Environmental Noise Reduction based on Deep Denoising Autoencoder

Aneeka Azmat
Institute of Human-Centered Computing, National Tsinghua University, Taiwan and Institute of Information Science, Academia Sinica, Taiwan
aneekaazmat89@iis.sinica.edu.tw

Imad Ali
Department of Computer Science
University of Swat, Pakistan
imadali@uswat.edu.pk

M. Gilvy Langgawan Putra
National Taiwan University of Science and Technology, Taiwan and Institut Teknologi Kalimantan, Indonesia
gilvy.langgawan@lecturer.itk.ac.id

Whenty Ariyanti
National Taiwan University of Science and Technology, Taiwan and Institute of Information Science, Academia Sinica, Taiwan
whenty@iis.sinica.ac.id

Talha Nadeem
COMSATS University
Islamabad, Pakistan
talhanadeem2397@gmail.com

Received: 4 August 2022 | Revised: 31 August 2022 | Accepted: 1 September 2022

Abstract-Speech enhancement plays an important role in Automatic Speech Recognition (ASR) even though this task remains challenging in real-world scenarios of human-level performance. To cope with this challenge, an explicit denoising framework called Deep Denoising Autoencoder (DDAE) is introduced in this paper. The parameters of DDAE encoder and decoder are optimized based on the backpropagation criterion, where all denoising autoencoders are stacked up instead of recurrent connections. For better speech estimation in real and noisy environments, we include matched and mismatched noisy and clean pairs of speech data to train the DDAE. The DDAE has the ability to achieve optimal results even for a limited amount of training data. Our experimental results show that the proposed DDAE outperformed the baseline algorithms. The DDAE shows superior performances based on three-evaluation metrics in noisy and clean pairs of speech data compared to three baseline algorithms.

Keywords- DDAE; limited data; noise reduction; autoencoders

I. INTRODUCTION

Speech enhancement is a key part of Automatic Speech Recognition (ASR) [1]. It has been implemented in many commercial products for decades, but in order to achieve human-level performance, it requires further improvement in real-world scenarios [2], e.g. when we talk to a person or a voice recognition system like an ATM verification call, many environmental noises are added and transmitted, posing difficulties to the system. To deal with such noises, speech signal processing algorithms have been developed to improve the quality and intelligibility.

Neural network-based algorithms are more efficient in learning high-order statistical information automatically by using nonlinear processing units and are efficient for noise reduction [3, 4]. Many algorithms are proposed to train deep neural networks efficiently for speech enhancement and noise reduction [5-7]. Deep learning algorithms have been used to extract speech features and for acoustic modeling [2]. The denoising autoencoder has been used for image processing and other classification applications to extract robust features when the input is a binary masked feature for each autoencoder. For speech feature extraction, a recurrent denoising autoencoder for ASR was proposed for noise reduction in [8]. There are many conventional methods for speech signal processing. For speech enhancement, the most used algorithms are Log Minimum Mean Square Error (logMMSE) [9], Karhunen-Loeve Transforms (KLT) [10], and Robust Principal Component Analysis (RPCA) [11], which are designed with specific additive background noise based on statistical speech properties and noisy signals. However, these methods do not work in real settings when people talk in a noisy environment. Consequently, it is essential to build a framework to deal with the various types of noise in real-world environments.

This article introduces an explicit denoising framework called the Deep Denoising Autoencoder (DDAE), a variant of the denoising autoencoder. However, unlike previous works, where the input is noisy speech and the output is clean speech

Corresponding author: Aneeka Azmat

for training, the DDAE is trained on datasets of matched noisy and clean pairs of speech and mismatched noisy and clean pairs of speech for speech estimation in real and live environments. In the DDAE, the parameters in the encoder and decoder (formed by multiple layers of neural networks) are optimized based on the backpropagation criterion, where all denoising autoencoders are stacked up instead of recurrent connections. The DDAE has the ability to achieve optimal results even with a limited amount of training data. Our experimental results of DDAE were compared with the speech enhancement algorithms RPCA, KLT, and logMMSE. Based on three evaluation metrics, the DDAE outperforms these three baseline algorithms in matched and mismatched noisy and clean speech.

II. THE DDAE FRAMEWORK

This article uses the DDAE framework presented in [12], which removes the noise quickly and efficiently. The DDAE can achieve optimal results even with inadequate training data. It uses the noisy signal to produce clean speech features efficiently, even under mismatched testing conditions. Unlike [13], where the denoising autoencoder was trained using clean speech only, we trained the denoising autoencoder with both noisy and clean speech data. Figure 1 shows a single-layer hidden neural autoencoder, trained on clean speech with noisy and speech data as input. It consists of one nonlinear and one linear encoding and decoding stage:

$$\left.\begin{aligned} h(\boldsymbol{y}_i) &= \sigma(\boldsymbol{W}_1\boldsymbol{y}_i + \boldsymbol{b}) \\ \hat{\boldsymbol{x}}_i &= \boldsymbol{W}_2 h(\boldsymbol{y}_i) + \boldsymbol{c} \end{aligned}\right\} \quad (1)$$

where $\boldsymbol{W}_1$ and $\boldsymbol{W}_2$ are the weight matrices for encoding and decoding neural network connections respectively. Typically, regularization takes place when $\boldsymbol{W}_1 = \boldsymbol{W}_2^{\boldsymbol{T}} = \boldsymbol{W}$. Also, $\boldsymbol{b}$ is the input layer's bias vector, whereas $\boldsymbol{c}$ represents the output layer's bias vector. Similarly, $\boldsymbol{y}_i$ is the noisy speech signal equivalent to the clean signal $\boldsymbol{x}_i$. The nonlinear logistic function utilized by the hidden neuron is $\sigma(\boldsymbol{x}) = (1 + e^{-x})^{-1}$. By optimizing the objective function in (2), the parameters are determined as follows:

$$L(\theta) = \sum_i ||\boldsymbol{x}_i - \hat{\boldsymbol{x}}_i||_2^2 \quad (2)$$

where the set of the parameters is defined as $\theta = \{\boldsymbol{W}, \boldsymbol{b}, \boldsymbol{c}\}$. In addition to using $\boldsymbol{W}_1 = \boldsymbol{W}_2^{\boldsymbol{T}} = \boldsymbol{W}$, we apply regularization on weights and hidden neural output. This improves the generalization and prevents overfitting, which is defined as follows:

$$J(\theta) = L(\theta) + \alpha\, ||\boldsymbol{W}||_2^2 + \beta_p(h(\boldsymbol{y}_i)) \quad (3)$$

where $||\boldsymbol{W}||_2^2 = \sum_{ij} w_{ij}^2 . p\left(h(\boldsymbol{y}_i)\right)_{ij}^2$ is a regularization function defined on the output neurons of the hidden layers, and α and β are its weight coefficients. Thus, the parameters may be derived as follows:

$$\theta^* \triangleq \arg\min_\theta J(\theta) \quad (4)$$

There are numerous unconstrained optimization algorithms to solve (4) for the estimation of $\boldsymbol{W}^*, \boldsymbol{b}^*, \boldsymbol{c}^*$. In this framework, the linear search-based algorithm of quasi-Newton is applied.

Figure 1 depicts the DDAE training procedure using noisy and clean voice samples. The DDAE can be constructed by stacking up several autoencoders. For the training of the DDAE, a dense layer of wised pretraining and fine-tuning is employed. When another hidden layer is introduced during the pretraining phase, the output of the previous hidden layer is used as the input of the subsequent autoencoder. The transformed noisy and clean speech data are used for training in denoising. As illustrated in Figure 1, the training pair for the first autoencoder is $\boldsymbol{y}$ and $\boldsymbol{x}$, and followed by $h\boldsymbol{y}_i$ and $h\boldsymbol{x}_i$ or the subsequent autoencoder. During the fine-tuning step, the initial network parameters are fixed as those obtained from the pretraining step. These training stages may produce a better overall result than training the DDAE with random initialization.

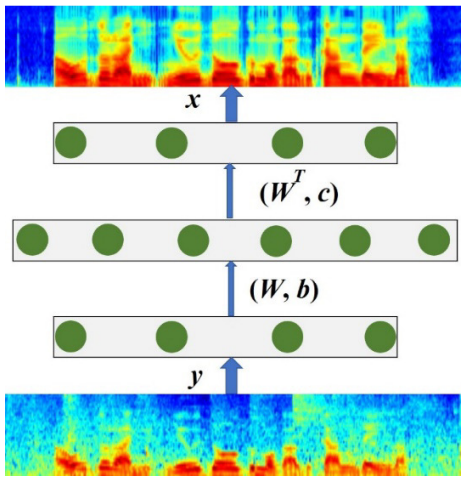

Fig. 1. The training process of the DDAE with noisy and clean speech data.

III. EXPERIMENTAL SETTING

This section evaluates the proposed DDAE framework and the baseline algorithms for speech enhancement tasks. A clean, continuous Taiwan Mandarin Hearing In Noise Test (TMHINT) data set with 100 utterances was used for training by adding 5 noise types (machine, babble, party crowd, restaurant, vacuum cleaner) and 5 signal-to-noise ratios (SNR). Therefore, the total training dataset for noisy speech contains 2500 utterances. The first testing scenario is a matching case in which 40 utterances are added with the same two types of noises (babble and party crowd) with 5 SNR levels. Thus, the total number of match case utterances are 400. More details of the dataset can be found in [14].

We have investigated the accuracy of noise reduction, speech enhancement, and intelligibility using 3 different metrics with a defined standard range set. The range for PESQ is between –0.5 to 4.5. As the value of PESQ increases, the quality of speech enhancement will increase. Similarly, the range for Short-Time Objective Intangibility (STOI) is between 0 and 1. As the value of STOI increases, the number of words identified by the listener will increase. The Speech Distortion Index (SDI) is the opposite of the first two metrics, ranging between 0 and 1. As the value of SDI decreases, so does the quality of speech enhancement. Most conventional speech enhancement algorithms are designed to filter out noisy speech

using the gain function estimation. This article compares 3 traditional algorithms, logMMSE, KLT, and RPCA, with the proposed DDAE. The unprocessed speech is named Noisy. The experimental results were evaluated using 2 matched noise types with two SNR conditions (6dB, -6dB) and 4 mismatched noise types with 3 SNR conditions (-2, 0, 2), as shown in Table I. The DDAE was trained using 5 layers, each containing 2048 neurons..

TABLE I. EXPERIMENTAL SETUP OF THE TMHINT DATASET

Train set	2500 utterances Noise type: Machine, cafeteria babble, crowd party, restaurant, vacuum cleaner SNR levels: -6dB, -3dB, 3dB, 6dB, 10dB
Test set	**Matched condition** 400 utterances Noise type: Cafeteria babble, crowd party SNR levels: -6dB, 6dB **Mismatched condition** 800 utterances Noise type: Applause, baby cry, grocery store, pink noise SNR levels: -2dB, 0dB, 2dB

IV. RESULTS

In this section, we evaluate and compare the results of the baseline algorithms with DDAE based on 3 evaluation criteria using different scenarios and SNR levels as mentioned in Table I.

Table II shows the average PESQ results, normalized across the different SNRs. Compared to the baseline algorithms, the DDAE framework has a better PESQ score, i.e. improved speech quality, for both matched and mismatched noise types. Table II demonstrates that the DDAE framework (5 hidden layer configuration, each layer containing 2048 hidden neurons) outperformed the 3 baselines algorithms, with the exception of most pink noise, where the logMMSE utperformed the DDAE framework and the other algorithms under mismatched stationery noise conditions. In addition, the DDAE exhibited a better PESQ score of speech enhancement than logMMSE, RPCA, and KLT by maintaining high scores for STOI. For instance, the average PSESQ score of the DDAE is 2.55, while the PESQ scores of RPCA, logMMSE, and KLT are 2.18, 2.15, and 1.90 respectively. Compared to the baseline algorithms, on average, the PESQ score of the DDAE is higher than RPCA, logMMSE, and KLT by 16.97 %, 18.60%, and 34.21%, respectively.

TABLE II. DDAE COMPARISON WITH THE BASELINE ALGORITHMS BASED ON THE AVERAGE PESQ SCORE

Noise type	**Framework**				
	Noisy	**KLT**	**logMMSE**	**RPCA**	**DDAE**
Matched noise (seen noise)					
Babble	2.26	1.89	2.21	2.25	2.58
Crowd party	2.26	1.85	2.15	2.24	2.61
Mismatched noise (unseen noise)					
Applause	2.13	1.76	1.89	1.99	2.95
Baby cry	2.17	1.77	1.97	2.02	2.50
Grocery	2.19	1.80	2.09	2.18	2.26
Pink noise	2.32	2.38	2.58	2.40	2.38
AVG	2.23	1.90	2.15	2.18	2.55

Table III illustrates the average STOI score for the DDAE and baseline algorithms, with 6 noise types with different SNRs. The Table shows that the DDAE has a better STOI score for all types of noise. Moreover, the DDAE demonstrated better average speech enhancement capabilities by maintaining high STOI scores. Compared to the baseline algorithms, the average STOI score of the DDAE is higher than RPCA, logMMSE, and KLT by 16.39%, 65.11%, and 10.93% respectively.

TABLE III. ALGORITHM COMPARISON BASED ON AVERAGE STOI SCORE

Noise Type	**Framework**				
	Noisy	**KLT**	**logMMSE**	**RPCA**	**DDAE**
Matched Noise (seen noise)					
Babble	0.58	0.58	0.38	0.55	0.68
Crowd party	0.63	0.61	0.46	0.66	0.71
Mismatched Noise (unseen noise)					
Applause	0.69	0.68	0.57	0.697	0.73
Baby cry	0.71	0.71	0.54	0.71	0.73
Grocery	0.45	0.58	0.29	0.45	0.60
Pink noise	0.52	0.68	0.31	0.52	0.74
AVG	0.59	0.64	0.43	0.61	0.71

Table IV shows the average SDI results averaged over the different SNRs for the 6 considered noise types. The Table shows that the DDAE performs better than the other algorithms for all types of noise. In addition, the DDAE demonstrated better speech enhancement capabilities. Compared to the baseline algorithms, the SDI score of the DDAE is 22.64%, 24.07%, and 12.96%, lower than the RPCA, logMMSE, and the performance of DDAE significantly outperformed the them. Figure 2 depicts the spectrogram of a test utterance contaminated with non-stationary noise applause at SNR = -2dB achieved by different models. For comparison, the spectrograms of clean and noisy speech signals are also presented in Figure 2(a)-(b). The spectrograms of the test utterance enhanced by logMMSE, KLT, and RPCA algorithms are shown in Figures 2(c)-(e), while Figure 2(f) depicts the spectrogram of the enhanced speech signals produced by the DDAE. Figure 2 clearly demonstrates that the DDAE framework efficiently restores clean speech under very challenging conditions (non-stationery noise, -2dB SNR) and effectively suppresses noise components from the noisy signal (Figure 2(b)). DDAE suppresses noise components more effectively and produces better speech quality than the previous approaches.

TABLE IV. ALGORITHM COMPARISON BASED ON THE AVERAGE SDI SCORE

Noise Type	**Framework**				
	Noisy	**KLT**	**logMMSE**	**RPCA**	**DDAE**
Matched noise (seen noise)					
Babble	0.61	0.45	0.48	0.50	0.40
Crowd party	0.61	0.48	0.48	0.50	0.38
Mismatched noise (unseen noise)					
Applause	0.61	0.55	0.585	0.58	0.41
Baby cry	0.61	0.56	0.59	0.56	0.38
Grocery	0.61	0.52	0.55	0.53	0.50
Pink noise	0.61	0.25	0.57	0.53	0.39
AVG	0.61	0.47	0.54	0.53	0.41

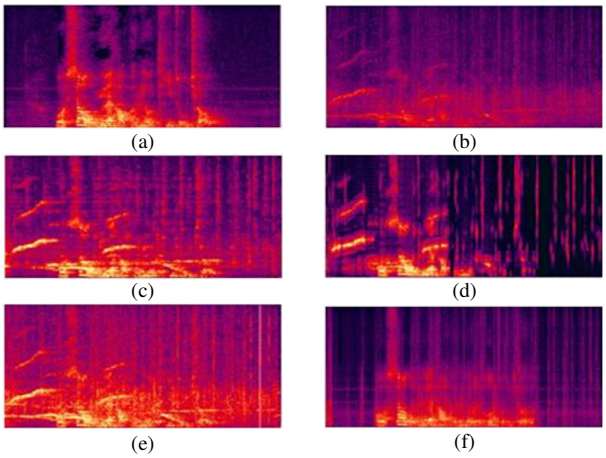

Fig. 2. Spectrograms of (a) clean, (b) noisy, (c) logMMSE, (d) KLT, (e) RPCA, and (f) DDAE. A noise of -2dB was added to the test utterance.

V. DISCUSSION AND CONCLUSION

In this study, the DDAE framework for speech enhancement of environmental noise is presented. The proposed framework provides a more understandable speech to the listener. The major contributions of this study are the introduction of the DDAE framework, which enhances speech intelligibility over 4 different baseline methods, and the confirmation, based on experimental data, that this proposed framework can improve speech performance. These contributions allow our proposed architecture to serve as a straightforward and efficient means of bridging the acoustic mismatch condition.

The DDAE framework provides better generalization and a universal approximation capability. For mismatch and non-stationery conditions, DDAE has better generalization performance compared to the conventional speech enhancement techniques. The experimental results demonstrate that even with insufficient training data, our proposed framework outperformed KLT, logMMSE, and RPCA under both matched and mismatched noise conditions at varying SNR levels. This is confirmed by the evaluation results, which show that DDAE outperforms the other algorithms on all levels, except for stationary noise where the logMMSE approach performs better. In addition, the comparative findings demonstrate that our method provides superior enhancement performance based on 3 widely used objective metrics. It is essential to find the ideal compromise between the parameter efficiency and the speech enhancement performance of the model. However, the majority of the baseline models cannot deliver great parameter efficiency. The experimental results suggest that the proposed model is superior to other speech enhancement frameworks in terms of trade-off.

REFERENCES

[1] W. Helali, Z. Hajaiej, and A. Cherif, "Real time speech recognition based on PWP thresholding and MFCC using SVM," *Engineering, Technology & Applied Science Research*, vol. 10, no. 5, pp. 6204-6208, Oct., 2020, https://doi.org/10.48084/etasr.3759.

[2] G. E. Dahl, D. Yu, L. Deng, and A. Acero, "Context-Dependent Pre-Trained Deep Neural Networks for Large-Vocabulary Speech Recognition," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 1, pp. 30–42, Jan. 2012, https://doi.org/10.1109/TASL.2011.2134090.

[3] X. Lu, M. Unoki, S. Matsuda, C. Hori, and H. Kashioka, "Controlling Tradeoff Between Approximation Accuracy and Complexity of a Smooth Function in a Reproducing Kernel Hilbert Space for Noise Reduction," *IEEE Transactions on Signal Processing*, vol. 61, no. 3, pp. 601–610, Oct. 2013, https://doi.org/10.1109/TSP.2012.2229991.

[4] Y. Bengio, "Learning Deep Architectures for AI," *Foundations and Trends® in Machine Learning*, vol. 2, no. 1, pp. 1–127, Nov. 2009, https://doi.org/10.1561/2200000006.

[5] A. A. Alasadi, T. H. Aldhayni, R. R. Deshmukh, A. H. Alahmadi, and A. S. Alshebami, "Efficient Feature Extraction Algorithms to Develop an Arabic Speech Recognition System," *Engineering, Technology & Applied Science Research*, vol. 10, no. 2, pp. 5547–5553, Apr. 2020, https://doi.org/10.48084/etasr.3465.

[6] A. Samad, A. U. Rehman, and S. A. Ali, "Performance Evaluation of Learning Classifiers of Children Emotions using Feature Combinations in the Presence of Noise," *Engineering, Technology & Applied Science Research*, vol. 9, no. 6, pp. 5088–5092, Dec. 2019, https://doi.org/10.48084/etasr.3193.

[7] M. Ranzato, F. J. Huang, Y.-L. Boureau, and Y. LeCun, "Unsupervised Learning of Invariant Feature Hierarchies with Applications to Object Recognition," in *2007 IEEE Conference on Computer Vision and Pattern Recognition*, Minneapolis, MN, USA, Jun. 2007, https://doi.org/10.1109/CVPR.2007.383157.

[8] A. L. Maas, Q. V. Le, T. M. O'Neil, O. Vinyals, P. Nguyen, and A. Y. Ng, "Recurrent Neural Networks for Noise Reduction in Robust ASR," in *Proceedings of The International Conference on Acoustics, Speech, & Signal Processing*, Dec. 2012.

[9] Y. Ephraim and D. Malah, "Speech enhancement using a minimum mean-square error log-spectral amplitude estimator," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 33, no. 2, pp. 443–445, Apr. 1985, https://doi.org/10.1109/TASSP.1985.1164550.

[10] U. Mittal and N. Phamdo, "Signal/noise KLT based approach for enhancing speech degraded by colored noise," *IEEE Transactions on Speech and Audio Processing*, vol. 8, no. 2, pp. 159–167, Mar. 2000, https://doi.org/10.1109/89.824700.

[11] E. J. Candès, X. Li, Y. Ma, and J. Wright, "Robust principal component analysis?," *Journal of the ACM*, vol. 58, no. 3, pp. 11:1-11:37, Mar. 2011, https://doi.org/10.1145/1970392.1970395.

[12] X. Lu, Y. Tsao, S. Matsuda, and C. Hori, "Speech enhancement based on deep denoising autoencoder," in *Interspeech 2013*, Aug. 2013, pp. 436–440, https://doi.org/10.21437/Interspeech.2013-130.

[13] P. Vincent, H. Larochelle, I. Lajoie, Y. Bengio, and P.-A. Manzagol, "Stacked Denoising Autoencoders: Learning Useful Representations in a Deep Network with a Local Denoising Criterion," *The Journal of Machine Learning Research*, vol. 11, pp. 3371–3408, Sep. 2010.

[14] L. L. N. Wong, S. D. Soli, S. Liu, N. Han, and M.-W. Huang, "Development of the Mandarin Hearing in Noise Test (MHINT)," *Ear and Hearing*, vol. 28, no. 2 Suppl, pp. 70S-74S, Apr. 2007, https://doi.org/10.1097/AUD.0b013e31803154d0.