한국음향학회지 제38권 제6호 pp. 670~677 (2019)
The Journal of the Acoustical Society of Korea Vol.38, No.6 (2019)
https://doi.org/10.7776/ASK.2019.38.6.670

pISSN : 1225-4428
eISSN : 2287-3775

Combining multi-task autoencoder with Wasserstein generative adversarial networks for improving speech recognition performance

음성인식 성능 개선을 위한 다중작업 오토인코더와 와설스타인식 생성적 적대 신경망의 결합

Chao Yuan Kao[1] and Hanseok Ko[1 †]

(고조원,[1] 고한석[1†])

[1]Department of Electronics and Computer Engineering, Korea University Anam Campus

(Received October 22, 2019; accepted November 11, 2019)

ABSTRACT: As the presence of background noise in acoustic signal degrades the performance of speech or acoustic event recognition, it is still challenging to extract noise-robust acoustic features from noisy signal. In this paper, we propose a combined structure of Wasserstein Generative Adversarial Network (WGAN) and Multi-Task AutoEncoder (MTAE) as deep learning architecture that integrates the strength of MTAE and WGAN respectively such that it estimates not only noise but also speech features from noisy acoustic source. The proposed MTAE-WGAN structure is used to estimate speech signal and the residual noise by employing a gradient penalty and a weight initialization method for Leaky Rectified Linear Unit (LReLU) and Parametric ReLU (PReLU). The proposed MTAE-WGAN structure with the adopted gradient penalty loss function enhances the speech features and subsequently achieve substantial Phoneme Error Rate (PER) improvements over the stand-alone Deep Denoising Autoencoder (DDAE), MTAE, Redundant Convolutional Encoder-Decoder (R-CED) and Recurrent MTAE (RMTAE) models for robust speech recognition.

Keywords: Speech enhancement, Wasserstein Generative Adversarial Network (WGAN), Weight initialization, Robust speech recognition, Deep Neural Network (DNN)

PACS numbers: 43.60.Bf, 43.60.Uv

초 록: 음성 또는 음향 이벤트 신호에서 발생하는 배경 잡음은 인식기의 성능을 저하시키는 원인이 되며, 잡음에 강인한 특징을 찾는데 많은 노력을 필요로 한다. 본 논문에서는 딥러닝을 기반으로 다중작업 오토인코더(Multi-Task AutoEncoder, MTAE) 와 와설스타인식 생성적 적대 신경망(Wasserstein GAN, WGAN)의 장점을 결합하여, 잡음이 섞인 음향신호에서 잡음과 음성신호를 추정하는 네트워크를 제안한다. 본 논문에서 제안하는 MTAE-WGAN는 구조는 구배 페널티(Gradient Penalty) 및 누설 Leaky Rectified Linear Unit (LReLU) 모수 Parametric ReLU (PReLU)를 활용한 변수 초기화 작업을 통해 음성과 잡음 성분을 추정한다. 직교 구배 페널티와 파라미터 초기화 방법이 적용된 MTAE-WGAN 구조를 통해 잡음에 강인한 음성특징 생성 및 기존 방법 대비 음소 오인식률(Phoneme Error Rate, PER)이 크게 감소하는 성능을 보여준다.

핵심용어: 음성인식, 와설스타이식 생성적 적대 신경망, 직교 구배 페널티, 초기화, 딥러닝

I. Introduction

With rapid advancement of deep learning, acoustic event recognition and Automatic Speech Recognition

†**Corresponding author:** Hanseok Ko (hsko@korea.ac.kr)
Department of Electronics and Computer Engineering, Korea University Anam Campus, 145 Anam-ro, Seongbuk-gu, Seoul 02841, Republic of Korea
(Tel: 82-2-3290-3239, Fax: 82-2-3291-2450)

(ASR) technologies have been widely used in our daily lives such as in intelligent virtual assistants, mobile devices and other electronic devices. However, presence of various types of noise in speech or intended acoustic signal degrades the performance of such recognition systems. Speech enhancement is considered a very crucial technique because it can reduce the impact of noise and improve recognition accuracy. There have been many approaches such as traditional speech enhancement approaches include Wiener filter,[1] Short Time Spectral Amplitude-Minimum Mean Square Error (STSA-MMSE)[2] and nonnegative matrix factorization.[3] Deep learning approaches include Deep Denoising AutoEncoder (DDAE), Deep Neural Network (DNN),[4] Convolutional Neural Network (CNN),[5] or Recurrent Neural Network (RNN)[6] have been applied for speech enhancement in past few years, and they can be divided into a regression method (mapping-based targets)[1,5,7] and a classification method (masking-based targets).[8,9] Although these methods have attained an acceptable level for speech enhancement, there is still room for improvement.

In recent years, Generative Adversarial Network (GAN) has been widely used across many applications of deep learning, from image generation[10] to video and sequence generation,[11,12] and has achieved better performance. Speech Enhancement GAN (SEGAN) is the first GAN-based model used for speech enhancement.[13] GAN is considered hard to train and sensitive to hyper-parameters. Also, the training loss type (L1 or L2) affects the enhancement performance as it has been noticed by Pandey and Wang, where the adversarial loss training in SEGAN does not achieve better performance than L1 loss training.[14] In addition, Donahue, et al. proposed Frequency-domain SEGAN (FSEGAN)[15] for robust attention-based ASR system,[16] and achieved lower Word Error Rate (WER) than WaveNet[17] based SEGAN. Afterward, Michelsanti proposed a state-of-the-art CNN based Pix2Pix framework[18] and Mimura et al. proposed a Cycle-GAN-based acoustic feature transformation[19] for robust ASR model.

These studies using many kinds of GAN framework demonstrated improved performances for speech enhancement tasks. Nonetheless,[13,15,19] compared their methods with conventional methods. Therefore, it is hard to demonstrate the advantage of adversarial loss training over L1 loss training for speech enhancement. In this work, we illustrate the effectiveness of the adversarial loss training by comparing our proposed Multi-Task AutoEncoder-Wasserstein Generative Adversarial Network-Gradient Penalty (MATAE-WGAN-GP) and a single generator based on MTAE.[20] To summarize, our contribution is to propose an architecture that combines MTAE and Wasserstein GAN for separating speech and noise signals into one network. This structure combines the advantages of multi-tasking learning and GAN, and result in improving PER performance. We also propose a weights initialization method based on He[21] for Leaky Rectified Linear Unit (LReLU) and Parametric ReLU (PReLU). As a result, loss becomes more stable during learning process, thereby avoiding possible exploding gradients problem in a deep network.

In summary, by adopting GP loss function, our proposed integrated model (MTAE-WGAN-GP) achieves lower PER over other state-of-the-art CNN and RNN for robust ASR system. This paper is organized as follows. In Section II, we present the proposed model structure and weights initialization. We then describe the experimental settings in Section III. The results are discussed and evaluated in Section IV and finally, conclusions are provided in Section V.

II. Proposed Approaches

2.1 Combining MTAE-WGAN-GP

Our proposed MTAE-WGAN-GP is composed of one generator and two critics as shown in Fig. 1. The generator is a fully connected MTAE and is intended to produce estimates of not only speech but also noise from noisy speech input. Speech estimate critic (C_{sc}) and noise estimate critic (C_{nc}) are both fully connected DNNs, tasked with determining if a given sample is real (s and n) or fake

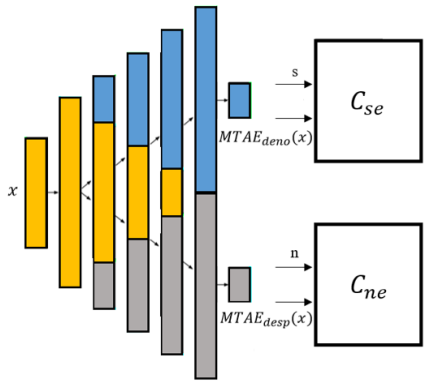

Fig. 1. MTAE-WGAN-GP structure: blue and yellow parts are the denoising autoencoder. Gray and yellow parts are a despeeching autoencoder. The yellow parts in the middle are shared weights and biases by two autoencoders.

[$MTAE_{deno}(x)$ and $MTAE_{desp}(x)$]. After training, we use a single MTAE based generator for our speech enhancement task. The loss function for generator composed of adversarial loss and L1 loss is represented by

$$\begin{aligned} L_{MTAE} = & -\lambda_1 E_{x\sim P_z}[C_{se}(MTAE_{deno}(x),x)] \\ & -(1-\lambda_1)E_{x\sim P_z}[C_{se}(MTAE_{desp}(x),x)] \\ & +\lambda_2[\lambda_{L1}\|MTAE_{deno}(x)-s\| \\ & +(1-\lambda_{L1})\|MTAE_{desp}(x)-n\|], \end{aligned} \tag{1}$$

where $MTAE_{deno}(x)$ and s are the estimated speech and the target clean speech respectively. $MTAE_{desp}(x)$ and n are the estimated noise and the target noise respectively. λ_1, λ_2 and λ_{L1} are hyper-parameters. By experiment, we set λ_1 = 0.5, λ_2 = 100 and λ_{L1} = 0.5 for the best performance in our system. Our model adopts Wasserstein distance as a continuous and almost differentiable function within the range restricted by 1-Lipschitz constraint. The loss function for the critics are represented by

$$\begin{aligned} L_{C_{se}} = & E_{x\sim P_z}[C_{se}(MTAE_{deno}(x),x)] \\ & - E_{y_s\sim P_{data}}[C_{se}(y_s,x)] \\ & + \lambda_{gp}E_{\hat{y_s}\sim P_{\hat{y_s}}}[(\|\nabla_{\hat{y_s}}C_{se}(\hat{y_s})\|_2-1)^2]. \end{aligned} \tag{2}$$

$$\begin{aligned} L_{C_{ne}} = & E_{x\sim -P_z}[C_{ne}(MTAE_{desp}(x),x)] \\ & - E_{y_n\sim P_{data}}[C_{ne}(y_n,x)] \\ & + \lambda_{gp}E_{\hat{y_n}\sim P_{\hat{y_n}}}[(\|\nabla_{\hat{y_n}}C_{ne}(\hat{y_n})\|_2-1)^2]. \end{aligned} \tag{3}$$

The generator consists of 5 hidden layers and 1024 units were set in the first layer. Then, as described in[20] the denoising exclusive units, the shared units and the despeeching exclusive units for each layer from the 1st to the 5th are (0, 1024, 0), (256, 768, 256), (512, 512, 512), (768, 256, 768) and (1024, 0, 1024), respectively. Additionally, in[18] we modify $E[x_l^2]$ term in Eq. (9) to $E[x_l^2]=\frac{1+\alpha_{Negative}^2}{2}Var[y_{l-1}]$ with LReLU activation function as our weight initialization (subsection 2.2).

The critics feed as not only real and fake data but also the input data x. The pairs are (s, x) and ($MTAE_{deno}(x)$, x) for speech estimate critic, and (n, x) and ($MTAE_{desp}(x)$, x) for noise estimate critic. The speech estimate critic network is composed of 4-layers with 1024, 768, 512, and 256 units, while the noise estimate critic is composed of 3-layers with 512 units per layer where both models use LReLU as activation function.

2.2 Initialization of weights for leaky and parametric rectified linear unit

Network parameter initialization plays a considerably significant part in the network training where inappropriate initialization could lead to poor results.[22] We briefly describe the initialization methods proposed by Xavier[23] and He,[21] and propose a modified initialization approach for LReLU and PReLU based activation. The method has been shown to be particularly effective when the number of network layers becomes large.

The response representation for DNN is:

$$Y_l = W_l X_l + B_l. \tag{4}$$

$$X_l = f(Y_{l-1}), \tag{5}$$

where W_l and B_l are weight and bias matrix. f is the activation and we use l to index a layer.

The idea of He initialization[21] is based on Xavier initialization[23] in that it preserves the same variance of the response input throughout the layers. As in,[23] by ini-

tializing the elements of W_l to be independent and identically distributed (i.i.d.), we assume that X_l elements are i.i.d. and both X_l and W_l are independent from each other. Then we can obtain:

$$Var[y_l] = n_l Var[w_l x_l], \tag{6}$$

where, y_l, w_l, and x_l are random variables of elements in Y_l, W_l, and X_l, respectively. n_l is the number of nodes. By setting w_l to have zero mean, variance of the product of independent variables can be written as:

$$Var[y_l] = n_l Var[w_l] E[x_l^2]. \tag{7}$$

Since ReLU function is not linear and does not have a zero mean, by initializing w_{l-1} to have a symmetric distribution around zero and setting b_{l-1} to zero, y_{l-1} will also have a symmetric distribution with zero mean.[21] Thus, the expectation of x_l can be written as: $E[x_l^2] = \frac{1}{2} Var[y_{l-1}]$, when ReLU is used as an activation function.[21] However, in the case of LReLU or PReLU being used as an activation function, $E[x_l^2]$ should be considered when x_l is less than zero.

Suppose the activation function is a linear transformation with slope α and zero intercept. Standard deviation (σ) and variance of y_{l-1} will become α and α^2 respectively.

In the case of two different alphas from zero mean, such as LReLU, we can calculate the mean defined as:

$$E[x_l^2] = \frac{\alpha_{Positive}^2 + \alpha_{Negative}^2}{2} Var[y_{l-1}], \quad ,for\ all\ x_l \tag{8}$$

where $\alpha_{Positive}$ is the slope for $x_l \geq 0$, and $\alpha_{Negative}$ is the slope for $x_l < 0$ of LReLU or PReLU.

For LReLU or PReLU, $\alpha_{Positive}$ is equal to 1. Thus, we can rewrite it as:

$$E[x_l^2] = \frac{1 + \alpha_{Negative}^2}{2} Var[y_{l-1}], for\ all\ x_l. \tag{9}$$

By substituting Eq. (9) into Eq. (7), we obtain:

$$Var[y_l] = \frac{1 + \alpha_{Negative}^2}{2} n_l Var[w_l]\, Var[y_{l-1}]. \tag{10}$$

And with L layers, we get:

$$Var[y_L] = Var[y_l](\prod_{l=2}^{L} \frac{1 + \alpha_{Negative}^2}{2} n_l Var[w_l]). \tag{11}$$

Finally, a sufficient condition is:

$$\frac{1 + \alpha_{Negative}^2}{2} n_l Var[w_l] = 1, \quad \forall l. \tag{12}$$

Therefore, the proposed initialization method in Eq. (12) leads to zero-mean Gaussian distribution and σ equal to $\sqrt{(2/n_l(1+\alpha_{Negative}^2)}$ where, b is initialized as zero. For the first layer (l = 1), the sufficient condition will be $n_l Var[w_l] = 1$, since there is no activation function applied to the input. The initial value of $\alpha_{Negative}$ for LReLU is set to 0.5 in this paper.

III. Experimental Setup

Two sets of experiments are conducted to evaluate our proposed model and initialization method. Firstly, we evaluate the effectiveness of proposed MTAE-WGAN-GP against state-of-the-art methods. Secondly, we compare the initial output variance and convergence of our proposed initialization against Xavier and He initialization.

3.1 Dataset

For training the proposed model, we used the Texas Instruments/Massachusetts Institute of Technology (TIMIT) training dataset which contains 3696 utterances from 462 speakers. The training utterance is augmented by 10 types of noise (2 artificial and 8 from YouTube.com: pink noise, red noise, classroom, laundry room, lobby, playground, rain, restaurant, river, and street). Each signal and back-

ground noise added together with three Signal to Noise Ratio (SNR) levels (5 dB, 15 dB, and 20 dB). The obtained dataset for training the proposed model contains 9 % of clean speech to ensure the effectiveness of the model even in clean environment. Wen has shown the effectiveness of using synthetic noise during training for speech enhancement task.[24]

TIMIT testing set that contains 192 utterances from 24 speakers is corrupted by 3 types of unseen noise (café, pub, and schoolyard), collected from ETSI EG 202 396-1 V1.2.2 (2008-09) with three different SNR levels (5 dB, 15 dB, and 20 dB). The augmentation for the dataset is conducted using ADDNOISE MATLAB.[25]

3.2 Preprocessing

Kaldi toolkit is used for training the ASR model using a Hybrid System (Karel's DNN) on a clean TIMIT Acoustic-Phonetic Continuous Speech Corpus training data. The sampling rate for the audio signals was at 16 kHz and features are extracted by means of short-time Fourier transform with window size of 25 ms and 10 ms window step. Here, we applied 23 Mel-filter banks, with Mel-scale from 20 Hz to 7800 Hz.

The proposed model (MTAE-WGAN-GP) and MTAE were trained by setting the data with concatenated 16 contiguous frames of 13-dimensional MFCCs (13x16). The same data format was used to conduct both experiments. All features are normalized per utterance within the range of [-1, 1]. All networks are trained using Root Mean Square Propagation (RMSprop) optimizer with a batch size of 100. For DDAE and MTAE architecture LReLU activation function is used except in the output layer which has no activation function.

IV. Results

4.1 Experiment 1

DDAE vs. MTAE vs. RNN vs. CNN vs. MTAE-WGAN-GP

We adopt L1 loss for all used training models. DDAE,[26] MTAE,[20] Recurrent MTAE (RMTAE) and Redundant Convolutional Encoder-Decoder (R-CED)[5] are used as baseline models to compare performance of the proposed model in terms of PER. Hence, by incorporating a typical ASR model, performance is evaluated by measuring how well the system recognizes noisy speech after the speech enhancement. The RMTAE model consists of 3 LSTM layers followed by 2 fully-connected layers with 256 units and LReLU as activation function except for the output layer. To avoid exploding gradients problem, we use a gradient clipping from -1 to 1.[27] The results are reported in Table 1.

Table 1 reports the performance of these models. It can be observed that over three SNR conditions and three unseen noise, the proposed method consistently improved the recognition accuracy by 19.6 %, 8.1 %, 6.9 %, 3.6 %, and 1.8 % relative to non-enhanced features (None),

Table 1. Performance comparison between non-enhanced features (None), DDAE, RMTAE (RNN), CNN (R-CED) and MTAE-WGAN-GP on 3 types of unseen noise with three SNR conditions.

SNR	PER (%)				
	Enhancement model	Cafe	Pub	School yard	Average
20 dB	None	28.4 %	30.3 %	32.5 %	30.4 %
	DDAE	27.8 %	27.6 %	28.3 %	27.9 %
	MTAE	27.8 %	27.0 %	28.5 %	27.8 %
	RMTAE (RNN)	**26.0 %**	**24.8 %**	**26.2 %**	**25.7 %**
	R-CED (CNN)	27.6 %	25.6 %	27.0 %	26.7 %
	MTAE-WGAN-GP	**25.9 %**	25.4 %	**26.5 %**	**25.9 %**
15 dB	None	34.9 %	36.5 %	39.9 %	37.1 %
	DDAE	30.7 %	30.9 %	33.9 %	31.8 %
	MTAE	30.7 %	30.2 %	33.2 %	31.4 %
	RMTAE (RNN)	**28.9 %**	**28.5 %**	31.8 %	29.7 %
	R-CED (CNN)	29.7 %	**28.3 %**	32.2 %	30.1 %
	MTAE-WGAN-GP	**28.9 %**	**28.1 %**	**30.4 %**	**29.1 %**
5 dB	None	52.8 %	57.7 %	59.8 %	56.8 %
	DDAE	45.5 %	49.1 %	49.9 %	48.2 %
	MTAE	44.0 %	48.7 %	49.3 %	47.3 %
	RMTAE (RNN)	42.5 %	47.2 %	48.3 %	46.0 %
	R-CED (CNN)	42.4 %	47.0 %	49.3 %	46.2 %
	MTAE-WGAN-GP	**40.3 %**	**44.9 %**	**46.8 %**	**44.0 %**

DDAE, MTAE, R-CED (CNN) and RMTAE (RNN). Especially at low SNR scenarios, the improvement becomes more apparent. Additionally, we observe that at high SNR condition (20 dB) the RMTAE (RNN) has a competitive performance compare to our proposed method. However, performance is degraded obviously when SNR becomes lower (15 dB and 5 dB).

MTAE-WGAN-GP achieves lower PER compare to a single generator MTAE. This demonstrates the effectiveness of adversarial loss training is better than using L1 loss alone.

4.2 Experiment 2

Xavier initialization vs. He initialization vs. Our initialization

We adopt a 10 layer-MTAE to compare with Xavier[23] and He initialization.[21] By increasing units linearly, the denoising exclusive units, the shared units, and the despeeching exclusive units are (0, 1200, 0) and (1200, 0, 1200) for 1st and 10th layers, respectively. Fig. 2 shows the histograms of the output distribution in each layer before training. We can observe that as the number of layers increases, the variance in He initialization increases dramatically while the variance in Xavier initialization gradually decreases toward zero. However, our proposed initialization keeps the output distribution and variance steady through each layer, as shown in Fig. 3.

Next, we compare our proposed initialization with He and Xavier on a 25 layers MTAE using the obtained loss of the model. Fig. 4 shows the loss of convergence during training. We can observe that in training our proposed initialization converges faster and more stable than Xavier initialization, while He initialization cannot converge and can easily suffer from exploding gradient problem during training with deep network. This illustrates the advantage of using the proposed initialization when training a deep network with LReLU and PReLU.

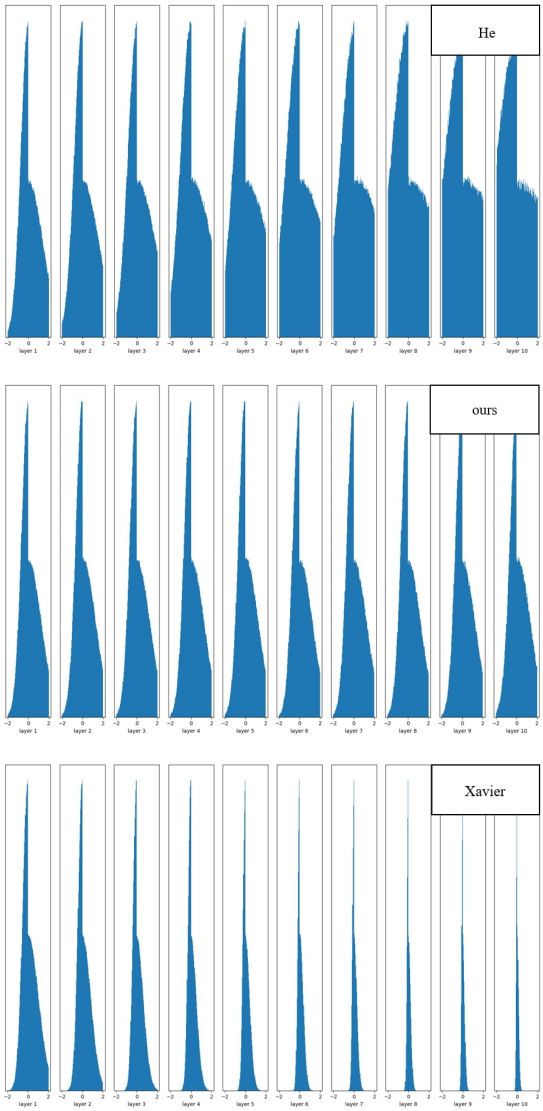

Fig. 2. The illustration of the distribution of output values in each layer. From top to bottom are He, Xavier, and our proposed initialization, respectively.

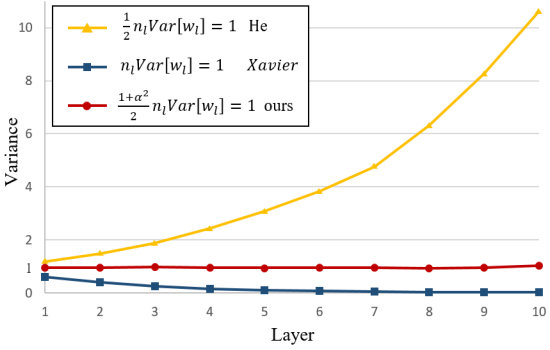

Fig. 3. The initial output variance in each layer.

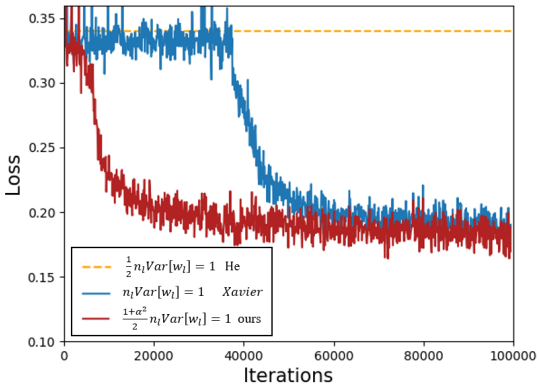

Fig. 4. The convergence of a 25-layer MTAE model. We use LReLU activation function in all layers except in the output layer. Our proposed initialization converges faster than "Xavier", while "He" cannot converge.

V. Conclusions

We proposed MTAE-WGAN combination as an architecture that integrates MTAE with WGAN and demonstrated improvement in ASR performance. Additionally, we proposed an initialization of weights for LReLU and PReLU and demonstrated that it converges faster with more stable than Xavier and He initialization. The results show that MTAE-WGAN-GP achieves 8.1 %, 6.9 %, 3.6 %, and 1.8 % PERs improvement relative to DDAE, MTAE, R-CED (CNN) and RMTAE (RNN) model, respectively.

Acknowledgements

This research is funded by the Ministry of Environment supported by the Korea Environmental Industry & Technology Institute's environmental policy-based public technology development project (2017000210001).

References

1. P. Scalart and J. V. Filho "Speech enhancement based on a priori signal to noise estimation," Proc. IEEE ICASSP. 629-632 (1996).
2. Y. Ephraim and D. Malah, "Speech enhancement using a minimum meansquare error short-time spectral amplitude estimator," IEEE Trans. Acoust. Speech Signal Process. **32**, 1109-1121 (1984).
3. N. Mohammadiha, P. Smaragdis, and A. Leijion, "Supervised and unsupervised speech enhancement using nonnegative matrix factorization," IEEE Trans. Audio, Speech Lang. Process. **21**, 2140- 2151 (2013).
4. Y. Xu, J. Du, L. -R. Dai, and C. -H. Lee, "A regression approach to speech enhancement based on deep neural networks," IEEE Trans. Audio, Speech Lang. Process. **23**, 7-19 (2015).
5. S. R. Park and J. W. Lee, "A fully convolutional neural network for speech enhancement," Proc. Interspeech, 1993-1997 (2017).
6. 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," Proc. Interspeech, 22-25 (2012).
7. X. Feng, Y. Zhang, and J. Glass, "Speech feature denoising and dereverberation via deep autoencoders for noisy reverberant speech recognition," Proc. IEEE ICASSP. 1759-1763 (2014).
8. B. Li and K. C. Sim, "A spectral masking approach to noise-robust speech recognition using deep neural networks," IEEE Trans. Audio, Speech Lang. Process. **22**, 1296-1305 (2014).
9. D. Wang and J. Chen, "Supervised speech separation based on deep learning: An overview," IEEE/ACM Trans. Audio, Speech Lang. Process. **26**, 1702-1726 (2018).
10. D. Berthelot, T. Schumm, and L. Metz, "Began: Boundary equilibrium generative adversarial networks." arXiv preprint arXiv:1703.10717 (2017).
11. S. Tulyakov, M. -Y. Liu, X. Yang, and J. Kautz, "Mocogan: Decomposing motion and content for video generation," Proc. the IEEE conference on computer vision and pattern recognition, 1526-1535 (2018).
12. L. Yu, W. Zhang, J. Wang, and Y. Yu, "Seqgan: Sequence generative adversarial nets with policy gradient." Thirty-First AAAI Conference on Artificial Intelligence, 2852-2858 (2017).
13. S. Pascual, A. Bonafonte, and J. Serra, "SEGAN: Speech enhancement generative adversarial network," Proc. Interspeech, 3642-3646 (2017).
14. A. Pandey and D. Wang, "On adversarial training and loss functions for speech enhancement," Proc. IEEE ICASSP. 5414-5418 (2018).
15. C. Donahue, B. Li, and R. Prabhavalkar, "Exploring speech enhancement with generative adversarial networks for robust speech recognition," Proc. IEEE ICASSP. 5024-5028 (2018).
16. W. Chan, N. Jaitly, Q. V. Le, and O. Vinyals, "Listen, attend and spell: A neural network for large voca-

bulary conversational speech recognition." Proc. IEEE ICASSP. 4960-4964 (2016).
17. A. Oord, S. Dieleman, H. Zen, K. Simonyan, O. Vinyals, A. Graves, N. Kalchbrenner, A. Senior, and K. Kavukcuoglu, "Wavenet: A generative model for raw audio." arXiv preprint arXiv:1609.03499 (2016).
18. D. Michelsanti and Z. H. Tan, "Conditional generative adversarial networks for speech enhancement and noise-robust speaker verification," Proc. Interspeech, 2008-2012 (2017).
19. M. Mimura, S. Sakai, and T. Kawahara, "Cross-domain speech recognition using nonparallel corpora with cycle-consistent adversarial networks," Proc. IEEE Automatic Speech Recognition and Understanding Workshop, 134-140 (2017).
20. H. Zhang, C. Liu, N. Inoue, and K. Shinoda, "Multi-task autoencoder for noise-robust speech recognition," Proc. IEEE ICASSP. 5599-5603 (2018).
21. K. He, X. Zhang, S. Ren, and J. Sun, "Delving deep into rectifiers: Surpassing human-level performance on imagenet classification," Proc. the IEEE International Conference on Computer Vision, 1026-1034 (2015).
22. M. D. Zeiler, M. Ranzato, R. Monga, M. Mao, K. Yang, Q.V. Le, P. Nguyen, A. Senior, V. Vanhoucke, J. Dean, and G. E. Hinton, "On rectified linear units for speech processing," Proc. IEEE ICASSP. 3517-3521 (2017).
23. X. Glorot and Y. Bengio, "Understanding the difficulty of training deep feedforward neural networks," Proc. the thirteenth international conference on artificial intelligence and statistics, 249-256 (2010).
24. S. X. Wen, J. Du, and C. -H. Lee, "On generating mixing noise signals with basis functions for simulating noisy speech and learning dnnbased speech enhancement models," Proc. IEEE International Workshop on MLSP. 1-6 (2017).
25. ITU-T, Rec. *P. 56: Objective Measurement of Active Speech Level*, 2011.
26. X. Lu, Y. T. Sao, S. Matsuda, and C. Hori, "Speech enhancement based on deep denoising autoencoder," Proc. Interspeech, 436-440 (2013).
27. R. Pascanu, T. Mikolov, and Y. Bengio, "On the difficulty of training recurrent neural networks," Proc. 30th ICML. 2347-2355 (2013).

Profile

▸ Chao Yuan Kao (고조원)

He received the B.S. degree in Physics from National Taiwan Normal University in 2013. During 2014–2017, he joined the ASUSTeK Computer Inc in Taipei, Taiwan (R.O.C.), where he worked in several Asus ZenFone and ZenPad series projects. From 2017, he has been in the M.S. course in Department of Electrical and Computer Engineering from Korea University. His interests include speech enhancement and Generative Adversarial Networks

▸ Hanseok Ko (고한석)

He received the B.S. degree from Carnegie Mellon University in 1982, M.S. degree from the Johns Hopkins University in 1988, and Ph.D. degree from the Catholic University of America in 1992, all in electrical engineering. At the onset of his career, he was with the WOL, Maryland, where his work involved signal and image processing. In March of 1995, he joined the faculty of the Electronics and Computer Engineering at Korea University, where he is currently Professor. His professional interests include speech/image signal processing for pattern recognition, multimodal analysis, and intelligent data fusion.