Subjective Feedback-based Neural Network Pruning for Speech Enhancement

Fuqiang Ye*, Yu Tsao†, Fei Chen*
*Department of Electrical and Electronic Engineering, Southern University of Science and Technology
†Research Center for Information Technology Innovation, Academic Sinica
Email: fchen@sustech.edu.cn Tel: +86-755-88018554

***Abstract*— Speech enhancement based on neural networks provides performance superior to that of conventional algorithms. However, the network may suffer owing to redundant parameters, which demands large unnecessary computation and power consumption. This work aimed to prune the large network by removing extra neurons and connections while maintaining speech enhancement performance. Iterative network pruning combined with network retraining was employed to compress the network based on the weight magnitude of neurons and connections. This pruning method was evaluated using a deep denoising autoencoder neural network, which was trained to enhance speech perception under nonstationary noise interference. Word correct rate was utilized as the subjective intelligibility feedback to evaluate the understanding of noisy speech enhanced by the sparse network. Results showed that the iterative pruning method combined with retraining could reduce 50% of the parameters without significantly affecting the speech enhancement performance, which was superior to the two baseline conditions of direct network pruning with network retraining and iterative network pruning without network retraining. Finally, an optimized network pruning method was proposed to implement the iterative network pruning and retraining in a greedy repetition manner, yielding a maximum pruning ratio of 80%.**

I. INTRODUCTION

Speech enhancement has been widely used in speech communication, automatic speech recognition, and speech coding. It aims to estimate clean speech from noisy sound with an acceptable speech quality and intelligibility. Many monaural speech enhancement methods, which use only single-channel speech information have been proposed, such as Wiener filtering, minimum mean square error (MMSE) based estimation, and subspace method [1]. Most speech enhancement algorithms are derived based on the prior distribution assumptions of the noisy speech and explore the statistical difference between the clean speech and noise signal. In real-world scenarios, those speech enhancement algorithms are usually less effective under nonstationary noise conditions. Recently with the development of the deep neural network (DNN) in signal processing [2-3], many neural network based speech enhancement algorithms have been proposed, which employ nonlinear processing units to learn higher order statistical information automatically [4-7]. For instance, an objective function for DNN-based speech enhancement was proposed to match human auditory perception. The proposed objective function helped to compute the gradients based on a perceptually motivated non-linear frequency scale and alleviated the over-smoothness of the estimated speech [8]. Furthermore, a speech enhancement algorithm based on deep denoising autoencoder (DDAE) was shown to provide superior performance to the traditional MMSE-based estimation [4, 5]. The DDAE-based speech enhancement combined with a noise classifier could potentially be integrated into an embedded signal processor to overcome the degradation of speech perception caused by noise [9, 10]. However, the superior performance of the neural network is at the cost of high computational complexity and power consumption, making it difficult to deploy neural network based speech enhancement to mobile and embedded devices. Hence, many recent studies have focused on designing approaches to compress DNN structures. Effective approaches include quantization, sparse or low-rank compressions, and network pruning [11].

Specially, the network pruning method has been studied for decades [12-14]. Early pruning approaches included optimal brain damage (OBD) and optimal brain surgeon (OBS), which reduced the number of network connections based on the Hessian of the loss function [12, 13]. It was shown that such pruning methods were more effective and accurate than the magnitude-based pruning method [15]; however, the necessary second-order derivatives required additional computational resources. Liu et al. compared the OBD-based pruning method with the magnitude-based pruning method for DNN-based speech classification accuracy and speech recognition performance. The classification accuracy and word error rate (WER) results showed that the OBD-based pruning method was superior for highly pruned network [14]. However, the accuracy and WER showed a slight difference between the OBD-based pruning and magnitude-based pruning methods. The magnitude-based pruning method gained more attention because it could be simply and efficiently implemented. Recently, Han et al. proposed a deep compression method that combined magnitude-based pruning, quantization and Huffman coding. They removed the redundant network connections and learned only those connections that are important [16-17]. This magnitude-based pruning method was shown to reduce the number of parameters by 9× and 13× on the AlexNet and VGG-16, respectively [16]. Motivated by this success in image processing, this study employed the magnitude-based pruning method for DNN-

based speech enhancement. Although the magnitude-based pruning method has been established in theory and evaluated in many experiments, the application to DNN-based speech enhancement is still questionable. This is largely because the full-connected speech enhancement network is very complex, particularly when dealing with nonstationary background noise, and the accurate perceptual evaluation of the enhanced speech is difficult.

Accurate evaluation of enhanced speech has long been a challenge for speech enhancement studies [18-23]. Many objective speech quality/intelligibility indices, e.g., speech-transmission index [19], normalized covariance metric [20], short-time objective intelligibility metric [21] and across-band envelope correlation metric [22], have been developed. However, these evaluation metrics could hardly predict the intelligibility of enhanced speech containing various non-linear distortion, caused by the nonlinear processing in speech enhancement [18]. Compared to objective speech intelligibility evaluation, subjective listening tests require human listeners to recognize speech signals, and these generally have the most accurate results for speech intelligibility. The word correct rate (WCR) was often used as the subjective evaluation criteria, which is calculated by dividing the number of correctly identified words by the total words for each test condition [23]. Hence, this study utilized WCR as a subjective feedback index to evaluate speech enhanced by the pruned network. In summary, the major goal of this study was to prune the network based on the weight magnitude of each connection and neuron and use WCR as a subjective feedback index to evaluate the speech enhancement performance of the sparse network during network pruning.

II. Network Pruning

A. DDAE-based Speech Enhancement

DDAE has been used to build a DNN architecture for speech enhancement [4, 5]. The basic structure of DDAE-based speech enhancement is shown in Fig. 1.

This network can be regarded as a multiple hidden layer neural associator with noisy speech as input and clean speech as output. The fast Fourier transform (FFT) is applied to the input signal to compute the spectrum of each overlapping windowed frame. A set of noisy-clean speech pairs are converted into a Mel-frequency power spectrum as the input features Y_m and output features X_m during the training phase. The frame in the FFT is denoted by m. For a DDAE model with D hidden layers, it can be obtained that:

$$h^1(Y_m) = \sigma(W^0 Y_m + b^0) \quad \dots \quad (1)$$

$$h^D(Y_m) = \sigma(W^{D-1} h^{D-1}(Y_m) + b^{D-1}), \quad (2)$$

$$\hat{X}_m = W^D h^D(Y_m) + b^D \quad (3)$$

where $\{W^0 \dots W^D\}$ and $\{b^0 \dots b^D\}$ are the matrices of the connections and the bias vectors for the DDAE model. $\hat{X}_m$ is the vector of enhanced speech corresponding to the noisy counterpart Y_m, and the activation function is given by $\sigma(t) = (1 + e^{-t})^{-1}$. The final parameters are determined by optimizing the following objective functions:

$$\theta^* = argmin_\theta (F(\theta) + \eta^0 ||W^0||_2^2 + \dots + \eta^D ||W^D||_2^2), \quad (4)$$

$$F(\theta) = \frac{1}{M} \sum_{m=1}^{M} ||X_m - \hat{X}_m||_2^2 \quad (5)$$

Here M is the total number of noisy-clean pairs. In the test phase, an inverse transform is performed to synthesize the restored speech waveforms with phase information of the corresponding noisy speech. The speech feature was extracted from frames with a 16 ms Hamming window and frame shifting of 8 ms. More detailed information regarding the DDAE-based speech enhancement can be found in studies [4, 5].

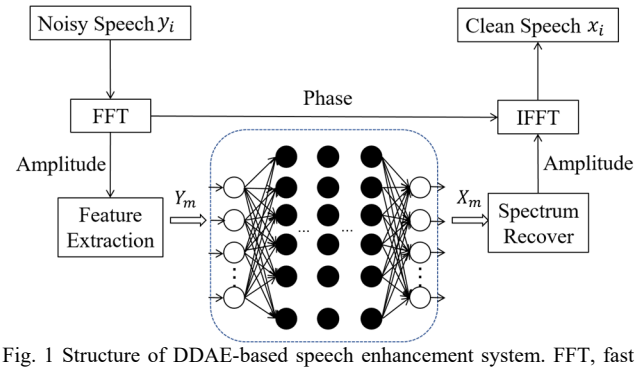

Fig. 1 Structure of DDAE-based speech enhancement system. FFT, fast Fourier transform; IFFT, inverse fast Fourier transform.

B. Iterative Network Pruning Method

The network pruning method proposed in this study is based on the magnitude of parameter weights. Pruning converts a dense neural network into a sparse one and reduces the number of parameters and computations while adequately preserving speech enhancement performance.

A block diagram of the pruning network method is shown in Fig. 2. Firstly, the network is trained normally to obtain the original parameter set Θ. The next step is to prune the network based on the magnitude of weights. The absolute value is employed as a simple index to determine the relative importance of the weight. Weights with absolute values below the pruning threshold are removed by setting them to zero. During this step, a mask matrix M is utilized to implement the network pruning. Weights below the pruning threshold have a corresponding mask of zero; otherwise, the value of the mask is one. The pruning network is realized by

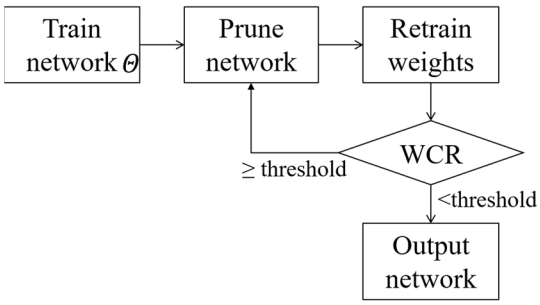

Fig. 2 Block diagram of the iterative pruning method with retraining.

computing the dot product between the original parameter set Θ and the mask matrix M.

The third step is to retrain the sparse network to obtain the final weights. Note that the initialization network for the retraining is the sparse network from the second step rather than the initial random network. The sparse network has converged, and so, keeping the surviving parameters provides better performance when retraining the sparse network. During the retraining phase, the learning rate needs to be adjusted. This is because the weights have already attained local minima during the training network phase. Usually, the retraining learning rate is reduced by one or two orders of magnitude.

The final step is to evaluate the performance of the pruned network using the WCR. The threshold is usually the WCR of speech enhanced by the original network. If the WCR is obviously lower than the threshold, the critical sparse network is obtained. Otherwise, it returns to the "prune network" step and prunes more weights. The pruning step combined with the one-time retraining is one iteration, and the maximum pruning ratio of parameters could be found by pruning the network progressively after several such iterations. This pruning method is called iterative pruning method.

Compared to the iterative pruning method, direct pruning method removes the weights of the original network globally according to the pruning ratio instead of performing progressive pruning. The optimized pruning method repeats the iterative pruning and retraining in a greedy way, which reconverges the pruned network. The optimized iterative pruning method is expected to provide a higher network pruning ratio.

III. Experiments

A. Databases and Settings

Experiments were conducted using utterances excerpted from the Mandarin Chinese version of hearing in noise test (MHINT) [24], which were pronounced by a male native speaker with a fundamental frequency ranging from 75-180 Hz, and recorded with a sampling rate of 16 kHz. This study focuses on challenging noisy conditions; hence two types of nonstationary noise were utilized, i.e., babble noise and construction jackhammer (CJ) noise. Half of the clean MHINT utterances were corrupted by the corresponding noise at –10, –5, 0, 5, and 10 dB input signal-to-noise ratios (SNRs) to form the training set. The other half of the clean utterances were corrupted by two noises at 0 dB input SNR to form the test set.

The DDAE-based speech enhancement model consisted of three layers, with 500 neurons in each hidden layer. The number of trained network parameters was 581,580. Then the original network was pruned iteratively to several ratios. For subjective feedback, listening experiments were conducted with 20 subjects having normal hearing to obtain the WCR of the speech enhanced by the different pruned networks.

B. Iterative Pruning

The trade-off curves between the pruning ratio and WCR under the babble and CJ noise conditions are shown in Fig. 3 and Fig. 4, wherein pruning ratio refers to proportion of parameters removed. The three network pruning methods were compared, including 1) iterative pruning with retraining, 2) direct pruning with retraining, and 3) iterative pruning without retraining. In Fig. 3 and Fig. 4, it can be seen from the red line that the iterative pruning method with retraining could prune 50% of the parameters without significantly affecting the WCR of enhanced speech while the maximum pruning ratio of iterative pruning without retraining is only up to 25% (the dashed line). In addition, the direct pruning method with retraining performs better than iterative pruning method without retraining, but its maximum pruning ratio (the dotted line) is lower than that of the iterative pruning method (the solid line).

C. Optimized Iterative Pruning

Trade-off curves of the optimized pruning method between the pruning ratio and WCR under the babble and CJ noise conditions compared with the iterative pruning method with

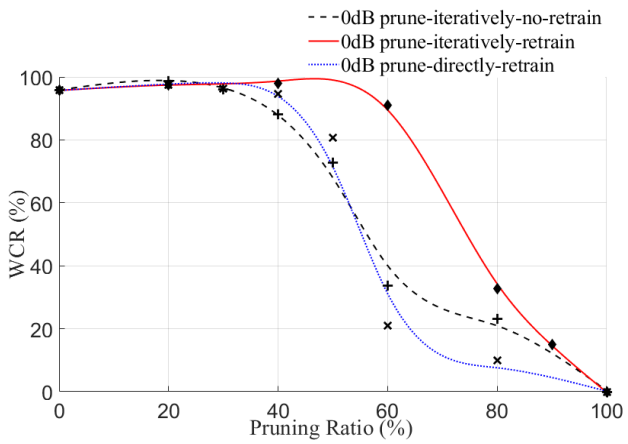

Fig. 3 Trade-off curves between pruning ratio and WCR of three pruning methods under the babble noise condition.

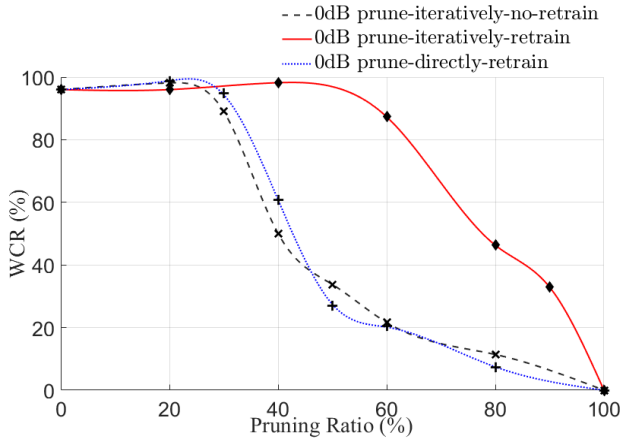

Fig. 4 Trade-off curves between pruning ratio and WCR of three pruning methods under the CJ noise condition.

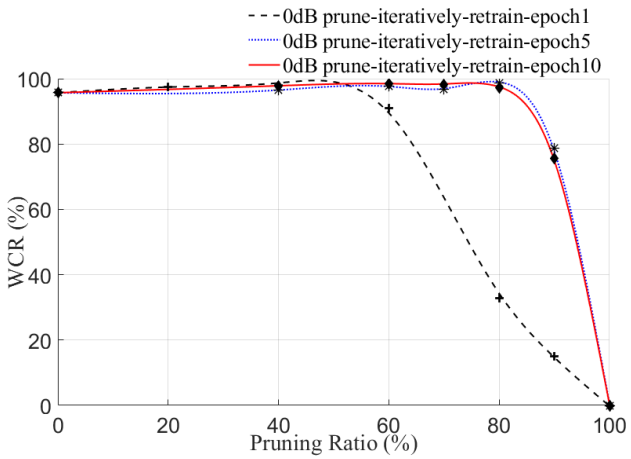

Fig. 5 Trade-off curves between pruning ratio and WCR of the optimized pruning method under the babble noise condition.

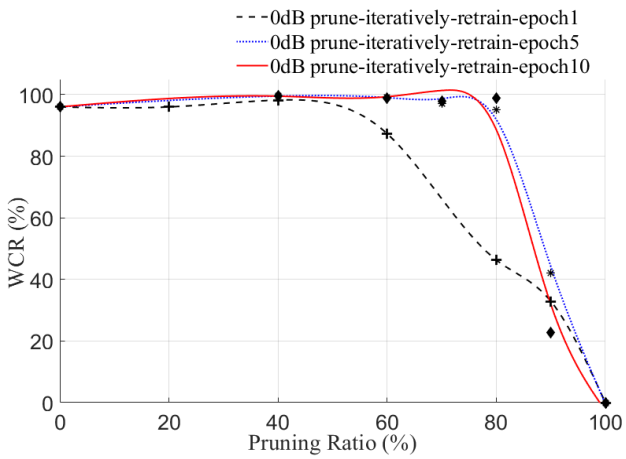

Fig. 6 Trade-off curves between pruning ratio and WCR of the optimized pruning method under the CJ noise condition.

retraining, are shown in Fig. 5 and Fig. 6. Under the 0 dB CJ noise condition, the optimized pruning method, which repeated the pruning and retraining 5 times (the dotted line) or 10 times (the solid line), could remove 75% of the parameters without affecting the subjective speech perception performance of the pruned network. Under the babble noise condition, the maximum pruning ratio could reach up to 80%, equivalent to reducing the network parameters by 5×. The optimized pruning method performs better than iterative pruning method with retraining (the dashed line). In addition, there is little difference between the two optimized pruning methods that repeat the iterative pruning and retraining steps 5 times and 10 times.

IV. Discussion And Conclusions

DNNs have been widely applied in speech signal processing for classification (automatic speech and speaker recognition) and regression (speech separation and enhancement) tasks. The DNN-based speech enhancement system automatically learns the nonlinear kernel space from noisy-clean speech pairs and performs well even under mismatched noise types; hence it can potentially be implemented in embedded and mobile speech processors. However, the network parameters are highly redundant, leading to large memory requirements and undesirable computational burden to embedded devices. This study evaluated magnitude-based network pruning methods to reduce network redundancy without significant degeneration in the speech enhancement performance. To dates, the accurate evaluation of the enhanced speech containing non-linear distortions arising from speech enhancement processing is still a challenging task. Hence, the present work employed WCR as the subjective intelligibility feedback index to evaluate the performance of the sparse networks after network pruning.

Experimental results related to the DDAE-based speech enhancement network in this work showed that the iterative pruning method with retraining could remove 50% of the network parameters without affecting the network performance in subjective speech perception. This result is superior to other implementations of the iterative pruning without retraining and direct pruning with retraining. Furthermore, by repeating the iterative pruning and retraining steps 5 times, the maximum pruning ratio of the network could be raised up to 80%, equivalent to a compression rate of 5:1. Future work will assess the efficacy of the optimized iterative pruning method on other speech enhancement networks with higher complexity or more challenging listening environments, and combine it with other compression strategies (e.g., quantization and hardware acceleration) to further reduce the network redundancy.

Acknowledgments

This work was supported by the Research Foundation of Department of Science and Technology of Guangdong Province (Grant No. 2018A050501001), and the Shenzhen High-level Overseas Talent Program (Grant No. KQJSCX20180319114453986).

References

[1] P. C. Loizou, *Speech Enhancement: Theory and Practice*, CRC Press, Boca Raton, FL, USA, 2007.

[2] G. E. Hinton and R. R. Salakhutdinov, “Reducing the dimensionality of data with neural networks,” *Science*, vol. 313, pp. 504–507, 2006.

[3] F. J. Huang, Y. L. Boureau, and Y. LeCun, “Unsupervised learning of invariant feature hierarchies with applications to Object Recognition,” *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1–8, 2007.

[4] X. Lu, Y. Tsao, S. Matsuda, et al., “Speech enhancement based on deep denoising autoencoder,” *Interspeech*, pp. 436–440, 2013.

[5] X. Lu, Y. Tsao, S. Matsuda, et al., “Ensemble modeling of denoising autoencoder for speech spectrum restoration,” *Interspeech*, pp.885–889, 2014.

[6] 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. Audio Speech Lang. Process.,* vol. 23, no. 1, pp. 7–19, 2015.

[7] D. Wang and J. Chen, "Supervised speech separation based on deep learning: an overview," *IEEE/ACM Trans. Audio Speech Lang. Process.,* vol. 26, no. 10, pp. 1702–1726, 2018.
[8] T. G. Kang, J.W. Shin, and N. S. Kim, "DNN-based monaural speech enhancement with temporal and spectral variations equalization," *Digit. Signal Process.*, vol. 74, pp. 102–110, 2018.
[9] Y.-H. Lai, F. Chen, S.-S. Wang, et al., "A deep denoising autoencoder approach to improving the intelligibility of vocoded speech in cochlear implant simulation," *IEEE Trans. Biomed. Eng.*, vol. 64, no.7, pp. 1568–1578, 2017.
[10] Y.-H. Lai, Y. Tsao, X. Lu, et al., "Deep learning–based noise reduction approach to improve speech intelligibility for cochlear implant recipients," *Ear & Hearing*, vol. 39, pp. 795–809, 2018.
[11] Z. Zhuang, M. Tan, B. Zhuang, et al., "Discrimination-aware channel pruning for deep neural networks," *Adv. Neural Inform. Process. Syst.*, pp. 883–894, 2018.
[12] Y. LeCun, J. S. Denker, and S. A. Solla, "Optimal brain damage," *Adv. Neural Inform. Process. Syst.*, pp. 598–605, 1990.
[13] B. Hassibi and D. G. Stork, "Second order derivatives for network pruning: optimal brain surgeon," *Adv. Neural Inform. Process. Syst.,* pp. 164–171, 1993.
[14] C. Liu, Z. Zhang, and D. Wang, "Pruning deep neural networks by optimal brain damage," *Interspeech*, pp. 1092–1095, 2014.
[15] S. J. Hanson and L. Y. Pratt, "Comparing biases for minimal network construction with back-propagation," *Adv. Neural Inform. Process. Syst.*, pp. 177–185, 1989.
[16] S. Han, J. Pool, J. Tran, et al., "Learning both weights and connections for efficient neural network," *Adv. Neural Inform. Process. Syst.*, pp. 1135–1143, 2015.
[17] S. Han, H. Mao, and W. J. Dally, "Deep compression: compressing deep neural networks with pruning, trained quantization and huffman coding," *International Conference on Learning Representations (ICLR)*, 2016.
[18] P. C. Loizou, and J. F. Ma, "Extending the articulation index to account for non-linear distortions introduced by noise-suppression algorithms," *J. Acoust. Soc. Am.* vol. 130, no. 2, pp. 985–995, 2011.
[19] H. Steeneken and T. Houtgast, "A physical method for measuring speech transmission quality," *J. Acoust. Soc. Am.*, vol.67, no. 1, pp. 318–326, 1980.
[20] J. F. Ma, Y. Hu, and P. C. Loizou, "Objective measures for predicting speech intelligibility in noisy conditions based on new band-importance functions," *J. Acoust. Soc. Am.,* vol.125, no. 5, pp. 3387–3405, 2009.
[21] C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen, "A short time objective intelligibility measure for time-frequency weighted noisy speech," *in Proc. ICASSP*, pp. 4214–4217, 2010.
[22] F. Chen, "Predicting the intelligibility of noise-corrupted speech non-intrusively by across-band envelope correlation," *Biomed. Signal Proces.,* vol. 24, pp. 109–113, 2016.
[23] F. Chen and P. C. Loizou, "Predicting the intelligibility of vocoded and wideband Mandarin Chinese," *J Acoust Soc Am*., vol. 129, no. 5, pp.3281–3290, 2011.
[24] L. L. Wong, S. D. Soli, S. Liu, et al., "Development of the Mandarin hearing in noise test (MHINT)," *Ear & hearing*, vol. 28, no. 2, pp. 70–74, 2007.