NOISE-ROBUST SPEECH RECOGNITION WITH 10 MINUTES UNPARALLELED IN-DOMAIN DATA

Chen Chen[1], *Nana Hou*[1], *Yuchen Hu*[1], *Shashank Shirol*[2], *Eng Siong Chng*[1] *

[1]School of Computer Science and Engineering, Nanyang Technological University, Singapore
[2]Manipal Institute of Technology, Manipal, India
chen1436@e.ntu.edu.sg

ABSTRACT

Noise-robust speech recognition systems require large amounts of training data including noisy speech data and corresponding transcripts to achieve state-of-the-art performances in face of various practical environments. However, such plenty of in-domain data is not always available in the real-life world. In this paper, we propose a generative adversarial network to simulate noisy spectrum from the clean spectrum (Simu-GAN), where only 10 minutes of unparalleled in-domain noisy speech data is required as labels. Furthermore, we also propose a dual-path speech recognition system to improve the robustness of the system under noisy conditions. Experimental results show that the proposed speech recognition system achieves 7.3% absolute improvement with simulated noisy data by Simu-GAN over the best baseline in terms of word error rate (WER).

Index Terms— Generative adversarial network, contrastive learning, automatic speech recognition

1. INTRODUCTION

Noise-robust automatic speech recognition (ASR) is a challenging task as huge training data and corresponding transcripts are required to achieve state-of-the-art word error rate (WER) performances [1–4]. However, such plenty of noisy data is not always available under some practical scenarios as well as collecting target data and transcribing them are also time-consuming and labor-intensive.

To address the problem of lacking enough noisy data, prior work [5] proposes to train the ASR system with large amounts of clean data and then finetunes with limited noisy in-domain data. Another work [6] proposes to extract the pure noise segments from the limited noisy in-domain data and then recursively adds them to the large amounts of clean data to generate the simulated noisy data. Such "mixup" simulated noisy data are then utilized for the subsequent ASR training.

*This research is supported by the Air Traffic Management Research Institute of Nanyang Technological University.

Recent works [7–9] in the vision field propose to utilize the generative adversarial network (GAN) to simulate the target-domain images. Furthermore, with only a few minutes of real speech as labels, GAN is also successfully applied in voice conversion task to generate high-quality speech [10]. The prior works are the inspiration source of this paper.

In this paper, we propose a generative adversarial network to simulate noisy data (Simu-GAN) with only 10 minutes of unparalleled in-domain noisy data for supervision, which provides a promising solution for the noise-robust ASR system with limited in-domain training data. When such clean-to-noisy mapping in the Simu-GAN is trained well, it can generate a large amount of simulated in-domain data at the run-time inference.

Specifically, the proposed Simu-GAN consists of a generator and a discriminator. The generator aims to map the clean spectrum to the noisy spectrum and the discriminator is introduced to distinguish the simulated noisy spectrum from the real noisy spectrum [11]. At the training stage, the multi-layer patch-wise contrastive loss [12] is utilized to learn the mutual information (*i.e.*, speech content) between the clean spectrum and the simulated noisy spectrum. The discriminator aims to narrow the differences between the simulated noisy spectrum and the real noisy spectrum (*i.e.*, background noises). At the run-time reference, only the generator is required to generate the simulated noisy data.

In addition, we propose a dual-path ASR system to evaluate the effectiveness of the proposed Simu-GAN, where both the noisy data and clean data are utilized in the training process. The experiment shows that the proposed Simu-GAN could significantly improve the recognition accuracy and achieve comparable performances compared with the upper-bound baseline.

2. SIMU-GAN ARCHITECTURE

Given the adequate clean data $X = \{x \in \mathcal{X}\}$ in the source domain and the limited noisy data $Y = \{y \in \mathcal{Y}\}$ in the target domain, we hope to generate the simulated noisy speech $\hat{X}$ that is close to the real noisy speech Y. To this end, we pro-

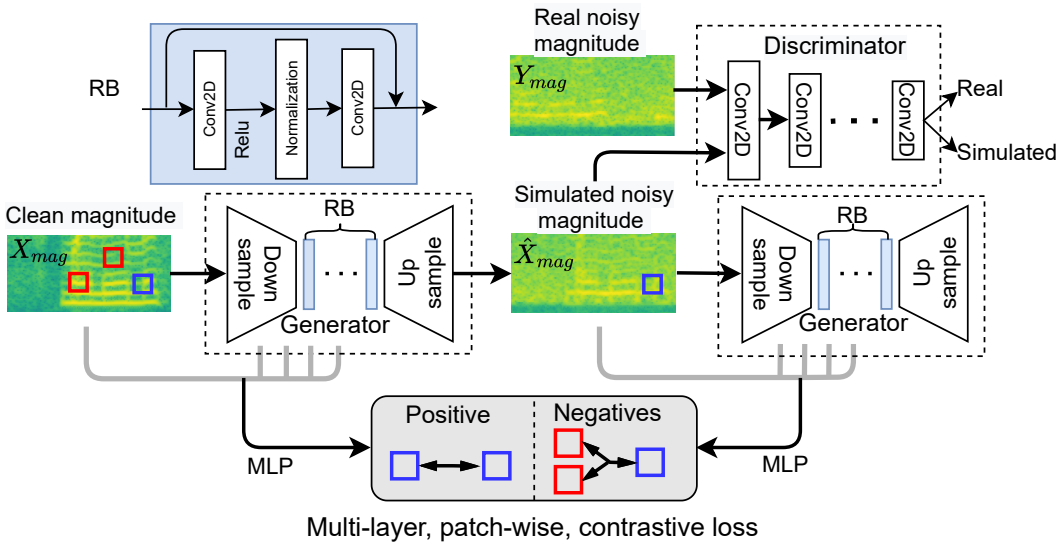

Fig. 1. The block diagram of proposed Simu-GAN structure. X, $\hat{X}$ and Y are the clean features, simulated noisy features and real noisy features, respectively. "RB" denotes the residual blocks, and "MLP" denotes two linear layers followed by the ReLU activation function.

pose the Simu-GAN architecture, where the clean speech X and the noisy speech Y are not required to be parallel.

2.1. Generator and Discriminator

The generator G is designed to map the clean magnitude X_{mag} to the real noisy magnitude Y_{mag} as shown in Fig.1. We first feed the clean features X_{mag} into two 2-D downsample convolutional layers with the kernel size of (3×3) and the stride of (2×2) as an encoder to learn the embeddings of the input features. Such encoded embeddings are then inputted to nine residual blocks (RB) to learn deep representations. Each residual block includes two convolutional layers with the kernel size of (3×3) and the stride of (1×1) followed by one dropout layer. Finally, two transposed convolutional layers with the kernel size of (3×3) and the stride of (2×2) act as a decoder to upsample the deep representations to the simulated noisy features $\hat{X}_{mag}$.

We now introduce the discriminator D to distinguish where the inputs come from (*i.e.*, simulated or real). The simulated magnitude features $\hat{X}_{mag}$ and the real noisy magnitude features Y_{mag} are fed into the discriminator, consisting of five 2-D convolutional layers with the kernel size of (4×4) followed by the LeakyReLU activation function. The stride takes (2×2) for the first three convolutional layers and (1×1) for the last two convolutional layers. At the training stage, the adversarial loss [7] is employed as:

$$\mathcal{L}_{GAN}(G, D, X, Y) = \mathbb{E}_{y\sim Y} \log D(y) + \mathbb{E}_{x\sim X} \log(1 - D(G(x))). \tag{1}$$

By minimizing the adversarial loss, the simulated features $\hat{X}_{mag}$ learn to be visually like the real noisy features Y_{mag}. Although the speech content of the simulated and noisy features are different, the discriminator mainly distinguishes the two features by the background noises [13].

2.2. Multi-layer Patch-wise Contrastive Loss

As the speech content of the clean and noisy features are different, we introduce the multi-layer patch-wise contrastive loss between the clean features X_{mag} and the simulated features $\hat{X}_{mag}$ to learn the mutual information (*i.e.*, speech content information).

Multi-layer and patch-wise. As shown in Figure 1, we reuse the generator G to learn the deep representations of the simulated features $\hat{X}_{mag}$. We first set one small patch in the simulated representations as "query" and then select the corresponding patch in the clean representations as the positive sample and random 256 patches as the negative samples. Such patches are reshaped via two linear layers with 256 units followed by the ReLU activation function. The contrastive loss is conducted between the positive sample and negative samples from 5 interested layers.

Contrastive Loss. We now introduce the contrastive learning to learn the mutual information by calculating the cross-entropy loss between the "query" and the positive/negative patches, formulated as:

$$\mathcal{L}_{MPC}(G, X) = \sum_{l=1}^{L} \sum_{i=1}^{I} -\log \left[\frac{e^{(\hat{z}_l^i \cdot z_l^i/\tau)}}{e^{(\hat{z}_l^i \cdot z_l^i/\tau)} + \sum_{j=1}^{J} e^{(\hat{z}_l^i \cdot z_l^j/\tau)}} \right] \tag{2}$$

where z_l^i and $\hat{z_l^i}$ denote the i^{th} positive patches in clean representations and simulated representations of the l^{th} layers in the generator, respectively. $\hat{z_l^j}$ is the j^{th} negative patches in simulated representations of the l^{th} layers in the generator. τ presents a temperature parameter in the contrastive learning [14].

Additionally, we also apply the multi-layer patch-wise contrastive loss $\mathcal{L}_{MPC}(G, Y)$ to the real noisy features Y_{mag} to prevent the generator from making unnecessary

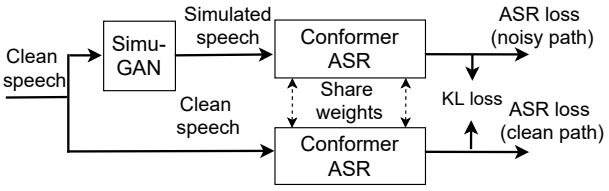

Fig. 2. The augmented dual-path ASR architecture. Two data flow are fed into Conformer-based ASR network as two independent batches, and the KL loss is computed between the outputs of decoder.

changes [12]. Therefore, the total loss function $\mathcal{L}_{total}$ of the Simu-GAN is formulated as:

$$\mathcal{L}_{total} = \mathcal{L}_{GAN}(G, D, X, Y) + \lambda \mathcal{L}_{MPC}(G, X) + \omega \mathcal{L}_{MPC}(G, Y) \tag{3}$$

where λ and ω are both set as 1 in this work.

3. DUAL-PATH ASR SYSTEM

To improve the robustness of the ASR systems under noisy conditions, we further propose a dual-path ASR system as shown in Figure 2. Specifically, we input the simulated speech generated by the Simu-GAN model and the corresponding clean speech as dual-path inputs into the conformer-based ASR system. We first introduce two ASR losses $\mathcal{L}^c_{asr}$ and $\mathcal{L}^n_{asr}$ calculated for the noisy path and clean path. We then propose a KL divergence-based consistency loss between the two decoder outputs of the noisy path and clean path as $\mathcal{L}_{kl}(X_{dec}, \hat{X}_{dec})$. Therefore, the loss of the dual-path ASR system $\mathcal{L}_{dp}$ is formulated as:

$$\mathcal{L}_{dp} = \alpha \mathcal{L}_{kl}(X_{dec}, \hat{X}_{dec}) + \beta \mathcal{L}^c_{asr} + (1 - \beta)\mathcal{L}^n_{asr} \tag{4}$$

where $\alpha, \beta \in [0, 1]$ are the parameters to balance the ASR loss $\mathcal{L}^{c/n}_{asr}$ and the KL loss $\mathcal{L}_{kl}(X_{dec}, \hat{X}_{dec})$. At the run-time inference, we only utilize the noisy path to evaluate the test set as the clean data of the test set is usually not available.

4. EXPERIMENTS AND RESULTS

4.1. Database

We conduct experiments on the dataset from robust automatic transcription of speech (RATS) program [15], which is recorded with a push-to-talk transceiver by playing back the clean Fisher data. The RATS has eight channels and could provide clean speech, noisy speech, and corresponding transcripts for various training goals. In this work, we select the data in channel A as the in-domain noisy data. The channel A includes 44.3-hours of training data, 4.9-hours of validation data, and 8.2-hours of testing data.

At the training stage for the proposed Simu-GAN, we only utilize small amounts of unparalleled clean/noisy data from the channel A. At the run-time reference for the Simu-GAN, we use the full clean data of the channel A to generate the simulated noisy data for subsequent ASR training. Such clean data, corresponding simulated noisy data and corresponding transcripts are utilized in the ASR training.

4.2. Experimental setup

At the training stage for Simu-GAN, the clean magnitude features were cut into the segments with the dimension of 129 $\times$ 128. The network was optimized by the Adam algorithm [16] and the learning rate started from 0.002.

At the training stage for all ASR baselines, the Conformer-based ASR system takes 80-dimensional log-mel feature as the input, where the encoder includes 12 Conformer layers and the decoder consists of 6 Transformer layers [17]. The byte-pair-encoding (BPE) [18] is utilized as the output token with a size of 994 and the shallow fusion [19] is employed to train a language model with the corresponding transcripts. For the dual-path ASR system, we set the hyper-parameters α and β to 0.4 and 0.7, respectively.

4.3. Reference Baselines

To evaluate the effectiveness of the proposed Simu-GAN and the dual-path ASR system, we built 5 baselines for comparison.

- Clean-ASR [17]: we train the single-path Conformer-based ASR system only with the clean RATS data.
- SpecAugment [20]: we swap the one frequency bin and one frame as a data augmentation approach for the clean-ASR system training.
- Finetune [5]: we tune the pre-trained clean-ASR system with the 10 minutes RATS channel A data.
- Mixup [6]: we generate the simulated noisy data by adding the noises segments extracting from full RATS channel A data and then train the single-path Conformer-based ASR system with the "mixup" simulated data.
- Noisy-ASR [17]: we train the single-path Conformer-based ASR system with the real RATS channel A data, which is the upper-bound performances with full in-domain data.

4.4. Results

4.4.1. Effect of the different amount of data for Simu-GAN training

We first analyze and summarize the WER performances of Simu-GAN with different amounts of training data on the single-path ASR system. Specifically, we utilize the different amounts of clean/noisy unparalleled data for the Simu-GAN training, and then the approximate 50-hours clean data

Table 1. The comparative study of the different amount of data for the Simu-GAN training. The WER (%) performances are evaluated on the single-path ASR system using the real test set of RATS channel A. "Clean speech" denotes the amount of the clean data of RATS channel A and "Noisy speech" presents the amount of the RATS channel A data. No speed perturbation is utilized in the following experiments.

System No.	Data for Simu-GAN training		WER (%)
	Clean speech	Noisy speech	
1	1 min	1 min	87.2
2	2 min	2 min	74.2
3	5 min	5 min	72.7
4	10 min	10 min	72.4
5	1 h	10 min	**68.9**

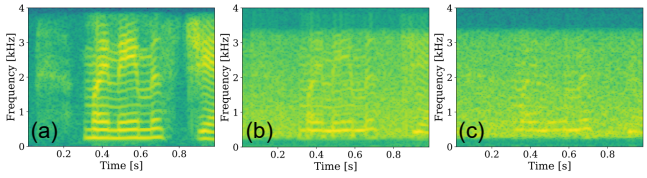

Fig. 3. The magnitude of a sample (10315_21016.wav) for (a) clean magnitude, (b) simulated noisy magnitude by Simu-GAN, and (c) real noisy magnitude from RATS channel A (ground-truth).

(44.3+4.9 hours) of RATS channel A are utilized to generate the simulated in-domain data for the subsequent ASR training and evaluation.

From Table 1, we observe that the performances significantly improve as the amount of clean/noisy training data increases. We obtain the best WER of 68.9% for the single-path ASR system with 1-hour clean training data and 10 minutes of noisy training data. To further show the contribution of the proposed Simu-GAN approach (system 5), we illustrate the magnitude spectrum of an example as shown in Figure 3. We can see that the proposed Simu-GAN approach can produce approximately the same spectrum (seen in Figure 3(b)) with the real RATS channel A sample (seen in Figure 3(c)). More listening samples are available at Github[1].

4.4.2. Effect of the proposed dual-path ASR system

We further report the effect of the proposed dual-path ASR system trained by the simulated noisy data, corresponding clean data, and transcripts shown in Table 2. Such simulated noisy data are generated by the proposed Simu-GAN with different amounts of training data. We observe that the proposed dual-path ASR systems could further improve the performances compared with the single-path ASR systems under the same amounts of training data. We obtain the best WER of 60.7% on the proposed dual-path ASR system with simulated data by Simu-GAN, which is trained by 1-hour clean training data and 10 minutes unparalleled noisy training data. We adopt this setting for Simu-GAN and the dual-path ASR system hereafter.

[1]https://chrisole.github.io/ICASSP2022-demo/

Table 2. The comparative study of the proposed dual-path ASR systems. "# ASR path" presents the number of path of the ASR systems (single-path or dual-path). "S.P." denotes the speed perturbation with $\{\times 0.9, \times 1.0, \times 1.1\}$.

System No.	Data for Simu-GAN training		# ASR path	S.P.	WER (%)
	Clean data	Noisy data			
4	10 min	10 min	Single	×	72.4
	10 min	10 min	Single	√	67.0
5	1 h	10 min	Single	×	68.9
	1 h	10 min	Single	√	63.0
6	10 min	10 min	Dual	×	65.4
	10 min	10 min	Dual	√	61.5
7	1 h	10 min	Dual	×	65.4
	1 h	10 min	Dual	√	**60.7**

Table 3. The comparative study of other competitive techniques. Speed perturbation with $\{\times 0.9, \times 1.0, \times 1.1\}$ are utilized in the following experiments.

Method Name	All data requirements for training			WER (%)
	Clean speech	Real noisy speech	Simulated noisy speech	
Clean-ASR	44.3 h	-	-	93.4
SpecAugment [20] (slight)	44.3 h	-	-	84.7
Finetune [5]	44.3 h	10 min with label	-	73.0
Mixup [6]	44.3 h	44.3 h without label	44.3 h	68.0
Simu-GAN (ours)	44.3 h	10 min without label	44.3 h	**60.7**
Noisy-ASR	-	44.3 h with label	-	49.5

4.4.3. Benchmark against other competitive methods

Table 3 summarizes the comparison between the proposed Simu-GAN and other competitive techniques in terms of the WER (%). We observe that the proposed Simu-GAN obtained the best performance. Comparing with the "Mixup" methods, the proposed Simu-GAN achieves the 7.3% absolute WER improvements. In addition, our best performance is also close to the upper-bound baseline trained by real in-domain data.

5. CONCLUSION

We propose a generative adversarial network to simulate the noisy in-domain speech from the clean speech (Simu-GAN) with 10 minutes of real noisy samples as labels to address the problem that the in-domain data is limited. We also propose a dual-path ASR system to improve the robustness of the ASR systems under noisy conditions. Experimental results show that the proposed Simu-GAN achieves the 7.3% absolute WER improvements on the dual-path ASR system compared with the best baseline.

6. REFERENCES

[1] A. Hannun, C. Case, J. Casper, B. Catanzaro, G. Diamos, E. Elsen, R. Prenger, S. Satheesh, S. Sengupta, A. Coates *et al.*, "Deep speech: Scaling up end-to-end speech recognition," *arXiv preprint arXiv:1412.5567*, 2014.

[2] W. Chan, N. Jaitly, Q. Le, and O. Vinyals, "Listen, attend and spell: A neural network for large vocabulary conversational speech recognition," in *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2016, pp. 4960–4964.

[3] C.-C. Chiu, T. N. Sainath, Y. Wu, R. Prabhavalkar, P. Nguyen, Z. Chen, A. Kannan, R. J. Weiss, K. Rao, E. Gonina *et al.*, "State-of-the-art speech recognition with sequence-to-sequence models," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 4774–4778.

[4] H. Hu, T. Tan, and Y. Qian, "Generative adversarial networks based data augmentation for noise robust speech recognition," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 5044–5048.

[5] D. Yu, L. Deng, and G. Dahl, "Roles of pre-training and fine-tuning in context-dependent dbn-hmms for real-world speech recognition," in *Proc. NIPS Workshop on Deep Learning and Unsupervised Feature Learning*, 2010.

[6] D. Ma, G. Li, H. Xu, and E. S. Chng, "Improving code-switching speech recognition with data augmentation and system combination," in *2019 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*. IEEE, 2019, pp. 1308–1312.

[7] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," *Advances in neural information processing systems*, vol. 27, 2014.

[8] A. Creswell, T. White, V. Dumoulin, K. Arulkumaran, B. Sengupta, and A. A. Bharath, "Generative adversarial networks: An overview," *IEEE Signal Processing Magazine*, vol. 35, no. 1, pp. 53–65, 2018.

[9] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros, "Image-to-image translation with conditional adversarial networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 1125–1134.

[10] H. Kameoka, T. Kaneko, K. Tanaka, and N. Hojo, "Stargan-vc: Non-parallel many-to-many voice conversion using star generative adversarial networks," in *2018 IEEE Spoken Language Technology Workshop (SLT)*. IEEE, 2018, pp. 266–273.

[11] N. Hou, C. Xu, E. S. Chng, and H. Li, "Domain adversarial training for speech enhancement," in *2019 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*. IEEE, 2019, pp. 667–672.

[12] T. Park, A. A. Efros, R. Zhang, and J.-Y. Zhu, "Contrastive learning for unpaired image-to-image translation," in *European Conference on Computer Vision*. Springer, 2020, pp. 319–345.

[13] N. Hou, C. Xu, E. S. Chng, and H. Li, "Learning disentangled feature representations for speech enhancement via adversarial training," in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021, pp. 666–670.

[14] Z. Wu, Y. Xiong, S. X. Yu, and D. Lin, "Unsupervised feature learning via non-parametric instance discrimination," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 3733–3742.

[15] D. Graff, K. Walker, S. M. Strassel, X. Ma, K. Jones, and A. Sawyer, "The rats collection: Supporting hlt research with degraded audio data." in *LREC*. Citeseer, 2014, pp. 1970–1977.

[16] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[17] D. Ma, N. Hou, V. T. Pham, H. Xu, and E. S. Chng, "Multitask-based joint learning approach to robust asr for radio communication speech," *arXiv preprint arXiv:2107.10701*, 2021.

[18] T. Kudo and J. Richardson, "Sentencepiece: A simple and language independent subword tokenizer and detokenizer for neural text processing," *arXiv preprint arXiv:1808.06226*, 2018.

[19] S. Karita, N. Chen, T. Hayashi, T. Hori, H. Inaguma, Z. Jiang, M. Someki, N. E. Y. Soplin, R. Yamamoto, X. Wang *et al.*, "A comparative study on transformer vs rnn in speech applications," in *2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*. IEEE, 2019, pp. 449–456.

[20] D. S. Park, W. Chan, Y. Zhang, C.-C. Chiu, B. Zoph, E. D. Cubuk, and Q. V. Le, "Specaugment: A simple data augmentation method for automatic speech recognition," *arXiv preprint arXiv:1904.08779*, 2019.