LETTER

Orthogonal Gradient Penalty for Fast Training of Wasserstein GAN Based Multi-Task Autoencoder toward Robust Speech Recognition

Chao-Yuan KAO[†], **Sangwook PARK**[††], **Alzahra BADI**[†], **David K. HAN**[†††], ***Nonmembers***, ***and*** **Hanseok KO**[†a)], ***Member***

SUMMARY Performance in Automatic Speech Recognition (ASR) degrades dramatically in noisy environments. To alleviate this problem, a variety of deep networks based on convolutional neural networks and recurrent neural networks were proposed by applying L1 or L2 loss. In this Letter, we propose a new orthogonal gradient penalty (OGP) method for Wasserstein Generative Adversarial Networks (WGAN) applied to denoising and despeeching models. WGAN integrates a multi-task autoencoder which estimates not only speech features but also noise features from noisy speech. While achieving 14.1% improvement in Wasserstein distance convergence rate, the proposed OGP enhanced features are tested in ASR and achieve 9.7%, 8.6%, 6.2%, and 4.8% WER improvements over DDAE, MTAE, R-CED(CNN) and RNN models.

key words: *speech enhancement, generative adversarial networks, deep learning, robust speech recognition*

1. Introduction

Automatic Speech Recognition (ASR) has been in wider usage in recent years including mobile devices, home assistants, and other electronic devices. Accuracies of these ASRs depend highly on the level of noise present in the input audio. For good ASR performance, speech enhancing preprocessing is considered critical. The enhancement approaches can generally be categorized into a regression approach (mapping based target) [1]–[3] or a classification approach (masking based target) [4]. It has been observed that masking based method performs better in terms of Short-Time Objective Intelligibility (STOI), while regression approaches have comparable performances based on Perceptual Evaluation of Speech Quality (PESQ) scores [5]. With the emergence of deep learning, methods such as Autoencoders or Generative Adversarial Networks (GANs) have been adopted for audio enhancement to increase speech intelligibility. Deep learning based methods such as Deep Denoising Autoencoder (DDAE), DNN [1], CNN [3], or RNN [6] achieved performance improvements over the conventional methods. Speech Enhancement GAN (SEGAN) first proposed in [7] has also been used for speech enhancement. Pandey and Wang found that training the generator alone with L1 loss performs better than adversarial training in SEGAN [8]. Donahue, et al. proposed Frequency-domain SEGAN (FSEGAN) that works in time-frequency representation without phase information in contrast to SEGAN which works with the waveform [9]. FSGAN in this implementation achieved lower word error rates (WERs) than that of SEGAN.

Michelsanti and Tan proposed a CNN based Pix2Pix framework, which outperformed a classical STSA-MMSE algorithm and a DNN based model for speech enhancement [10]. Mimura et al. proposed a Cycle-GAN-based acoustic feature transformation and showed its effectiveness in noisy speech recognition and speaking style adaption [11]. In [12] an adversarial training-based mask estimation has shown to capture speech and noise signals without supervised data.

Although these methods based on deep learning framework demonstrated improved performances over the traditional methods, their effectiveness is still limited when applied in noisy environments. To overcome the noise problem, we hinge on the idea of separating noise and speech components in a single network frame. We developed a speech-noise separation method based on the Multi-Task AutoEncoder (MTAE) [13].

Our contribution in this Letter is to develop a novel Orthogonal Gradient Penalty (OGP) in Wasserstein GAN (WGAN) architecture, combines the advantages of multi-task learning and WGAN for separating noise and speech contents in a single network, that results in more rapid convergence during training and achieves improved WER performance.

In summary, our proposed model (MTAE_WGAN_OGP) enables fast training but achieves better performance over the state-of-art CNN and RNN models for speech enhancement. The rest of the paper is organized as follows. Section 2 introduces the proposed orthogonal gradient penalty on WGAN model structure. The experimental settings are in Sect. 3. Finally, we evaluate the results in Sect. 4.

2. Proposed Method

2.1 Wasserstein GAN Based Multi-Task Autoencoder (MTAE_WGAN)

Figure 1 shows the WGAN structure we used for develop-

Manuscript received October 10, 2019.
Manuscript revised November 27, 2019.
Manuscript publicized January 27, 2020.

[†]The authors are with the School of Electrical Engineering, Korea University, Seoul, 02841, Korea.
[††]The author is with the School of Electrical and Computer Engineering, Johns Hopkins University, Baltimore, MD 21218, USA.
[†††]The author is with the US Army Research Laboratory (ARL), Adelphi, Maryland, USA.
a) E-mail: hsko@korea.ac.kr

DOI: 10.1587/transinf.2019EDL8183

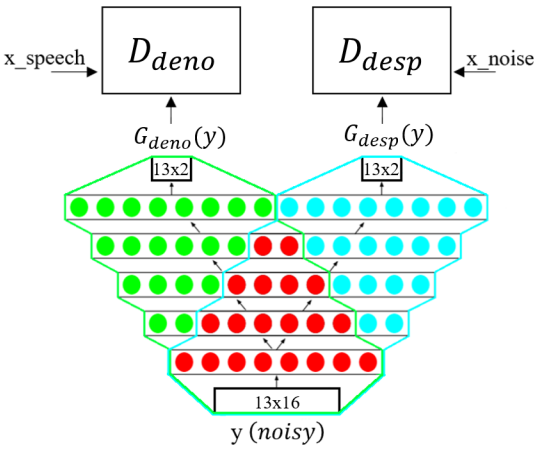

Fig. 1 MTAE structure: the left side is a denoising autoencoder (green lines), contains green and red nodes. The right side is a despeeching autoencoder (blue lines), contains red and blue nodes. The red nodes in the middle are shared (weights and biases) by two autoencoders.

ing the proposed orthogonal gradient penalty. WGAN is composed of one generator based on MTAE and two discriminators for denoising (D_{deno}) and despeeching (D_{desp}). The generator makes both enhanced speech and estimated noise from noisy input. The two discriminators try to distinguish real samples, *x_speech* and *x_noise* from fake samples, $G_{deno}(y)$ and $G_{desp}(y)$, respectively. If these networks are optimized, we can use the generator as a network for speech enhancement.

According to adversarial training, the objective function for MTAE based generator is represented by (1) as follows.

$$\begin{aligned} L_G = & -\lambda_1 E_{y\sim P_z}[D_{deno}(G_{deno}(y), y)] \\ & -(1-\lambda_1) E_{y\sim P_z}[D_{desp}(G_{desp}(y), y)] \\ & +\lambda_2 L_{MTAE} \end{aligned} \tag{1}$$

where λ_1 and λ_2 are hyper-parameters, and L_{MTAE} is the L1 loss. By experiment, we set $\lambda_1 = 0.5$ and $\lambda_2 = 100$. The objective functions for the discriminators (e.g. denoising, despeeching) are represented by (2) and (3) respectively as follows.

$$\begin{aligned} L_{D_{deno}} = & E_{y\sim P_z}[D_{deno}(G_{deno}(y), y)] \\ & - E_{x_s\sim P_{data}}[D_{deno}(x_s, y)] \\ & + \lambda_{gp} E_{\hat{x}_s\sim P_{\hat{x}_s}}[(\|\nabla_{\hat{x}_s} D_{deno}(\hat{x}_s)\|_2 - 1)^2] \end{aligned} \tag{2}$$

$$\begin{aligned} L_{D_{desp}} = & E_{y\sim P_z}[D_{desp}(G_{desp}(y), y)] \\ & - E_{x_n\sim P_{data}}[D_{desp}(x_n, y)] \\ & + \lambda_{gp} E_{\hat{x}_n\sim P_{\hat{x}_n}}[(\|\nabla_{\hat{x}_n} D_{desp}(\hat{x}_n)\|_2 - 1)^2] \end{aligned} \tag{3}$$

In the implementation, the generator consists of 5 hidden layers with 1024, 1280, 1536, 1792, and 2048 units, the number of hidden units increase linearly as reported in [13]. To estimate a concatenated 2 contiguous frames of speech and noise, a concatenated 16 contiguous frames of 13-dimensional MFCCs (13x16) noisy features are used as input. Two discriminators feed as real and fake pairs

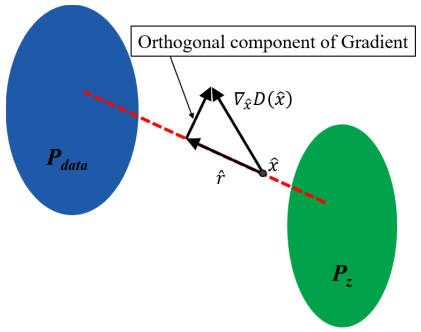

Fig. 2 Orthogonal gradient penalty between P_{data} and P_z.

(x_speech, y), $(G_{deno}(y), y)$ and (x_noise, y), $(G_{desp}(y), y)$, respectively as shown in Fig. 1. The denoising discriminator network is 4-layers with 1024, 768, 512, and 256 units, and 3-layers with 512 units per layer for despeeching discriminator.

2.2 Orthogonal Gradient Penalty (OGP)

Although a gradient penalty denoted in (4) is widely used for satisfying Lipschitz condition in WGAN [14]–[16], a hyper-parameter λ_{gp} is too sensitive to train network. H. Petzka et al. proposed a modified gradient penalty for resolving the problem [17]. These methods only considered a magnitude of the gradient to satisfy the Lipschitz condition without regard to the direction of the gradient. In theory, the direction of gradient always suggests the best way for optimization. But, Stochastic Gradient Descent (SGD) obtained by averaging gradients on each mini-batch point is practically applied to train deep networks in many ways [18]. Thus, we investigate and seek the best direction of gradient for learning and propose a new form named Orthogonal Gradient Penalty (OGP). The method is motivated by the optimal discriminator with loss function penalty derived from considering straight lines connecting coupled points from generator distribution P_z and the data distribution P_{data}. Therefore, the loss function should not only consider the magnitude as in the previously proposed method in [19] but also the direction toward the target data distribution (from P_z to P_{data}). We believe by adding the direction into the penalty term will ensure the model to converge faster and finds a better solution. Figure 2 illustrates the proposed penalty by the orthogonal direction components.

From the original gradient penalty in Eq. (4), let $E_{x\sim P_{data}}[D(x, y)] = \boldsymbol{p}$ and $E_{z\sim P_z}[D(G(y), y)] = \boldsymbol{q}$, we can write a unit direction vector of the straight line, $\hat{r}$ as in Eq. (5).

$$GP = \lambda_{gp} E_{\hat{x}\sim P_{\hat{x}}}[(\|\nabla_{\hat{x}} D(\hat{x})\|_2 - 1)^2] \tag{4}$$

$$\hat{r} = \frac{\boldsymbol{p} - \boldsymbol{q}}{\sqrt{(\boldsymbol{p} - \boldsymbol{q})^T (\boldsymbol{p} - \boldsymbol{q})}} \tag{5}$$

where, $\boldsymbol{p}$ and $\boldsymbol{q}$ represent the sample of the real and the generated data that have been obtained from the MTAE. Next, by subtracting a unit direction vector, we can constrain on

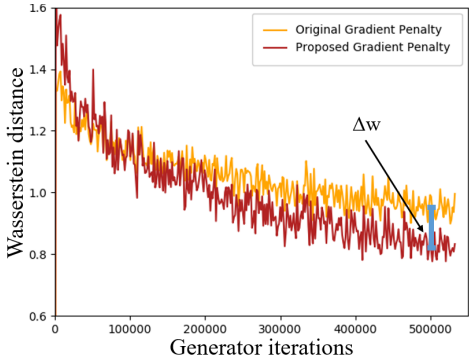

Fig. 3 The denoising discriminator Wasserstein distance of our model, where our proposed gradient penalty converges faster when training.

an orthogonal component of the gradient to the shortest path for optimization in (6). Note that λ_{gp} is set to 100 in all experiments.

$$OGP = \lambda_{gp} E_{\hat{x}\sim P_{\hat{x}}}[(\|\nabla_{\hat{x}} D(\hat{x}) - \hat{\boldsymbol{r}}\|_2)^2] \tag{6}$$

This modification enables the discriminator to direct the enhanced data (fake) to the shortest direction to the real data. As demonstrated in Fig. 3, our proposed penalty term in (6) results in lower error rates and achieves 14.1% improvement in Wasserstein distance convergence rate after 500k generator iterations in comparison to the original penalty in (4). Δw indicates the distance gain at $500k^{th}$ generator iteration.

3. Experiment Setup

3.1 Dataset

The TIMIT database is used in the experiment. Models are trained on TIMIT training utterances corrupted by 10 types of background noise (2 artificial: pink and red noise; and 8 from YouTube.com: classroom, laundry room, lobby, playground, rain, restaurant, river, and street) with four signal-to-noise ratio (SNR) levels (5, 10, 15, and 20 dB), and 9% of clean speeches are retained. Synthetic noise has been shown to be effective for speech enhancement task [20].

Validation set consists of TIMIT validation utterances corrupted by 10 types of noise used for training but with different samples unseen by the training. The testing set is corrupted with 4 types of unseen noise (café, pub, schoolyard, and shopping center), collected from ETSI EG 202 396-1 V1.2.2 (2008-09). The validation and test sets are all grouped at four SNR levels (5, 10, 15, and 20 dB) under all the aforementioned noise conditions by using ADDNOISE MATLAB [21].

3.2 Preprocessing

For the ASR model, we used the KALDI toolkit to train a Combination SGMM and Dans DNN model on a clean TIMIT training data. The features used from audio signals are sampled at 16 kHz and extracted by applying a short-time Fourier transform with a window size of 25 ms and 10 ms window step. Then, we apply 23 Mel-filter banks, with Mel-scale from 20 Hz to 7800 Hz.

For training the denoising MTAE_WGAN_OGP, the features are normalized in the range of $[-1, 1]$ per utterance. We apply the RMSprop optimizer with the learning rate of 0.0001. The models are trained with a batch size of 100. LReLU activation function is used in all layers except in the final output layer for DDAE and MTAE architecture.

4. Results

To compare effectiveness of the proposed MTAE_WGAN_OGP, we used MTAE architecture in [13], DDAE [22], R-CED(CNN) [3] and RNN as baseline models. Performance is evaluated using word error rates (WERs) obtained from ASR model. The RNN model consists of 3 layers of LSTM and one fully connected layer as output layer. The number of LSTM cells and FCN nodes are 512 in each layer. To deal with exploding gradients problem, we use a gradient clipping from −1 to 1 [23]. The results are summarized in Table 1.

From Table 1, it is demonstrated that the MTAE_WGAN_OGP consistently gives lower WERs compared to that of baseline models. The MTAE_WGAN_OGP achieve 9.7%, 8.6%, 6.2%, and 4.8% overall improvement of the WERs relative to DDAE, MTAE, R-CED(CNN) and RNN model. WER improvements become more pronounced at low SNR conditions. In addition, we can observe that the RNN model performs well at high SNR conditions. However, as SNR becomes lower, the performance becomes worse. For speech enhancement tasks, frames that are too far apart may not be necessary for denoising tasks. Thus, concatenated only contiguous frames in Feedforward Neural Networks seem to work sufficiently.

The proposed gradient penalty gives a slight improvement in the WER over the original penalty on average. More significantly, the proposed gradient penalty converges faster in training as shown in Fig. 3. Since training GAN is usually slow and unstable, the proposed orthogonal penalty alleviates these difficulties.

5. Conclusion

We proposed a new orthogonal gradient penalty (OGP) method for WGAN that outperforms and converges faster compared to the magnitude-based gradient penalty. The results have shown that the proposed MTAE_WGAN_OGP achieved 9.7%, 8.6%, 6.2%, and 4.8% WER improvements relative to DDAE, MTAE, R-CED(CNN) and RNN model, respectively, while the training achieved 14.1% convergence rate improvement.

Acknowledgments

The authors of Korea University are funded by the Ministry

Table 1 WER comparison between DDAE, CNN(R-CED), RNN, MTAE_WGAN_GP and MTAE_WGAN_OGP on 4 types of unseen background noise with four SNR conditions.

snr	Model	WER (%)				
		Café	Pub	Schoolyard	Shop	Avg
20 dB	DDAE	27.5%	27.4%	28.0%	25.4%	27.1%
	MTAE	27.6%	26.8%	28.3%	25.5%	27.0%
	R-CED (CNN)	27.5%	25.4%	26.8%	25.3%	26.2%
	RNN	26.2%	**24.9%**	26.5%	**24.6%**	**25.5%**
	MTAE_WGAN_GP	**25.8%**	25.3%	**26.4%**	**24.6%**	**25.5%**
	MTAE_WGAN_OGP	**25.6%**	25.3%	**26.1%**	**24.4%**	**25.3%**
15 dB	DDAE	30.3%	30.6%	33.6%	26.5%	30.2%
	MTAE	30.4%	29.9%	33.0%	26.5%	29.9%
	R-CED (CNN)	29.5%	**28.1%**	31.9%	**25.1%**	28.6%
	RNN	29.2%	28.6%	32.0%	**25.1%**	28.7%
	MTAE_WGAN_GP	28.8%	**28.0%**	30.2%	25.9%	28.2%
	MTAE_WGAN_OGP	**28.0%**	**28.1%**	**29.4%**	25.6%	**27.8%**
10 dB	DDAE	37.5%	38.5%	41.4%	31.9%	37.3%
	MTAE	36.2%	37.3%	40.0%	31.6%	36.3%
	R-CED (CNN)	35.3%	35.0%	38.9%	30.1%	34.8%
	RNN	35.3%	35.4%	39.3%	31.2%	35.3%
	MTAE_WGAN_GP	**32.7%**	**33.9%**	37.4%	**28.9%**	**33.2%**
	MTAE_WGAN_OGP	**32.7%**	**33.9%**	**36.2%**	29.3%	**33.0%**
5 dB	DDAE	44.9%	48.7%	49.4%	36.1%	44.7%
	MTAE	43.8%	48.4%	49.1%	35.7%	44.2%
	R-CED (CNN)	42.0%	46.5%	49.0%	34.8%	43.0%
	RNN	42.1%	46.2%	48.0%	35.2%	42.8%
	MTAE_WGAN_GP	40.2%	44.9%	46.5%	**33.5%**	41.3%
	MTAE_WGAN_OGP	**39.7%**	**44.5%**	**45.4%**	**33.5%**	**40.7%**

of Environment supported by the Korea Environmental Industry & Technology Institute's environmental policy-based public technology development project (2017000210001). David Han's contribution is supported by the US Army Research Laboratory.

References

[1] Y. Xu, J. Du, L.-R. Dai, and C.-H. Lee, "A regression approach to speech enhancement based on deep neural networks," IEEE/ACM Trans. Audio, Speech, Language Process., vol.23, no.1, pp.7–19, Jan. 2015.

[2] X. Feng, Y. Zhang, and J. Glass, "Speech feature denoising and dereverberation via deep autoencoders for noisy reverberant speech recognition," IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp.1759–1763, May 2014.

[3] S.R. Park and J.W. Lee, "A fully convolutional neural network for speech enhancement," Proc. Annual Conference of the International Speech Communication Association, INTERSPEECH, pp.1993–1997, 2017.

[4] B. Li and K.C. Sim, "A spectral masking approach to noise-robust speech recognition using deep neural networks," IEEE Trans. Audio, Speech, Language Process., vol.22, no.8, pp.1296–1305, Aug. 2014.

[5] D. Wang and J. Chen, "Supervised speech separation based on deep learning: An overview," IEEE/ACM Trans. Audio, Speech, Language Process., vol.26, no.10, pp.1702–1726, Oct. 2018.

[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," 13th Annual Conference of the International Speech Communication Association 2012, INTERSPEECH 2012, pp.22–25, 2012.

[7] S. Pascual, A. Bonafonte, and J. Serrà, "SEGAN: Speech enhancement generative adversarial network," Proc. Annual Conference of the International Speech Communication Association, INTERSPEECH, vol.2017-August, no.D, pp.3642–3646, 2017.

[8] A. Pandey and D. Wang, "On adversarial training and loss functions for speech enhancement," IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp.5414–5418, Sept. 2018.

[9] C. Donahue, B. Li, and R. Prabhavalkar, "Exploring speech enhancement with generative adversarial networks for robust speech recognition," IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp.5024–5028, Sept, 2018.

[10] D. Michelsanti and Z.-H. Tan, "Conditional generative adversarial networks for speech enhancement and noise-robust speaker verification," Proc. Annual Conference of the International Speech Communication Association, INTERSPEECH, pp.2008–2012, 2017.

[11] M. Mimura, S. Sakai, and T. Kawahara, "Cross-domain speech recognition using nonparallel corpora with cycle-consistent adversarial networks," 2017 IEEE Automatic Speech Recognition and Understanding Workshop, ASRU 2017, pp.134–140, 2018.

[12] T. Higuchi, K. Kinoshita, M. Delcroix, and T. Nakatani, "Adversarial training for data-driven speech enhancement without parallel corpus," 2017 IEEE Automatic Speech Recognition and Understanding Workshop, ASRU 2017, pp.40–47, Dec. 2018.

[13] H. Zhang, C. Liu, N. Inoue, and K. Shinoda, "Multi-task autoencoder for noise-robust speech recognition," IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp.5599–5603, April 2018.

[14] I. Gulrajani, F. Ahmed, M. Arjovsky, V. Dumoulin, and A. Courville, "Improved training of Wasserstein GANs," Advances in Neural Information Processing Systems, pp.5768–5778, 2017.

[15] C. Donahue, B. Li, and R. Prabhavalkar, "Exploring speech enhancement with generative adversarial networks for robust speech recognition," IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp.5024–5028, 2018.

[16] X. Wei, B. Gong, Z. Liu, W. Lu, and L. Wang, "Improving the improved training of Wasserstein GANs: A consistency term and its dual effect," International Conference on Learning Representations (ICLR), 2018.

[17] H. Petzka, A. Fischer, and D. Lukovnicov, "On the regularization of Wasserstein GANs," International Conference on Learning Representations (ICLR), 2018.

[18] L. Bottou, "Large-scale machine learning with stochastic gradient descent," Proc. COMPSTAT'2010, eds. Y. Lechevallier and G. Saporta, ch. Large-Scal, pp.177–186, Physica-Verlag HD, Heidelberg, 2010.

[19] M. Arjovsky, S. Chintala, and L. Bottou, "Wasserstein GAN," International Conference on Machine Learning (ICML), Jan. 2017.

[20] S.-X. Wen, J. Du, and C.-H. Lee, "On generating mixing noise signals with basis functions for simulating noisy speech and learning DNN-based speech enhancement models," IEEE International Workshop on Machine Learning for Signal Processing (MLSP), pp.1–6, Dec. 2017.

[21] ITU, "P.56 recommendation: Objective measurement of active speech level," 2011.

[22] X. Lu, Y. Tsao, S. Matsuda, and C. Hori, "Speech enhancement based on deep denoising autoencoder," Proc. Annual Conference of the International Speech Communication Association, INTERSPEECH, pp.436–440, 2013.

[23] R. Pascanu, T. Mikolov, and Y. Bengio, "On the difficulty of training recurrent neural networks," 30th International Conference on Machine Learning, ICML 2013, pp.2347–2355, 2013.