ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) | 978-1-7281-7605-5/20/$31.00 ©2021 IEEE | DOI: 10.1109/ICASSP39728.2021.9413499

NEURAL KALMAN FILTERING FOR SPEECH ENHANCEMENT

Wei Xue, Gang Quan, Chao Zhang, Guohong Ding, Xiaodong He, Bowen Zhou

JD AI Research

ABSTRACT

Conventional learning-based speech enhancement methods usually utilize existing building blocks to design the deep neural networks (DNNs), while how to effectively integrate the statistical signal processing based schemes, which are expert-knowledge driven and could ameliorate the over-fitting problem, into the network design remains an open issue. In this paper, we extend the conventional Kalman filtering (KF) and propose a supervised-learning based neural Kalman filter (NKF) for speech enhancement. Similar to KF, the proposed method first obtains a prediction from the speech evolution model and then integrates the short-term instantaneous observation by linear weighting, and the weights are calculated by comparing between the speech prediction residual error and the environmental noise level. An end-to-end network is designed to convert the speech linear prediction model in KF to non-linear, and to compact all other conventional linear filtering operations. Different with other DNN based methods, the proposed method provides a specialized network design inspired from the conventional signal processing, the backpropagation can be directly applied on the linear filtering operations integrated from KF. We conduct experiments in different noisy conditions, and the results demonstrate that the proposed method outperforms the baseline methods which are based on either signal processing or DNNs.

Index Terms— Speech Enhancement, Kalman filtering, Deep Neural Network

1. INTRODUCTION

Speech enhancement aims to suppress the environmental noise without distorting the target speech, and has wide applications in systems such as communication, hearing aids and automatic speech recognition.

Compared with the conventional signal processing based methods which use simple linear statistic models for speech and noise, such as Wiener filtering (WF) [1–3], subspace estimation [4–6] and Kalman filtering (KF) [7–11], the speech enhancement performance has been dramatically improved by deep neural networks (DNNs) which learn the non-linear mapping from the noisy feature to the clean target. Generally, the networks are designed based on either the classical building blocks for DNN such as feed-forward network (FNN) [12–14], convolutional neural network (CNN) [15–17], recurrent neural network (RNN) [18–20], or concatenation of these building blocks such as UNet [21–23] and convolutional recurrent neural network (CRNN) [24, 25]. Although the architectures are designed to effectively model the different time-frequency dependencies of speech and noise, there always lacks an explicit criterion for the model design, which makes it hard to interpret and optimize the intermediate representations, and also makes the performance highly rely on the diversity of the training data. In addition, many achievements in conventional signal processing based methods, which integrate the expert knowledge to derive the optimization steps and optimal filters given the statistics of speech and noise, have not fully been exploited by the supervised-learning based speech enhancement.

This paper proposes a neural Kalman filter (NKF) to fully integrate the statistical signal processing into DNN, by extending the KF to the supervised learning scheme. We note that [26] uses a DNN to obtain a preprocessed speech for linear prediction (LP) model estimation in KF but uses the conventional KF for speech enhancement. An end-to-end network is designed in this paper to convert the speech LP model as well as the noise estimation model in KF to non-linear, and to compact all other conventional linear filtering operations. Clean speech estimates from RNN and WF are obtained, and are linearly combined by an NKF gain to yield the NKF output. The signal processing operations can serve as regularization for the network to avoid overfitting, and the backpropagation (BP) can be straightforwardly applied to the linear filtering processes. Moreover, the proposed method also overcomes the problem of unrealistic linear model assumptions in KF. We conduct experiments in different noisy conditions, and evaluation results demonstrate the effectiveness of the proposed method.

2. SIGNAL MODEL AND KF

We present the modulation-domain KF [7, 8], since it has shown superior performances than both time-domain and short-time Fourier transform (STFT)-domain KFs [27]. The modulation-domain KF regards the amplitude of speech in each frequency bin as a time-varying signal, and adopts a

This work was supported by JD-BAAI joint project.

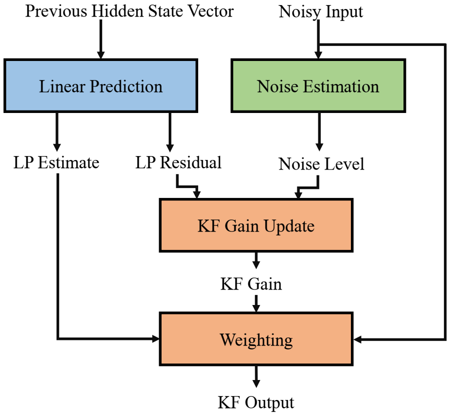

Fig. 1. Diagram of KF based speech enhancement.

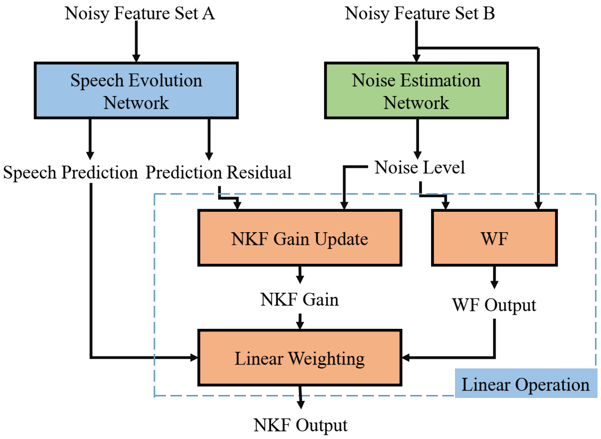

Fig. 2. Diagram of the proposed NKF.

signal model as:

$$|Y(t,f)| = |X(t,f)| + |V(t,f)|, \tag{1}$$

where $Y(t,f)$, $X(t,f)$, and $V(t,f)$ represent the STFT signals of the noisy speech, clean speech and noise, respectively, and $|\cdot|$ takes the amplitude. The clean speech amplitude $|X(t,f)|$ can be further expressed by a P-order LP model as:

$$\mathbf{x}(t,f) = \mathbf{A}(f)\mathbf{x}(t-1,f) + \mathbf{u}W(t,f), \tag{2}$$

where $\mathbf{x}(t,f) = [|X(t,f),|X(t-1,f)|,...,|X(t-P+1,f)|]^T$ is the hidden state vector, $\mathbf{A}(f)$ is the state transmission matrix defined in [7] according to the LP coefficients of speech, $\mathbf{u} = [1,0,...,0]^T$ is a $P \times 1$ vector, and $W(t,f)$ is the LP residual. In practice, the unknown LP coefficients are estimated via LP analysis on the output of WF.

As shown in Fig. 1, the KF takes two stages for speech enhancement: predicting and updating. Given the hidden state $\mathbf{x}(t-1|t-1,f)$ that consists of the clean speech estimates in the previous frame, the LP estimation of clean speech $\mathbf{x}(t|t-1,f)$ is first obtained using (2) as:

$$\mathbf{x}(t|t-1,f) = \mathbf{A}(f)\mathbf{x}(t-1|t-1,f), \tag{3}$$

and then used to update the KF estimate by incorporating the noisy observation in the current frame:

$$\begin{aligned}&\hat{\mathbf{x}}(t|t,f)\\&= [\mathbf{I} - \mathbf{G}(t,f)\mathbf{u}^T]\hat{\mathbf{x}}(t|t-1,f) + \mathbf{G}(t,f)|Y(t,f)|,\end{aligned} \tag{4}$$

where the $\mathbf{G}(t,f)$ is the KF gain determined by comparing between the noise variance $\sigma_v^2 = E\{V(t,f)V^*(t,f)\}$ and the variance matrix of the LP residual $\mathbf{R}_{ee}(t|t-1,f)$, as

$$\mathbf{G}(t,f) = \frac{\mathbf{R}_{ee}(t|t-1,f)\mathbf{u}}{\sigma_v^2 + \mathbf{u}^T\mathbf{R}_{ee}(t|t-1,f)\mathbf{u}}. \tag{5}$$

We observe that (5) takes a similar form to WF. It is also shown in [9–11] that KF can be seen as introducing the speech evolution into WF.

3. PROPOSED METHOD

The KF gain in (4) is important to KF since it serves as a weighting factor such that the KF output adaptively approximates the LP estimate in noise-dominated TF bins, and approximates the instantaneous observation when the noise level is low. It is also worth noting that, essentially the output is a combination of $\hat{\mathbf{x}}(t|t-1,f)$ and $|Y(t,f)|$, while how to determine the weight remains a problem. The KF introduces the *LP residual* to select the KF gain. With the LP residual, it is shown that the KF gain can be optimally determined and updated under the minimum mean-squared error (MMSE) criterion, based on the statistics of LP estimate and noise.

The above discussion motivates an integration of signal processing into the learning-based speech enhancement. Although networks have been proposed which generally contain modules with different expected functions, they usually rely on a large amount of diversified data to obtain the optimal network weights. However, by using the expert knowledge to obtain the optimal solutions from conventional signal processing, it is possible to impose constraints or define the mathematical formulations for the weights of the connections between different modules. In this way, the signal processing operations can be seen as the regularization for the network, which is helpful to reduce the reliance on the training data and cope with the over-fitting problem. In addition, since the signal processing operations are usually linear and differentiable, the network that integrates the signal processing operations can be straightforwardly optimized using the BP algorithm.

In this section, we extend the above KF to the supervised learning scheme and propose an NKF, whose structure is shown in Fig. 2, and the linear operations without learnable parameters are highlighted. The proposed NKF takes a similar form to the conventional KF, which first predicts the clean speech from a speech evolution model, and then updates the estimation by incorporating short-term information from the observation. Different from the KF, neural networks are used to replace the LP model of speech with a more realistic non-linear model which is learned from data, and also integrates the noise estimation process into the network. We note that the proposed NKF combines the network prediction with the WF output rather than the raw observation to obtain the final output, for which the reason will be explained later. The analytical expression of the NKF gain is defined according to the optimal KF gain derived under the MMSE criterion.

3.1. Non-linear Speech Evolution

The KF utilizes an auto-regression (AR) model for speech evolution which is usually not adequate to represent the temporal characteristics of speech in reality. A better strategy would be using a network to learn the non-linear mappings from the clean speech signals (hidden variables in the KF) in previous frames to the clean speech in the current frame.

We note that the concept of hidden variable has been widely used by the RNN, which has shown a strong capability of sequential modelling and yield superior performances for the speech enhancement task [18–20]. Thus it provides a natural choice for the NKF to model the non-linear speech evolution. Here, a long short-term memory (LSTM) network is constructed as in Fig. 3, which learns two targets simultaneously. Similar to many conventional LSTM based speech enhancers, the network takes the noisy features as input, and predicts the clean amplitude of speech. Moreover, in accordance with KF, the prediction residual is also estimated, which will be used to determine the optimal NKF gain.

Specifically, in each frame, a feature vector is formed using the noisy amplitudes in all frequency bins, and a feature sequence can be obtained by stacking the feature vectors over a set of continuous frames. The feature sequence is first fed into the LSTM layers to model the temporal evolution of speech, and two separate fully-connected output layers are used to convert the LSTM outputs into the clean amplitude prediction and the prediction residual, respectively.

We note that unlike the KF that adopts a feedback loop to feed the KF estimation of the previous frame into the LP model, the LSTM uses the noisy amplitude spectrum as input. This is because we believe that the speech evolution has already been modelled by the hidden state propagation in the LSTM, and the additional noisy observation in each frame can help the LSTM to achieve more accurate predictions.

3.2. Incorporating WF

The speech amplitude estimated by the non-linear LSTM will be updated by incorporating the instantaneous observation in the current frame. In the conventional KF, by using the feedback loop to update the statistics of LP estimate and residual, it can be shown that incorporating the noisy input to update the KF output is equivalent to the WF [9] when LP is excluded. Since the feedback loop is not used by the proposed NKF, in this section, based on the derivations in the KF, the WF is directly applied to incorporate the instantaneous information. The WF filters the signals only in the current frame, instead of performing prediction using several previous frames as is done by the LSTM.

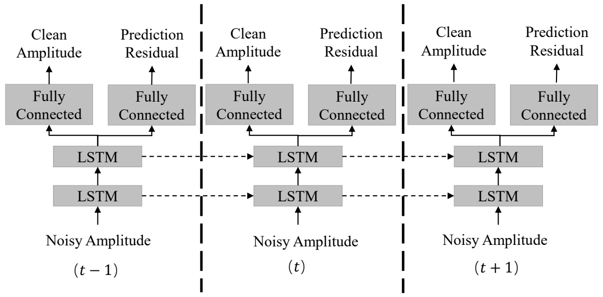

Fig. 3. Structure of LSTM prediction network.

Under the MMSE criterion, the optimal WF $H_{\text{Wiener}}(t,f)$ is given by

$$H_{\text{Wiener}}(t,f) = \frac{\sigma_x^2(t,f)}{\sigma_y^2(t,f)} = 1 - \frac{\sigma_v^2(t,f)}{\sigma_y^2(t,f)}, \tag{6}$$

where $\sigma_y^2(t,f)$, $\sigma_x^2(t,f)$ and $\sigma_v^2(t,f)$ are the variances of the noisy speech, clean speech and additive noise, respectively. In practice, $\sigma_y^2(t,f)$ is computed from the noisy observation, and as shown in Fig. 2, a noise estimation network is constructed based on the simple ReLU-activated FNN. The network takes the noisy amplitudes and variances for frames within a left-side context window as input, and predicts the noise variances. Then the WF output is obtained as

$$|\hat{X}(t,f)|_{\text{Wiener}} = H_{\text{Wiener}}(t,f)|Y(t,f)|. \tag{7}$$

3.3. Linear Weighting

The clean amplitude estimates from LSTM and WF are finally combined by linear weighting to yield the NKF output $|\hat{X}(t,f)|_{\text{NKF}}$. With KF gain, we similarly define an NKF gain:

$$G_{\text{NKF}} = \frac{\sigma_r^2(t,f)}{\sigma_v^2(t,f) + \sigma_r^2(t,f)}, \tag{8}$$

where $\sigma_r^2(t,f)$ is the variance of LSTM prediction residual. Then the NKF output is calculated by:

$$|\hat{X}|_{\text{NKF}} = (1 - G_{\text{NKF}})|\hat{X}|_{\text{LSTM}} + G_{\text{NKF}}|\hat{X}|_{\text{Wiener}}, \tag{9}$$

where $|\hat{X}|_{\text{LSTM}}$ denotes the LSTM estimation, the TF bin index "(t,f)" is omitted for simplicity.

Since the KF and WF are all derived under the MMSE criterion, the NKF network is optimized by taking the MSE between the clean amplitude and the NKF estimation as the loss function. The time-domain speech is recovered by inverse STFT which uses the phase of the noisy speech.

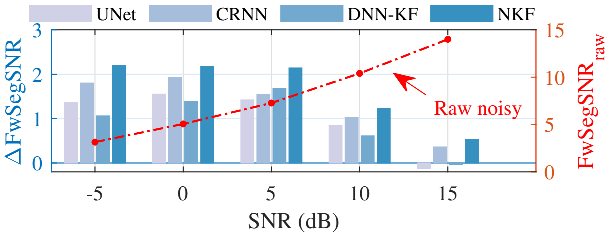

Fig. 4. Improvements of FwSegSNR for different SNRs.

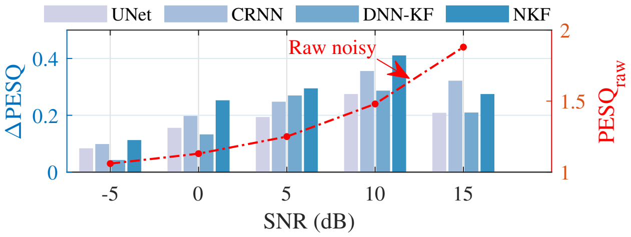

Fig. 5. Improvements of PESQ for different SNRs.

4. EXPERIMENT

4.1. Experimental Setup

We use the clean speech database, Librispeech [28], and two noise databases including PNL-100Nonspeech-Sounds (PNL) [29] and the noise subset of MUSAN corpus (MUSAN-Noise) [30] to prepare the training and testing data.

Unmatched noisy conditions for training and testing are generated. A 100-hour training set and a 10-hour development set are obtained by mixing the "CLEAN-360" subset of Librispeech and the MUSAN-Noise, under a speech-to-noise ratio (SNR) randomly chosen from $\{-6 : 3 : 21\}$ dB. The "TEST-CLEAN" subset of Librispeech and the PNL noise set are used to produce a 10-hour test set at SNR levels of $\{-5 : 5 : 15\}$ dB. The sample rate of all signals is 16 kHz, and the analysis window for STFT and feature extraction is 256 samples with a 75% overlap.

For the proposed NKF, the LSTM prediction network has two 1024-node hidden LSTM layers, and the noise estimation FNN has three layers with 1024 nodes in the hidden layer. The left-side context window for the noise estimation FNN is 30 and $\sigma_y^2(t, f)$ in (6) is computed using previous 20 frames.

Three baseline methods working in the TF domain are used for comparison, and they include a) an UNet which is based on CNN and uses the encoder-decoder framework with skip connections; b) a CRNN which adopts an LSTM into the UNet to capture the temporal characteristics; c) a DNN-KF which follows [26] to resolve the LP coefficient and noise estimation problem by the neural networks, and uses the conventional KF for speech enhancement. During training, the batch size of all methods is 16, and the sequence length is 2048 frames. The networks are trained by 20 epochs.

Three objective speech quality measures including the frequency-weighted segmental SNR (FwSegSNR), perceptual evaluation of speech quality (PESQ) and short-Time Objective Intelligibility (STOI) [31] are used for evaluation. For all metrics, higher value indicates better performance.

4.2. Results

Results of different methods for different SNRs are depicted from Fig. 4 to Fig. 6. In Fig. 4, it can be observed that the proposed NKF can consistently yield the highest improvement on the FwSegSNR over the noisy speech, which indicates the effectiveness of the NKF to suppress noise. The NKF also has the best capability to preserve speech when performing noise reduction, which is shown in Fig. 5 and Fig. 6 that the highest improvements of PESQ and STOI are achieved.

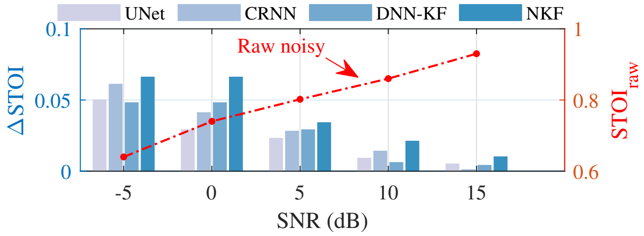

Fig. 6. Improvements of STOI for different SNRs.

We can see that generally different methods yield the most significant improvements when the SNR is 0 dB and 5 dB. The speech features would be highly contaminated by noise when the noise level is extremely high, making the speech enhancement problem more difficult. On the other hand, since the objective measures for speech in high SNR cases are already high, there is less potential for improvement. It is also worth noting that the DNN-KF which relies on an external processor to estimate the LP coefficients and the noise level achieves the worst performances in low SNR conditions, which is due to that the inaccurate information provided by the external processor, and that the assumption of linear speech evolution model is not satisfied.

5. CONCLUSION

An NKF based speech enhancement method is proposed by integrating the DNN with the statistical signal processing, following the framework of conventional KF. The statistical signal processing components can be seen as providing priori expert knowledge and serve as a regularization for the network. Experimental results in different noisy conditions show the effectiveness of the proposed method.

6. REFERENCES

[1] N. Wiener, *The Extrapolation, Interpolation and Smoothing of Stationary Time Series*. New York, NY, USA: John Wiley & Sons, Inc., 1949.

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

[3] J. Chen, J. Benesty, Y. Huang, and S. Doclo, "New insights into the noise reduction Wiener filter," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 14, no. 4, pp. 1218–1234, Jul. 2006.

[4] P. S. K. Hansen, "Signal subspace methods for speech enhancement," Ph.D. dissertation, Lyngby, Sep. 1997.

[5] Y. Hu and P. C. Loizou, "A subspace approach for enhancing speech corrupted by colored noise," *IEEE Signal Process. Lett.*, vol. 9, no. 7, pp. 204–206, Jul. 2002.

[6] Y. Ephraim and H. L. Van Trees, "A signal subspace approach for speech enhancement," in *Proc. IEEE Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, vol. 2, Apr. 1993, pp. 355–358.

[7] S. So and K. K. Paliwal, "Modulation-domain Kalman filtering for single-channel speech enhancement," *Speech Communication*, vol. 53, no. 6, pp. 818–829, Jul. 2011.

[8] Y. Wang and M. Brookes, "Model-based speech enhancement in the modulation domain," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 26, no. 3, pp. 580–594, mar 2018.

[9] W. Xue, A. H. Moore, M. Brookes, and P. A. Naylor, "Multichannel Kalman filtering for speech enhancement," in *Proc. IEEE Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, Calgary, Canada, Apr. 2018.

[10] ——, "Modulation-domain multichannel Kalman filtering for speech enhancement," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 26, no. 10, pp. 1833–1847, 2018.

[11] ——, "Modulation-domain parametric multichannel Kalman filtering for speech enhancement," in *Proc. European Signal Processing Conf. (EUSIPCO)*, Roma, Italy, Sep. 2018.

[12] Y. Xu, J. Du, L.-R. Dai, and C.-H. Lee, "An experimental study on speech enhancement based on deep neural networks," *IEEE Signal Processing Letters*, vol. 21, no. 1, pp. 65–68, jan 2014.

[13] C. Schüldt and P. Händel, "Decay rate estimators and their performance for blind reverberation time estimation," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 22, no. 8, pp. 1274–1284, Aug. 2014.

[14] M. Tu and X. Zhang, "Speech enhancement based on deep neural networks with skip connections," in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, mar 2017.

[15] N. Mamun, S. Khorram, and J. H. Hansen, "Convolutional neural network-based speech enhancement for cochlear implant recipients," in *Proc. Conf. of Intl. Speech Commun. Assoc. (INTERSPEECH)*, 2019.

[16] Z. Ouyang, H. Yu, W.-P. Zhu, and B. Champagne, "A fully convolutional neural network for complex spectrogram processing in speech enhancement," in *Proc. IEEE Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, may 2019.

[17] S. R. Park and J. W. Lee, "A fully convolutional neural network for speech enhancement," in *Interspeech 2017*. ISCA, aug 2017.

[18] F. Weninger, H. Erdogan, S. Watanabe, E. Vincent, J. L. Roux, J. R. Hershey, and B. Schuller, "Speech enhancement with LSTM recurrent neural networks and its application to noise-robust ASR," in *Latent Variable Analysis and Signal Separation*. Springer International Publishing, 2015, pp. 91–99.

[19] T. Gao, J. Du, L.-R. Dai, and C.-H. Lee, "Densely connected progressive learning for LSTM-based speech enhancement," in *Proc. IEEE Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, apr 2018.

[20] L. Sun, J. Du, L.-R. Dai, and C.-H. Lee, "Multiple-target deep learning for LSTM-RNN based speech enhancement," in *2017 Hands-free Speech Communications and Microphone Arrays (HSCMA)*. IEEE, 2017.

[21] A. Pandey and D. Wang, "A new framework for supervised speech enhancement in the time domain," in *Interspeech*, 2018, pp. 1136–1140.

[22] C.-H. Yang, J. Qi, P.-Y. Chen, X. Ma, and C.-H. Lee, "Characterizing speech adversarial examples using self-attention u-net enhancement," in *Proc. IEEE Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, may 2020.

[23] H.-S. Choi, J.-H. Kim, J. Huh, A. Kim, J.-W. Ha, and K. Lee, "Phase-aware speech enhancement with deep complex U-net," in *International Conference on Learning Representations*, 2018.

[24] K. Tan and D. Wang, "A convolutional recurrent neural network for real-time speech enhancement." in *Interspeech*, 2018, pp. 3229–3233.

[25] M. Strake, B. Defraene, K. Fluyt, W. Tirry, and T. Fingscheidt, "Fully convolutional recurrent networks for speech enhancement," in *Proc. IEEE Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, may 2020.

[26] 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 *2019 IEEE International Symposium on Circuits and Systems (ISCAS)*. IEEE, may 2019.

[27] W. Xue, A. H. Moore, M. Brookes, and P. A. Naylor, "Speech enhancement based on modulation-domain parametric multichannel Kalman filtering," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 29, pp. 393–405, 2021.

[28] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: An ASR corpus based on public domain audio books," in *Proc. IEEE Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, apr 2015.

[29] G. Hu and D. Wang, "A tandem algorithm for pitch estimation and voiced speech segregation," vol. 18, no. 8, pp. 2067–2079, nov 2010.

[30] D. Snyder, G. Chen, and D. Povey, "MUSAN: A Music, Speech, and Noise Corpus," 2015, arXiv:1510.08484v1.

[31] P. C. Loizou, *Speech Enhancement: Theory and Practice, Second Edition*. CRC Press, Feb. 2013.