A Deep Neural Network Based Kalman Filter for Time Domain Speech Enhancement

Hongjiang Yu[1], Zhiheng Ouyang[1], Wei-Ping Zhu[1], Benoit Champagne[2], and Yunyun Ji[1]

[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, z_ouyan*}*@encs.concordia.ca, weiping@ece.concordia.ca, benoit.champagne@mcgill.ca, jyy@ece.concordia.ca*

***Abstract*—In this paper, we present a novel deep neural network (DNN) based Kalman filter (KF) algorithm for speech enhancement, where DNN is applied for estimating key parameters in the KF, namely, the linear prediction coefficients (LPCs). By training the DNN with a large database and making use of the powerful learning ability of DNN, our proposed DNN-KF algorithm is able to estimate LPCs from noisy speech more accurately and robustly, leading to an improved performance as compared to traditional KF based approaches in speech enhancement. Experimental results demonstrate that our DNN-KF method outperforms two existing KF based speech enhancement methods in terms of both speech quality and intelligibility.**

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

I. Introduction

Speech enhancement (SE) has been extensively applied in a wide range of fields such as speech recognition, wireless communications, hearing aids and smart home devices, where the received input speech signals are often corrupted by different kinds of noises. The main purpose of SE is to improve speech quality and intelligibility, so as to obtain better user experience in those applications. Several methods have been put forward over the past decades, among which frequency-domain and time-domain algorithms are two prominent categories.

Frequency-domain algorithms nomally involve transforming the noisy speech into the frequency-domain via the discrete Fourier transform (DFT), and then approximating the clean speech spectrum based on the observed noisy spectrum. Finally, the estimated clean spectrum is then converted back to the time-domain by the inverse DFT. Within this category, Wiener filtering [1] and minimum mean square error (MMSE) amplitude estimators [2] are two of the most well-known techniques. Both of them are capable of estimating the spectral amplitude of clean speech, but their enhanced speech output usually suffers from musical and residual noise, due in part to the combination of the noisy phases with the estimated clean amplitudes during the reconstruction process.

In time-domain algorithms, the SE problem is viewed as a filtering problem, in which the enhancement filter is designed to reduce the additive noise corrupting the speech without introducing noticeable distortion in the enhanced speech output. Unlike most of the frequency-domain algorithms which only enhance the amplitude, time-domain algorithms implicitly enhance both amplitude and phase at the same time. The Kalman filter (KF) based on MMSE criterion [3] is a well-known time-domain SE method, in which the speech is modeled as an autoregressive process and the enhanced speech is obtained by Kalman filtering. In this context, the linear prediction coefficients (LPCs) are important parameters for the implementation of KF.

Early KF based algorithms such as [3] were limited to the reduction of white Gaussian noise, while the LPCs were predicted from the clean speech, which is, however, not accessible in practical applications. To overcome this obstacle, Koo and Gibson [4] suggested an approximate expectation-maximization (EM) algorithm that iterates between Kalman filtering of noisy samples and estimation of the speech parameters. To further improve the accuracy of the estimated parameters, several methods [5]–[8] were advanced recently, such as LPC-based formant enhancement [5] and codebook based KF approach [6]. These methods, which attempt to reduce the sensitivity of LPC prediction in the presence of noise, lead to better SE results.

Subband based KF algorithms have also been studied to further improve the SE performance. In [9], a subband iterative KF method was proposed wherein the noisy speech is decomposed into high-frequency (HF) and low-frequency (LF) subband components. An iterative KF is then applied in the HF subband, while the LF subband unprocessed, on the basis that the LF subband mainly contains the intelligible speech components; consequently, this approach cannot reduce non-negligible noise contained in the LF component. In [10], to address this limitation, a voice activity detection based adaptive threshold scheme is applied to each subband frame as pre-processing, and an iterative KF is then employed in each subband for further noise reduction. Experiments have shown that these subband KF algorithms can outperform their fullband counterparts.

In recent years, the deep neural network (DNN) based methods have produced solutions to complex problems that were previously unattainable with traditional signal processing techniques. In particular, the application of DNN to the SE problem has led to significant breakthroughs [11]–[13]. DNN based SE algorithms operate mostly in the frequency domain, and reconstruct the enhanced speech by combining the estimated speech magnitude with the phase of the noisy observations.

In this paper, a DNN based KF is proposed for time-

978-1-7281-0397-6/19/$31.00 ©2019 IEEE

domain speech enhancement, where DNN is trained to learn the relationship between the line spectrum frequencies (LSFs) of the noisy speech and those of the clean speech.. The estimated LSFs of the clean speech are transformed to LPCs and then applied into a KF for SE. In contrast to the aforementioned DNN based algorithms, our method operates in the time-domain and enhances both speech magnitude and phase. Apparently, the use of DNN as opposed to traditional processing methods allows a more accurate estimation of the clean speech's LPCs. Experimental results under various conditions of noise show that our proposed DNN based KF method can yield better speech quality and intelligibility than previous subband iterative KF based algorithms.

II. Kalman Filter for Speech Enhancement

A. Speech Models

Consider a time-domain noisy speech $y(n)$ as given by

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

where $s(n)$ and $w(n)$ represent the clean speech and the additive noise respectively. In KF based SE algorithms, the clean speech signal $s(n)$ is usually considered as the output of an autoregressive (AR) process. The p-th order AR speech model is represented as follows,

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

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

To facilitate the presentation of KF based SE, we make use of matrix notation and introduce the following vectors: the speech state vector $\mathbf{u}(n)$, noisy speech vector $\mathbf{y}(n)$, additive noise vector $\mathbf{w}(n)$ and driving noise vector $\mathbf{v}(n)$, which are respectively given by

$$\begin{aligned} \mathbf{u}(n) &= [s(n-p+1), \ldots, s(n-1), s(n)]^T \\ \mathbf{y}(n) &= [y(n-p+1), \ldots, y(n-1), y(n)]^T \\ \mathbf{w}(n) &= [w(n-p+1), \ldots, w(n-1), w(n)]^T \\ \mathbf{v}(n) &= [v(n-p+1), \ldots, v(n-1), v(n)]^T \end{aligned} \tag{3}$$

using the transition matrix F as defined by

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

A speech signal is then expressed as

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

where H is a pth order identity matrix and $G = [0, \cdots, 0, 1]^T \in R^p$.

B. Process Equations

The process of recovering the desired signal from the observed noisy speech can be seen as a Kalman filtering problem [3], which can be 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 MMSE estimate of the state vector $\mathbf{u}(n)$ given the past observations $y(1), \ldots, y(n-1)$, $P(n|n)$ the filtered state error covariance matrix, and $P(n|n-1)$ the predicted state error correlation matrix. The speech estimate at discrete-time n can finally be given by

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

It can be seen from the above equations that several parameters should be calculated before performing 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.

III. Proposed Speech Enhancement System

The overall block diagram of our DNN based SE system with KF is depicted in Fig.1. It consists of two stages, namely: training stage and enhancement stage. In the training stage, a DNN is trained to learn the mapping from the noisy LSFs to the clean ones. In the enhancement stage, a KF with the the DNN-based estimated parameters is applied to the noisy speech to obtain the enhanced speech.

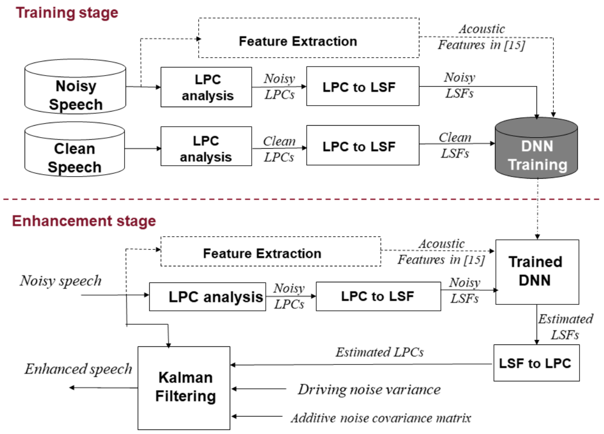

Fig. 1. A bolck diagram of proposed speech enhancement system.

A. Training Stage

LPCs are calculated using both noisy and clean speech databases, and then converted into LSFs. The noisy LSFs are used as input features to the DNN, while the clean LSFs are used as output targets for the DNN. The reason for using LSFs instead of LPCs in our algorithm is that LSFs have a well-behaved dynamic range, while LPCs have a large dynamic range of values. Therefore, the stability of the training stage is easier to guarantee in the LSF domain [14].

In order to better explore the relationship between the noisy and clean LSFs, we would first like to investigate the use of other possible acoustic features in combination with the LSFs to form an extended input feature set. In [15], the following four feature types are shown to have good performance when acting as input to DNN. They are 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 filter-bank energies (GF) and their deltas. Then, we will investigate the performance when these feature types are combined with LSFs as our input feature set.

For supervised training, the architecture adopted in our method is a feedforward neural network with many levels of non-linear units to represent a highly non-linear regression function that maps noisy LSFs to clean ones. As Fig. 2 depicts, our DNN is composed of one input layer, one output layer and three hidden layers with 1024 units in each layer. This structure has been verified to yield the best results in [14]. The rectified linear unit (ReLU) model is employed for the hidden layers, while the linear model is used for the output layer.

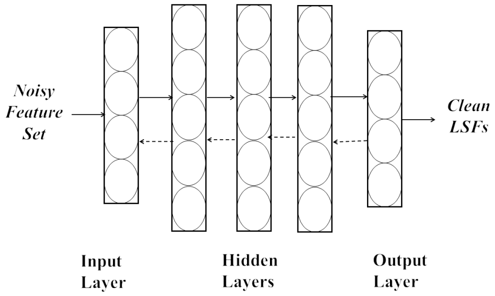

Fig. 2. Structure of the proposed DNN for LSF estimation..

Back propagation with the MMSE-based cost function between the estimated clean LSFs and the reference clean LSFs is adopted to train the DNN. During the training, our DNN can automatically learn the complex mapping from noisy LSFs to clean LSFs given sufficient training samples. The well-trained DNN will be used in the enhancement stage to obtain estimated clean LSFs from the noisy LSFs.

B. Enhancement Stage

Considering an unknown input noisy speech, the Kalman filtering parameters need to be calculated beforehand to employ our DNN based KF speech enhancement method.

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

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

Then according to [7], 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 $\mathbf{r_y} = E\left[\mathbf{y}\left(n\right)y^T\left(n\right)\right]$, $\hat{\mathbf{a}}$ is the LPC vector and σ_w^2 is the variance of additive noise.

The proposed DNN based KF algorithm is described as follows: estimating clean LSFs from noisy LSFs with the proposed DNN, and then converting them to LPCs to form the state transition matrix F. Then we compute the covariance matrix of the measured noise using (8), and the driving noise variance using (9). At last, performing the Kalman filtering defined in (4) to obtain $\hat{\mathbf{u}}\left(n|n\right)$. Finally, The enhanced speech is given by: $\hat{s}\left(n\right) = G^T\hat{\mathbf{u}}\left(n|n\right)$.

IV. Experimental Results

A. Experimental Setup

In this study, the clean speech is selected from the IEEE corpus [16]. We choose 670 utterances for training and 50 utterances for testing. Six types of noises are picked from NOISEX-92 database [17]. Among them four types (babble, white, street, factory) are regarded as seen noises, and the other two (pink, car) as unseen noise. In the training stage, the noisy speech are obtained by mixing clean training speech with seen noises at four levels (-5dB, 0dB, 5dB, 10dB) of signal-to-noise rates (SNRs) , which results in 10720 utterances. In the enhancement stage, both seen noises and unseen noises are mixed with clean testing speeches at the above four SNR levels. The number of noisy utterances used in enhancement stage is 800 for seen noise and 400 for unseen noise, respectively. The sampling frequency for the speech and noise signals is set to 16kHz. A rectangular window is used to divide the audio signals into 20 ms frames with no overlap. In the implementation of the KF algorithm we initialize with $\mathbf{u}(0|0) = \mathbf{0}$ and $P(0|0)$ be identity matrix, and set the speech AR order $p = 12$.

To assess the enhancement performance, two objective metrics are adopted in our experiment: the perceptual evaluation of speech quality (PESQ) [18] and the short-time objective intelligibility (STOI) [19]. PESQ focuses on evaluating speech quality while STOI provides a measure of speech intelligibility.

B. Feature Set

At first, we investigate three different input feature sets in order to learn the best possible mapping between the noisy and clean LSFs. Specifically, we consider the following: the LSF-only set, the multi-feature set (AMS+RASTA-PLP+MFCC+GFCC) in [15], and the joint set, which is formed by combining the LSF-only set with the multi-feature set. The objective assessment results of the enhanced speech

TABLE I
OBJECTIVE RESULTS WITH DIFFERENT FEATURE SETS

		-5dB	0dB	5dB	10dB
PESQ	Noisy	1.26	1.53	1.79	2.12
	LSF-only Set	1.53	1.97	2.30	2.56
	Multi-feature Set	1.58	2.01	2.37	2.61
	Joint Set	**1.62**	**2.03**	**2.41**	**2.65**
STOI	Noisy	0.62	0.72	0.82	**0.89**
	LSF-only Set	0.63	0.74	0.82	0.87
	Multi-feature Set	0.66	0.76	**0.84**	0.88
	Joint Set	**0.67**	**0.77**	**0.84**	**0.89**

when using the above three feature sets as DNN input are shown in Table I.

It is observed that the enhanced speech from the joint set achieves the highest PESQ and STOI scores; this is because the joint set with more acoustic features contains more information about the speech, where each feature has its own advantages. As a result, the joint set is selected to be our feature set in the remaining experiments.

C. Results and Comparison

Controlled experiments are conducted to evaluate our proposed DNN-KF algoirthm. Two existing KF based SE algoirthms are adopted as reference methods for comparison, i.e.: the subband iterative KF (S-IKF) [9] and the adaptive threshold iterative KF (AT-IKF) method [10].

a) Seen noises: Table II gives the average objective score of different KF based SE algorithms on seen noises. Obviously, the DNN-KF outperforms the other two KF algorithms in most cases. It can be inferred that using DNN for predicting clean LPCs contributes to improving the performance of KF based SE algorithms. However, the S-IKF gives the best STOI score in the case of input SNR 10dB. This improvement of speech intelligibility is achieved by the subband processing in the S-IKF. Although our DNN-KF does not employ subband processing, it still yields good STOI scores.

TABLE II
OBJECTIVE RESULTS ON SEEN NOISY SPEECHES

		-5dB	0dB	5dB	10dB
PESQ	Noisy	1.26	1.53	1.79	2.12
	S-IKF	135	1.67	1.97	2.28
	AT-IKF	1.37	1.69	2.04	2.37
	DNN-KF	**1.62**	**2.03**	**2.41**	**2.66**
STOI	Noisy	0.62	0.72	0.82	0.89
	S-IKF	0.62	0.73	0.82	**0.90**
	AT-IKF	0.54	0.65	0.75	0.82
	DNN-KF	**0.67**	**0.77**	**0.84**	0.89

b) Unseen noises: Table III shows the average objective scores of different KF based SE algorithms on unseen noises. Firstly, the highest results demonstrate that the DNN-KF gives the best performance even under the unseen noise environment. Compared to seen noise, the improvements of the objective score on the enhanced speech decrease a bit, because the prediction error of LPCs increases under the case of unseen noise. Benefiting from the large training database, it can also be found that the performance of DNN-KF is relatively stable under different background noises in comparison with other two KF methods. In addition, the S-IKF achieves better performance than AT-IKF in these two unseen noises because the threshold of AT-IKF can not be accurately set, especially at low SNR. At last, it should be mentioned that the noisy speech has the best STOI score when the input SNR is 10dB, which means the KF based SE algorithms introduce distortion to the desired speech while reducing noises. Although the STOI scores of DNN-KF drop for unseen noise, they still remain high, indicating a good speech intelligibility.

TABLE III
OBJECTIVE RESULTS ON UNSEEN NOISY SPEECHES

		-5dB	0dB	5dB	10dB
PESQ	Noisy	1.30	1.54	1.84	2.19
	S-IKF	1.38	1.62	1.93	2.23
	AT-IKF	1.25	1.60	1.95	2.30
	DNN-KF	**1.57**	**2.04**	**2.39**	**2.63**
STOI	Noisy	0.60	0.70	0.81	**0.89**
	S-IKF	0.60	0.71	0.81	**0.89**
	AT-IKF	0.51	0.64	0.74	0.81
	DNN-KF	**0.63**	**0.74**	**0.82**	0.88

V. CONCLUSION

In this paper, DNN has been introduced to improve the performance of traditional KF based SE algorithm. It has been employed to estimate clean LPCs from noisy speech feature set. Due to DNN's powerful learning ability, our DNN-KF algorithm achieves better speech quality as well as intelligibility compared to the existing S-IKF and AT-IKF algorithms. In addition, with the help of large training data, our DNN-KF is more robust when the speech is corrupted by different noises.

ACKNOWLEDGMENT

The authors acknowledge the support from China Scholarships Council (CSC No.201606270200), NSERC of Canada under a CRD project sponsored by Microchip in Ottawa, Canada, and the National Natural Science Foundation of China (Grant No. 61601248).

REFERENCES

[1] J. S. Lim and A. V. Oppenheim, "Enhancement and bandwidth compression of noisy speech," *Proceedings of the IEEE*, vol. 67, no. 12, pp. 1586–1604, 1979.

[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] 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.

[4] 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.

[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] 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.

[7] Y. Xia and J. Wang, “Low-dimensional recurrent neural network-based Kalman filter for speech enhancement,” *Neural Networks*, vol. 67, pp. 131–139, 2015.

[8] 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.

[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] M. Zhao and W.-P. Zhu, “Adaptive wavelet packet thresholding with iterative Kalman filter for speech enhancement,” *in Proc. of IEEE Global Conference on Signal and Information Processing (GlobalSIP)*, pp. 71–75, 2017.

[11] Y. Xu, J. Du, L.-R. Dai, and C.-H. Lee, “A regression approach to speech enhancement based on deep neural networks,” *IEEE/ACM Trans. on Audio, Speech and Language Processing (TASLP)*, vol. 23, no. 1, pp. 7–19, 2015.

[12] A. Narayanan and D. Wang, “Ideal ratio mask estimation using deep neural networks for robust speech recognition,” *in Proc. of IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 7092–7096, 2013.

[13] Z. Ouyang, H. Yu, W.-P. Zhu, and B. Champagne, “A deep neural network based harmonic noise model for speech enhancement,” *Proc. Interspeech 2018*, pp. 3224–3228, 2018.

[14] C. J. Chun, S. H. Jeong, S. Y. Park, and H. K. Kim, “Extension of monaural to stereophonic sound based on deep neural networks,” in *Audio Engineering Society Convention 139*, 2015.

[15] Y. Wang, K. Han, and D. Wang, “Exploring monaural features for classification-based speech segregation,” *IEEE Trans. on Audio, Speech, and Language Processing (TASLP)*, vol. 21, no. 2, pp. 270–279, 2013.

[16] IEEE Subcommittee, “Ieee recommended practice for speech quality measurements,” *IEEE Trans. on Audio and Electroacoustics*, vol. 17, pp. 225–246, 1969.

[17] 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.

[18] 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.

[19] 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.