Speech Enhancement Based on Deep Denoising Autoencoder

Xugang Lu[1]*, Yu Tsao*[2]*, Shigeki Matsuda*[1]*, Chiori Hori*[1]

1. National Institute of Information and Communications Technology, Japan
2. Research Center for Information Technology Innovation, Academic Sinica, Taiwan

Abstract

We previously have applied deep autoencoder (DAE) for noise reduction and speech enhancement. However, the DAE was trained using only clean speech. In this study, we further introduce an explicit denoising process in learning the DAE. In training the DAE, we still adopt greedy layer-wised pretraining plus fine tuning strategy. In pretraining, each layer is trained as a one hidden layer neural autoencoder (AE) using noisy-clean speech pairs as input and output (or transformed noisy-clean speech pairs by preceding AEs). Fine tuning was done by stacking all AEs with pretrained parameters for initialization. The trained DAE is used as a filter for speech estimation when noisy speech is given. Speech enhancement experiments were done to examine the performance of the trained denoising DAE. Noise reduction, speech distortion, and perceptual evaluation of speech quality (PESQ) criteria are used in the performance evaluations. Experimental results show that adding depth of the DAE consistently increase the performance when a large training data set is given. In addition, compared with a minimum mean square error based speech enhancement algorithm, our proposed denoising DAE provided superior performance on the three objective evaluations.

Index Terms: Deep autoencoder learning, autoencoder, noise reduction, speech enhancement.

1. Introduction

Estimating clean speech from noisy ones is very important for many real applications of speech technology, such as automatic speech recognition (ASR), and hearing aids. Many noise reduction and speech enhancement methods have been proposed, such as Wiener filtering, minimum mean square error (MMSE) based estimation, and signal subspace method [1]. Most of them focused on exploring the statistical difference (mainly focus on the second order statistical structure) between speech and noise. The performance improvement is guaranteed if noise and speech is separable in the explored space. High order statistical information exploration for noise reduction was also proposed in which a function approximation in a reproducing kernel Hilbert space method was applied for speech estimation [2]. However, the kernel function was manually given which may not be efficient for speech processing.

Neural network with nonlinear processing units can be used to learn high order statistical information automatically and can be used for noise reduction. In order to efficiently learn the statistical information, it is believed that a deep network (with multiple hidden layers) is preferred than a shallow network (with single or less hidden layers) [3]. In order to efficiently train a deep network, many training algorithms were proposed [4, 5, 6]. The basic strategy is to train a deep network with greedy layer wised pretraining plus fine tuning. With this strategy, deep learning was successfully applied in speech feature extraction and acoustic modeling [8]. Different from their applications to acoustic modeling, we have applied deep autoencoder (DAE) for noise reduction and speech enhancement [7]. In our previous study, the DAE was trained only using clean speech data set. Both the input and output of the DAE are clean speech. When there comes a noisy speech, the denoising was done as projecting the noisy speech into the clean speech signal subspace (or basis functions) expanded by the DAE. In this case, the DAE is trained to only encode clean speech statistical information. In this study, we further advance our study by explicitly introducing a denoising process in training the DAE. In training, noisy speech is input to the DAE, and clean speech is set as the output. Based on this processing, the DAE explicitly learns the statistical difference between clean and noisy speech. The basis functions expanded by the DAE try to emphasize speech statistical information by considering the information from both speech and noise.

Denoising autoencoder was already used in image processing and other applications, particularly applied to extract noisy robust feature for classification [9]. In their study, the input to each AE was bit-masked or distorted version of clean features, such as binary masked features, which is not suitable for speech processing. For noise reduction and speech enhancement, we make noisy data set from clean ones by adding many types of noise to clean speech, and training each AE using noisy-clean speech pairs or transformed pairs. Based on denoising autoencoder concept, recurrent denoising autoencoder was proposed for reducing noise in speech feature extraction for ASR [10]. In our study, we focus on speech enhancement problem by simply stacking many denoising autoencoders without any recurrent connections, and evaluate the performance based on noise reduction, speech distortion, and perceptual evaluation of speech quality criteria.

The paper is organized as follows. Section 2 introduces the basic architecture of deep autoencoder with explicit denoising processing. Section 3 gives definitions of the evaluation criteria which will be extensively used in experiments. Section 4 showed detailed experimental results and evaluations. Discussions and conclusion are given in section 5.

2. Deep denoising autoencoder

Although restrict Boltzmann machine (RBM) was firstly introduced to build a deep belief network (DBN) [4], it is difficult for traditional optimization algorithms to be used for training the network. As a substitute, the neural autoencoder (AE) is an equivalent module to the RBM in building a DAE [5]. One of the advantages of using AE and DAE is that many traditional optimization algorithms are ready to be used in training. Previously, we adopted the DAE for noise reduction and speech enhancement [7]. However, the DAE was trained using clean speech data set. Different from the usage of denoising au-

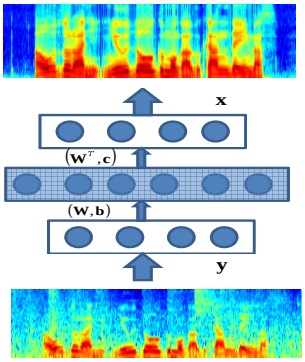

Figure 1: Training neural autoencoder with noisy-clean speech pairs.

toencoder in robust feature extraction [9], we use a noisy-clean speech pair to train the AE as shown in Fig. 1. This is a one hidden layer neural autoencoder trained with noisy speech as input and clean speech as output. It includes one nonlinear encoding stage and one linear decoding stage for real valued speech as:

$$\begin{aligned} h(\mathbf{y}_i) &= \sigma(\mathbf{W}_1\mathbf{y}_i + \mathbf{b}) \\ \hat{\mathbf{x}}_i &= \mathbf{W}_2 h(\mathbf{y}_i) + \mathbf{c}, \end{aligned} \tag{1}$$

where $\mathbf{W}_1$ and $\mathbf{W}_2$ are encoding and decoding matrix as the neural network connection weights, respectively. Usually, tied weight matrix, i.e., $\mathbf{W}_1 = \mathbf{W}_2{}^T = \mathbf{W}$, is used as one type of regularization. $\mathbf{b}$ and $\mathbf{c}$ are the vectors of biases of input and output layers, respectively. The nonlinear function of hidden neuron is a logistic function defined as $\sigma(\mathbf{x}) = (1+\exp(-\mathbf{x}))^{-1}$. The parameters are determined by optimizing the following objective function as:

$$L(\Theta) = \sum_i \|\mathbf{x}_i - \hat{\mathbf{x}}_i\|_2^2, \tag{2}$$

where $\Theta = \{\mathbf{W}, \mathbf{b}, \mathbf{c}\}$ is the parameter set, and $\mathbf{x}_i$ is the clean speech corresponding to the noisy version $\mathbf{y}_i$.

Besides using tied weights, incorporating regularization on weights and hidden neural output can help for a better generalization in order to avoid overfitting. For example, the weight decay and sparse regularization on outputs of hidden neurons are formulated as:

$$J(\Theta) = L(\Theta) + \alpha\|\mathbf{W}\|_2^2 + \beta\rho(h(\mathbf{y})), \tag{3}$$

where $\|\mathbf{W}\|_2^2 = \sum_{i,j} w_{ij}^2$. $\rho(h(\mathbf{y}))$ is a regularization function on the hidden neural outputs. α and β are the regularization weighting coefficients. In our study, we set $\alpha = 0.0002$, and $\beta = 0$ (we will consider sparse regularization in our future work). Then the parameter set can be obtained as:

$$\Theta^* \triangleq \arg\min_{\Theta} J(\Theta) \tag{4}$$

The optimization of Eq. (4) can be solved by using many unconstrained optimization algorithms. In this study, a linear search based quasi-Newton optimization algorithm is used to estimate $(\mathbf{W}^*, \mathbf{b}^*, \mathbf{c}^*)$ [11].

By stacking several AEs, a DAE can be built. We adopt greedy layer wised pretraining plus fine tuning to train the DAE. In pretraing stage, when adding one more hidden layer, the input of the next AE is the output of the preceding hidden layer. In denoising case, the transformed noisy-clean speech pairs will be used for training. For example, as shown in Fig. 1, the training pair for the first AE is $\mathbf{y}$ and $\mathbf{x}$, and then the training pair for the next AE will be $h(\mathbf{y}_i)$ and $h(\mathbf{x}_i)$. After pretraining of each autoencoder in a layer by layer manner, all the layers are stacked to form a deep autoencoder for fine tuning. In fine tuning stage, the initial network parameters are fixed as the parameters obtained from pretraining stage. Based on these training procedures, it is possible that the final solution is better than training the DAE with a random initialization.

3. Evaluation criteria

We focus on the noise reduction and speech enhancement task. Therefore, in this study, we evaluate the performance of the neural network with the following three criteria which are widely used in speech enhancement literature, namely, noise reduction, speech distortion, and perceptual evaluation of speech quality (PESQ) [1]. Since we will use them extensively in our experiments, we briefly give their definitions in this section. The measure of noise reduction is defined as:

$$\text{Reduct} \triangleq \frac{1}{N*d}\sum_{i=1}^{N}|\hat{\mathbf{x}}_i - \mathbf{y}_i| \tag{5}$$

The measure of speech distortion is defined as:

$$\text{Dist} \triangleq \frac{1}{N*d}\sum_{i=1}^{N}|\hat{\mathbf{x}}_i - \mathbf{x}_i| \tag{6}$$

In these two definitions, the average of absolute difference between estimated signal and noisy or clean speech is used. N is the total number of testing data, and d is the dimension of the input data (size of the first layer of the DAE). Based on noise reduction criterion (it is denoted as "Reduct" in experiments), the larger the value, the better quality of the restored speech. However, reducing much noise inevitably causes speech distortion. Based on speech distortion measurement (it is denoted as "Dist" in experiments), the less the value, the better quality of the restored speech is.

In addition to these two objective criteria, perceptual evaluation of speech quality (PESQ), which is a mean opinion score (MOS) like objective evaluation, is also used to evaluate the quality of the restored speech. Although it is not exactly corresponding to subjective evaluation, it shows high correlation to MOS [1]. The feature used in training the DAE is Mel frequency power spectrum (MFP). However, the PESQ evaluation needs waveform for evaluation. After getting the restored MFP, we perform an inverse transform to synthesize the restored speech with phase information of noisy speech. For consistency in using MFP for measuring the PESQ, the reference signal is also inverse synthesized from clean MFP. The PESQ score ranges from -0.5 to 4.5 corresponding to low to high speech quality.

4. Experiments and evaluations

In this section, we evaluate the deep denoising autoencoder on speech enhancement task. A clean continuous Japanese speech data set with 350 utterances was used for training, and 50 utterances for testing. Noisy data set was made by adding two types of noises (factory and car noise signals) to the clean data set. Three levels of signal to noise ratio (SNR) were made as 0, 5, and 10 dB. The MFP with 40 filter bands was used as the feature. The feature was extracted from 16 ms windowed signal

Table 1: Effect of training data set size (hidsize 100)

Training set size	10 k	40 k	80 k
Reduct (dB)	**1.99**	1.94	1.93
Dist (dB)	0.60	0.48	**0.47**
PESQ	2.80	3.30	**3.33**

Table 2: Effect of training data set size (hidsize 300)

Training set size	10 k	40 k	80 k
Reduct (dB)	**2.01**	1.94	1.93
Dist (dB)	0.61	0.47	**0.44**
PESQ	2.77	3.32	**3.44**

with 8 ms frame shift. The inputs to the DAE are MFP spectral patches. Each patch is selected from several continuous frames of the spectrum. 80,000 MFP spectral patches from the training speech are randomly selected. Different from making noisy training data set as in [9], the noisy MFP spectral patches were selected according to the clean MFP spectral patches, i.e., exactly the same time locations in utterances.

In ASR application, one of the most important contributions from deep learning framework is that long temporal window data can be concatenated to train the model. In our experiments, we also have compared the speech enhancement performance based on models trained with different size of input spectral patches. We increased the sizes of spectral patches to be 3, 7, and 11 frames. Correspondingly, the dimensions of input to the autoencoder are 120, 280, and 440, respectively. In our experiments, we find that increasing input patch size consistently improved the speech enhancement performance but with the cost of increasing model complexity (large size of model parameters with large training patch size). In addition, when patch size is larger than 11 frames, there is no significant improvement any more (less than 0.01 dB improvement based on speech distortion measure, and no improvement based on PESQ measure). In our following experiments, 11-frame patch size was used.

4.1. Effect of training data set size

For a given AE network, if training data set size is small, the training may cause over-fittings, which result in bad generalization. Therefore, large amount of training data set is preferred. However, training with large amount of training data is time consuming, and the network may be updated slowly after it is trained in some degree. In order to examine the performance for speech restoration based on different training data set size, we trained a basic denoising AE (as shown in Fig. 1) with training data set size of 10 k, 40 k, and 80 k (MFP spectral patches), respectively. The factory noise with SNR 10 dB condition is considered. The performance of the restoration is measured based on the three criteria (refer to section 3), and the results are shown in tables 1, 2, and 3 for hidden layer size of 100, 300, and 500, respectively. From these three tables, we can see that increasing training data set size always helps in improving the quality of the restored speech based on "Dist" and "PESQ" criteria, but with a little decrease in noise reduction. By comparing the first columns in tables 1, 2 and 3, we can see that when training data set size is small, e.g., 10 k, increasing the number of hidden neurons does not help to improve the restoration performance. However, when large training data set is used, e.g., 80 k, increasing the number of hidden neurons helps a lot (by comparing the third columns in tables 1, 2 and 3).

Table 3: Effect of training data set size (hidsize 500)

Training set size	10 k	40 k	80 k
Reduct (dB)	**2.01**	1.94	1.93
Dist (dB)	0.62	0.47	**0.43**
PESQ	2.64	3.36	**3.52**

Table 4: Performance regarding to hidden layer size

hidsize	100	300	500
Reduct (dB)	**1.93**	**1.93**	**1.93**
Dist (dB)	0.47	0.44	**0.43**
PESQ	3.33	3.44	**3.52**

4.2. Effect of hidden layer size

Intuitively, increasing the number of hidden neurons helps to increase the capacity of the AE for function approximation. For a clear look at how the hidden layer size affect the restoration performance, we summarize the results in table 4 for training data set size of 80 k with differen size of hidden neurons. From this table, we can see that increasing the number of hidden neurons improved the speech restoration. However, as we have discussed in subsection 4.1, over-fitting may occur for large network since more parameters need to be trained in large network than in small network, particularly when training data set size is small. From the results in subsections 4.1, and 4.2, we can see that a tradeoff of between the size of training data set and size of hidden neurons must be considered when designing the denoising autoencoder.

4.3. Effect of depth

In most deep learning studies, the general conclusion is that increasing the depth of the neural network always helps in performance either for pattern classification or for encoding [3, 4, 12]. Similarly, we increase the depth of the network by stacking several AEs to form a DAE, and carry out speech denoising experiments. The experimental condition was set the same as in subsection 4.1. In addition, the numbers of hidden neurons 100 and 300 are investigated, and the depth is increased from 1 to 3. The results are shown in tables 5 and 6 (80 k training data set). From these tables, we can see that increasing the depth of the DAE improves the quality of the restored speech based on speech distortion and PESQ criteria, and with only a little decrease in noise reduction.

We further carried out experiments by setting the number of hidden neurons to 500, and increased the depth from 1 to 3. The results are shown in table 7. From this table, however, we can not see the same tendency as in tables 5 and 6. Only network with depth 2 improved the performance. Increasing depth to 3, however cannot improve on DAE with depth 2. One possible reason is that when increasing the depth, the training data set size is not sufficient to fully train the large number of network parameters (as discussed in subsection 4.1).

Table 5: Effect of depth in DAE

hidsize*layer	100*1	100*2	100*3
Reduct (dB)	**1.93**	**1.93**	**1.93**
Dist (dB)	0.47	0.44	**0.43**
PESQ	3.33	**3.39**	**3.39**

Table 6: Effect of depth in DAE

hidsize*layer	300*1	300*2	300*3
Reduct (dB)	**1.93**	1.92	1.92
Dist (dB)	0.44	**0.40**	**0.40**
PESQ	3.44	3.52	**3.61**

Table 7: Effect of depth in DAE

hidsize*layer	500*1	500*2	500*3
Reduct (dB)	**1.93**	1.91	1.92
Dist (dB)	0.43	**0.40**	0.42
PESQ	3.52	**3.61**	3.52

4.4. Comparison with traditional noise reduction algorithms

There are many speech enhancement algorithms [1], most of them are based on a gain function estimation for noisy speech filtering with a noise tracking algorithm. In our comparison, we took the MMSE plus improved minimum controlled recursive averaging (IMCRA) noise tracking algorithm [13].

Two types of noises (car and factory noises) and three SNR conditions (0, 5, and 10 dB) were tested. The DAE with depth 3 and hidden layer size 100 was examined. The DAE was trained for each noise type. First, we compared the quality of the restored speech visually on the spectrum. The restored spectrum for factory noise in SNR 10 dB condition is shown in Fig. 2.

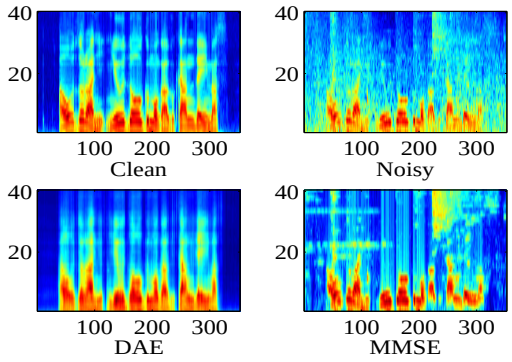

Figure 2: Horizontal axis: time frame index, vertical axis: Mel filter band index; clean speech (upper-left), and noisy speech (upper-right); restored speech based on DAE (lower-left) and MMSE (lower-right).

Comparing the two restored spectrum, we can see that more severe speech distortion as well as more noise residues in restored spectrum by the MMSE method than by the DAE. We can expect a better quality improvement by using the DAE than using the MMSE. We further quantitatively compared the restoration quality based on the three criteria defined in section 3. The comparisons are shown in tables 8, 9, and 10. From these three tables, we can see that speech restoration based on DAE significantly outperformed that of based on the MMSE, only with the exception of car noise condition based on noise reduction criterion.

Table 8: Evaluation based on noise reduction (dB).

Evaluations	Noise reduction			
SNR (dB)	Factory noise		Car noise	
	MMSE	DAE	MMSE	DAE
0	2.35	**2.72**	**1.05**	0.83
5	2.08	**2.32**	**0.92**	0.63
10	1.84	**1.93**	**0.82**	0.47

Table 9: Evaluation based on speech distortion (dB).

Evaluations	Speech distortion			
SNR (dB)	Factory noise		Car noise	
	MMSE	DAE	MMSE	DAE
0	1.56	**0.59**	0.63	**0.27**
5	1.28	**0.47**	0.59	**0.24**
10	1.05	**0.43**	0.57	**0.21**

Table 10: Evaluation based on PESQ.

Evaluations	PESQ			
SNR (dB)	Factory noise		Car noise	
	MMSE	DAE	MMSE	DAE
0	1.22	**2.82**	2.90	**3.98**
5	1.73	**3.19**	3.05	**4.09**
10	2.15	**3.39**	3.17	**4.18**

5. Conclusion and discussions

Deep learning has been successfully applied in pattern classification and signal processing, particularly in acoustic modeling for ASR. Based on the same idea, we have applied the DAE for noise reduction and speech enhancement [7]. In this study, we further introduced a denoising processing in training the AE by using noisy-clean speech pairs. The advantage of this method is that the DAE automatically learns the statistical difference between speech and noise which helps to separate speech and noise for speech enhancement. In our experiments, we confirmed that increasing depth of the DAE helps for speech enhancement. In addition, compared with traditional speech enhancement methods, the DAE can explore nonlinear and high order statistical information for speech enhancement. It is similar as projecting noisy speech in a nonlinear kernel space for a better separation of noise and speech by using high order statistical information. However, the nonlinear kernel space explored by the DAE is automatically learned from noisy-clean speech pairs which is much more suitable for denoising than using a given kernel function.

Many issues need to be further investigated. The first one is how to effectively incorporate prior knowledge in modeling the DAE. For example, speech signal has many well structured, multi-scale temporal-frequency patterns and transitions. It can be introduced in a hierachical deep network structure for speech enhancement. The second is concerned with how to make the DAE generalize well. We have introduced regularization techniques in section 2. Considering the sparse distribution property of speech, sparse regularization can be a promising regularization technique for DAE [14]. In our future work, we will design a proper sparse regularization technique for DAE. Lastly, in experiments, only two types of noise conditions were tested. In the future, more noise conditions as well as large data set will be examined.

6. References

[1] Loizou, P. C., Speech Enhancement: Theory and Practice, CRC Press, 2007.

[2] Lu, X., Unoki, M., Matsuda, S., Hori, C., Kashioka, H., "Controlling tradeoff between approximation accuracy and complexity of a smooth function in a reproducing kernel Hilbert space for noise reduction," IEEE Trans. on Signal Processing, 61 (3): 601-610, 2013.

[3] Bengio, Y., "Learning deep architectures for AI," Foundations and Trends in Machine Learning, 2(1): 1-127, 2009.

[4] Hinton, G. E., and Salakhutdinov, R., "Reducing the Dimensionality of Data with Neural Networks," Science, 313: 504-507, 2006.

[5] Bengio, Y., Lamblin, P., Popovici, D., and Larochelle, H., "Greedy layer-wise training of deep networks," In Advances in Neural Information Processing Systems, 19: 153-160, MIT Press, Cambridge, 2007.

[6] Ranzato, M. A., Huang, F. J., Boureau, Y. L., LeCun, Y., "Unsupervised Learning of Invariant Feature Hierarchies with Applications to Object Recognition," IEEE conference on Computer Vision and Pattern Recognition, 1-8, 2007.

[7] Lu, X. Matsuda, S., Hori, C., Kashioka, H., "Speech restoration based on deep learning autoencoder with layer-wised learning," INTERSPEECH, Portland, Oregon, Sept., 2012.

[8] Dahl, G., Yu, D., Deng, L., Acero, A., "Context-dependent pre-trained deep neural networks for large vocabulary speech recognition," IEEE Transactions on Audio, Speech, and Language Processing, 20 (1): 30-42, 2011.

[9] Vincent, P., Larochelle, H., Lajoie, I., Bengio, Y., Manzagol, P., "Stacked Denoising Autoencoders: Learning Useful Representations in a Deep Network with a Local Denoising Criterion," Journal of Machine Learning Research, 11(Dec): 3371-3408, 2010.

[10] Maas, A, Le, Q., O'Neil, T., Vinyals, O., Nguyen, P., Ng, A, "Recurrent Neural Networks for Noise Reduction in Robust ASR," Interspeech 2012, Portland, 2012.

[11] Schmidt, M., Van Den Berg, E., Friedl, M. P., Murphy, K., "Optimizing costly functions with simple constraints: A limited-memory projected quasi-newton algorithm," in Proc. of Conf. on Artificial Intelligence and Statistics, 456-463, 2009.

[12] Deng, L., Seltzer, M., Yu, D., Acero, A., Mohamed, A., Hinton, A., "Binary Coding of Speech Spectrograms Using a Deep Auto-encoder," in Proc. of Interspeech, 1692-1695, 2010.

[13] Cohen, I., "Noise spectrum estimation in adverse environments: improved minima controlled recursive averaging," IEEE Trans. Speech Audio Process. 11 (5): 466-475, 2003.

[14] Lee, H., Ekanadham, C., and Ng, A. Y., "Sparse deep belief net model for visual area V2," in Advances in Neural Information Processing Systems (NIPS), 20: 873-880, 2008.