$ the additive noise vector and $\mathbf{v}(n)$ the driving noise vector, which are respectively given by

$$\begin{aligned} \mathbf{u}(n) &= [s(n-p+1), \ldots, s(n-1), s(n)]^T \\ \mathbf{r}(n) &= [r(n-p+1), \ldots, r(n-1), r(n)]^T \\ \mathbf{w}(n) &= [w(n-p+1), \ldots, w(n-1), w(n)]^T \\ \mathbf{v}(n) &= [v(n-p+1), \ldots, v(n-1), v(n)]^T \end{aligned} \tag{11}$$

The denoising process of the standard Kalman filtering is summarized by the following estimation and updating equations

$$\begin{cases} \mathbf{e}(n) = \mathbf{r}(n) - \mathbf{G}^T\hat{\mathbf{u}}(n|n-1) \\ \mathbf{K}(n) = \mathbf{P}(n|n-1)\left(\mathbf{R}_w + \mathbf{P}(n|n-1)\right)^{-1} \\ \hat{\mathbf{u}}(n|n) = \hat{\mathbf{u}}(n|n-1) + \mathbf{K}(n)\mathbf{e}(n) \\ \mathbf{P}(n|n) = (\mathbf{I} - \mathbf{K}(n))\mathbf{P}(n|n-1) \\ \hat{\mathbf{u}}(n+1|n) = \mathbf{F}\hat{\mathbf{u}}(n|n) \\ \mathbf{P}(n+1|n) = \mathbf{F}\mathbf{P}(n|n)\mathbf{F}^T + \sigma_v^2\mathbf{G}\mathbf{G}^T \end{cases} \tag{12}$$

where $\mathbf{e}(n)$ is the innovation vector, $\mathbf{R}_w$ the covariance matrix of the additive noise, $\mathbf{K}(n)$ the Kalman gain matrix, $\hat{\mathbf{u}}(n|n)$ and $\hat{\mathbf{u}}(n|n-1)$ the filtered estimate and the MMSE estimate of state vector $\mathbf{u}(n)$, respectively, $\mathbf{P}(n|n)$ and $\mathbf{P}(n|n-1)$ the filtered and the predicted state error correlation matrix, respectively. The denoised speech $d(n)$ is finally given by

$$d(n) = \mathbf{G}^T\hat{\mathbf{u}}(n|n) \tag{13}$$

The enhancement performance is dependent on the accuracy of the parameter estimation in the Kalman filter. These parameters include the driving noise variance σ_v^2, the covariance matrix of the additive noise $\mathbf{R}_w$, and the transition matrix $\mathbf{F}$ which contains the LPCs of the speech signal model.

2.7 Parameter estimation

Since the transition matrix $\mathbf{F}$ is obtained from the DNN, only the other two parameters, i.e. σ_v^2 and $\mathbf{R}_w$, need to be estimated before the Kalman filtering of the reconstructed speech. As usual, the covariance matrix can be estimated during the speech-absent frames. Thus, estimation accuracy of $\mathbf{R}_w$ is highly dependent on the ability to detect the voice and unvoiced parts of the noisy speech. Here, the voice activity detector (VAD) algorithm [15] based on speech energy and spectral flatness is adopted for this purpose. Different from our previous work [43], the VAD is applied to the reconstructed speech $r(n)$ rather than the noisy speech, which helps make a correct decision of the unvoiced parts as seen in Fig. 3, and in turn, improve the estimation accuracy of the noise covariance matrix.

For the estimation of the variance of the driving noise $v(n)$, we solve the Yule-Walker equations for the linear prediction model of the reconstructed speech, instead of using the estimation algorithm given in [43]. The comparison of the estimated variance σ_v^2 is shown in Fig. 4, which shows that the new algorithm achieves a better performance.

2.8 Summary of proposed system

The main processing steps of the proposed hybrid system are summarized as follows:

1. Estimating clean LSFs and magnitudes from noisy features with the proposed DNNs.
2. Synthesising the reconstructed speech $r(n)$ with the estimated magnitude and the noisy phase spectra.
3. Converting LSFs to LPCs to form the state transition matrix $\mathbf{F}$.

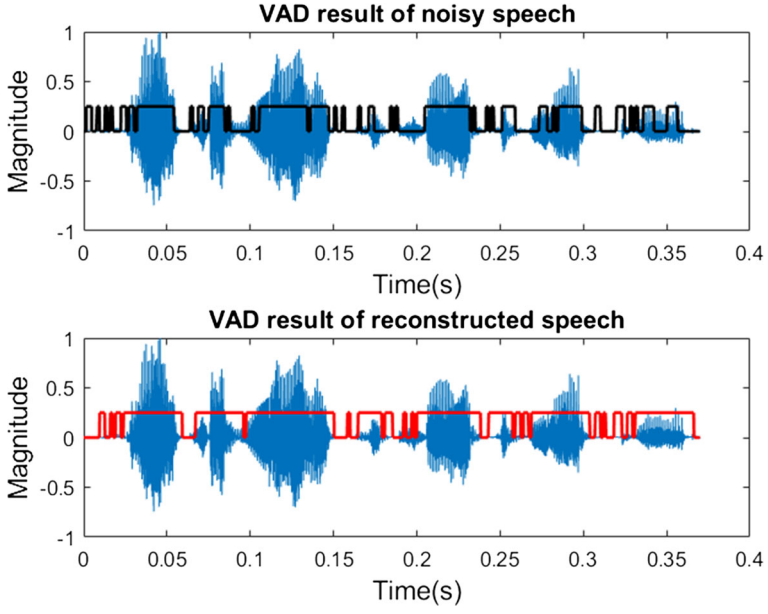

Fig. 3 The VAD results of noisy and reconstructed speech. The blue waveform is the original clean speech. The decision line represents an unvoiced part when its value equals to 0, and a voiced part otherwise. The noisy speech is corrupted with pink noise at -3 dB

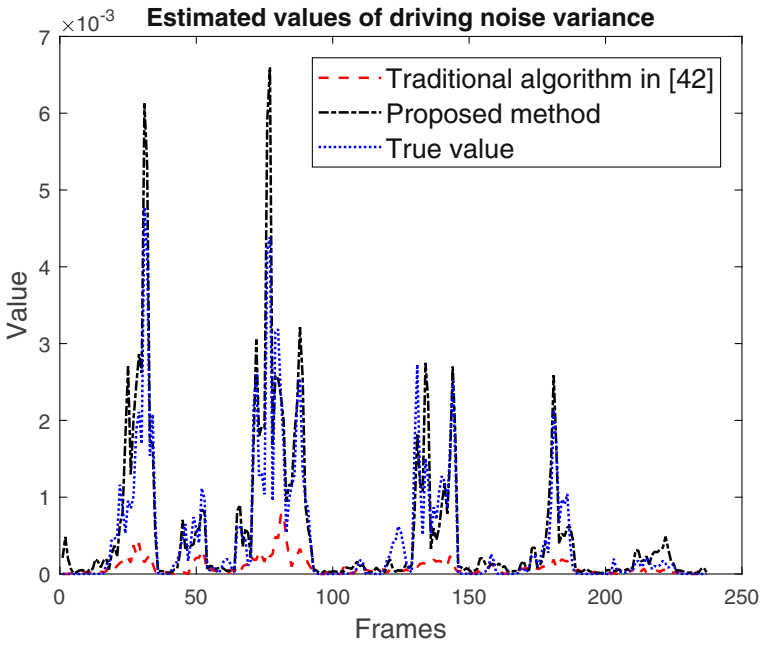

Fig. 4 The estimation of driving noise variance through different methods. Our algorithm (black) is closer to the true one (blue). The noisy speech is corrupted with pink noise at -3 dB

4. Computing the covariance matrix of the additive noise and the driving noise variance from the reconstructed speech.
5. Performing Kalman filtering of the reconstructed speech (12) to obtain $\hat{\mathbf{u}}(n|n)$, and the final enhanced speech $d(n)$ as given by (13).

3 Experimental results

3.1 Experimental setup

3.1.1 Databases

The clean speech is selected from the IEEE sentence database[1], as the latter contains phonetically balanced sentences with relatively low word-context predictability [6]. The corpus is comprised of 72 lists, each of which containing 10 sentences. We choose the first 67 lists (670 utterances) for the training part and the remaining 5 lists (50 utterances) for the enhancement part. The noises are selected from the NOISEX-92 database [32], which contains white noise and a variety of non-stationary noises[2]. Each noise signal has a duration of approximate 4 minutes. Four types of noises (babble,white, street and factory) are regarded as seen noise, and another four types (pink, buccaneer2, destroyerengine and hfchannel) as unseen noise. The spectrograms of the noise signals used in our experiments are shown in

[1]Available at website https://www.crcpress.com/downloads/K14513/K14513_CD_Files.zip
[2]More details can be found at http://mi.eng.cam.ac.uk/comp.speech/Section1/Data/noisex.html

Fig. 5. In the training part, the noisy speech is obtained by mixing clean training utterances with seen noise at four different levels of SNRs, i.e., -3dB, 0dB, 3dB and 6dB, which results in 10720 utterances. In the enhancement part, both seen and unseen noises are mixed with clean testing utterances at the above mentioned four SNR levels. The number of noisy utterances used in the enhancement part is 800 for both seen and unseen noises. The sampling frequency for the speech and noise signals is set to 16kHz.

3.1.2 Reference methods

To evaluate the performance of the proposed new system, we choose several existing approaches for comparison, which include one traditional Kalman filtering algorithm: Iter-KF; and four recent DNN based methods, i.e.: DNN-MAG, DNN-IRM, FSEGAN and our previous work DNN-KF. These are introduced briefly in the following.

Iter-KF [3] The enhanced speech is obtained by iteratively performing conventional Kalman filtering, in which the LPCs are updated in each iteration.

DNN-MAG A DNN is employed to directly explore the mapping from the noisy magnitude spectrum to the clean one. The enhanced speech is synthesised with the estimated magnitude and noisy phase spectra. This method is similar to the DNN-LPS [41], the only difference between these two methods being that DNN-LPS uses the log-power spectrum as features.

DNN-IRM [37] A DNN is trained for better predicting the IRM. The estimated IRM is then applied to the noisy magnitude spectrogram to reduce the noise part, and the enhanced speech is then reconstructed from the masked magnitude and noisy phase spectra.

DNN-KF [43] A DNN is used to predict the LSFs for Kalman filtering. The noisy speech is processed by the Kalman filter to obtain the enhanced speech. We note that the DNN-MAG and DNN-IRM are frequency-domain speech enhancement methods, while the DNN-KF is a time-domain method.

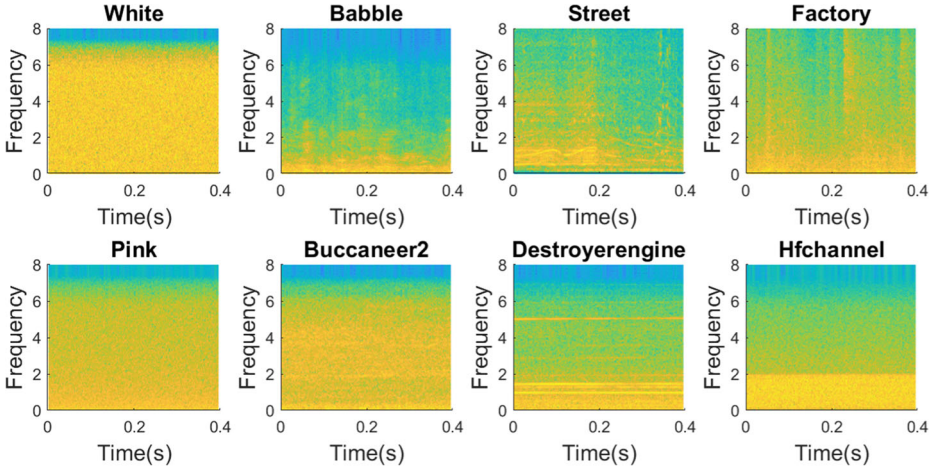

Fig. 5 Spectrograms of different noises. The first line (white, babble,street, factory) depicts seen noise in the training part, while the other depicts unseen noise in the enhancement part

FSEGAN [1] A least-square GAN is utilized to generate the clean speech magnitude spectrogram from the noisy one. The enhanced speech is reconstructed from the generated clean magnitude and noisy phase spectra.

In order to fairly evaluate the performance of the method proposed in this paper, we use the same DNN configuration in all the methods except FSEGAN. For FSEGAN, we adopt the settings provided in [1] and adjust other network parameters to optimize performance. For the remaining DNN based methods, we use the standard feed-forward network configuration which comprises one input layer, one output layer and three hidden layers with 1024 units. For DNN-MAG and DNN-IRM, the Hamming window is selected to divide each utterance into 20 ms frames with an 10 ms frame shift (50% overlap). A 320-point DFT is then computed for each frame resulting into 161 samples. For DNN-KF, a rectangular window is used to divide the audio signals into 20 ms frames with no overlap. For the proposed hybrid system, the STFT setting used in the magnitude spectrogram computation is the same as that of DNN-MAG, and the framing process in the LSFs estimation is the same as that of DNN-KF. In the implementation of the Kalman filtering algorithm, we set $\mathbf{u}(0|0) = \mathbf{0}$, $\mathbf{P}(0|0)$ as an identity matrix, and the speech AR order as $p = 12$.

3.1.3 Objective metrics

To evaluate the enhancement performance, two objective metrics are selected: the perceptual evaluation of speech quality (PESQ) measure [7] and the short-time objective intelligibility (STOI) measure [30]. PESQ and STOI evaluate the processed speech from two different aspects: speech quality and intelligibility, and are widely adopted in speech-related applications.

PESQ is proposed in the ITU-T recommendation P.862. It measures the distortion between the original and processed signal. Firstly the signals are equalized to a standard listening level, then aligned in time to correct for time delays, and then processed through an auditory transform to obtain the loudness spectra. The difference between the loudness spectra of the processed signal and that of original signal is computed and averaged over time and frequency to produce the prediction of subjective mean opinion score (MOS). 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. It extracts short-time blocks of the clean and processed signals to compute the average of the correlations across blocks, and the average correlation is then taken as the intelligibility score. The STOI yields high correlation with subjective intelligibility score.

3.2 Results and discussions

Tables 1 and 2 show the average objective scores of the different speech enhancement algorithms on both seen and unseen noises respectively. In general, for the seen noise, the overall objective scores achieved by DNN-IRM and the proposed hybrid method are close, and are superior to the remaining methods. For the unseen noise, the overall objective scores clearly show that the proposed hybrid method performs better than the other three DNN based methods in most cases, except for the STOI score of DNN-IRM at 6dB SNR. A more detailed analysis of the results is provided in the following.

Table 1 Objective scores of different methods on seen noise

	PESQ				STOI			
Methods	-3dB	0dB	3dB	6dB	-3dB	0dB	3dB	6dB
Unprocessed	1.41	1.52	1.68	1.86	0.66	0.72	0.78	0.83
Iter-KF	1.55	1.79	2.01	2.25	0.66	0.72	0.79	0.84
DNN-MAG	1.89	2.13	2.34	2.55	0.75	0.82	0.86	0.88
DNN-IRM	2.01	**2.28**	**2.47**	**2.67**	**0.80**	**0.84**	**0.88**	**0.91**
DNN-KF	1.71	1.93	2.13	2.30	0.71	0.77	0.81	0.85
FSEGAN	1.85	2.02	2.19	2.35	0.70	0.75	0.80	0.84
Proposed	**2.05**	2.23	2.44	2.61	0.79	**0.84**	**0.88**	0.90

The bold entries show the best score of the results of different methods

3.2.1 Seen noise

In the case of seen noise (Table 1), Iter-KF achieves the worst performance among all tested methods, which is mainly caused by the inaccurate estimation of the AR parameters. We also note that the performance of FSEGAN is quite limited on our tested database. One possible reason could be that the generative model requires a larger amount of training data to learn the underlying distribution of the target features; otherwise mode collapse may happen in the training part [29].

Next, we compare the performances of the other four DNN related approaches, i.e., DNN-MAG, DNN-IRM, DNN-KF and the proposed hybrid system. Clearly, DNN-IRM shows the best overall performance, especially in the high SNR region for PESQ. One possible reason for this outcome under matched condition is the use of different targets: for DNN-MAG and DNN-KF, the targets (clean magnitudes or LSFs) are the same across different noises and SNRs, and thus the DNN has to learn a many-to-one mapping; whereas for DNN-IRM, the targets (ideal ratio masks) depend on the noise type and SNR, and thus the DNN is faced with the simpler task of learning a one-to-one mapping [37].

Our system also exhibits better performance than DNN-KF and DNN-MAG. For DNN-KF, although the use of DNN in LSFs estimation improves the performance of Kalman

Table 2 Objective scores of different methods on unseen noise

	PESQ				STOI			
Methods	-3dB	0dB	3dB	6dB	-3dB	0dB	3dB	6dB
Unprocessed	1.38	1.51	1.66	1.83	0.66	0.72	0.78	0.84
Iter-KF	1.64	1.84	2.04	2.26	0.68	0.75	0.81	0.85
DNN-MAG	1.73	1.92	2.13	2.32	0.71	0.78	0.83	0.87
DNN-IRM	1.81	2.05	2.29	2.51	0.75	0.81	**0.86**	**0.90**
DNN-KF	1.73	2.01	2.21	2.38	0.71	0.77	0.82	0.85
FSEGAN	1.74	1.95	2.16	2.35	0.69	0.76	0.82	0.85
Proposed	**1.96**	**2.16**	**2.36**	**2.52**	**0.77**	**0.83**	**0.86**	0.89

The bold entries show the best score of the results of different methods

filtering, two other parameters, i.e., the variance of the driving noise, σ_v^2 , and the covariance matrix of the additive noise, $\mathbf{R}_w$, are not accurately predicted from the noisy speech, which brings distortion to the final output speech. For DNN-MAG, the quality of the enhanced speech is hindered by the residual noise, especially at lower SNR. Our hybrid system, which can be regarded as a combination of DNN-MAG and DNN-KF, leads to a better enhanced speech because it employs the reconstructed speech as the input of Kalman filtering, and thus can provide more accurate estimates of σ_v^2 and $\mathbf{R}_w$, which in turn helps the Kalman filter better reduce the residual noises in the reconstructed speech.

Compared to DNN-IRM, our system achieves about the same level of performance. The PESQ score is slightly better than DNN-IRM at -3dB SNR and a little worse at higher SNR, while the STOI scores for both methods are quite close at all SNRs. Hence, in the case of seen noise, our proposed hybrid system and DNN-IRM achieve the best performance among all the evaluated methods.

3.2.2 Unseen noise

We first investigate the generalization capability of the tested methods by considering unseen noise. Upon comparison of the results in Tables 1 and 2, we note that all the methods suffer from a performance degradation. Comparing the results in Tables 1 and 2, we find that at high SNR, the performance of Iter-KF remains at a similar level as it belongs the class of unsupervised methods. In contrast, the objective scores of FSEGAN, DNN-IRM and DNN-MAG suffer a noticeable decrease, suggesting that the trained DNNs cannot achieve the same prediction accuracy under unseen noise. However, such a decrease in objective scores is not observed with DNN-KF, whose PESQ scores now exceed those of DNN-MAG for SNR $\geq$ 0dB. This may be explained by the fact that the use of DNN in DNN-KF is limited to the LSFs estimation, while the core processing function, i.e. Kalman filtering, is a conventional method and therefore its performance should remains at a similar level whether in seen or unseen noise situations. While the performance of our proposed system drops slightly in the case of unseen noise, this degradation is not as significant as that observed with the DNN-MAG and DNN-IRM methods.

The overall performance of the proposed hybrid system is significantly better than the other methods in terms of both PESQ and STOI scores, except for the STOI scores of DNN-IRM at high SNRs. However, at high SNR, intelligibility is less of a concern, as it is not difficult to understand the speech in this case, while the speech quality remains our major concern, which is well handled by the proposed system as reflected by PESQ scores. At low SNR, sthe speech intelligibility is severely impacted by the additive noise and should be our priority task. Clearly, the proposed hybrid method gives better STOI scores in low SNR situations. In conclusion, the proposed hybrid system achieves the best overall performance in unseen noise, after considering the various aspects of objective evaluation metrics.

We also characterize the enhancement performances of our proposed system on the different types of noise. The objective scores of the processed speech on each unseen noise at 0dB SNR are given in Figs. 6 and 7, respectively.

As can be seen from the results in Figs. 6 and 7, the overall performance of the processed speech on pink and buccaneer noises is better than that obtained on destroyerengine and hfchannel noises for all the methods. This is because the former two unseen noises share more similarities with some of the training noises and exhibit a less complex structure when compared to the latter two noises, so that the DNN can output a more accurate prediction. This finding indicates that the performance of DNN based methods indeed varies with different noises. According to the PESQ scores in Fig. 6, the proposed hybrid system produces

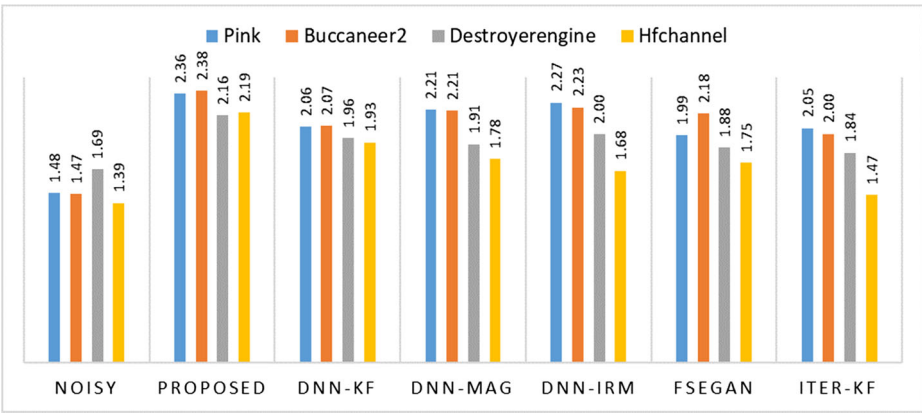

Fig. 6 PESQ scores on different noises at 0dB SNR

enhanced speech with better quality for all test noises. Further, for the STOI scores, the proposed hybrid system still achieves the highest scores on all noises.

3.3 Further look at enhanced speech

In order to better understand the characteristics of the enhanced speech signals resulting from the methods under evaluation, illustrative waveforms and spectrograms are plotted and compared. The noisy speech is obtained by mixing a selected clean speech utterance with hfchannel noise at 3dB SNR.

Figure 8 shows the residual noises and the distortions existing in the enhanced speech in the time-domain. The processed speech from FSEGAN or DNN-MAG contains a large amount of residual noises, which is caused by the difficulty in learning the mapping from the noisy magnitude spectrogram to the clean one. Iter-KF and DNN-KF perform well in removing the additive noise, but they both bring distortion to the original speech. For example, the speech component after 0.3s is suppressed by Iter-KF while the magnitudes of the processed speech of DNN-KF is strongly attenuated. DNN-IRM and the proposed hybrid system achieve a better performance than the other methods, as they can remove more noise

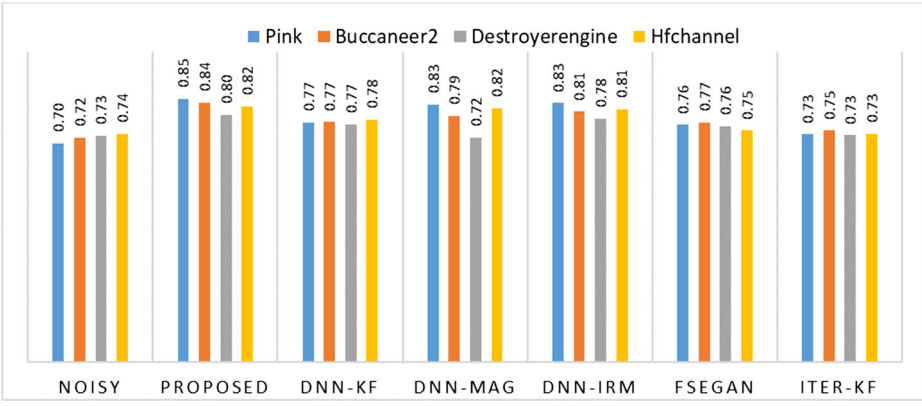

Fig. 7 STOI scores on different noises at 0dB SNR

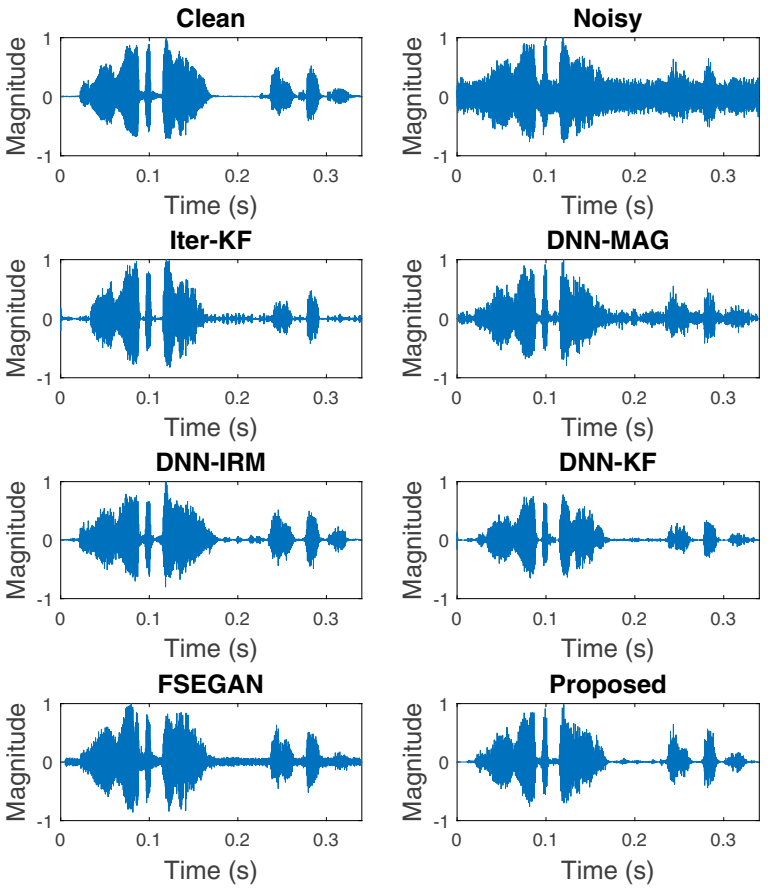

Fig. 8 Time domain waveforms of the clean, noisy and enhanced speech signals for different methods

without bringing significant distortions. Finally, for this particular experiment with unseen noise, our system is slightly better than DNN-IRM, as the residual noise is lower in the unvoiced part near the middle of the utterance.

Figure 9 demonstrates the effects of the residual noises and the distortions in the harmonic structures of the enhanced speech in the time-frequency domain. For Iter-KF, we can see the musical noise structure in the spectrogram in the region between 2kHz and 3kHz. The spectrogram of FSEGAN also exhibits some undesirable structures, which likely cause the degradation of performance. We make further comparison among the four DNN related methods. While the harmonic structures of the voiced parts with DNN-MAG are well preserved up to about 3kHz, a significant amount of residual noise is present during the unvoiced parts. The processed speech with DNN-IRM is affected by high-level residual broadband noise, which the method cannot adequately remove. While introducing less noise during the unvoiced parts, DNN-KF tends to suppress the high-frequency components of the voiced parts of speech, leading to a decrease of speech quality. For this example, the spectrogram of the enhanced speech with the proposed hybrid system seems to provide the best quality, i.e.: clearer harmonic structures of the voiced part, and the less residual noises during unvoiced parts.

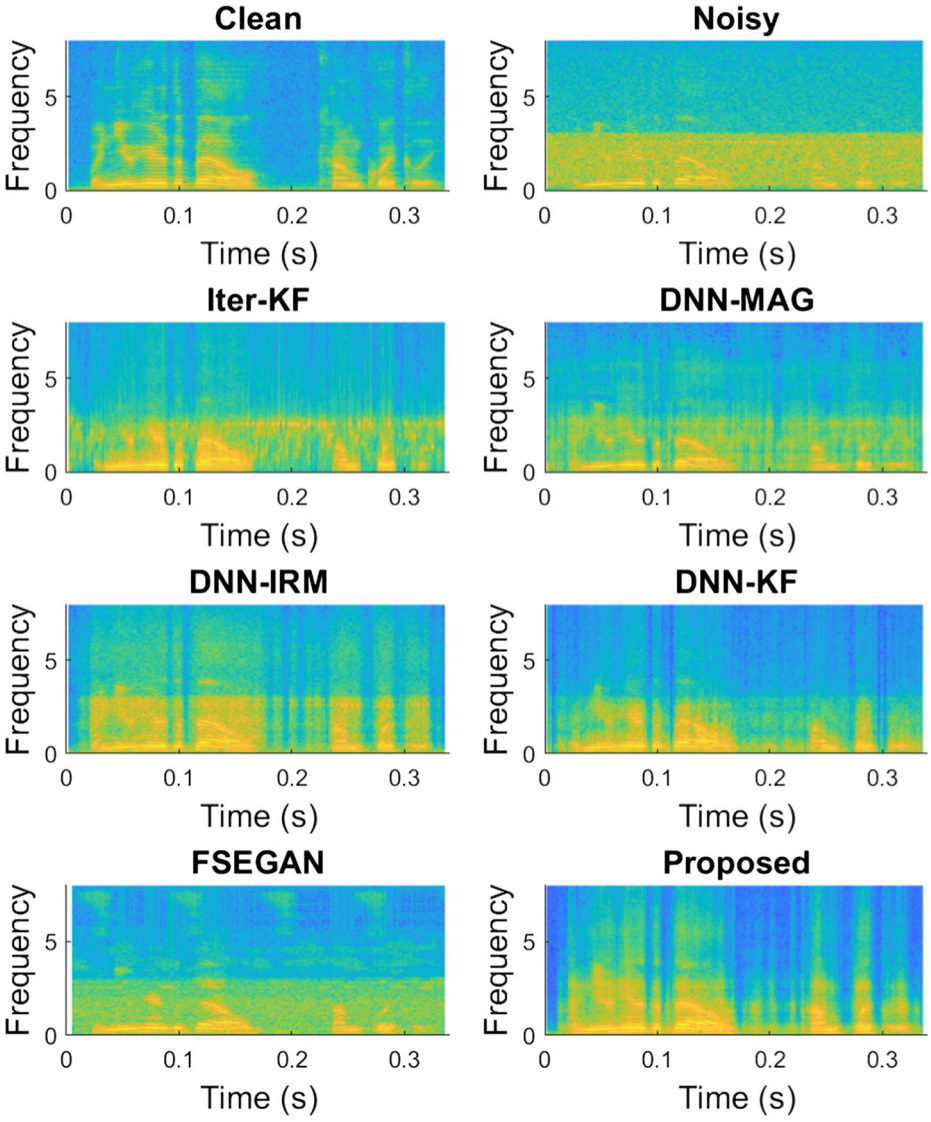

Fig. 9 Spectrograms of the clean, noisy and enhanced speech signals for different methods

4 Conclusion

In this paper, we have proposed a hybrid speech enhancement system consisting of the DNN based speech reconstruction followed by Kalman filtering, in order to improve enhancement performance under unseen noise conditions. Instead of focusing on training with as many kinds of noise types as possible to improve the generalization capability, our system first reconstructs the speech with the estimated magnitude spectrum from the DNN and the noisy phase spectrum. Kalman filtering is then applied to further remove the residual noise. By doing so, the proposed hybrid system is more capable to cope with unseen noise in

real-world environments. In addition, the use of DNN-based LSFs estimation along with the reconstructed speech provide more accurate parameters for Kalman filtering, thus leading to a better denoising performance. Finally, the hybrid system involves time domain as well as frequency domain processing, which could be regarded as as a form of joint estimation for both the magnitude and phase short-time spectra. Experiments show that the proposed hybrid system can achieve significant improvements in PESQ and STOI scores as compared with the traditional Kalman filtering, as well as more recent DNN and GAN based methods across different unseen noise conditions.

Acknowledgements The work was supported by NSERC of Canada under a CRD grant sponsored by Microchip in Ottawa, Canada. H. Yu also acknowledges the financial support from the China Scholarships Council (CSC No.201606270200).

References

1. Donahue C, Li B, Prabhavalkar R (2018) Exploring speech enhancement with generative adversarial networks for robust speech recognition. In: IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp 5024–5028
2. Fu SW, Hu T, Tsao Y, Lu X (2017) Complex spectrogram enhancement by convolutional neural network with multi-metrics learning. In: Machine Learning for Signal Processing (MLSP), pp 1–6
3. Gannot S, Burshtein D, Weinstein E (1998) Iterative and sequential Kalman filter-based speech enhancement algorithms. IEEE Transactions on Speech and Audio Processing 6(4):373–385
4. Gibson JD, Koo B, Gray SD (1991) Filtering of colored noise for speech enhancement and coding. IEEE Trans. Signal Process. 39(8):1732–1742
5. Han W, Zhang X, Min G, Sun M, Yang J (2016) Joint optimization of audible noise suppression and deep neural networks for single-channel speech enhancement. In: IEEE International Conference on Multimedia and Expo (ICME), pp 1–6
6. IEEE Subcommittee (1969) IEEE recommended practice for speech quality measurements. IEEE Transactions on Audio and Electroacoustics 17:225–246
7. 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 P862
8. Jeon KM, Park SY, Chun CJ, Park NI, Kim HK (2017) Multi-band approach to deep learning-based artificial stereo extension. ETRI J. 39(3):398–405
9. Kavalekalam MS, Christensen MG, Gran F, Boldt JB (2016) Kalman filter for speech enhancement in cocktail party scenarios using a codebook-based approach. In: IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp 191–195
10. Krawczyk M, Gerkmann T (2014) STFT phase reconstruction in voiced speech for an improved single-channel speech enhancement. IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP) 22(12):1931–1940
11. Lim JS, Oppenheim AV (1979) Enhancement and bandwidth compression of noisy speech. Proc. IEEE 67(12):1586–1604
12. Loweimi E, Barker J, Hain T (2017) Statistical normalisation of phase-based feature representation for robust speech recognition. In: IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp 5310–5314
13. McLoughlin IV (2008) Line spectral pairs. Signal processing 88(3):448–467
14. Mellahi T, Hamdi R (2015) LPC-based formant enhancement method in Kalman filtering for speech enhancement. AEU-International Journal of Electronics and Communications 69(2):545–554
15. Moattar MH, Homayounpour MM (2009) A simple but efficient real-time voice activity detection algorithm. In: European Signal Processing Conference (EUSIPCO), pp 2549–2553
16. Narayanan A, Wang D (2013) Ideal ratio mask estimation using deep neural networks for robust speech recognition. In: IEEE Int. Conference on Acoustics, Speech and Signal Processing (ICASSP), pp 7092–7096
17. Nicolson A, Paliwal KK (2019) Deep learning for minimum mean-square error approaches to speech enhancement. Speech Comm. 111:44–55
18. Nie S, Liang S, Liu B, Zhang Y, Liu W, Tao J (2018) Deep noise tracking network: A hybrid signal processing/deep learning approach to speech enhancement. In: Proceedings of Interspeech, pp 3219–3223

19. Nower N, Liu Y, Unoki M (2015) Restoration scheme of instantaneous amplitude and phase using Kalman filter with efficient linear prediction for speech enhancement. Speech Comm. 70:13–27
20. Ouyang Z, Yu H, Zhu WP, Champagne B (2018) A deep neural network based harmonic noise model for speech enhancement. In: Proceedings of Interspeech, pp 3224–3228
21. Ouyang Z, Yu H, Zhu WP, Champagne B (2019) A fully convolutional neural network for complex spectrogram processing in speech enhancement. In: IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp 5756–5760
22. Paliwal K, Basu A (1987) A speech enhancement method based on Kalman filtering. In: IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), vol 12, pp 177–180
23. Paliwal K, Wójcicki K, Shannon B (2011) The importance of phase in speech enhancement. Speech Comm. 53(4):465–494
24. Pascual S, Bonafonte A, Serra J (2017) SEGAN: Speech enhancement generative adversarial network. In: Proceedings of Interspeech, pp 3642–3646
25. Roy SK, Zhu WP, Champagne B (2016) Single channel speech enhancement using subband iterative Kalman filter. In: IEEE International Symposium on Circuits and Systems (ISCAS), pp 762–765
26. Shi G, Shanechi MM, Aarabi P (2006) On the importance of phase in human speech recognition. IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP) 14(5):1867–1874
27. So S, Wójcicki KK, Lyons JG, Stark AP, Paliwal KK (2009) Kalman fitler with phase spectrum compensation algorithm for speech enhancement. In: IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp 4405–4408
28. Soni MH, Shah N, Patil HA (2018) Time-frequency masking-based speech enhancement using generative adversarial network. In: IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp 5039–5043
29. Srivastava A, Valkov L, Russell C, Gutmann MU, Sutton C (2017) veegan: Reducing mode collapse in gans using implicit variational learning. In: Advances in Neural Information Processing Systems, pp 3308–3318
30. Taal CH, Hendriks RC, Heusdens R, Jensen J (2011) An algorithm for intelligibility prediction of time-frequency weighted noisy speech. IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP) 19(7):2125–2136
31. Tu M, Zhang X (2017) Speech enhancement based on deep neural networks with skip connections. In: IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp 5565–5569
32. Varga A, Steeneken HJ (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 Comm. 12(3):247–251
33. Wan EA, Nelson AT (1999) Networks for speech enhancement. Handbook of neural networks for speech processing 139:1
34. Wang D, Lim J (1982) The unimportance of phase in speech enhancement. IEEE Transactions on Acoustics, Speech, and Signal Processing 30(4):679–681
35. Wang Q, Muckenhirn H, Wilson K, Sridhar P, Wu Z (2019) VoiceFilter: Targeted voice separation by speaker-conditioned spectrogram masking. In: Proceedings of Interspeech, pp 2728–2732
36. Wang Y, Han K, Wang D (2013) Exploring monaural features for classification-based speech segregation. IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP) 21(2):270–279
37. Wang Y, Narayanan A, Wang D (2014) On training targets for supervised speech separation. IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP) 22(12):1849–1858
38. Williamson D, Wang D (2017) Time-frequency masking in the complex domain for speech dereverberation and denoising. IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP) 25(7)
39. Williamson DS, Wang Y, Wang D (2016) Complex ratio masking for monaural speech separation. IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP) 24(3):483–492
40. Xia Y, Wang J (2015) Low-dimensional recurrent neural network-based Kalman filter for speech enhancement. Neural Netw. 67:131–139
41. Xu Y, Du J, Dai LR, Lee CH (2013) An experimental study on speech enhancement based on deep neural networks. IEEE Signal Processing Letters 21(1):65–68
42. Xu Y, Du J, Dai LR, Lee CH (2015) A regression approach to speech enhancement based on deep neural networks. IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP) 23(1):7–19
43. Yu H, Ouyang Z, Zhu WP, Champagne B (2019) A deep neural network based Kalman filter fortime domain speech enhancement. In: IEEE International Symposium on Circuits and Systems (ISCAS), pp 1–5
44. Zheng N, Zhang XL (2018) Phase-aware speech enhancement based on deep neural networks. IEEE/ACM Transactions on Audio, Speech, and Language Processing (TASLP) 27(1):63–76

Hongjiang Yu received the B.E. and M.S. degrees from Wuhan University, Wuhan, China, in 2016. He is currently pursuing his Ph.D. degree in Department of Electrical and Computer Engineering, Concordia University, Montreal, Canada. His research interests include speech enhancement, speech/audio quality assessment, and machine learning to problems in speech and audio processing.

Wei-Ping Zhu (SM'97) received the B.E. and M.E. degrees from Nanjing University of Posts and Telecommunications, and the Ph.D. degree from Southeast University, Nanjing, China, in 1982, 1985, and 1991, respectively, all in electrical engineering. He was a Postdoctoral Fellow from 1991 to 1992 and a Research Associate from 1996 to 1998 with the Department of Electrical and Computer Engineering, Concordia University, Montreal, Canada. During 1993–1996, he was an Associate Professor with the Department of Information Engineering, Nanjing University of Posts and Telecommunications. From 1998 to 2001, he worked with hi-tech companies in Ottawa, Canada, including Nortel Networks and SR Telecom Inc. Since July 2001, he has been with Concordia's Electrical and Computer Engineering Department as a full-time faculty member, where he is presently a Full Professor. His research interests include digital signal processing and machine learning, speech and statistical signal processing, and signal processing for wireless communication with a particular focus on MIMO systems and cooperative communication.

Dr. Zhu served as an Associate Editor for the IEEE Transactions on Circuits and Systems Part I: Fundamental Theory and Applications during 2001-2003, an Associate Editor for Circuits, Systems and Signal Processing during 2006-2009, and an Associate Editor for the IEEE Transactions on Circuits and Systems Part II: Transactions Briefs during 2011-2015. He was also a Guest Editor for the IEEE Journal on Selected Areas in Communications for the special issues of: Broadband Wireless Communications for High Speed Vehicles, and Virtual MIMO during 2011-2013. He was an Associate Editor of Journal of The Franklin Institute (JFI) during 2015-2019. Since January 2020, he has been a Subject Editor of JFI. Dr. Zhu was the Secretary of Digital Signal Processing Technical Committee (DSPTC) of the IEEE Circuits and System Society during June 2012-May 2014, and the Chair of the DSPTC during June 2014-May 2016.

Zhiheng Ouyang received his B.E. degree from Hangzhou Dianzi University, Hangzhou, China, in 2017. He is pursuing his M.A.Sc. at Concordia University, Montreal, Canada. He is interested in machine learning for speech and audio processing.

Benoit Champagne received the B.Ing, in Engineering Physics from Ecole Polytechnique of Montreal (1983), the M.Sc. in Physics from University of Montreal (1985), and the Ph.D. in Electrical Engineering from University of Toronto (1990). From 1990 to 1999, he was Assistant and then Associate Professor at INRS-Telecommunications, Montreal. In 1999, he joined McGill University, where he is now a Full Professor in the ECE Department. His research interests include statistical signal processing and wireless communications, where he has coauthored more than 300 publications. He has been Associate Editor for the IEEE Signal Processing Letters and the IEEE Trans. on Signal Processing.