Multi-Task Multi-Network Joint-Learning of Deep Residual Networks and Cycle-Consistency Generative Adversarial Networks for Robust Speech Recognition

Shengkui Zhao, Chongjia Ni, Rong Tong, Bin Ma

Machine Intelligence Technology, Alibaba Group

{shengkui.zhao, ni.chongjia, rong.tong, b.ma}@alibaba-inc.com

Abstract

Robustness of automatic speech recognition (ASR) systems is a critical issue due to noise and reverberations. Speech enhancement and model adaptation have been studied for long time to address this issue. Recently, the developments of multi-task joint-learning scheme that addresses noise reduction and ASR criteria in a unified modeling framework show promising improvements, but the model training highly relies on paired clean-noisy data. To overcome this limit, the generative adversarial networks (GANs) and the adversarial training method are deployed, which have greatly simplified the model training process without the requirements of complex front-end design and paired training data. Despite the fast developments of GANs for computer visions, only regular GANs have been adopted for robust ASR. In this work, we adopt a more advanced cycle-consistency GAN (CycleGAN) to address the training failure problem due to mode collapse of regular GANs. Using deep residual networks (ResNets), we further expand the multi-task scheme to a multi-task multi-network joint-learning scheme for more robust noise reduction and model adaptation. Experiment results on CHiME-4 show that our proposed approach significantly improves the noise robustness of the ASR system by achieving much lower word error rates (WERs) than the state-of-the-art joint-learning approaches.

Index Terms: Robust speech recognition, convolutional neural networks, acoustic model, generative adversarial networks

1. Introduction

Automatic speech recognition (ASR) systems often need to deal with distorted target speech interfered by various ambient noise and reverberations in far-field noisy environments, therefore hardly achieve similar recognition accuracy as in close-talking scenarios. Many research works attempt to overcome the robustness issue through designing robust acoustic modeling or speech enhancement approaches. In the past decade, the robust acoustic modeling approaches that are built on the deep neural networks (DNNs) [1, 2, 3], the convolutional neural networks (CNNs) [4, 5, 6, 7] and the recurrent neural networks (RNNs) [8, 9, 10] have greatly improved the robust modeling capabilities. However, when the mismatch of training noise conditions and testing noise conditions increases, the performance of the above acoustic modeling approaches often degrades dramatically. On the other hand, the speech enhancement approaches such as the traditional statistical methods [11, 12] and the deep neural networks [13, 14, 15] are also adopted for easing the training and testing noise mismatches. However, the criteria of speech enhancement such as mean squared errors tend to distort the target speech and often lead to a suboptimal solution [16].

The multi-task joint-learning schemes [16, 17, 18] are developed to combine feature enhancement and acoustic modeling in a unified modeling framework for better ASR performance. The work [16] demonstrated that the joint training of a hybrid DNN architecture for feature mapping and acoustic modeling can greatly improve ASR performance of noisy speech. The work [17] proposed a shared DNN architecture with task-dependent target layers for feature enhancement and acoustic modeling, which also achieved improved ASR performance. To overcome the requirements of the paired clean-noisy training data of the above joint-learning approaches, the work [18] proposed to apply the generative adversarial networks (GANs) and the adversarial training method [19] in the join-learning framework. It was shown that the discriminator of GANs that is used to distinguish the enhanced samples from the true clean samples can guide the learning of the generator of GANs for enhancing noisy speech features. The discriminator does not need matched clean-noisy pairs in the adversarial training process. Therefore, the deployment of GANs in acoustic modeling has greatly simplified the data preparation task as well as the complex front-end design. Alternately, GANs can also be used to map both noisy and clean features into a common feature domain for robust ASR as done in [20]. Despite the fast developments of advanced GANs for computer visions, the above studies of GANs for ASR modeling only focused on the regular GANs. Since mapping speech features from one domain to another domain is highly under constrained [21], regular GANs easily suffer from model collapse problems which results in the failure of acoustic modeling.

In this work, we address the under-constrained issue of speech feature mapping by applying the more advanced cycle-consistency GAN (CycleGAN) [21] in the joint-learning framework. Different from the regular GANs, CycleGANs use both adversarial loss and cycle consistency loss to address the under-constrained issue. As a result, the generator of CycleGANs is forced to produce diverse results and maintain informative speech characteristics. It has been demonstrated that CycleGANs have achieved superior performance in image-domain translation tasks in comparison to regular GANs [21]. Our experimental results also show the CycleGAN approach can generate more reliable and diverse enhanced speech features than the regular GANs approach [18]. To further improve the modeling capabilities of our acoustic model, we apply the ensemble approach [22] with a multi-network structure for robust feature representations [23]. Our ensemble features are concatenations of the outputs of the generator of a CycleGAN and a parallel-designed deep residual network (ResNet) [24]. The ensemble features are channel-wisely recalibrated using the squeeze-and-excitation block [23]. Our proposal forms a multi-task multi-network joint-learning framework.

The remainder of this paper is organized as follows. We

firstly introduce the deep residual networks and the cycle-consistency generative adversarial networks in Section 2. We then present our proposed multi-task multi-network joint learning framework in Section 3. We provide experiments and results in Section 4. Section 5 concludes this paper.

2. Deep Residual Networks and Cycle-Consistency Generative Adversarial Networks

2.1. Deep residual networks

It is well known that the vanishing gradient introduces training difficulty for deeper networks [24]. The deep residual networks (ResNets) introduced by K. He [24] consist of "shortcut connections" to solve the convergence and the degradation problems when networks go deeper. In the residual blocks of ResNets, instead of directly learning a nonlinear mapping from input x to output $H(x)$, the residual mapping $F(x) := H(x) - x$ is learnt. The original mapping is then recast into $F(x) + x$. ResNets have greatly increased the depth of the trainable networks and achieved very good performance in many classification problems [24] and speech applications [25, 26].

2.2. Cycle-consistency generative adversarial networks

The cycle-consistency generative adversarial networks (CycleGANs) introduced by Zhu et al. [21] aims to learn the mapping between an input image to an output image when paired training data is not available. Since this mapping is highly under-constrained, the regular GANs introduced by I. Goodfellow [19] tend to fall in model-collapse. Different from the regular GANs, CycleGANs add an inverse mapping F: $Y \rightarrow X$ on top of the target mapping G: $X \rightarrow Y$. The cycle consistency loss enforces $F(G(X)) \approx X$. Therefore, the adversarial learning process of CycleGANs contains two types of loss optimizations: adversarial loss and cycle consistency loss.

The adversarial loss is used for optimizing the mapping function $G : X \rightarrow Y$, where the generator G tries to generate samples $G(x)$ that are to be indistinguishable from real samples y in domain Y, while the discriminator D aims to distinguish between $G(x)$ and real samples y. The cycle consistency loss enforces the existing of an inverse mapping, i.e., $x \rightarrow G(x) \rightarrow F(G(x)) \rightarrow x$ with the following objective:

$$\min_F V(F) = \frac{1}{2}\,\mathbb{E}_{x\sim p(x)}[\|F(G(x)) - x\|_1] \tag{1}$$

where the L_1 norm is used to enforce a one-to-one inverse mapping. Although regular GANs have been introduced to many speech applications such as speech conversation [27], speech synthesis [28], speech enhancement [29], robust ASR [18, 20], CycleGAN is rarely explored for robust ASR. We adopt CycleGAN architecture into our proposed approach.

3. The Proposed Multi-Task Multi-Network Joint Learning Framework

3.1. The proposed architecture

The overall flowchart of our proposed approach is illustrated in Fig. 1. The main task is to classify given frame-level input features into senone labels which are aligned labels for Hidden Markov Model (HMM) states of the acoustic model. The framework consists of a generator (G), a discriminator (D), an inverse generator (F), a dual network (M), a feature squeeze-and-excitation block (SE), and a senone label classifier (C). G, D, and F form the CycleGAN architecture. G aims to the generate enhanced features $\mathbf{x}$ given the noisy input features $\tilde{\mathbf{x}}$. F aims to map the enhanced features back to the noisy features by the cycle consistency. D aims to distinguish the enhanced features from the true clean features. M is a parallel network taking noisy input features. The output features of G and M are passed to the SE block, where the features are channel-wisely recalibrated and concatenated to form input features for C. C classifies the input features into senone labels. The whole acoustic model consists of the encoder of G, M, the SE block and C.

We adopt the encoder-decoder architecture [18] for both G and F. For ease of design, we use same networks and keep same input and output dimensions for G and F. The encoders and decoders are implemented with ResNets to increase network depth. In G and F, the decoding process mirrors the encoding process. In order to bypass low-level features for feature reconstruction, while maintaining the discriminative information in the condensed bottleneck representation for classification, we add skip connections from the residual blocks of the encoders to the residual blocks of the decoders in G and F. D is built to be a 2-class classifier for true clean or fake clean. It takes either a batch of clean features from the dataset or a batch of enhanced features from the generator. M has a ResNet structure and takes the noisy features as input and outputs the discriminative features for classification. Different from the features learnt by the encoder of G, M learns adaptive feature representations from only classification task while the encoder of G learns from the multi-task minimizations of C, D and F. The SE block is applied to the output features of the encoder of G and the output of M, which learns to use global information to selectively emphasize informative features and suppress less useful ones. In the SE block, all features are concatenated to form an input vector $\mathbf{h}$ of the classifier C, where the senone labels are classified. The whole networks are jointly optimized by the adversarial loss criterion, the cycle-consistency criterion, the L1-norm criterion and the softmax loss criterion.

3.2. The optimization objectives

We apply the least squares losses [30] for the adversarial training and the objective for the discriminator D is to minimize:

$$\begin{aligned}\min_D V(D) =& \frac{1}{2}\,\mathbb{E}_{\mathbf{x}\sim p(\mathbf{x})}\left[(D(\mathbf{x})-1)^2\right] \\ &+ \frac{1}{2}\,\mathbb{E}_{\tilde{\mathbf{x}}\sim p(\tilde{\mathbf{x}})}\left[\left(D\left(G(\tilde{\mathbf{x}})\right)\right)^2\right]\end{aligned} \tag{2}$$

The objective of the inverse generator F is to minimize the cycle consistency loss using L_1 norm:

$$\min_F V(F) = \frac{1}{2}\,\mathbb{E}_{\mathbf{x}\sim p(\mathbf{x})}\left[\|F\left(G(\tilde{\mathbf{x}})\right) - \tilde{\mathbf{x}}\|_1\right] \tag{3}$$

The objective for the classifier C is to minimize the classification cross-entropy loss based on the vector $\mathbf{h}$:

$$\min_C V(C) = \frac{1}{2}\,\mathbb{E}_{\mathbf{h}\sim p(\mathbf{h})}\left[-\log\left(C(k/\mathbf{h})\right)\right] \tag{4}$$

The adversarial G loss is to maximize the chances that the discriminator D classifies its output to be the true clean class:

$$\min_G V_{GAN}(G) = \frac{1}{2}\,\mathbb{E}_{\tilde{\mathbf{x}}\sim p(\tilde{\mathbf{x}})}\left[\left(D\left(G(\tilde{\mathbf{x}})\right)-1\right)^2\right] \tag{5}$$

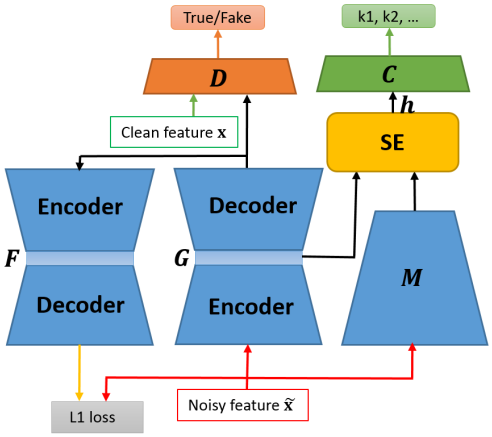

Figure 1: *The flowchart of the proposed multi-task multi-network joint-learning framework.*

With multi-task optimization, the full G loss is given by:

$$\min_{G} V(G) = V(C) + \alpha V_{GAN}(G) + \beta V(F) \qquad (6)$$

where α and β are hyper-parameters valued between 0 and 1. When $\beta = 0$ and $\alpha \neq 0$, the CycleGAN model G is equivalent to the regular GAN model G [18]. When $\alpha = 0$ and $\beta = 0$, the CycleGAN model G is equivalent to a CNN model.

Finally, the objective for the network M is to minimize the classification loss of C. Therefore, we have

$$\min_{M} V(M) = V(C) \qquad (7)$$

All the parameters of D, G, F, M and C are alternatively trained to fine-tune the model. The SE block is updated together with the minimization of C loss.

4. Experiments and Results

4.1. CHiME-4 dataset

The CHiME-4 dataset [31] was used to evaluate our proposed approach and compare with the baseline [18]. The dataset consists of real and simulated recordings of speech from the Wall Street Journal (WSJ0) corpus. The real recordings are from 6 channels under 4 types of noise environments: bus, cafe, pedestrian area, and street junction. The simulated data is generated from clean speech utterances and continuous background noise recordings. The overall dataset involves a training set of 1600 real and 7138 simulated utterances, a development set of 1640 real and 1640 simulated utterances, and a test set of 1320 real and 1320 simulated utterances. All data are sampled at 16 kHz. In our experiments, all 6 channels except channel 2 were pooled for model training and the official one-channel track lists were used for development and evaluation. For GANs training, the real samples were drawn from the original WSJ0 training data.

4.2. Experimental setup

In all the experiments, we used 40-dimensional filter-bank features with 19 frames of context features concatenated as 40×19 dimensional input. All filter-bank features were normalized to have zero mean on each dimension and were globally normalized to be -1 to 1. The senone labels for training data are obtained from force-alignments of a well-trained GMM-HMM system using Kaldi [32]. The posterior probabilities of the classifier C are decoded using Kaldi WFST decoder and the WSJ 5k trigram LM is used as language model.

Table 1: *WERs (%) of all compared models on the CHiME-4 corpus with single-channel evaluation. Models: (A) DNN×7, (B) GAN-CNN×4, (C) CNN×10, (D) ResNet×17, (E) CycleGAN-ResNet×17, (F) ResNet×33, (G) Dual-CycleGAN-ResNet×33.*

Models	(A)	(B)	(C)	(D)	(E)	(F)	(G)
dev_sim	19.38	17.43	16.24	15.55	15.4	15.4	**15.01**
dev_real	18.61	17.05	15.36	14.44	14.0	13.8	**13.35**
avg	19.00	17.24	15.8	15.00	14.7	14.59	**14.18**
test_sim	31.72	28.39	27.35	26.32	25.3	25.1	**24.9**
test_real	36.29	31.83	30.66	28.68	27.9	27.2	**26.58**
avg	34.01	30.11	29.01	27.50	26.61	26.15	**25.74**

In our proposed approach, we built a 17-layer ResNet with {channels × blocks} of {64×2, 128×2, 256×2, 512×2} for the encoders of G and F. We applied stride-2 convolutions with 3 × 3 filter kernels. For the decoders of G and F, we built a mirrored 17-layer ResNet with {channels × blocks} of {512×2, 256×2, 128×2, 64 × 2}. The transposed stride-2 convolutions were used with zero paddings. The network M was built as a 33-layer ResNet with {channels × blocks} of {64×3, 128×4, 256×6, 512×3}. We built D and C as 2-layer DNNs of 1024 neurons with input dimensions of 760×1 and 2048×1, respectively. In G, F, and M, the leaky rectified linear units (LeakyReLUs) were used as the nonlinear activations. In D and C, the ReLUs [33] were used as the nonlinear activations and softmax function was used for output layers. For all hidden layers, dropout with a probability of 0.3 was added and batch-normalization was performed. During training, we used a mini-batch of 128 samples and the learning rate was set to 0.0002. The Adam optimizer [34] was used for backpropagation optimization and the whole training algorithm was implemented with PyTorch [35]. We set the hyper-parameters $\alpha = 0.4$ and $\beta = 1.0$ for best performance. The acoustic model trained by our proposed approach was denoted as 'Dual-CycleGAN-ResNet×33'.

We chose three existing systems for comparison purpose. The first baseline system was a 7-layer DNN acoustic model (denoted as 'DNN×7'), which is the official Kaldi *chime4* recipe. The second baseline system was the regular GAN approach [18] (denoted as 'GAN-CNN×4'), which consists of a 4-layer CNN encoder, a 4-layer CNN decoder and a 2-layer DNN classifier. The third baseline system was a 10-layer CNN plus 4-layer DNN model [36] (denoted as 'CNN×10). We also separately trained acoustic models using the encoder of G plus C (denoted as 'ResNet×17'), the CycleGAN part with M discarded (denoted as 'CycleGAN-ResNet×17'), and the network M plus C with CycleGAN discarded (denoted as 'ResNet×33') for comparison. All methods use the same 40-dimensional filter-bank input features.

4.3. Results

We first report the WERs of all compared systems on the development and test set in Table 1. It is observed that all CNN-based approaches achieved better performance than the DNN-based 'DNN×7' approach, which agrees with the literature studies [36]. In the CNN-based approaches, the network depth is one of the dominant factors for improving the ASR performance. The deeper the CNN networks in the ASR systems

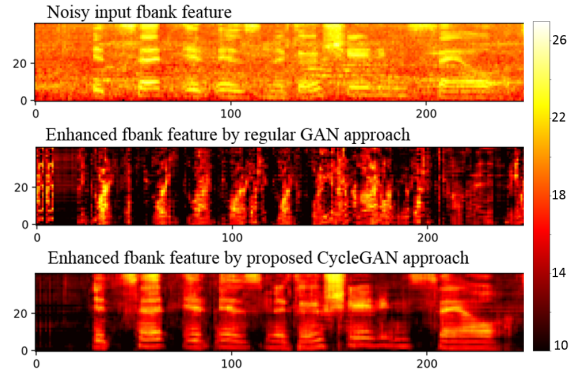

Figure 2: *A sample of noisy input fbank feature for sentence "it said it is negotiating the sale" enhanced by the regular GAN approach [18] and our proposed CycleGAN approach.*

were used, the better performance was obtained. Without using the joint-learning framework, the deeper 'ResNet×33' approach performs better than the 'CNN×10' and 'ResNet×17' approaches. After applying the multi-task joint-learning framework, the 'CycleGAN-ResNet×17' approach achieves large improvement over the 'ResNet×17' approach where both approaches have the same network structures and complexities during evaluation. Although the 'GAN-CNNx4' approach also significantly improves over the baseline 'DNN×7' approach, its performance still has a large gap compared to the 'CycleGAN-ResNet×17' approach and 'Dual-CycleGAN-ResNet×33' approach. Among all the compared approaches, the proposed 'Dual-CycleGAN-ResNet×33' approach achieves the best performance, which mainly owns to the deep ResNet architectures, the deep adversarial training using CycleGANs, and the dual-network architecture with ensemble features.

Fig. 2 illustrates GAN training failure case for the regular GAN approach [18] due to the highly under-constrained problem. The enhanced features were generated by the generators of the regular GAN approach and our approach after a complete training process. It is observed that the generator of the regular GAN approach fails to preserve the speech structures as shown in the original noisy speech. Our approach employing the cycle-consistency constraint successfully overcomes this problem. As shown in the Fig. 2, our approach works well on both noise reduction and the preservation of the speech structures.

5. Conclusions

In this paper, we proposed a multi-task multi-network joint-learning framework which integrates the latest CycleGAN and ResNets for robust speech recognition. Ensemble features from CycleGAN and ResNet show improved discriminative classification performance. Our proposed learning framework avoids paired clean-noisy training data and boosts the training stability and the noise robustness of ASR systems. Our evaluation on the CHiME-4 noisy test set showed significant performance improvement compared to the state-of-the-art systems.

6. References

[1] G. Hinton, L. Deng, D. Yu, G. Dahl, A.-R. Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, T. Sainath, and B. Kingsbury, "Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups," *IEEE Signal Processing Magazine*, vol. 29, no. 6, pp. 82–97, 2012.

[2] F. Seide, L. Deng, and D. Yu, "Conversational speech transcription using context-dependent deep neural networks," in *Proceedings of Interspeech*, 2011, pp. 437–440.

[3] G. Dahl, D. Yu, L. Deng, and A. Acero, "Context-dependent pre-trained deep neural networks for large-vocabulary speech recognition," *IEEE Transactions on Audio Speech and Language Processing*, vol. 20, no. 1, pp. 30–42, 2012.

[4] T. Sainath, A.-R. Mohamed, B. Kingsbury, and B. Ramabhadran, "Deep convolutional neural networks for lvcsr," in *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing*, May 2013, p. 8614–8618.

[5] T. Sainath and C. Parada, "Convolutional neural networks for smallfootprint keyword spotting," in *Proceedings of Interspeech*, 2015, p. 1478–1482.

[6] T. Sainath, B. Kingsbury, G. Saon, H. Soltau, A.-R. Mohamed, G. Dahl, and B. Ramabhadran, "Deep convolutional neural networks for large-scale speech tasks," *Neural Networks*, vol. 64, p. 39–48, 2015.

[7] Y. Qian, M. Bi, T. Tan, and K. Yu, "Very deep convolutional neural networks for noise robust speech recognition," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 24, p. 2263–2276, Aug 2016.

[8] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural computation*, vol. 9, no. 8, p. 1735–1780, 1997.

[9] H. Sak, A. Senior, and F. Beaufays, "Long short-term memory recurrent neural network architectures for large scale acoustic modeling," in *Proceedings of Interspeech*, 2014.

[10] V. Peddinti, Y. Wang, D. Povey, and S. Khudanpur, "Low latency acoustic modeling using temporal convolution and lstms," *IEEE Signal Processing Letters*, 2017.

[11] P. Vincent, H. Larochelle, Y. Bengio, and P.-A. Manzagol, "Extracting and composing robust features with denoising autoencoders," in *Proceedings of the 25th International Conference on Machine Learning*, 2008.

[12] B. Li and K. C. Sim, "Improving robustness of deep neural networks via spectral masking for automatic speech recognition," in *Proceedings of the IEEE Workshop on Automatic Speech Recognition and Understanding*, 2013.

[13] J. Du, Q. Wang, T. Gao, Y. Xu, L. Dai, and C.-H. Lee, "Robust speech recognition with speech enhanced deep neural networks," in *Proceedings of Interspeech*, 2014.

[14] K. Han, Y. He, D. Bagchi, E. Fosler-Lussier, and D. Wang, "Deep neural network based spectral feature mapping for robust speech recognition," in *Proceedings of Interspeech*, 2015.

[15] M. Mimura, S. Sakai, and T. Kawahara, "Speech dereverberation using long short-term memory," in *Proceedings of Interspeech*, 2015.

[16] T. Gao, J. Du, L. Dai, and C.-H. Lee, "Joint training of front-end and back-end deep neural networks for robust speech recognition," in *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing*, 2015.

[17] Z.-Q. Wang and D. L. Wang, "A joint training framework for robust automatic speech recognition," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 24, no. 4, p. 796–806, Apr 2016.

[18] B. Liu, S. Nie, Y. Zhang, D. Ke, S. Liang, and W. Liu, "Boosting noise robustness of acoustic model via deep adversarial training," in *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing*, 2018.

[19] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in *Proceedings of the Neural Information Processing Systems*, 2014.

[20] S. Sun, B. Zhang, L. Xie, and Y. Zhang, "An unsupervised deep domain adaptation approach for robust speech recognition," *Neurocomputing*, vol. 257, p. 79–87, 2017.

[21] J. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," in *International Conference on Computer Vision*, 2017.

[22] W. Xiong, J. Droppo, X. Huang, F. Seide, M. Seltzer, A. Stolcke, D. Yu, and G. Zweig, "The microsoft 2016 conversational speech recognition system," in *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing*, 2017.

[23] Y. Chen, J. Li, H. Xiao, X. Jin, S. Yan, and J. Feng, "Dual path networks," in *Proceedings of the Neural Information Processing Systems*, 2017.

[24] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2016.

[25] Y. Wang, X. Deng, S. Pu, and Z. Huang, "Residual convolutional ctc networks for automatic speech recognition," in *arXiv preprint arXiv: 1702.07793*, 2017.

[26] T. Tan, Y. Qian, H. Hu, Y. Zhou, W. Ding, and K. Yu, "Adaptive very deep convolutional residual network for noise robust speech recognition," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2018.

[27] C.-C. Hsu, H.-T. Hwang, Y.-C. Wu, Y. Tsao, and H.-M. Wang, "Voice conversion from unaligned corpora using variational autoencoding wasserstein generative adversarial networks," in *arXiv preprint arXiv:1704.00849*, 2017.

[28] B. Bollepalli, L. Juvela, and P. Alku, "Generative adversarial network-based glottal waveform model for statistical parametric speech synthesis," in *Proceedings of Interspeech*, 2017.

[29] S. Pascual, A. Bonafonte, and J. Serr, "Segan: Speech enhancement generative adversarial network," in *Proceedings of Interspeech*, 2017.

[30] X. Mao, Q. Li, H. Xie, R. Y. Lau, Z. Wang, and S. P. Smolley, "Least squares generative adversarial networks," in *IEEE International Conference on Computer Vision*, 2017.

[31] E. Vincent, S. Watanabe, A. A. Nugraha, J. Barker, and R. Marxer, "An analysis of environment, microphone and data simulation mismatches in robust speech recognition," *Computer Speech and Language*, 2016.

[32] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz, and et al., "The kaldi speech recognition toolkit," in *IEEE 2011 workshop on automatic speech recognition and understanding*, 2011.

[33] V. Nair and G. Hinton, "Rectified linear units improve restricted boltzmann machines," in *International Conference on Machine Learning*, 2010.

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

[35] "pytorch," https://github.com/pytorch/pytorch.

[36] Y. Qian, M. Bi, T. Tan, and K. Yu, "Very deep convolutional neural networks for noise robust speech recognition," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 24, p. 2263–2276, Aug 2016.