Raw Waveform-based Speech Enhancement by Fully Convolutional Networks

Szu-Wei Fu$^{12}$, Yu Tsao$^1$, Xugang Lu$^3$, Hisashi Kawai$^3$

$^1$ Research Center for Information Technology Innovation, Academia Sinica, Taipei, Taiwan
$^2$ Department of Computer Science and Information Engineering, National Taiwan University, Taipei, Taiwan
$^3$ National Institute of Information and Communications Technology, Kyoto, Japan

{jasonfu,yu.tsao}@citi.sinica.edu.tw; {xugang.lu, hisashi.kawai}@nict.go.jp

Abstract

This paper proposes a fully convolutional network (FCN) model for raw waveform-based speech enhancement. The proposed system performs speech enhancement in an end-to-end (i.e. waveform-in and waveform-out) manner, which differs from most existing denoising methods that process the magnitude spectrum (e.g. log power spectrum (LPS)) only. Because the fully connected layers, which are involved in deep neural networks (DNN) and convolutional neural networks (CNN), may not accurately characterize local information of speech signals, especially for high frequency components, we employed fully convolutional layers to model the waveform. More specifically, FCN only consists convolutional layers and hence the local temporal structures of speech signals can be efficiently and effectively preserved with a relatively small number of weights. Experimental results show that DNN and CNN based models have limited capability to restore high frequency components of waveforms, thus leading to imperfect intelligibility of enhanced speech. On the other hand, the proposed FCN model can not only well recover the waveforms but also outperform the LPS-based DNN baseline in terms of STOI and PESQ. In addition, the number of model parameters in FCN is roughly only 0.2% compared with that in DNN and CNN.

Index Terms: speech enhancement, fully convolutional network, raw waveform, phase

1. Introduction

Speech enhancement (SE) has been widely used as a preprocessor in speech-related applications, such as speech coding [1], hearing aids [2], automatic speech recognition (ASR) [3] and cochlea implants [4, 5]. In the past, various speech enhancement (SE) approaches have been developed. Notable examples include spectral subtraction [6], minimum-mean square error (MMSE) based spectral amplitude estimator [7], Wiener filtering [8], and non-negative matrix factorization (NMF) [9]. Recently, deep denoising autoencoder (DDAE) and deep neural networks (DNN) based SE models have also been proposed and extensively investigated [10-12]. In addition, to efficiently model the local temporal-spectral structures of spectrogram, convolutional neural networks (CNN) are also employed to further improve the SE performance [13, 14]. Most of these denoising models focus only on processing the magnitude spectrogram (e.g., log-power spectra, LPS) and leave the phase in its original noisy form. Because there is no clear structure in a phase spectrogram, it is difficult to precisely estimate clean phases from noisy counterparts [15].

Several recent researches have shown the importance of phase when spectrograms are resynthesized back into timedomain waveforms [16, 17]. For example, Paliwal et al. confirmed the importance of phase for perceptual quality in speech enhancement, especially when window overlap and length of the Fourier transform are increased [16]. To further improve the performance of speech enhancement, phase information is considered in some up-to-date researches [15, 18, 19]. Williamson et al. [15, 18] employed a DNN for estimating the complex ratio mask (CRM) from a set of complementary features, and the magnitude and phase can be jointly enhanced through CRM. Although being confirmed to provide satisfactory denoising performance, these methods still need to map features between time and frequency domains for analysis and resynthesizing through (inverse) Fourier transform.

In the field of ASR, several researches have shown that deep learning based models with raw waveform inputs can achieve higher accuracy than that with the hand-crafted features (e.g. MFCC) [20-25]. Because the acoustic patterns in time domain can appear in any positions, most of these methods employ CNN for efficiently detecting the useful information. On the other hand, in the field of speech enhancement, directly using the raw waveforms as system inputs has not been well studied yet. When compared to ASR, in addition to distinguishing speech patterns from noise, SE has to further generate the enhanced speech outputs. In the time domain, each estimated sample point has to cooperate with its neighbors to represent frequency components. This interdependency may make the model laborious to generate the high and low frequency components in the same time. Until recently, waveform [26] was proposed and successful models raw audio waveforms through sample wise prediction and dilated convolution.

In this paper, we intend to investigate the capability of different deep learning based SE methods with raw waveform inputs. We first note that the fully connected layers may not well preserve the local information to generate high frequency components. Therefore, we employ the fully convolutional network (FCN) model to enable each output sample locally depends on the neighboring input regions. FCN is very similar to conventional CNN except that the fully connected layers are removed [27]. Recently, FCN has been proposed for SE [28] to process the magnitude spectrum. Based on the unique property of FCN and the successful results in [28], we have adopted FCN to build our waveform-in and waveform-out system. Experimental results show that the proposed FCN model can not only well recover the waveform but also dramatically reduce the number of parameters compared with DNN and CNN.
2. Raw waveform speech enhancement

The goal of SE is to improve the intelligibility and quality of a noisy speech signal [29]. Since the properties in the log domain are more consistent with the human auditory system, conventionally, the log power spectrum is extracted from raw speech signal for deep learning-based denoising models [11, 12, 30-32]. However, employing LPS as features will bring two drawbacks. First, the phase components have not been well considered. In other words, when synthesize the enhanced speech signal back to time domain, the phase components are simply borrowed from the original noisy speech, which may degrade the perceptual quality of enhanced speech [16, 17]. Second, the (inverse) Fourier transform has to be applied for the mapping between time and frequency domains, resulting in the increase of unnecessary computation load. In this study, we intend to investigate deep learning based SE with raw waveform inputs and explore solutions to address possible issues.

2.1. The characteristics of raw waveform

Figure 1 shows the system for analyzing deep learning based SE with waveform-in and waveform-out. The characteristics of a signal represented in time domain is very different form that in the frequency domain. In the frequency domain, the value of a feature (frequency bin) represents the energy of the corresponding frequency component. However, in the time domain, a feature (sample point) alone does not carry much information, it has to cooperate with its neighbors to represent a certain frequency component. For example, a sample point has to be very different/similar to its neighbors to represent high/low frequency component. This interdependency may make the model laborious to represent high and low frequency components in the same time and cause many denoising models choose to work on frequency domain rather than in time domain [6-9, 11]. In addition, unlike the spectrogram of speech signal (e.g. the consonants usually only occupy high frequency bins while the repeated patterns of formants usually concentrate on low to middle frequency bins), the patterns in time domain can appear in any position, which imply that the convolution operation can efficiently find the useful locally acoustic information. Therefore, most researches employed the CNN model for modeling raw waveform [20-24, 26].

2.2. The problems in fully connected layers for modeling raw waveform

The output layer and last hidden layer in DNN and CNN are connected in a fully connected manner, as shown in Fig. 2. We argue that this connection manner makes it difficult to model high and low frequency components of waveform in the same time. The relation between output layer and last hidden layer can be represented as (bias is neglected here for simplicity):

$$ y = Wh $$ (1)

where $y = [y_1, ..., y_T]^T \in \mathbb{R}^{N \times 1}$ is the output sample points of estimated waveform, and $N$ is the number of points in a frame. $W = [w_1, ..., w_N]^T \in \mathbb{R}^{N \times H}$ is the weight matrix, $h$ is the number of nodes in the last hidden layer and $w_n \in \mathbb{R}^{H \times 1}$ is the weight vector that connects the hidden layer $h \in \mathbb{R}^{N \times 1}$ and the output sample $y_n$. In other words, each sample point can be represented as:

$$ y_t = w_t^T h $$ (2)

W

Clean

Waveform

Denoising

Model

Noisy

Waveform

Figure 1: Speech enhancement using raw waveform.

With fixed $h$, we consider two situations: (1) when $y_t$ is in the high frequency region, its value should be very different from its neighbors (e.g. $y_{t-1}$, $y_{t+1}$), which implies that $w_t$ and $(w_{t-1}, w_{t+1})$ cannot be very similar; (2) when $y_t$ is in the low frequency region, we can deduce that $w_t$ and $(w_{t-1}, w_{t+1})$ should be similar with each other. However, since $W$ is fixed after training, situations (1) and (2) cannot be satisfied in the same time. Therefore, we argue that it should be difficult for fully connected layers to generate high and low frequency parts of waveform in the same time. In fact, the hidden fully connected layers also have difficulty in modeling raw waveforms. We will discuss this problem in more detail in section 5.

3. Fully convolutional networks (FCN)

From the previous section, it is shown that fully connected layers may not model raw waveforms precisely. Therefore, in this paper, we try to apply fully convolutional networks (FCN), which do not contain any fully connected layers, to perform SE in the waveform domain. FCN is very similar to conventional CNN, except that all the fully connected layers are removed. This can bring several benefits and has obtained great success in the computer vision field for modeling raw pixel outputs [27]. Removing the fully connected layers can dramatically reduce the number of weights in the model, which can mitigate the storage burden in a portable device like smart phone. In addition, each output sample in FCN only locally depends on the neighboring input regions as shown in Fig. 3. This is different from fully connected layers where the local information and the spatial arrangement of the previous features cannot be well preserved.
To more specifically explain why FCN can model high and low frequency components of raw waveforms simultaneously, we start with the connections between output layer and last hidden layer. The relation between output sample \( y_t \) and the connected hidden nodes \( R_t \) (also called receptive field) can be simply represented as (bias is neglected for simplicity):

\[
y_t = F^T R_t
\]

(3)

where \( F \in \mathbb{R}^{f \times 1} \) denotes the learned filter, and \( f \) is the size of the filter. Please note that \( F \) is shared in the convolution operation and is fixed for every output samples. Therefore, if \( y_t \) is in the high frequency region, \( R_t \) and \( (R_{t-1}, R_{t+1}) \) should not be very similar with each other. Whether \( R_t \) is different with its neighbors depends on the filtered outputs of previous locally connected nodes (input) \( I_t \). For example, when the input \( I_t \) is in the high frequency region, and the filter \( G \) is a high-pass filter, then \( R_t \) (and hence \( y_t \)) may also be very different from its neighbors. This argument can also hold for the low frequency case. Therefore, FCN can well preserve the local input information and handle the difficulty of using fully connected layers to modeling high and low frequency components simultaneously. When comparing the locations of subscript \( t \) from (2) to (3), it can be observed that \( t \) is changed from the model \( (w_t) \) to connected nodes \( (R_t) \). This implies that in the fully connected case, the model has to deal with the interdependency between output samples, while in FCN, it is the connected nodes that handle the interdependency.

4. Experiments

4.1. Experimental setups

In our experiments, the TIMIT corpus [33] was used to prepare the training and test sets. For the training set, 600 utterances were randomly selected and corrupted with five noise types (Babble, Car, Jackhammer, Pink, and Street), at five SNR levels (-10 dB, -5 dB, 0 dB, 5 dB, and 10 dB). For the test set, we randomly selected another 100 utterances (different from those used in the training set). To make experimental conditions more realistic, both noise types and SNR levels of the training and test sets were mismatched. Thus, we adopted three other noise signals: (White Gaussian noise (WGN), a stationary noise) and (Engine and Baby cry, two non-stationary noises), with another five SNR levels: -12 dB, -6 dB, 0 dB, 6 dB, and 12 dB to form the test set. All the results reported in Section 4.2 were averaged across the three noise types.

In this work, 512 sample points were extracted from the waveforms to form a frame for the proposed SE model.

In addition, the 257 dimensional LPS were also obtained from the frame for the baseline system. The CNN in this experiment has four convolutional layers with padding (each layer consists 15 filters with filter size 11) and two fully connected layers (each 1024 nodes). FCN has the same structure as CNN except the fully connected layers are replaced with another convolutional layer. DNN has only four hidden layers (each layer consists 1024 nodes), since when it grows deeper, the performance starts to degrade due to optimization issue [34]. All the models employ parametric rectified linear units (PReLU) [35] as activation functions and are trained using adam [36] with batch normalization [37].

To evaluate the performance of proposed models, the perceptual evaluation of speech quality (PESQ) [38] and the short-time objective intelligibility (STOI) scores [39] were used to evaluate the speech quality and intelligibility, respectively.

4.2. Experimental results

4.2.1. Qualitative Comparison

In this section, we intend to investigate different deep learning models (DNN, CNN and FCN) for SE with raw waveform. Figure 4 shows an example of modeling high frequency signal by DNN and FCN. From this figure, it can be observed that this is difficult for DNN to produce the corresponding high frequency signal as FCN. As pointed out in section 2.2,
Table 1: Performance comparison of different systems in terms of STOI and PESQ.

| SNR (dB) | DNN-baseline (LPS) | DNN (waveform) | CNN (waveform) | FCN (waveform) |
|----------|-------------------|----------------|----------------|----------------|
| 12       | 0.814 2.334       | 0.737 2.548    | 0.788 2.470    | 0.874 2.718    |
| 6        | 0.778 2.140       | 0.715 2.396    | 0.753 2.302    | 0.833 2.346    |
| 0        | 0.717 1.866       | 0.655 2.118    | 0.673 2.011    | 0.758 1.995    |
| -6       | 0.626 1.609       | 0.549 1.816    | 0.561 1.707    | 0.639 1.719    |
| -12      | 0.521 1.447       | 0.429 1.573    | 0.441 1.453    | 0.506 1.535    |
| Avg.     | 0.691 1.879       | 0.617 2.090    | 0.643 1.989    | **0.722** 2.063 |

Figure 6: Four spectrograms of a TIMIT utterance corrupted by Engine noise at SNR=12dB: (a) clean speech (b) noisy speech (c) enhanced by DNN (using waveform) and (d) enhanced by FCN (using waveform).

Next, we present the spectrograms of a clean speech utterance, the same utterance corrupted by the Engine noise, DNN enhanced speech, and FCN enhanced speech, respectively, in Fig. 6 (a), (b), (c), and (d), where the raw waveform is used as the inputs for DNN and FCN. When comparing Fig. 6 (a) and (c), it can be clearly observed that the high frequency speech parts are missing in the spectrogram of DNN-enhanced waveform. This phenomenon can also be observed in CNN (not shown in this paper due to the limited space) but not as serious as the case in DNN. The missing high frequency components may lead to lower intelligibility scores (we will confirm this point in the next section). On the other hand, by comparing Fig. 6 (a) and (d), we can note that FCN-enhanced speech components are well preserved with noise being effectively removed, suggesting that FCN can generate enhanced speech with both high quality and intelligibility.

4.2.2. Quantitative Comparison

Finally we present the results of the average STOI and PESQ scores on the test set, among different models and features, which are summarized in Table 1. In the table, DNN-baseline (LPS) denoted the DNN model with the log power spectrogram inputs, as used in [11]; DNN (waveform) CNN (waveform), and FCN (waveform), respectively, denoted DNN, CNN, and FCN models with waveform inputs. From this table, we can see that the waveform-based DNN achieves the highest PESQ score, while the worst STOI score, suggesting that it cannot strike a good balance between the two goals of speech enhancement (improving the intelligibility and quality of a noisy speech signal). This can be owing to that the model removes too many speech components with noise in the same time. On the other hand, FCN can achieve the highest STOI score while yielding satisfactory PESQ score. It is worthy mentioning that since the fully connected layers are removed, the number of weights involved in FCN is roughly only 0.2% when comparing to that involved in DNN and CNN.

5. Discussion

We noted that the issue of missing high frequency components becomes more serious when the number of fully connected layers increased. In addition to the problem pointed out in section 2.2, the hidden fully connected layers actually also have difficulty in preserving the high frequency information. Unlike the filtered results by convolutional layer (as shown in Fig. 3), the mapped feature space by fully connected layer is abstract and does not retain the spatial arrangement of the previous features. On the other hand, in time domain waveform, the estimated sample points have to cooperate with its neighbors to
represent a certain frequency component. Therefore, fully connected layers destroy the dependency of neighboring features, making it difficult to model waveforms. This well explains why CNN has relatively minor issue on missing high frequency components when compared to DNN, since CNN contains less fully connected layers.

6. Conclusions

The contribution of this paper is two-fold. First, we investigate the capability of different deep learning based SE methods with raw waveform inputs. The results indicated that fully connected layers may not be necessary because: (1) they dramatically increase the model parameters; (2) they have limited capability to preserve the spatial arrangement of the features, which is extremely important for generating waveforms. Second, to overcome this problem, FCN is employed and confirmed to yield better results compared to DNN with LPS-inputs. In the future, we will further explore the SE performance with different features (e.g., LPS, complex spectrogram, and waveform) on various models (DNN, CNN and FCN, etc.) to compare their SE performance.

7. References

[1] J. Li, S. Sakamoto, S. Hongo, M. Akagi, and Y. Suzuki, “Two-stage binaural speech enhancement with Wiener filter for high-quality speech communication,” Speech Communication, vol. 53, pp. 677-689, 2011.

[2] H. Levitt, “Noise reduction in hearing aids: A review,” Journal of rehabilitation research and development, vol. 38, p. 111, 2001.

[3] J. Li, L. Deng, Y. Gong, and R. Haeb-Umbach, “An overview of noise-robust automatic speech recognition,” IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 22, pp. 745-777, 2014.

[4] F. Chen, Y. Hu, and M. Yuan, “Evaluation of noise reduction methods for sentence recognition by mandarin-speaking cochlear implant listeners,” Ear and hearing, vol. 36, pp. 61-71, 2015.

[5] Y.-H. Lai, F. Chen, S.-S. Wang, X. Lu, Y. Tsao, and C.-H. Lee, “A deep denoising autoencoder approach to improving the intelligibility of vocoded speech in cochlear implant simulation,” IEEE Transactions on Biomedical Engineering, 2016.

[6] S. F. Boll, “Suppression of acoustic noise in speech using spectral subtraction,” IEEE Transactions on Acoustics, Speech and Signal Processing, vol. 27, pp. 113-120, 1979.

[7] Y. Ephraim and D. Malah, “Speech enhancement using a minimum-mean square error short-time spectral amplitude estimator,” IEEE Transactions on Acoustics, Speech and Signal Processing, vol. 32, pp. 1109-1121, 1984.

[8] P. Scalart, “Speech enhancement based on a priori signal to noise estimation,” in International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2006, pp. 4029-4032.

[9] K. W. Wilson, B. Raj, P. Smaragdis, and A. Divakaran, “Speech denoising using nonnegative matrix factorization with priors,” in International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2016.

[10] X. Lu, Y. Tsao, S. Matsuda, and C. Hori, “Speech enhancement based on deep denoising autoencoder,” in INTERSPEECH, 2013, pp. 436-440.

[11] 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, pp. 65-68, 2014.

[12] Y. Xu, J. Du, L.-R. Dai, and C.-H. Lee, “A regression approach to speech enhancement based on deep neural networks,” IEEE/ACM Transactions on Audio, Speech, and Language Processing vol. 23, pp. 7-19, 2015.

[13] S.-W. Fu, Y. Tsao, and X. Lu, “SNR-aware convolutional neural network modeling for speech enhancement,” in INTERSPEECH, 2016, pp. 3768-3772.

[14] L. Hui, M. Cai, C. Guo, L. He, W.-Q. Zhang, and J. Liu, “Convolutional maxout neural networks for speech separation,” in 2015 IEEE International Symposium on Signal Processing and Information Technology (ISSPIT), 2015, pp. 24-27.

[15] D. S. Williamson, Y. Wang, and D. Wang, “Complex ratio masking for joint enhancement of magnitude and phase,” in 2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2016, pp. 5220-5224.

[16] K. Paliwal, K. Wojcicki, and B. Shannon, “The importance of phase in speech enhancement,” speech communication, vol. 53, pp. 465-494, 2011.

[17] J. Le Roux, “Phase-controlled sound transfer based on maximally-inconsistent spectrograms,” Signal, vol. 5, p. 10, 2011.

[18] D. S. Williamson, Y. Wang, and D. Wang, “Complex ratio masking for monaural speech separation,” IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 24, pp. 483-492, 2016.

[19] K. Li, B. Wu, and C.-H. Lee, “An iterative phase recovery framework with phase mask for spectral mapping with an application to speech enhancement,” in INTERSPEECH, 2016, pp. 3773-3777.

[20] D. Palaz, R. Collobert, and M. M. Doss, “Estimating phoneme class conditional probabilities from raw speech signal using convolutional neural networks,” in INTERSPEECH, 2013, pp. 1766-1770.

[21] D. Palaz and R. Collobert, “Analysis of cnn-based speech recognition system using raw speech as input,” in INTERSPEECH, 2015, pp. 11-15.

[22] D. Palaz, M. M.-. Doss, and R. Collobert, “Convolutional neural networks-based continuous speech recognition using raw speech signal,” in 2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2015, pp. 4295-4299.

[23] P. Golik, Z. Tüske, R. Schlüter, and H. Ney, “Convolutional neural networks for acoustic modeling of raw time signal in LVCSR,” in INTERSPEECH, 2015, pp. 26-30.

[24] W. Dai, C. Dai, S. Qu, J. Li, and S. Das, “Very deep convolutional neural networks for raw waveforms,” arXiv preprint arXiv:1610.00087, 2016.

[25] Z. Tüske, P. Golik, R. Schlüter, and H. Ney, “Acoustic modeling with deep neural networks using raw time signal for LVCSR,” in INTERSPEECH, 2014, pp. 890-894.

[26] A. v. d. Oord, S. Dieleman, H. Zen, K. Simonyan, O. Vinyals, A. Graves, et al., “Wavenet: a generative model for raw audio,” arXiv preprint arXiv:1609.03499, 2016.

[27] J. Long, E. Shelhamer, and T. Darrell, “Fully convolutional networks for semantic segmentation,” in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 3431-3440.

[28] S. R. Park and J. Lee, “A fully convolutional neural network for speech enhancement,” arXiv preprint arXiv:1609.07132, 2016.

[29] J. Benesty, S. Makino, and J. D. Chen, Speech enhancement Springer, 2005.

[30] Y. Xu, J. Du, L.-R. Dai, and C.-H. Lee, “Global variance equalization for improving deep neural network based speech enhancement,” in IEEE China Summit & International Conference on Signal and Information Processing (ChinaSIP), 2014, pp. 71-75.

[31] Y. Xu, J. Du, L.-R. Dai, and C.-H. Lee, “Dynamic noise aware training for speech enhancement based on deep neural networks,” in INTERSPEECH, 2014, pp. 2670-2674.

[32] T. Gao, J. Du, L.-R. Dai, and C.-H. Lee, “SNR-based progressive learning of deep neural network for speech enhancement,” in INTERSPEECH, 2016, pp. 3713-3717.

[33] J. Lyons, “DARPA TIMIT acoustic-phonetic continuous speech corpus,” National Institute of Standards and Technology, 1993.

[34] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 770-778.
[35] K. He, X. Zhang, S. Ren, and J. Sun, "Delving deep into rectifiers: Surpassing human-level performance on imagenet classification," in Proceedings of the IEEE International Conference on Computer Vision, 2015, pp. 1026-1034.

[36] D. Kingma and J. Ba, "Adam: A method for stochastic optimization," arXiv preprint arXiv:1412.6980, 2014.

[37] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in Proceedings of the 32nd International Conference on Machine Learning (ICML-15), 2015, pp. 448-456.

[38] A. Rix, J. Beerends, M. Hollier, and A. Hekstra, "Perceptual evaluation of speech quality (PESQ), an objective method for end-to-end speech quality assessment of narrowband telephone networks and speech codecs," ITU-T Recommendation, p. 862, 2001.

[39] C. H. Taal, R. C. Hendriks, R. Hessdens, and J. Jensen, "An algorithm for intelligibility prediction of time–frequency weighted noisy speech," IEEE Transactions on Audio, Speech, and Language Processing, vol. 19, pp. 2125-2136, 2011.