GANS FOR CHILDREN: A GENERATIVE DATA AUGMENTATION STRATEGY FOR CHILDREN SPEECH RECOGNITION

Peiyao Sheng, Zhuolin Yang, Yanmin Qian

MoE Key Lab of Artificial Intelligence
SpeechLab, Department of Computer Science and Engineering
Shanghai Jiao Tong University, Shanghai, China

ABSTRACT

Due to the high acoustic variability, children speech recognition suffers significant performance reduction on most ASR systems which are optimized mainly using adults speech with limited or even none children speech. One of the most straight ideas to solve this problem is to increase the children's speech data during training, however, it is restricted by the more difficult process and higher cost when collecting children's speech compared to adults'. In this work, we develop a generative adversarial network (GANs) based data augmentation method to increase the size of children's training data to improve speech recognition performance for children's speech. Two different types of GANs are explored under WGAN-GP training framework, including the unconditional GANs with an unsupervised learning framework and the conditional GANs using acoustic states as conditions. The proposed data augmentation approaches are evaluated on a Mandarin speech recognition task, with only 40-hour children speech or further including 100-hour adult speech in the training. The results show that more than relative 20% WER reduction can be obtained on children speech testset with the proposed method, and the generated children speech with GAN even can improve the adults' speech within our experimental setups.

Index Terms— children speech recognition, data augmentation, generative adversarial networks

1. INTRODUCTION

The development of deep learning significantly boosted the improvement of automatic speech recognition (ASR) systems[1, 2, 3]. However, children's speech recognition still fails to perform as expected because of the following two reasons. First of all, the performance degradation of children's ASR systems can be attributed to the high level of acoustic variability in the speech patterns of children[4, 5, 6], which directly results in the acoustic mismatch between childrens and adults speech. Secondly, it is more difficult and costly to collect children's speech so that most current advanced ASR systems were trained mainly with adults' dataset and limited children's data [7]. Along with the expansion of its application scope in all aspects of life such as education, internet of things, entertainment and security, it is of great importance to enhance the performance of children's ASR systems.

In order to address this problem, some studies try to reduce the mismatch between children's and adults' voice by algorithms like Vocal Tract Length Normalisation(VTLN), Stochastic Feature Mapping(SFM) and Prosody Usage Optimization [8, 9, 10, 11]. Another way to improve the performance of children's ASR is to introduce more childrens corpus in the process of training data-driven models. Some works get improved results by combining adults speech with childrens speech and find that female adult speech will bring more benefits[12, 13, 14, 15]. However, the improvement brought by deep learning models is restricted probably due to the lack of large amounts of childrens training data. Therefore, some studies utilize multi-task learning frameworks or transfer learning to adapt adults' speech to children's speech to overcome the limitation of data. And from this point, our work tends to develop a data augmentation method to improve children's speech recognition using generative models which are supposed to learn implicit feature distribution of children speech.

In this work, we mainly explores Generative Adversarial Networks (GANs)[16] for data augmentation to effectively increase the amount of training data, thereby improving the performance of children's ASR systems. GANs was first proposed in the field of computer vision and achieved huge success in high-resolution image generation, video prediction[17], sketch retrieval[18] etc., and then quickly applied to many other fields such as natural language processing [19] and speech recognition [20]. Although the research regarding GANs in the field of speech is not as much as image, there are many promising results in various topics when utilizing GANs, such as speech synthesis[21, 22], voice conversion[23], speech enhancement[24], spoken language identification[25] and acoustic scene classification

Yanmin Qian is the corresponding author.

This work was supported by the China NSFC projects (No. 61603252 and No. U1736202). Experiments have been carried out on the PI supercomputer at Shanghai Jiao Tong University.

(ASC)[26]. In recent work[20, 27], data augmentation for robust speech recognition using GANs was explored at the first time. In this work, we develop a data augmentation strategy utilizing WGAN-GP (Wassistain GAN with gradient penalty)[28] training procedure and explore both unconditional and conditional learning framework[29] to generate extra data, with which we will further train an augmented acoustic model. The experiments on Mandarin speech test set including children and adults' subset show that our proposed method will achieve a relative 10% to 20% Word Error Rate (WER) reduction on children's test set while still keeping great performance on adults' test set.

The rest of this paper is organized as follows. Section 2 briefly introduces the basic generative adversarial networks and the conditional version, together with two improved work, Wasserstein GAN with gradient penalty and conditional GAN with projection discriminator. The new proposed GAN-based data augmentation approach for children ASR is described in Section 3, where an unsupervised learning framework with the unlabeled generated data and a supervised learning framework with acoustic states as conditions are described. Section 4 introduces the experimental setting and shows the results and analysis, and Section 5 gives the conclusions.

2. GENERATIVE ADVERSARIAL NETWORKS

2.1. Generative adversarial network

Generative Adversarial Nets (GAN)[30] achieved much state-of-art progress in various generative tasks, by utilizing the adversarial learning process of two models: a generator (G) and a discriminator (D). The whole process can be regarded as a competition between D and G: Generator G aims at transforming a Gaussian noise $z \sim N(0,1)$ to a fake sample $\hat{x}$, such that the sample $\hat{x}$ can not be distinguished from the real data samples by Discriminator D. Discriminator is trained to make the fake samples and real samples distinguishable. The objective function for discriminator is:

$$\max_D \mathbb{E}_{x\sim P_r} \log D(x) + \mathbb{E}_z \log(1 - D(G(z))) \quad (1)$$

The discriminator D is trained to predict each datas validity 1 (true) and 0 (false). For the objective function of generator G is:

$$\min_G -\mathbb{E}_z \log D(G(z)) \quad (2)$$

Thus, G aims at generate samples which are classified as true by discriminator.

2.2. Wasserstein GAN

However, the original GANs training is unstable and many researches have tried to propose new training criterion to improve the stability and convergence of GANs training[31, 32, 33, 34]. Recently, Wasserstein Generative Adversarial Network (WGAN)[32] and Improved Training of Wasserstein GANs (WGAN-GP)[33], utilized Wasserstein distance between two distribution. Wasserstein distance, which is also called Earth-mover distance, is utilized as a distance estimator calculated by the discriminator in the form of gradient penalty loss due to its desirable property of being continuous and differentiable almost everywhere under mild assumptions. Specifically, improved WGAN using objective functions as follows:

$$\max_D \mathbb{E}_{x\sim P_r} D(x) - \mathbb{E}_z D(G(z)) - (||\nabla_{\hat{x}} D(\hat{x})|| - 1)^2 \quad (3)$$

$$\min_G -\mathbb{E}_z \log D(G(z)) \quad (4)$$

$$\hat{x} = \alpha x + (1-\alpha) G(z) \quad (5)$$

where α is a random number between 0 and 1. The Gradient Penalty (GP) term enforces the norm of gradients of D to 1. This formulation can provide more stable GAN training process.

2.3. cGAN

In order to embed the condition information into GAN's training process, the conditional GAN (cGAN)[35, 36, 37], an extension of original GAN, utilize condition information in both the generator and discriminator. By integrating additional condition information, cGAN can generate data under desired condition. The objective function in cGAN can be written as follows:

$$\max_D \mathbb{E}_{x\sim P_r} D(x,c) - \mathbb{E}_z D(G(z,c)) - (||\nabla_{\hat{x}} D(\hat{x},c)|| - 1)^2 \quad (6)$$

$$\min_G -\mathbb{E}_z \log D(G(z,c)) \quad (7)$$

In order to embed conditional information into GAN's training process, many researchers have tried some possible methods[35, 36, 37, 38]. We followed Miyato's work[37], which regards conditional information as a projection between Discriminator's intermediate features and conditional labels.

In this work, two generative frameworks are explored: the unconditional GANs and conditional GANs (cGANs). Both of them are implemented on the frame-level, which is the feature map on the children's speech spectrum extracted from the waveform of speech. Specifically, the filter bank (FBANK) features are used as the input of the discriminator and the output of the generator. The basic unit of to input and generate is a context sequence of frames whose concatenation is roughly at the scale of a syllable.

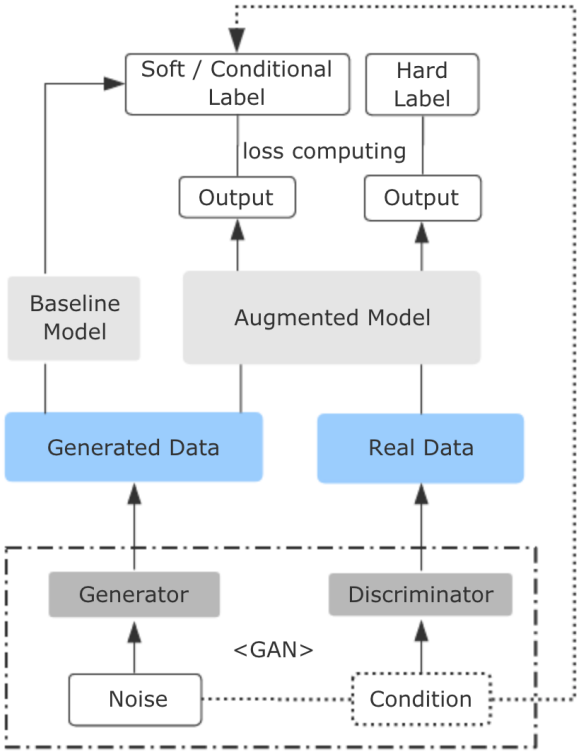

Fig. 1. The proposed framework of data augmentation with GANs for children ASR. First of all, GAN models are trained using real children's data. Secondly, acoustic baseline model and augmented models are trained. Furthermore, the output probability of baseline model in unconditional experiments and state information in conditional experiments are correspondingly utilized to get the labels.

3. DATA AUGMENTATION USING GANS FOR CHILDREN ASR

Based on the generative models described in Section 2, two types of data augmentation frameworks are explored in this work, the unconditional GANs and conditional GANs (cGANs). Both of them are implemented on the frame-level, which is the feature map on the children's speech spectrum extracted from the waveform of speech. Specifically, the filter bank (FBANK) features are used as the input of the discriminator and the output of the generator. The basic unit of to input and generate is a context sequence of frames whose concatenation is roughly at the scale of a syllable. Given those K-dimension FBANK features, we stack m of them to form a $m \times K$ matrix. In our experiments, we set $K = 40$ and $m = 20$.

3.1. unconditional GANs

As shown in Figure 1, with the original real data, firstly different generative models will be trained to generate extra augmented data. For the unconditional generative experiments, due to the lack of labels for generated data, an unsupervised learning strategy is developed where we make use of a baseline acoustic model to generate the posterior probability for each frame, which is referred as the soft label later. Assuming that the distributions between the real data and augmented data generated a well-trained GANs model own high similarity, the KL divergence is used as training criterion for the acoustic model, which can derive an optimization function:

$$J = \sum_{\boldsymbol{o}_t \in D_g} \sum_{s} p_{baseline}(s|\boldsymbol{o}_t) \log p_{aug}(s|\boldsymbol{o}_t) \quad (8)$$

$$+ \sum_{\boldsymbol{o}_t \in D_r} \sum_{s} p_{ref} \log p_{aug}(s|\boldsymbol{o}_t) \quad (9)$$

where $\boldsymbol{o}_t$ is input features, s is the acoustic state, p_{ref} is the original labels as the reference. We denote generated dataset and real speech dataset as D_g and D_r. The posterior probabilities from the baseline acoustic model and augmented acoustic model are denoted as $p_{baseline}(s|\boldsymbol{o}_t)$ and $p_{aug}(s|\boldsymbol{o}_t)$, mentioned as soft labels.

3.2. conditional GANs

For the conditional GANs experiments, acoustic states, i.e. the clustered senone labels in pre-trained acoustic systems will be utilized as specific conditional information, serving as the guidance for network training and data generation. The embedding method of state condition is similar to [39]. As for the generator, the state information is still prepared to be a one-hot vector for concatenation. For the discriminator, an inner product is required to be taken between the embedded condition vector and the feature vector to introduce the condition information into the model. In comparison to unconditional GANs, whose information in the generative process is only a random noise vector, these acoustic states can also directly be used when deriving labels for augmented data generated from the generator in cGANs. With these data and labels, a new augmented acoustic model can be obtained by jointly training using the real data and generated data.

4. EXPERIMENTS

4.1. Experimental setup and baseline models

In the experiments, three types of dataset are used: 1) a 100-hour hand-transcribed Mandarin adult corpus including 120K utterances with an average duration of 3 seconds. 2) a 40-hour hand-transcribed Mandarin children corpus including 47k utterances. 3) a testing set containing four subsets of children speech, 16k utterances in total and two subsets of adults

Table 1. WER(%) comparison of acoustic modeling with different training data. B-01 is baseline model trained by children's data only and B-02 is baseline model trained by both children's and adults' data. G-01 and G-02 are unconditional generative models with different network architecture. G-03 trains the model same as G-01 but with extra adults' data. And CG-01 and CG-02 are conditional generative models trained with different labels.

ID	Training set	Children					Adults		
		A	B	C	D	AVG	A	B	AVG
B-01	40h Children	45.92	58.77	20.31	24.81	31.99	54.77	48.85	53.44
G-01	+ 40h GANs	**41.45**	49.82	16.18	22.37	27.85	39.10	33.28	37.80
G-02	+ 40h GANs (conv)[1]	41.48	50.85	**15.69**	**21.95**	**27.65**	38.74	34.35	37.55
B-02	+ 100h Adults	55.23	45.31	17.43	31.51	34.48	19.20	18.56	18.83
G-03	+ + 40h GANs	49.67	43.75	16.90	27.55	31.46	**18.86**	17.67	18.59
CG-01	+ + 40h-cGANs	50.00	43.34	16.95	27.53	31.54	18.97	17.81	18.71
CG-02	+ + 40h-cGANs (comb)[2]	49.84	**43.21**	16.76	27.40	31.38	18.87	**17.63**	**18.59**

[1] 'conv' means using convolutional layers.
[2] 'comb' means using the combination of soft labels and conditional labels for the generated children speech.

speech, 8k utterances in total. For children testing dataset it contains 4 different sub-datasets(A, B, C, D) which sampled from different environments while 2 sub-datasets(A, B) in Adults testing dataset. There are obvious differences among these sets of data, including collecting devices, domains.

Gaussian mixture model based hidden Markov models (GMM-HMM) is first built with Kaldi toolkit[40] using the standard recipe, consisting of 9663 clustered states trained using maximum likelihood estimation. With the well-trained GMM-HMM model, state level labels can be derived by performing a forced-alignment over the 100-hour real adults speech and 40-hour real children speech. All deep neural network (DNN) acoustic models are built with Kaldi using cross entropy criterion and asynchronous stochastic gradient descent (ASGD) based back propagation (BP) algorithm. 95% of training data are used for training and the rest 5% are used for validation. The standard testing pipelines in the Kaldi recipes are used for decoding and scoring.

The baseline models in our experiments contains 5 hidden layers with 2048 units in each layer, and the ReLU activation function is used after each layer; The input layer has 1320 units since we use 40-dimension filter bank features with Δ and $\Delta\Delta$, and context expansion with 5 frames on each side. The output layer consisted of 9663 units corresponding to GMM-HMM clustered states. For the better comparison, two baseline models with two experimental setups (B-01 and B-02) are trained with the same architecture but different training sets. B-01 is trained using only children's speech and B-02 is trained with both children's and adults' speech. Word error rate (WER) of two baseline models is listed in Table 1. It is observed that 1) only with the limited children speech for the system building, the performance is very bad for both chiledren's or adults' speech. 2) adding more adults data can dramatically improve the accuracy on adults' speech, but still limited effect for children's speech (or even degradation on some sets). The children speech is much more difficult to be recognized than adult speech in the conventional ASR.

4.2. Generative models

All the GANs models for data augmentation used here are implemented with PyTorch[41]. For unconditional GANs (G-01 and G-03), we use 4 layers fully connected network strucutre ($800 \rightarrow 1024 \rightarrow 768 \rightarrow 256 \rightarrow 1$) with ReLU activation function. Likewise, generator use the reverse which also contains four fully connected layers together with a sigmoid function for output. In G-02, we also introduce convolutional layers for better analysis on the structure configuration: For the discriminator, there are three convolutional layers with channels $\{128, 256, 512\}$ and strides $\{(1, 2), (3, 3), (3, 3)\}$, followed with a Leaky ReLU activation function after each layer and a fully connected layer at the end. Similar to the discriminator, there is a fully connected layer to transfer the input random noise, and then the generator uses three transposed convolutional layers to generate the feature maps. The input of generator is a random noise with dimension 256 sampled from centered isotropic multivariate Gaussian.

Based on the architecture of G-01 and G-03, our cGANs model takes conditional information as extra input in the form of one-hot vector and projects to 256 dimension vector v_c through a fully connected layer. Meanwhile the original feature input will also projects to a 256-dimension vector v_f through another fully connected layer. After that, v_f will be used by the rest of networks to compute adversarial loss and we add the inner product of v_c and v_f as the conditional loss.

In the experiment of CG-01, we use conditional information as direct labels for cGANs training. For CG-02, we use a linear combination of soft and conditional hard labels as the training labels by assigning a hyper-parameter $\beta \in [0, 1]$. More specifically, the new labels can be derived as follows:

$$p_{comb} = \lambda p_{baseline}(s|\boldsymbol{o}_t) + (1 - \beta)p_{condition} \qquad (10)$$

During the training process, the discriminator D is updated 5 times then followed one time update on the genera-

tor G in each mini-batch. The gradient penalty parameter λ is set to 10. The networks are trained using Adam, and the mini-batch size is set to 64.

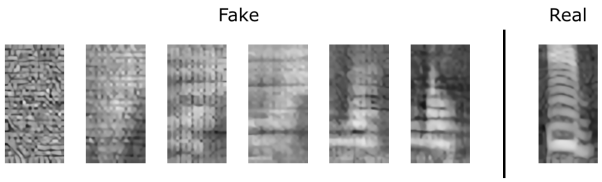

Fig. 2. Generated children's feature maps from GANs on different training epochs. The seven figures on the left is fake feature maps and the figure on the most right is the real one.

4.3. Evaluation and analysis

Visualization on generated data To better understand the feature samples generated from GANs, we visualize and compare the feature maps of the real children's speech data samples with the generated samples from the process of the model training in figure 2. According to the figure we can find that with the increased training epochs, the feature quality generated from the same noise vector are gradually improved with the convergence of the model, and the final generated sample has the high similarity with the real feature sample of children's. Comparing the different units in the generated feature map, it can be seen that the well-convergent model can generate features with diversity using different random noises.

Exploration on the generative models In order to study how the generative models will affect the quality of generated features and the improvement they can bring for ASR systems, we firstly test generative models with different network architectures using basic GAN or conditional GAN, and the experimental results using the proposed approach are shown in Table 1. 1) Setup#1 only with the children speech in the original training set: after adding the same amount of generated children's data as the real data, `G-01` can significantly reduce WER relatively 10% $\sim$ 20% on the children test set compared to `B-01`. Using convolutional layers in the GAN model in `G-02` can further improve the results. With such a data-limited setup, the generated children speech is also very helpful for recognizing the adults' speech. 2) Setup#2 with both adults' and children's speech in the original training set: it is obvious that WER on aults' speech can be dramatically reduced while the performance on children's speech becomes worse, and the adults' data seems useless for recognizing children speech. Using the proposed GAN or cGAN based children data augmentation method, we can still obtain a large WER reduction on children's testing set, which is also consistent with the observation in setup#1. Moreover, the generated children data is also helpful for the adults' speech, and even a slight improvement on adult speech is achieved. Both data generated from GANs (`G-03`) and cGAN (`CG-01`, `CG-02`) can improve the acoustic models to get better results on children's testing set, which indicates that both soft labels and conditional labels successfully guide the model training and the combination of them can achieve a better performance.

Exploration on the amount of data In the experiment, we also explored whether the amount of augmented data will have big influence on acoustic modeling. With the fixed 40-hour children's speech + 100-hour adults' speech in the second setup, we compare the systems using different amount of augmented children data, i.e. from 20 hours to 80 hours, and the results are shown in Table 2. At first, the WER is reduced as the generated data increases, but such improvement will be close to saturation when it is close to the real children data size used to train the generative models.

Table 2. Average Children WER(%) comparison of systems with different amounts of generated children data in the experimental setup#2. G-03 model is used here, which is shown in Table 1.

Data amount	0h	20h	40h	60h	80h
AVG WER(%)	34.48[1]	32.97	31.46	31.23	31.31

[1] The experiments are based on `B-02`

5. CONCLUSION

To sum up, our work has shown that with the limited children data and labels, our unsupervised and combined frameworks are able to produce powerful generative models. Via experiments across a diverse set of model settings, we have shown that introducing the augmented data generated by GANs lead to a significant enhancement for children's ASR system. The final system can reached a more than relative 20% WER reduction on children speech, and the newly GAN based generated children speech even can improve the adults' speech under some conditions.

This is the first work to use GAN on data generation for children speech recognition, and the comprehensive exploration on the different forms of GANs is performed. For the future work, we will extend the proposed framework to other attempts, such as generative transformation and sequential generation to further explore the effect of data augmentation. Also, it is interesting to explore the differentiation of speech recognition caused by age, gender and other key parameters using generative models.

6. REFERENCES

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

[2] Li Deng, Jinyu Li, Jui-Ting Huang, Kaisheng Yao, Dong Yu, Frank Seide, Michael Seltzer, Geoff Zweig, Xiaodong He, Jason Williams, et al., "Recent advances in deep learning for speech research at microsoft," in *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, 2013, pp. 8604–8608.

[3] Yanmin Qian, Mengxiao Bi, Tian Tan, and Kai Yu, "Very deep convolutional neural networks for noise robust speech recognition," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 24, no. 12, pp. 2263–2276, 2016.

[4] Alexandros Potamianos and Shrikanth Narayanan, "Robust recognition of children's speech," *IEEE Transactions on speech and audio processing*, vol. 11, no. 6, pp. 603–616, 2003.

[5] Sungbok Lee, Alexandros Potamianos, and Shrikanth Narayanan, "Acoustics of childrens speech: Developmental changes of temporal and spectral parameters," *The Journal of the Acoustical Society of America*, vol. 105, no. 3, pp. 1455–1468, 1999.

[6] Matteo Gerosa, Diego Giuliani, and Shrikanth Narayanan, "Acoustic analysis and automatic recognition of spontaneous children's speech," in *Ninth International Conference on Spoken Language Processing*, 2006.

[7] Anton Batliner, Mats Blomberg, Shona D'Arcy, Daniel Elenius, Diego Giuliani, Matteo Gerosa, Christian Hacker, Martin Russell, Stefan Steidl, and Michael Wong, "The pf_star children's speech corpus," in *Ninth European Conference on Speech Communication and Technology*, 2005.

[8] Rohit Sinha and Syed Shahnawazuddin, "Assessment of pitch-adaptive front-end signal processing for childrens speech recognition," *Computer Speech & Language*, vol. 48, pp. 103–121, 2018.

[9] Shweta Ghai, *Addressing pitch mismatch for children's automatic speech recognition*, Ph.D. thesis, 2011.

[10] Prashanth Gurunath Shivakumar, Alexandros Potamianos, Sungbok Lee, and Shrikanth Narayanan, "Improving Speech Recognition for Children using Acoustic Adaptation and Pronunciation Modeling," in *Proc. of the Workshop on Child Computer Interaction (WOCCI)*, 2014, vol. 5, pp. 15–19.

[11] Chenda Li and Yanmin Qian, "Prosody usage optimization for children speech recognition with zero resource children speech.," in *INTERSPEECH*, 2019.

[12] Hank Liao, Golan Pundak, Olivier Siohan, Melissa K. Carroll, Noah Coccaro, Qi Ming Jiang, Tara N. Sainath, Andrew Senior, Francoise Beaufays, and Michiel Bacchiani, "Large vocabulary automatic speech recognition for children," *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, vol. 2015-Janua, pp. 1611–1615, 2015.

[13] Mengjie Qian, Ian McLaughlin, Wu Quo, and Lirong Dai, "Mismatched training data enhancement for automatic recognition of children's speech using DNN-HMM," *Proceedings of 2016 10th International Symposium on Chinese Spoken Language Processing, ISCSLP 2016*, , no. m, 2017.

[14] Joachim Fainberg, Peter Bell, Mike Lincoln, and Steve Renals, "Improving children's speech recognition through out-of-domain data augmentation," *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, vol. 08-12-Sept, pp. 1598–1602, 2016.

[15] Yao Qian, Xinhao Wang, Keelan Evanini, and David Suendermann-Oeft, "Improving DNN-Based Automatic Recognition of Non-native Children ' s Speech with Adult Speech," *Proceedings of WOCCI*, pp. 40–44, 2016.

[16] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio, "Generative adversarial nets," in *Advances in neural information processing systems*, 2014, pp. 2672–2680.

[17] Xiaodan Liang, Lisa Lee, Wei Dai, and Eric P Xing, "Dual motion gan for future-flow embedded video prediction," in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 1744–1752.

[18] Antonia Creswell and Anil Anthony Bharath, "Adversarial training for sketch retrieval," in *European Conference on Computer Vision*. Springer, 2016, pp. 798–809.

[19] Scott Reed, Zeynep Akata, Xinchen Yan, Lajanugen Logeswaran, Bernt Schiele, and Honglak Lee, "Generative adversarial text to image synthesis," *arXiv preprint arXiv:1605.05396*, 2016.

[20] Hu Hu, Tian Tan, and Yanmin Qian, "Generative Adversarial Networks Based Data Augmentation for Noise

Robust Speech Recognition," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. apr 2018, pp. 5044–5048, IEEE.

[21] Takuhiro Kaneko, Hirokazu Kameoka, Nobukatsu Hojo, Yusuke Ijima, Kaoru Hiramatsu, and Kunio Kashino, "Generative adversarial network-based postfilter for statistical parametric speech synthesis," in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2017, pp. 4910–4914.

[22] Yuki Saito, Shinnosuke Takamichi, and Hiroshi Saruwatari, "Statistical parametric speech synthesis incorporating generative adversarial networks," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 26, no. 1, pp. 84–96, 2017.

[23] Chin-Cheng Hsu, Hsin-Te Hwang, Yi-Chiao Wu, Yu Tsao, and Hsin-Min Wang, "Voice conversion from unaligned corpora using variational autoencoding wasserstein generative adversarial networks," *arXiv preprint arXiv:1704.00849*, 2017.

[24] Santiago Pascual, Antonio Bonafonte, and Joan Serrà, "Segan: Speech enhancement generative adversarial network," *arXiv preprint arXiv:1703.09452*, 2017.

[25] Peng Shen, Xugang Lu, Sheng Li, and Hisashi Kawai, "Conditional generative adversarial nets classifier for spoken language identification.," in *INTERSPEECH*, 2017, pp. 2814–2818.

[26] Seongkyu Mun, Sangwook Park, David K Han, and Hanseok Ko, "Generative adversarial network based acoustic scene training set augmentation and selection using svm hyper-plane," *Proc. DCASE*, pp. 93–97, 2017.

[27] Peiyao Sheng, Zhuolin Yang, Hu Hu, Tian Tan, and Yanmin Qian, "Data augmentation using conditional generative adversarial networks for robust speech recognition," in *2018 11th International Symposium on Chinese Spoken Language Processing (ISCSLP)*. IEEE, 2018, pp. 121–125.

[28] Ishaan Gulrajani, Faruk Ahmed, Martin Arjovsky, Vincent Dumoulin, and Aaron C Courville, "Improved training of wasserstein gans," in *Advances in Neural Information Processing Systems*, 2017, pp. 5767–5777.

[29] Mehdi Mirza and Simon Osindero, "Conditional generative adversarial nets," *arXiv preprint arXiv:1411.1784*, 2014.

[30] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio, "Generative adversarial nets," in *Advances in neural information processing systems*, 2014, pp. 2672–2680.

[31] Xi Chen, Yan Duan, Rein Houthooft, John Schulman, Ilya Sutskever, and Pieter Abbeel, "Infogan: Interpretable representation learning by information maximizing generative adversarial nets," in *Advances in neural information processing systems*, 2016, pp. 2172–2180.

[32] Martin Arjovsky, Soumith Chintala, and Léon Bottou, "Wasserstein gan," *arXiv preprint arXiv:1701.07875*, 2017.

[33] Ishaan Gulrajani, Faruk Ahmed, Martin Arjovsky, Vincent Dumoulin, and Aaron C Courville, "Improved training of wasserstein gans," in *Advances in Neural Information Processing Systems*, 2017, pp. 5767–5777.

[34] Junbo Zhao, Michael Mathieu, and Yann LeCun, "Energy-based generative adversarial network," *arXiv preprint arXiv:1609.03126*, 2016.

[35] Mehdi Mirza and Simon Osindero, "Conditional generative adversarial nets," *arXiv preprint arXiv:1411.1784*, 2014.

[36] Augustus Odena, Christopher Olah, and Jonathon Shlens, "Conditional image synthesis with auxiliary classifier gans," in *Proceedings of the 34th International Conference on Machine Learning-Volume 70*. JMLR. org, 2017, pp. 2642–2651.

[37] Takeru Miyato and Masanori Koyama, "cgans with projection discriminator," *arXiv preprint arXiv:1802.05637*, 2018.

[38] Scott Reed, Zeynep Akata, Xinchen Yan, Lajanugen Logeswaran, Bernt Schiele, and Honglak Lee, "Generative adversarial text to image synthesis," *arXiv preprint arXiv:1605.05396*, 2016.

[39] Takeru Miyato and Masanori Koyama, "cgans with projection discriminator," *arXiv preprint arXiv:1802.05637*, 2018.

[40] Daniel Povey, Arnab Ghoshal, Gilles Boulianne, Lukas Burget, Ondrej Glembek, Nagendra Goel, Mirko Hannemann, Petr Motlicek, Yanmin Qian, Petr Schwarz, et al., "The kaldi speech recognition toolkit," Tech. Rep., IEEE Signal Processing Society, 2011.

[41] Adam Paszke, Sam Gross, and Soumith Chintala, "Pytorch," 2017.