High-frequency Component Restoration for Kalman Filter Based Speech Enhancement

Hongjiang Yu[1], Wei-Ping Zhu[1], and Benoit Champagne[2]

[1]Dept. of Electrical and Computer Engineering, Concordia University, Montreal, Canada
[2]Dept. of Electrical and Computer Engineering, McGill University, Montreal, Canada
Email: *ho_yu@encs.concordia.ca, weiping@ece.concordia.ca, benoit.champagne@mcgill.ca*

***Abstract*—In this paper, we present a deep neural network (DNN) based algorithm to restore the high-frequency (HF) component of the enhanced speech processed by Kalman filtering, where the DNN is applied for estimating the magnitude of HF component from the low-frequency (LF) counterpart. The complete HF component is then computed with the estimated magnitude given by the DNN and the phase of the Kalman filtered speech. By incorporating our restoration algorithm into Kalman filter based speech enhancement method, our new speech enhancement system is able to recover the HF component with better perceptual quality and less distortion. Experimental results demonstrate that the proposed method outperforms the state-of-the-art Kalman filter based method in terms of both speech quality and intelligibility.**

***Index Terms*—speech enhancement, Kalman filter, deep neural network, speech bandwidth expansion**

I. Introduction

To achieve better speech quality and improve user experience in speech processing related applications, such as speech recognition, hearing aids and smart home devices, speech enhancement has been often adopted as pre-processing to remove the background noises. Various methods have been proposed in the past few decades, among which Kalman filtering has been of great interest since it is able to process non-stationary noisy speech and produce enhanced speech without musical noise.

In Kalman filter based methods, the auto-regressive model is widely adopted as the speech production model and is incorporated in the Kalman recursion equations. As such, the performance of Kalman filtering is considerably dependent on the accuracy of the parameter estimation, i.e., linear prediction coefficients (LPCs) and excitation variance. A piece of pioneering work employing Kalman filter to remove white Gaussion noise was found in [1], which achieves better performance as compared with the Wiener filter [2]. It should be pointed out that the parameters involved in the Kalman filter in [1] are extracted from the clean speech rather than the noisy speech, which is not available in practical applications. On the other hand, directly computing the parameters from the corrupted speech would be inaccurate and unreliable, leading to a performance degradation. As such, several algorithms have been proposed to estimate the parameters.

The authors in [3]–[5] have proposed to estimate online the speech parameters and enhanced speech simutaneously from the noisy observations. In particular, the expectation-maximization (EM) algorithm in [3] is the most typical procedure, that iterates between Kalman filtering of noisy samples and estimation of the speech parameters. In each iteration the Kalman filter enhances the speech to obtain better parameter estimation, and this method generally improves the final results after a few iterations. Other researchers such as those in [6]–[8] used an off-line training approach to predict the speech parameters based on a training database beforehand. Recently in [8], a deep neural network (DNN) is utilized to explore the relationship between the speech parameters of the noisy speech and those of clean speech. By taking advantages of the powerful learning capability of a deep model and a large training database, the DNN based Kalman filtering achieves significant improvement over conventional iterative Kalman filter [9].

Despite the performance gain from the Kalman filter based methods, it is found that the enhanced speech suffers from the loss or attenuation of its high-frequency component. To address this problem, subband Kalman filters have been investigated in [9], [10], wherein the noisy speech is decomposed into high-frequency (HF) and low-frequency (LF) components. The iterative Kalman filters with different parameters are then applied into the HF subband and LF subband separately. Experimental results demonstrate that the subband Kalman filter algorithm outperforms the fullband counterpart. However, the HF component is still suppressed relative to the LF component. In other words, the desired speech in the HF subband is removed together with the noise when conducting Kalman filtering.

In this paper, we propose a HF component restoration algorithm for Kalman filter based speech enhancement to further improve the performance. Inspired by the speech bandwidth expansion [11], the Kalman filtering denoised speech is first divided into HF and LF components. The LF component, which is considered to be of good quality, is then used to restore the HF component. At last, the enhanced speech is resynthesised by the LF component of the Kalman filter denoised speech and the recovered HF component. For the HF component restoration, a DNN is trained to learn the relationship between the log-magnitude of the LF component and that of the HF component. The full HF component is reconstructed with the estimated magnitude and the phase of

978-1-7281-3320-1/20/$31.00 ©2020 IEEE

the Kalman filter denoised speech. Experimental results show that our improved Kalman filter based speech enhancement method yields better speech quality and intelligibility than the one without HF component restoration.

II. Kalman Filter Based Speech Enhancement

A. Autoregressive Speech Model

Before introducing Kalman filtering, speech models should be defined first. The first one is the noisy speech model $y(n)$, which can be regarded as a mixture of the clean speech $s(n)$ and the additive noise $w(n)$,

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

where n is the discrete time index.

Secondly, the clean speech $s(n)$ is usually represented by a source-filter model in Kalman filter based method. A widely-adopted model is the autoregressive (AR) form,

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

where a_i are LPCs of the speech, p the order of the model, and $v(n)$ the driving white noise with variance σ^2.

B. Kalman Filtering

The speech models are expressed in matrix and vector notations to facilitate the presentation of the Kalman filter based speech enhancement, namely,

$$\begin{cases} \mathbf{u}(n) = F\mathbf{u}(n-1) + G\mathbf{v}(n) \\ \mathbf{y}(n) = H\mathbf{u}(n) + \mathbf{w}(n) \end{cases} \tag{3}$$

where the transition matrix F is given by

$$F = \begin{bmatrix} 0 & 1 & 0 & \cdots & 0 & 0 \\ 0 & 0 & 1 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & 0 & 1 \\ a_p & a_{p-1} & a_{p-2} & \cdots & a_2 & a_1 \end{bmatrix}, \tag{4}$$

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

$$\begin{aligned} \mathbf{u}(n) &= [s(n-p+1), \ldots, s(n-1), s(n)]^T \\ \mathbf{y}(n) &= [y(n-p+1), \ldots, y(n-1), y(n)]^T \\ \mathbf{w}(n) &= [w(n-p+1), \ldots, w(n-1), w(n)]^T \\ \mathbf{v}(n) &= [v(n-p+1), \ldots, v(n-1), v(n)]^T \end{aligned} \tag{5}$$

The denoising process of the Kalman filtering is summarized by the following equations

$$\begin{cases} e(n) = y(n) - G^T \hat{\mathbf{u}}(n|n-1) \\ K(n) = P(n|n-1)\left(R_w + P(n|n-1)\right)^{-1} \\ \hat{\mathbf{u}}(n|n) = \hat{\mathbf{u}}(n|n-1) + K(n)\,\mathbf{e}(n) \\ P(n|n) = (I - K(n))\,P(n|n-1) \\ \hat{\mathbf{u}}(n+1|n) = F\hat{\mathbf{u}}(n|n) \\ P(n+1|n) = FP(n|n)F^T + \sigma_v^2 GG^T \end{cases} \tag{6}$$

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

$$d(n) = G^T \hat{\mathbf{u}}(n|n) \tag{7}$$

C. Parameter Estimation

Several parameters in the above equations should be estimated accurately in order to achieve excellent performance of Kalman filtering. Those parameters include the driving noise variance σ^2, the covariance matrix of the additive noise R_w, and the transition matrix F which contains the LPCs of the speech signal model.

At first, we adopte the DNN based algorithm presented in [8] for the LPCs prediction, in which LPCs are calculated from speech databases and converted into line spectrum frequencies (LSFs). A DNN is then trained off-line for learning the relationship from the noisy feature set (including noisy LSFs and other four acoustic features) to clean LSFs. The estimated LSFs are predicted by the well-trained DNN and transformed back to LPCs for performing Kalman filtering.

Secondly, the covariance matrix can be approximately estimated during the speech-absent frames:

$$R_w = E\left[\mathbf{w}(n)\,\mathbf{w}^T(n)\right] \tag{8}$$

Finally, according to [4], the variance of the driving noise $v(n)$ can be estimated by means of:

$$\sigma_v^2 = E\left[y(n)^2\right] - \mathbf{r_y}^{\mathbf{T}}\hat{\mathbf{a}} - \hat{\sigma}_{\mathbf{w}}^{\mathbf{2}} \tag{9}$$

where $\mathbf{r_y} = E\left[\mathbf{y}(n)\,y^T(n)\right]$, $\hat{\mathbf{a}}$ is the LPC vector and σ_w^2 is the variance of additive noise.

III. Proposed Speech Enhancement System

The overall block diagram of our speech enhancement system with Kalman filtering and HF component restoration is depicted in Fig.1. The system is composed of the off-line training stage and the enhancement stage. In the training stage, a DNN is trained to learn the mapping from the log-magnitude of the LF component to that of HF component for clean speech. In the enhancement stage, the noisy speech is first processed by a Kalman filter to obtain denoised speech. Subband analysis is followed to decompose the denoised speech into HF and LF components. Then, the estimated HF component is recovered with the estimated magnitude predicted by the well-trained DNN and the phase from the denoised speech. Finally, the enhanced speech is obtained by an combination of the estimated HF component and the LF component of the denoised speech.

A. Training Stage

Before training the DNN, the clean speech $s(n)$ is divided into HF component $s_H(n)$ and LF component $s_L(n)$ by subband analysis. The corresponding short-time Fourier transform (STFT) spectrograms are defined by $S_H(k,l)$ and

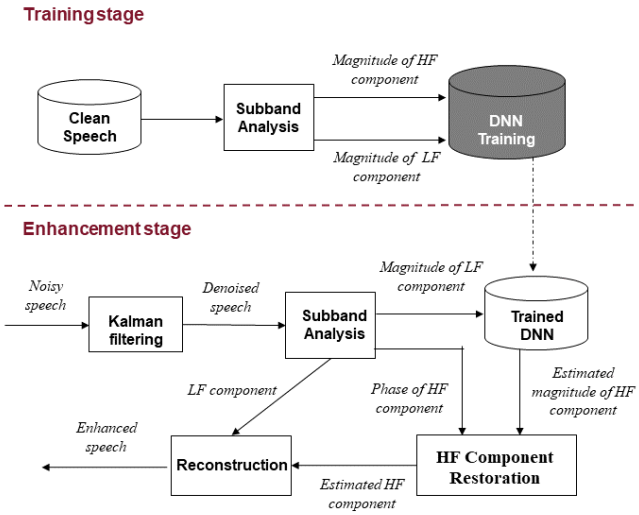

Fig. 1. A block diagram of proposed speech enhancement system.

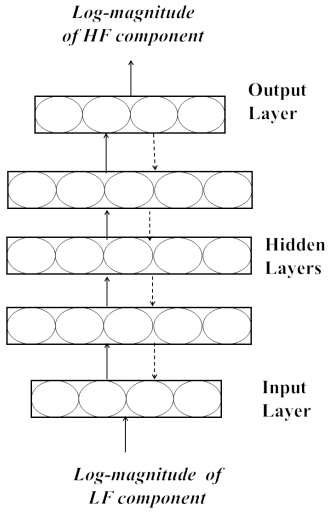

Fig. 2. Structure of the proposed DNN for HF component magnitude estimation.

$S_L(k,l)$, with k and l denote the frequency bin and frame index, respectively. For simplification, the index k and l will be omitted in the remaining discussion. The magnitude of S_L is extracted as the input feature of DNN, while that of S_H is set as the training target. Since the magnitude spectrum usually has a very large dynamic range, the log-function and normalization are adopted to compress both the feature and target for better training.

Besides the feature of the current frame, the neighbouring time frames are incorporated to form an extended feature set, in order to make full use of the temporal information of the speech. As such, the feature vector centred at the kth frame is defined by $\tilde{\mathbf{F}}(k) = [\mathbf{F}(k-p), \ldots, \mathbf{F}(k), \ldots, \mathbf{F}(k+p)]$, with p donating the number of neighbouring frames involved on each side.

A fully-connected feed-forward DNN as depicted in Fig.2 is employed for training. The DNN has three hidden layers with 1024 units in each layer between the input layer and the output layer. The activation function used in the hidden layer is the rectified linear unit (ReLU), while a linear function is used in the output layer

To update weights and biases until the network is able to achieve good performance, back propagation following a gradient-based optimization algorithm is commonly adopted. Back propagation computes the gradient, whereas stochastic gradient descent uses the gradients to train the DNN model, so as to minimize the value of the cost function, which is defined as the mean square error between the reference and the estimated log-magnitude spectrogram of the HF component

$$MSE = \frac{1}{M}\sum_{m=1}^{M}\left[\left(ln|\hat{S}_H| - ln|S_H|\right)^2\right] \qquad (10)$$

where M denotes the speech frames, $|\hat{S}_H|$ the estimated magnitude and $|S_H|$ the reference one. The well-trained DNN will be used in the enhancement stage to obtain the estimated magnitude of the HF component from that of the LF component.

B. Enhancement Stage

The procedure of the proposed speech enhancement system can be briefly summarized as following steps. Firstly, Kalman filtering introduced in Section II is applied to the noisy speech $y(n)$ for denosing in time-domain. The denoised speech $d(n)$ is then decomposed into HF component $d_H(n)$ and LF component $d_L(n)$ by subband analysis.

Secondly. the aforementioned DNN based HF component restoration algorithm is required to compensate the distortion in $d_H(n)$. Here, the LF component of the denoised speech is employed as input in restoration for the reason that the LF component is of high quality after Kalman filtering. The STFT spectrogram $\hat{D}_H$ of the recovered HF component is reconstructed with the estimated magnitude given by the well-trained DNN and the phase of $d_H(n)$, i.e., $\hat{D}_H = |\hat{D}_H|e^{j\phi_{dh}}$. The inverse STFT is performed to achieve the estimated HF component $\hat{d}_H(n)$.

Finally, the enhanced speech $\hat{s}(n)$ is synthesised with the unprocessed LF component $d_L(n)$ and the recovered HF component $\hat{d}_H(n)$, i.e., $\hat{s}(n) = d_L(n) + \hat{d}_H(n)$.

IV. Experimental Results

A. Experimental Setup

The IEEE corpus [12] is selected as clean speech database, in which 670 utterances are used for training and 50 utterances for testing. Four types of noises (pink, buccaneer2, destroyerengine, hfchannel) are picked from NOISEX-92 database [13] and mixed with the clean speech at four levels (-3dB, 0dB, 3dB, 6dB) of signal-to-noise ratio (SNR). In training stage, the DNN is trained only on clean speech database to explore the relationship between its LF and HF components. Since 670 utterances are not enough for deep learning, we repeat them

for 16 times to get 10720 utterances. In the enhancement stage, four noises are mixed with 50 clean testing speeches at the above four SNR levels. Thus, the number of noisy utterances in enhancement stage is 800. The sampling frequency for the speech and noise signals is set to 16 kHz. The window size of STFT is 320 with 50% overlap.

We compare the proposed new system with the conventional subband iterative Kalman filter (denoted as S-IKF) [9] and the recent DNN based Kalman filter [8] (denoted as DNN-KF), to verify the benefit of incorporating the HF component restoration. For fair comparison, the configuration in the Kalman filtering part of the new system is kept the same as in [8]. As such, the proposed new system can be regraded as the combination of the DNN-KF and HF component restoration. Two objective metrics are adopted in our experiment to assess the enhancement performance: the perceptual evaluation of speech quality (PESQ) [14] and the short-time objective intelligibility (STOI) [15]. PESQ is widely adopted in evaluating speech quality, while STOI is proved to be highly related to speech intelligibility.

B. Results and Comparison

Table I shows the average objective scores of the processed speeches. First of all, both the DNN-KF and proposed method show better overall performance than the S-IKF, and the improvement reflects the superiority of using DNN to predict LPC for Kalman filtering. In addition, it is shown that the speeches from the proposed system have better PESQ and STOI scores than DNN-KF, which demonstrates the advantage of introducing the HF component restoration. Finally, by comparing the results between the enhanced speeches from the proposed method with respect to input SNRs, it can be found that the improvement is greater at high SNRs. One possible reason for this phenomenon is that the quality of the denoised speech at high SNRs is better than the one at low SNRs, which is beneficial to the restoration of the HF component.

TABLE I
OBJECTIVE RESULTS ON NOISY SPEECHES

		-3dB	0dB	3dB	6dB
PESQ	Noisy	1.37	1.51	1.65	1.82
	S-IKF	1.52	1.68	1.83	2.00
	DNN-KF	1.73	2.01	2.21	2.38
	Proposed	**1.74**	**2.03**	**2.25**	**2.43**
STOI	Noisy	0.65	0.72	0.78	0.83
	S-IKF	0.66	0.73	0.79	0.84
	DNN-KF	0.71	0.77	0.82	0.85
	Proposed	**0.72**	**0.79**	**0.84**	**0.87**

In order to better illustrate the benefits of the HF restoration, the spectrograms of the enhanced speeches resulting from the DNN-KF and proposed method are plotted and compared in Fig.3. Comparing with DNN-KF, whose HF component is significantly suppressed, the proposed method gives better spectrogram of the processed speech that is similar to the original one in HF component. The high clarity of the harmonics in the HF component of our enhanced speech also indicates the superiority of HF component restoration.

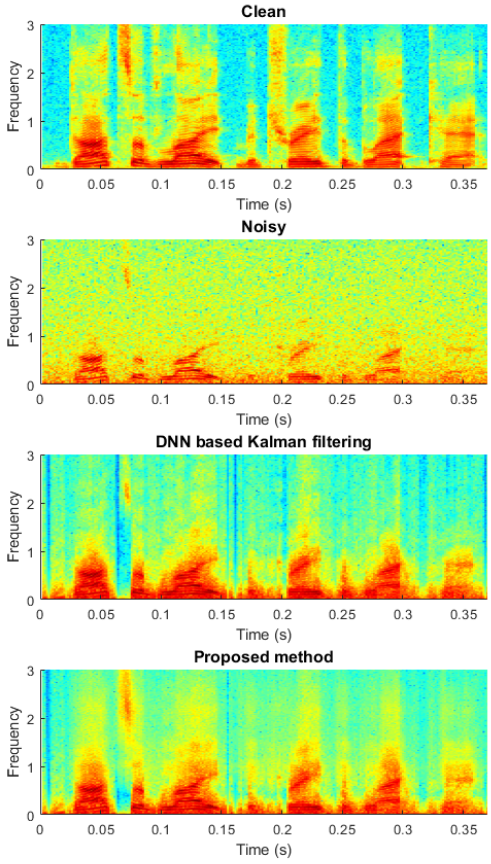

Fig. 3. Spectrograms of the clean, noisy and enhanced speeches.

V. CONCLUSION

In this paper, a DNN based HF component restoration algorithm has been proposed to improve the performance of Kalman filter based speech enhancement method, where the DNN is employed to explore the relationship between the magnitude of the LF component and that of the HF component. Benefiting from the HF component restoration, our new speech enhancement system is able to reduce the distortion in the HF component of the denoised speech, leading to a better speech quality as well as intelligibility compared to the existing DNN-KF method.

ACKNOWLEDGMENT

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

[1] K. Paliwal and A. Basu, "A speech enhancement method based on Kalman filtering," *in Proc. of IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 177–180, 1987.

[2] Y. Ephraim and D. Malah, "Speech enhancement using a minimum-mean square error short-time spectral amplitude estimator," *IEEE Trans. on Acoustics, Speech, and Signal Processing (TASLP)*, vol. 32, no. 6, pp. 1109–1121, 1984.
[3] J. D. Gibson, B. Koo, and S. D. Gray, "Filtering of colored noise for speech enhancement and coding," *IEEE Trans. on signal processing*, vol. 39, no. 8, pp. 1732–1742, 1991.
[4] Y. Xia and J. Wang, "Low-dimensional recurrent neural network-based Kalman filter for speech enhancement," *Neural Networks*, vol. 67, pp. 131–139, 2015.
[5] T. Mellahi and R. Hamdi, "LPC-based formant enhancement method in Kalman filtering for speech enhancement," *AEU-International Journal of Electronics and Communications*, vol. 69, no. 2, pp. 545–554, 2015.
[6] N. Nower, Y. Liu, and M. Unoki, "Restoration of instantaneous amplitude and phase using Kalman filter for speech enhancement," *in Proc. of IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 4633–4637, 2014.
[7] M. S. Kavalekalam, M. G. Christensen, F. Gran, and J. B. Boldt, "Kalman filter for speech enhancement in cocktail party scenarios using a codebook-based approach," *in Proc. of IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 191–195, 2016.
[8] H. Yu, Z. Ouyang, W.-P. Zhu, B. Champagne, and Y. Ji, "A deep neural network based kalman filter for time domain speech enhancement," *in Proc. of IEEE International Symposium on Circuits and Systems (ISCAS)*, pp. 1–5, 2019.
[9] S. K. Roy, W.-P. Zhu, and B. Champagne, "Single channel speech enhancement using subband iterative Kalman filter," *in Proc. of IEEE International Symposium on Circuits and Systems (ISCAS)*, pp. 762–765, 2016.
[10] W.-R. Wu and P.-C. Chen, "Subband Kalman filtering for speech enhancement," *IEEE Transactions on Circuits and Systems II: Analog and Digital Signal Processing*, vol. 45, no. 8, pp. 1072–1083, 1998.
[11] K. Li and C.-H. Lee, "A deep neural network approach to speech bandwidth expansion," *in Proc. of IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 4395–4399, 2015.
[12] IEEE Subcommittee, "IEEE recommended practice for speech quality measurements," *IEEE Trans. on Audio and Electroacoustics*, vol. 17, pp. 225–246, 1969.
[13] A. Varga and H. J. Steeneken, "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 Communication*, vol. 12, no. 3, pp. 247–251, 1993.
[14] ITU-R, "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*, 2001.
[15] C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen, "An algorithm for intelligibility prediction of time–frequency weighted noisy speech," *IEEE Trans. on Audio, Speech, and Language Processing (TASLP)*, vol. 19, no. 7, pp. 2125–2136, 2011.