SELF-SUPERVISED DENOISING AUTOENCODER WITH LINEAR REGRESSION DECODER FOR SPEECH ENHANCEMENT

Ryandhimas E. Zezario[12], *Tassadaq Hussain*[3], *Xugang Lu*[4], *Hsin-Min Wang*[5], *Yu Tsao*[1]

[1]Research Center for Information Technology Innovation, Academia Sinica, Taiwan
[2]Department of Computer Science and Information Engineering, National Taiwan University, Taiwan
[3]Taiwan International Graduate Program in Social Network and Human-Centered Computing, Institute of Information Science, Academia Sinica, Taiwan
[4]National Institute of Information and Communications Technology, Japan
[5]Institute of Information Science, Academia Sinica, Taiwan

ABSTRACT

Nonlinear spectral mapping-based models based on supervised learning have successfully applied for speech enhancement. However, as supervised learning approaches, a large amount of labelled data (noisy-clean speech pairs) should be provided to train those models. In addition, their performances for unseen noisy conditions are not guaranteed, which is a common weak point of supervised learning approaches. In this study, we proposed an unsupervised learning approach for speech enhancement, i.e., denoising autoencoder with linear regression decoder (DAELD) model for speech enhancement. The DAELD is trained with noisy speech as both input and target output in a self-supervised learning manner. In addition, with properly setting a shrinkage threshold for internal hidden representations, noise could be removed during the reconstruction from the hidden representations via the linear regression decoder. Speech enhancement experiments were carried out to test the proposed model. Results confirmed that the proposed DAELD could achieve comparable and sometimes even better enhancement performance as compared to the conventional supervised speech enhancement approaches, in both seen and unseen noise environments. Moreover, we observe that higher performances tend to achieve by DAELD when the training data cover more diverse noise types and signal-to-noise-ratio (SNR) levels.

Index Terms— *speech enhancement, deep denoising autoencoder, unsupervised learning.*

1. INTRODUCTION

Speech enhancement aims to retrieve clean speech signals from noisy ones and serves as an important pre-processor in many speech-related tasks, such as automatic speech recognition (ASR) [1–3], assistive listening [4–8], speech coding [9, 10], and speaker recognition [11, 12] systems. In the past, numerous speech enhancement approaches have been proposed. Notable examples include minimum-mean-square-error (MMSE) short-time spectral amplitude estimator [13], spectral subtraction [14], Karhunen–Loeve transform [15], and Wiener filter [16]. These traditional approaches are derived based on statistical assumptions of speech and noise signals. When the assumption fails, the speech enhancement performance can be notably degraded. More recently, various machine learning (ML)-based speech enhancement methods have been proposed. Among them, deep-learning-based methods have caught great attention in recent years [17-29]. Generally, the deep-learning-based methods use a deep-neural-network (DNN) model to perform noisy-to-clean speech transformation in a supervised learning fashion. More specifically, to train a high-performance DNN-based speech enhancement system, a large amount and a wide variety of noisy-clean training pairs is required. When the training samples are insufficient, the model cannot be trained well, and thus the enhancement performance can be limited. Meanwhile, when operating in an environment of unseen noise types or speakers, the enhancement performance can be notably degraded.

Different from supervised learning, unsupervised learning does not require labelled training data and thus can handle the data dependency problem. In the literature, deep learning models trained with unsupervised learning can extract essential representations from the salient structure of the input data. One notable unsupervised learning model is autoencoder, which consists of an encoder and a decoder. The encoder transforms the input physical data into latent features, which are then reconstructed to the original data by the decoder [30]. By stacking multiple layers of encoder-decoder architectures, we can establish a deep autoencoder (DAE) model. In previous studies, the DAE model has been used to perform dimensionality reduction [30], face recognition [31] and natural language processing [32].

Several DAE-based unsupervised learning frameworks have been proposed for the speech enhancement task. For instance, a DAE was successfully implemented and trained using clean magnitude spectrum [33]. In that system, clean speech signals are placed in both input and output to train the DAE model. During testing, the DAE tries to "recall" clean components from noisy speech utterances. In [34], the authors proposed a two-staged approach by stacking a pair of supervised and unsupervised DAE for robust speech enhancement. The first DAE acts as a regular speech enhancement model, while the second DAE performs as a purity checker in an unsupervised manner during the testing stage [34]. More recently, the authors in [35] followed a similar strategy by implementing a DAE as a selection model to choose the most suitable for speech enhancement using reconstruction error. Although DAE has shown notable improvement for robust speech enhancement performance, it mostly built as an additional system to support the supervised speech enhancement systems. Recently, a self-supervised speech denoising system was proposed [36]. The system used a dual-microphone setup and considers two noisy realizations of a clean speech signal as the input and the output target.

In the past, we have proposed a series of speech enhancement works based on the deep denoising encoder-decoder architecture (DDAE) [17, 37, 38]. Both the encoder and decoder of DDAE are formed by multiple layers of neural networks. The parameters in the encoder and decoder are optimized based on the backpropagation (BP) criterion in a supervised learning fashion (require paired noisy-clean training data). Although DDAE has been confirmed to achieve satisfactory performance in the speech enhancement task, its applicability is confined by a well-known deep and supervised learning issue: the requirement of a large amount of paired noisy-clean training data. The goal of the present study is to address the above limitation. We first investigate to use a linear regression function for the decoder in DDAE (termed DAELD in this study) to simply the overall model architecture. Next, we attempt to train the DAELD model in an unsupervised learning fashion.

In this study, two types of DAELD have been established, and the difference lies in the criteria used to estimate the parameters in the encoder. The first DAELD model trains the encoder by BP, and thus termed DAELD$_{(BP)}$; the second model builds the encoder by stacking multiple sparse hierarchical extreme learning machine (HELM) [39] based autoencoders, and the model is termed DAELD$_{(SAE)}$. We tested the proposed DAELD in both supervised learning and unsupervised learning (self-supervised learning) modes on two datasets, namely Aurora-4 [40] and TIMIT [41]. Experimental results confirm that DAELD$_{(SAE)}$ achieves a higher perceptual evaluation of speech quality (PESQ) [42] as compare to DAELD$_{(BP)}$ in unseen noise types while DAELD$_{(SAE)}$ and DAELD$_{(BP)}$ achieve comparable short-time objective intelligibility (STOI) [43] scores in both seen and unseen noise environments. Moreover, compared to the supervised DAE based SE system [17], the proposed DAELD framework achieves better speech quality (PESQ) and intelligibility (STOI) when the training data cover more diverse noise types and signal-to-noise-ratio (SNR) levels.

The remainder of this paper is organized as follows. Section II introduces the proposed DAELD. Section III describes the experimental setup and results. Section IV concludes our findings.

2. THE PROPOSED DAELD SYSTEM

The proposed DAELD follows the same encoder-decoder structure as DDAE to perform speech enhancement, which is shown in Fig. 1. The DAELD model adopts the encoder and decoder layers to convert speech signals to high-dimensional feature representations and convert the representations back to speech signals, respectively. Different from the previous DAE based SE systems [17], which uses clean speech signals as the output target, we calculate the weights in the encoder in an unsupervised self-learning training criterion. The overall DAELD model consists of two stages, namely offline and online stages. In the offline stage, we estimate the parameters of the encoder and decoder that are formed by non-linear and linear functions, respectively. In the online stage, the noisy speech signals are first processed by the encoder to obtain high-dimensional feature representations, which are then transformed to obtain enhanced speech signals by the linear transformation estimated in the offline stage. More details will be given in the following discussion.

2.1 The offline stage

In the offline stage, given speech feature ($\boldsymbol{Y} = [\boldsymbol{y}_1, \dots, \boldsymbol{y}_n \dots, \boldsymbol{y}_N]^T$), the DAELD approach intends to estimate a mapping function:

$$\hat{\boldsymbol{x}}_{\boldsymbol{n}} = F(\boldsymbol{y}_{\boldsymbol{n}}), \tag{1}$$

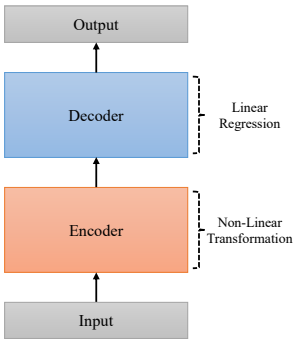

Fig. 1: The architecture of the DAELD model, which is a DDAE model using a linear regression function as the decoder.

where $\hat{\boldsymbol{x}}_{\boldsymbol{n}}$ and $F(.)$ indicate n-th enhanced speech feature and the DAELD transformation, respectively; N is the total number of training samples. As mentioned earlier, we build two types of DAELD models, namely DAELD $_{(SAE)}$ and DAELD $_{(BP)}$.

For DAELD $_{(SAE)}$, we compute the encoder by hierarchical layers of sparse autoencoders based on a fast-iterative shrinkage-threshold algorithm (FISTA) [44] in an unsupervised learning manner. The goal of the encoder in DAELD $_{(SAE)}$ is to extract representative features through the self-supervised learning process. In the decoder layer of DAELD $_{(SAE)}$, the linear transformation $\boldsymbol{\beta}_{SAE}$ is linearly estimated by performing Moore-Penrose (MP) pseudoinverse:

$$\boldsymbol{\beta}_{SAE} = \left(\delta \boldsymbol{I} + \boldsymbol{H}_{SAE}{}^T \boldsymbol{H}_{SAE}\right)^{-1} \boldsymbol{H}_{SAE}{}^T \boldsymbol{Y}, \tag{2}$$

where $\boldsymbol{I}, \boldsymbol{H}_{SAE}$ and $\boldsymbol{Y}$ indicate identity matrix, hidden layer output ($\boldsymbol{H}_{SAE} = \left[\left[\boldsymbol{h}_{SAE_1}, \dots, \boldsymbol{h}_{SAE_n} \dots, \boldsymbol{h}_{SAE_N}\right]^T \alpha \boldsymbol{1}\right]$, $\boldsymbol{h}_{SAE_n}$ is the hidden layer output of the n-th input feature vector, α is scaling factor, and $\boldsymbol{1}$ is an all-one vector), and training target ($\boldsymbol{Y} = [\boldsymbol{y}_1, \dots, \boldsymbol{y}_n \dots, \boldsymbol{y}_N]^T$), respectively. As compared to supervised methods, our proposed approach does not require clean speech features as the output for fine-tuning the model's weights.

On the other hand, DAELD $_{(BP)}$ first performs BP algorithm [45] with the minimum mean square error criterion to fine-tune the parameters in the encoder. The noisy speech feature $\boldsymbol{Y}$ is used as the input and target to train DAELD $_{(BP)}$. Once the model has fully optimized, the updated parameter $\{\boldsymbol{W}^0 \dots \boldsymbol{W}^J; \boldsymbol{b}^0 \dots \boldsymbol{b}^J\}$ is applied to estimate the hidden layer output. Finally, based on the generated hidden layer output, the linear regression $\boldsymbol{\beta}_{BP}$ of DAELD $_{(BP)}$ can be estimated as:

$$\boldsymbol{\beta}_{BP} = \left(\delta \boldsymbol{I} + \boldsymbol{H}_{BP}{}^T \boldsymbol{H}_{BP}\right)^{-1} \boldsymbol{H}_{BP}{}^T \boldsymbol{Y}. \tag{3}$$

where $\boldsymbol{H}_{BP} = \left[\left[\boldsymbol{h}_{BP_1}, \dots, \boldsymbol{h}_{BP_n} \dots, \boldsymbol{h}_{BP_N}\right]^T \alpha \boldsymbol{1}\right]$, and $\boldsymbol{h}_{BP_n}$ is the hidden layer output of the n-th input feature vector.

2.2 Online stage

In the online stage of DAELD, given noise speech feature $\overline{\boldsymbol{Y}}$, we obtain hidden layer output $\overline{\boldsymbol{H}}$ by the encoder whose parameters are trained in the unsupervised manner. Based on the estimated linear transformation, $\boldsymbol{\beta}$ (either $\boldsymbol{\beta}_{SAE}$ from Eq. (2) or $\boldsymbol{\beta}_{BP}$ from Eq. (3), the enhanced speech spectral can be estimated as:

$$\hat{\boldsymbol{X}} = \overline{\boldsymbol{H}}\boldsymbol{\beta}, \tag{4}$$

where $\hat{\boldsymbol{X}}$ is the enhanced data, ($\hat{\boldsymbol{X}} = [\hat{\boldsymbol{x}}_1, \dots, \hat{\boldsymbol{x}}_n \dots, \hat{\boldsymbol{x}}_N]^T$).

3. EXPERIMENTS

3.1 Experimental setup

We evaluated the proposed DAELD on Aurora-4 [40] and TIMIT [41] datasets. For the Aurora-4 task, all noisy training utterances from the original Aurora-4 dataset were used to train the speech enhancement models. The training set consisted of 2676 noisy utterances that were corrupted by six types of noises (babble, car, restaurant, street, airport, and train) at SNR levels varying from 10 to 20 dB. For testing, noisy utterances (contaminated with babble and car noises) at SNR levels varying from 5 to 15 dB, were used as the test data.

For the TIMIT dataset, we used the 4620 training utterances and contaminated them by injecting 90 types of noises at eight SNR levels (from -10 dB to 25 dB with steps of 5 dB) into the clean training utterances. More specifically, each utterance was corrupted by one noise type at a particular SNR level. The testing utterances consisted of 100 randomly selected clean utterances (from the TIMIT test set) that were later contaminated with four unseen (two stationary and two non-stationary) noise types under five SNR levels (-12 dB, -6 dB, 0dB, 6dB and 12 dB) to test the enhancement performance. Both training and testing for all the experiments, we used 80-dimensional Mel frequency power spectrum (MFP) as the acoustic feature (the same as used in our previous studies [17, 39]).

We first prepared two DAELD models trained in a self-supervised learning manner, termed DAELD$_{(SAE)}$(u) and DAELD$_{(BP)}$(u). Both models were formed by a three-layered architecture with [1000 1000 16000] hidden nodes. For comparison, we used the same model architecture to train another two DAELD models in a supervised fashion, termed DAELD$_{(SAE)}$(s) and DAELD$_{(BP)}$(s). For comparison purpose, we trained a DDAE [17] using the same architecture. Please note that for DAELD$_{(BP)}$(s) and DDAE, paired noisy-clean training data were used to calculate both encoder and decoder where the decoders for DAELD$_{(BP)}$(s) and DDAE are linear and nonlinear transformations, respectively. For DAELD$_{(SAE)}$(s), the paired noisy-clean training data were used to calculate the linear regression function (in Eq. (2)), and encoder is still trained in a noisy-noisy self-supervised manner. Moreover, DDAE used nonlinear transformations for both encoder and decoder; the DAELD models used a nonlinear transformation for encoder and linear transformation for decoder. In addition, a traditional MMSE speech enhancement algorithm was also tested for comparison.

3.2 Objective evaluation results

We first report the objective evaluation results on the Aurora-4 task. The PESQ and STOI results are shown in Figs 2 and 3, respectively, where the average results over different SNR levels were reported. The results of unprocessed noisy speech are also listed and denoted as Noisy. From Fig. 2, we first note that supervised-learning, DAELD$_{(SAE)}$(s) and DAELD$_{(BP)}$(s), and unsupervised-learning, DAELD$_{(SAE)}$(u) and DAELD$_{(BP)}$(u), achieved improvements over Noisy and the traditional MMSE method for both Babble (a non-stationary noise) and Car (a stationary noise) conditions. Next, both DAELD$_{(SAE)}$(s) and DAELD$_{(BP)}$(s) outperform DDAE and MMSE systems in Babble noise, showing the effectiveness of using a linear transformation for the decoder in DAELD. Meanwhile, the unsupervised-learning methods (DAELD$_{(SAE)}$(u) and DAELD$_{(BP)}$(u)) can yield improved results over Noisy while slightly underperformed DDAE and the supervised-learning counterparts. For the Car noise, DAELD$_{(SAE)}$(u) and DAELD$_{(BP)}$(u) achieved higher performances as compared with the their supervised-learning counterparts (DAELD$_{(SAE)}$(s) and DAELD$_{(BP)}$(s)).

From Fig. 3, we first noted that the supervised-learning methods (DAELD$_{(BP)}$(s), DAELD$_{(SAE)}$(s)) achieved a significantly better speech intelligibility in both Babble and Car noise conditions as compared with unsupervised-learning counterparts and DDAE. We also noted that DAELD$_{(BP)}$(u) and DAELD$_{(SAE)}$(u)) achieve low STOI scores while dealing with non-stationary noise environments.

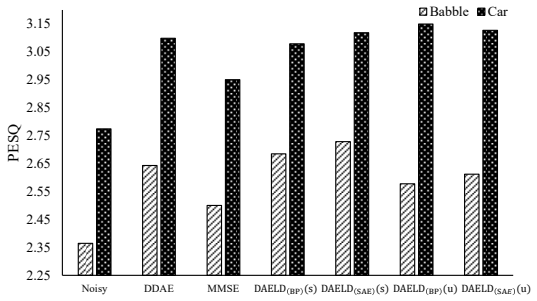

Fig. 2: PESQ performance comparison of noisy, DDAE, MMSE, and four DAELD systems for Aurora 4 dataset.

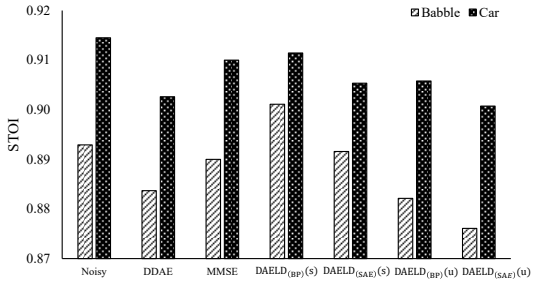

Fig. 3: STOI performance comparison of noisy, DDAE, MMSE and four DAELD systems for Aurora 4 dataset.

Table 1. PESQ comparison of Noisy, DDAE, MMSE and four DAELD systems under stationary and non-stationary noise conditions.

	12	6	0	-6	-12	Ave
Stationary Noises (Car and Engine)						
Noisy	2.45	1.95	1.60	1.39	1.30	1.74
DDAE	2.53	2.18	1.79	1.47	1.32	1.86
MMSE	2.78	2.24	1.81	1.53	1.36	1.94
DAELD$_{(BP)}$(s)	2.63	2.27	1.**89**	1.53	1.35	1.93
DAELD$_{(SAE)}$(s)	2.64	2.27	1.87	1.52	1.37	1.94
DAELD$_{(BP)}$(u)	2.78	2.27	1.86	1.57	1.40	1.98
DAELD$_{(SAE)}$(u)	**2.80**	**2.31**	**1.89**	**1.58**	**1.40**	**2.00**
Non-stationary Noises (Babble and Restaurant)						
Noisy	2.50	2.03	1.71	1.48	1.37	1.82
DDAE	2.61	2.27	1.89	1.58	1.40	1.95
MMSE	2.61	2.10	1.71	1.46	1.26	1.83
DAELD$_{(BP)}$(s)	2.70	2.35	1.98	1.65	1.46	2.03
DAELD$_{(SAE)}$(s)	**2.75**	**2.40**	**2.01**	**1.68**	**1.48**	**2.06**
DAELD$_{(BP)}$(u)	2.70	2.21	1.85	1.59	1.42	1.95
DAELD$_{(SAE)}$(u)	2.73	2.24	1.87	1.59	1.42	1.97

For the TIMIT task, the noises used to prepare the testing data were not involved in the training set. Thus, the enhancement became more challenging since the testing noises were unseen. We listed the PESQ and STOI results Tables 1 and 2, respectively, where the average results for specific SNR levels were reported. The average PESQ and STOI scores are also listed and denoted as "Ave". From Table 1, we can note that all of the four DAELD systems outperform

Noisy, DDAE, and MMSE in terms of average PESQ scores for both stationary and non-stationary noise types. Moreover, when comparing to DAELD$_{(BP)}$(s) and DAELD$_{(SAE)}$(s), DAELD$_{(BP)}$(u) and DAELD$_{(SAE)}$(u) achieve higher PESQ scores in the stationary noise conditions, while lower PESQ scores in the non-stationary noise conditions.

Table 2. STOI comparison of Noisy, DDAE, MMSE and four DAELD systems under stationary and non-stationary noise conditions.

	12	6	0	-6	-12	Ave
Stationary Noise (Car and Engine)						
Noisy	**0.91**	**0.82**	0.68	0.54	**0.43**	**0.67**
DDAE	0.81	0.75	0.65	0.51	0.37	0.62
MMSE	**0.91**	**0.82**	0.68	0.53	0.40	**0.67**
DAELD$_{(BP)}$(s)	0.82	0.76	0.67	0.54	0.40	0.64
DAELD$_{(SAE)}$(s)	0.82	0.76	0.67	**0.55**	0.42	0.65
DAELD$_{(BP)}$(u)	0.89	0.81	0.68	0.54	0.41	**0.67**
DAELD$_{(SAE)}$(u)	0.88	0.81	**0.69**	0.54	0.41	**0.67**
Non-stationary Noise (Babble and Restaurant)						
Noisy	**0.93**	**0.85**	**0.74**	0.62	0.52	**0.73**
DDAE	0.82	0.78	0.71	0.61	0.50	0.68
MMSE	0.92	0.84	0.73	0.61	0.50	0.72
DAELD$_{(BP)}$(s)	0.83	0.79	0.73	**0.63**	0.52	0.70
DAELD$_{(SAE)}$(s)	0.82	0.79	0.72	**0.63**	**0.53**	0.70
DAELD$_{(BP)}$(u)	0.90	0.83	0.73	0.61	0.50	0.72
DAELD$_{(SAE)}$(u)	0.89	0.83	0.73	0.61	0.50	0.71

Next, from Table 2, we noted that the four DAELD models could provide comparable average STOI scores as compared to Noisy. We also noted that under low SNR conditions, such as 0, -6 and -12 SNR levels, DAELD methods are more effective as compared to DDAE and MMSE. The results from Tables 1 and 2 show that the DAELD using either BP or SAE to prepare the encoder, and trained in both supervised and unsupervised fashion, can yield notable speech quality improvements while maintain comparable speech intelligibility as compared to unprocessed noisy speech, DDAE, and MMSE speech enhancement methods.

3.2 Hidden layer analysis

In previous sections, we have reported the evaluation results of DAELD speech enhancement systems in terms of objective evaluation metrics. In this section, we intend to analyze the behaviors of the DAELD models BP and SAE-based encoders by visualizing the hidden layer outputs. Fig. 4 shows the hidden layer representations of a clean utterance (left panels) and its noisy version (right panels). The top two panels present the features from the SAE-based encoders; the bottom two panels show the features from the BP-based encoders. When comparing the top two panels, we note that the representations obtained by clean and noisy utterances show very similar patterns. The results showed that the SAE-based encoders can already suppress noise components from the noisy inputs. Please see the red rectangular regions in both clean and noisy utterances. We observed very similar results from comparing the bottom two panels (please also see the red rectangular regions) in Fig. 4, thus also confirming the BP-based encoders can already suppress noise components. However, the top two and bottom two panels are very different, showing the encoders rained by BP and SAE are rather different.

3.3 Spectrogram analysis

Next, we present the spectrograms to visually compare the speech enhancement results obtained by DAELD and DDAE. Fig. 5 shows

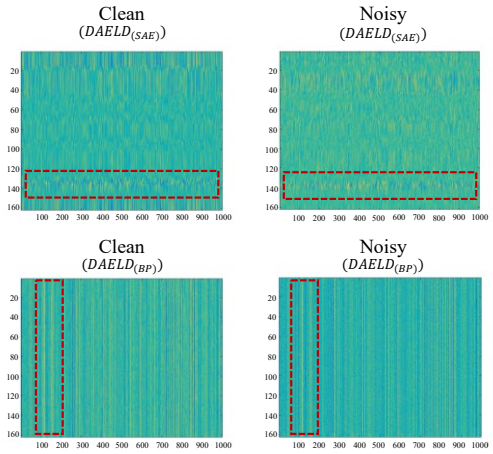

Fig. 4: Hidden layer representation of clean and noisy utterances at DAELD$_{(SAE)}$ and DAELD$_{(BP)}$ SE models.

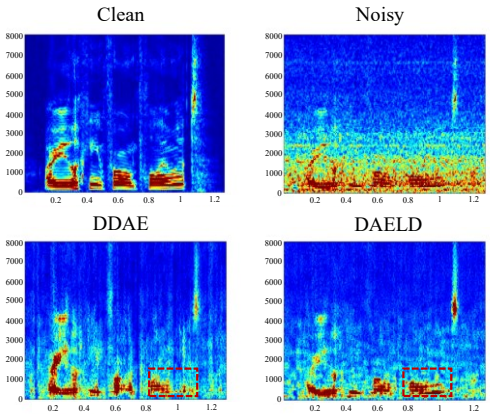

Fig. 5: Spectrograms of a clean utterance (Clean), along with its noisy version (engine noise at 0 dB SNR) (Noisy), and the DDAE and DAELD enhanced ones.

the spectrograms of clean utterance, its noisy version (under engine-noise at 0 dB SNR) and the DDAE and DAELD enhanced ones. From the figure, we can observe that DDAE effectively reduced the noise components, while the DAELD enhanced utterance shows clearer speech structures, as shown in the red rectangular regions.

4. CONCLUSION

The main contribution of this study is two-fold. First, we investigated to use a linear regression function to form the decoder of the DDAE model (termed DAELD) and tested the DAELD model on two speech enhancement tasks (Aurora-4 and TIMIT). Experimental results showed that DAELD achieved comparable and sometimes even better performance in terms of PESQ and STOI as compared to DDAE, where the decoder is a nonlinear transformation. Second, we have investigated the performance of the DAELD system trained in a self-supervised learning fashion. Experimental results showed that the self-supervised trained DAELD can still achieve notable improvements over unprocessed noise and traditional MMSE method, representing a significant step toward the realization of unsupervised speech enhancement using deep learning models. In the future, we will further test DAELD's capability in other speech-processing tasks, such as dereverberation, or multimodal (audio-visual) speech enhancement tasks.

5. REFERENCES

[1] 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, no. 4, pp. 745–777, 2014.

[2] J. Li, L. Deng, R. Haeb-Umbach, and Y. Gong, "Robust automatic speech recognition: a bridge to practical applications." *Academic Press*, 2015.

[3] Z.-Q. Wang and D. Wang, "A joint training framework for robust automatic speech recognition," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 24, no. 4, pp. 796–804, 2016.

[4] P. C. Loizou, "Speech enhancement: theory and practice," *CRC Press*, 2007.

[5] D. Wang, "Deep learning reinvents the hearing aid," *IEEE Spectrum*, vol. 54, no. 3, pp. 32–37, 2017.

[6] G. S. Bhat and C. K. Reddy, "Smartphone based real-time super gaussian single microphone speech enhancement to improve intelligibility for hearing aid users using formant information," *in Proc. EMBC*, pp. 5503–5506, 2018.

[7] H. Levitt, "Noise reduction in hearing aids: an overview," *Journal of Rehabilitation Research and Development*, vol. 38, pp. 111–121, 2001.

[8] 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, no. 1, pp. 61-71, 2015.

[9] R. Martin and R. V. Cox, "New speech enhancement techniques for low bit rate speech coding," *in Proc. IEEE Workshop on Speech Coding,* pp. 165–167, 1999.

[10] Z. Zhao, H. Liu, and T. Fingscheidt, "Convolutional neural networks to enhance coded speech," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 27, no. 4, pp. 663–678, 2019.

[11] S. Shon, H. Tang, and J. Glass, "Voiceid loss: speech enhancement for speaker verification," *arXiv preprint arXiv:1904.03601*, 2019.

[12] M. Kolbk, Z.-H. Tan, and J. Jensen, "Speech enhancement using long short-term memory based recurrent neural networks for noise-robust speaker verification," *in Proc. SLT*, pp. 305–311, 2016.

[13] Y. Ephraim and D. Malah, "Speech enhancement using a minimum mean-square error short-time spectral amplitude estimator," *IEEE Transactions on Acoustic, Speech, and Signal Processing,* vol. 32, pp. 1109–1121, 1984.

[14] S. Boll, "Suppression of acoustic noise in speech using spectral subtraction," *IEEE Transactions on Acoustic, Speech, and Signal Processing,* vol. 27, no. 2, pp. 113–120, 1979.

[15] A. Rezayee and S. Gazor, "An adaptive KLT approach for speech enhancement," *IEEE Transactions on Speech and Audio Processing,* vol. 9, no. 2, pp. 87–95, Feb. 2001.

[16] P. Scalart and J. V. Filho, "Speech enhancement based on a priori signal to noise estimation," *in Proc. ICASSP*, pp. 629–632, 1996.

[17] X. Lu, Y. Tsao, S. Matsuda, and C. Hori, "Speech enhancement based on deep denoising autoencoder," *in Proc. INTERSPEECH*, pp. 436–440, 2013.

[18] 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, no. 1, pp. 7–9, 2015.

[19] L. Sun, J. Du, L.-R. Dai, and C.-H. Lee, "Multiple-target deep learning for LSTM-RNN based speech enhancement," *in Proc. HSCMA*, pp. 136–140, 2017.

[20] M. Kolbæk, Z.-H. Tan, and J. Jensen, "Speech intelligibility potential of general and specialized deep neural network based speech enhancement systems," *IEEE/ACM Transactions on Audio, Speech and Language Processing*, vol. 25, no. 1, pp. 153–167, 2017.

[21] D. Wang and J. Chen, "Supervised speech separation based on deep learning: An overview," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 26, no. 10, pp.1702-1726, 2018.

[22] B. Xia and C. Bao, "Wiener filtering based speech enhancement with weighted denoising auto-encoder and noise classification," *Speech Communication*, vol. 60, pp. 13–29, 2014.

[23] S.-W. Fu, Y. Tsao, and X. Lu, "SNR-aware convolutional neural network modeling for speech enhancement," *in Proc. INTERSPEECH*, pp. 3768–3772, 2016.

[24] Z. Chen, S. Watanabe, H. Erdogan, and J. R. Hershey, "Speech enhancement and recognition using multi-task learning of long short-term memory recurrent neural networks," *in Proc. INTERSPEECH*, pp. 3274–3278, 2015.

[25] H. Erdogan, J. R. Hershey, S. Watanabe, and J. Le Roux, "Phase-sensitive and recognition-boosted speech separation using deep recurrent neural networks," *in Proc. ICASSP*, pp. 708–712, 2015.

[26] J. Chen and D. L. Wang, "Long short-term memory for speaker generalization in supervised speech separation," *in Proc. INTERSPEECH*, pp. 3314–3318, 2016.

[27] A. Pandey and D. Wang, "On adversarial training and loss functions for speech enhancement," *in Proc. ICASSP*, pp. 5414-5418, 2018.

[28] K. Qian, Y. Zhang, S. Chang, X. Yang, D. X., Florêncio, and M. Hasegawa-Johnson, "Speech enhancement using Bayesian Wavenet," *in Proc. INTERSPEECH*, pp. 2013-2017, 2017.

[29] S. Wang, K. Li, Z. Huang, S. M. Siniscalchi, and C.-H. Lee, "A transfer learning and progressive stacking approach to reducing deep model sizes with an application to speech enhancement," *in Proc. ICASSP*, pp. 5575-5579, 2017.

[30] G. E. Hinton and R. R. Salakhutdinov, "Reducing the dimensionality of data with neural networks," *Science*, vol. 313, no. 5786, pp. 504–507, 2006.

[31] G. E. Hinton, A. Krizhevsky, S. D. Wang, "Transforming auto-encoder," *in Proc. ICANN*, pp. 44-51, 2011.

[32] C.-Y. Liou, W.-C. Cheng, J.-W. Liou, D.-R. Liou, "Transforming autoencoder for words," *Neurocomputing*, vol.139, pp. 84–96, 2014.

[33] X. Lu, S. Matsuda, C. Hori and H. Kashioka, "Speech restoration based on deep learning autoencoder with layer-wised learning," *in. Proc. INTERSPEECH*, pp. 1504-1507, 2012.

[34] M. Kim and P. Smaragdis, "Adaptive denoising autoencoders: a fine-tuning scheme to learn from test mixtures," *in Proc. LVA/ICA*, pp. 100-107, 2015.

[35] M. Kim, "Collaborative deep learning for speech enhancement: A run-time model selection method using autoencoders," *in Proc. ICASSP*, pp. 76–80, 2017.

[36] N. Alamdari, A. Arian, and K. Nasser, "Self-supervised deep learning-based speech denoising," *arXiv preprint arXiv: 1904.12069*, 2019.

[37] X. Lu, Y. Tsao, S. Matsuda, and C. Hori, "Ensemble modeling of denoising autoencoder for speech spectrum restoration," *in Proc. INTERSPEECH*, pp. 885-889, 2014.

[38] C. Yu, R. E. Zezario, J. Sherman, Y.-Y. Hsieh, X. Lu, H.-M. Wang, and Y. Tsao, "Speech enhancement based on denoising autoencoder with multi-branched encoders," *arXiv preprint arXiv: 2001.01538*, 2020.

[39] T. Hussain, S. M. Siniscalchi, C.-C. Lee, S.-S. Wang, Y. Tsao, and W.-H. Liao, "Experimental study on extreme learning machine applications for speech enhancement," *IEEE Access,* vol. 5, pp. 25542 – 25554, 2017.

[40] N. Parihar, J. Picone, D. Pearce, and H.-G. Hirsch, "Performance analysis of the Aurora large vocabulary baseline system," *in Proc. EUSIPCO*, pp. 553–556, 2004.

[41] J. W. Lyons, "DARPA TIMIT acoustic-phonetic continuous speech corpus," *National Institute of Standards and Technology*, 1993.

[42] A. Rix, J. Beerends, M. Hollier, and A. Hekstra, "Perceptual evaluation of speech quality (PESQ)-a new method for speech quality assessment of telephone networks and codecs," *in Proc. ICASSP*, pp. 749–752, 2001.

[43] C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen, "An algorithm for intelligibility prediction of time-frequency weighted noisy speech," *IEEE/ACM Transactions on Audio, Speech and Language Processing*, vol.19, no.7, pp.2125–2136, 2011.

[44] A. Beck and M. Teboulle, "A fast iterative shrinkage-thresholding algorithm for linear inverse problems," *SIAM Journal on Imaging Sciences,* vol. 2, no. 1, pp. 183–202, 2009.

[45] R. Hecht-Nielsen, "Theory of the backpropagation neural network," *in Proc. IJCNN*, pp. 593–605, 1989.