Speech Communication 114 (2019) 1–9

Contents lists available at ScienceDirect

Speech Communication

journal homepage: www.elsevier.com/locate/specom

Data augmentation using generative adversarial networks for robust speech recognition

Yanmin Qian[a,*], Hu Hu[b], Tian Tan[a]

[a] MoE Key Lab of Artificial Intelligence SpeechLab, Department of Computer Science and Engineering, Shanghai Jiao Tong University, Shanghai, China
[b] School of Electrical and Computer Engineering, Georgia Institute of Technology, Atlanta, USA

ARTICLE INFO

Keywords:
Robust speech recognition
Generative adversarial networks
Conditional generative adversarial networks
Data augmentation
Very deep convolutional neural network

ABSTRACT

For noise robust speech recognition, data mismatch between training and testing is a significant challenge. Data augmentation is an effective way to enlarge the size and diversity of training data and solve this problem. Different from the traditional approaches by directly adding noise to the original waveform, in this work we utilize generative adversarial networks (GAN) for data generation to improve speech recognition under noise conditions. In this paper we investigate different configurations of GANs. Firstly the basic GAN is applied: the generated speech samples are based on spectrum feature level and produced frame by frame without dependence among them, and there is no true labels. Thus, an unsupervised learning framework is proposed to utilize these untranscribed data for acoustic modeling. Then, in order to better guide the data generation, condition information is introduced into GAN structures, and the conditional GAN is utilized: two different conditions are explored, including the acoustic state of each speech frame and the original paired clean speech of each speech frame. With the incorporation of specific condition information into data generation, these conditional GANs can provide true labels directly, which can be used for later acoustic modeling. During the acoustic model training, these true labels are combined with the soft labels which make the model better. The proposed GAN-based data augmentation approaches are evaluated on two different noisy tasks: Aurora4 (simulated data with additive noise and channel distortion) and the AMI meeting transcription task (real data with significant reverberation). The experiments show that the new data augmentation approaches can obtain the performance improvement under all noisy conditions, which including additive noise, channel distortion and reverberation. With these augmented data by basic GAN / conditional GAN, a relative 6% to 14% WER reduction can be obtained upon an advanced acoustic model.

1. Introduction

In recent years, significant progress has been observed in automatic speech recognition (ASR) due to the introduction of deep neural networks based acoustic models (Hinton et al., 2012; Seide et al., 2011b; Dahl et al., 2012). These works showed promising performance improvement over traditional Gaussian mixture models (GMMs). However, these systems still do not work well under noisy environments (e.g., scenarios with additive noise, channel distortion and reverberation (Wang and Gales, 2012; Pearce, 2002; Hain et al., 2012a), where the strength of speech signal is lower, leading to low SNR and making the systems susceptible to additive noise and reverberation.

In order to solve the robustness problem in acoustic modeling, there are some effective solutions. Some previous works focused on the structure of acoustic models, such as convolutional neural networks (CNNs) (Abdel-Hamid et al., 2012; 2014; 2013; Qian et al., 2016), time delay neural networks (TDNNs) (Lang et al., 1990), CNN-LSTM-DNNs (CLDNNs) (Sainath et al., 2015a; 2015b), etc. These newly proposed advanced acoustic model structures can significantly improve the performance in noisy environments of ASR systems. Acoustic modeling adaptation is also an effective way (Seide et al., 2011a; Liao, 2013). It can significantly boost the performance corresponding to the specific scenario or speaker. In addition, speech enhancement is another strategy which is applied on the front-end signal or acoustic feature (Yu et al., 2008; Narayanan and Wang, 2013; Yoshioka and Gales, 2015). The speech enhancement technologies can often remove the noise from target speech before the recognition process and lead to a better recognition results. Moreover, the back-end technologies on acoustic modeling and the front-end technologies on denoising or dereverberation can be combined to get the better performance (Tan et al., 2018; Li et al., 2014a; Gong, 1995).

* Corresponding author.
E-mail addresses: yanminqian@sjtu.edu.cn (Y. Qian), huhu@gatech.edu (H. Hu), tantian@sjtu.edu.cn (T. Tan).

https://doi.org/10.1016/j.specom.2019.08.006
Received 4 January 2019; Received in revised form 23 July 2019; Accepted 14 August 2019
Available online 19 August 2019
0167-6393/

The main point to improve noise robustness of speech recognition is to solve the mismatch problem between the training and testing. Due to the large quantity of noise types in real scenarios, it is impossible to collect enough data covering all noise conditions in real world. Thus, data augmentation is an effective strategy to increase the quantity of training data and the diversity of noisy types, which can improve the model robustness. Multi-style training (MTR) has been widely adopted for a long time (Lippmann et al., 1987). In the traditional data augmentation methods, the noise is directly added to the original clean speech (Ko et al., 2015; 2017; Kim et al., 2017), which can get the simulated noisy data manually. Also, some neural network based methods are investigated for data augmentation (Cui et al., 2015). By these ways, although a performance gain is observed, there are still two main limitations: (1) the diversity of generated data is dependent on the existed speech and noise data; (2) artificial noises may have problems, such like an unrealistic stationarity, unrealistic repeating of the same noise, and too simplified room acoustics simulations.

In recent years, generative adversarial network (GAN) has received some great interests in computer vision communities (Goodfellow et al., 2014; Salimans et al., 2016; Shrivastava et al., 2016). A GAN consists of two networks: a discriminator to distinguish natural and generated samples, and a generator to deceive the discriminator. By adversarial training, it can learn generative models using two-player zero-sum game, which produces samples from the real data distribution. More recently, the condition information is introduced to GAN, named conditional GAN, and successfully applied to image generation and style transfer tasks (Mirza and Osindero, 2014; Odena et al., 2017; Isola et al., 2016). Additional condition information can guide the generation type for generator, which we can obtain specific samples with the assigned type. As for speech-related tasks, the application of GAN is still limited. There are several preliminary attempts, mainly focus on speech synthesis (Kaneko et al., 2017a; Saito et al., 2017), voice conversion (Hsu et al., 2017; Kaneko et al., 2017b) and speech enhancement (Pascual et al., 2017; Higuchi et al., 2017; Donahue et al., 2017). Some works focus on other tasks such as spoken language identification (Peng Shen and Kawai, 2017) and acoustic scene classification (Seongkyu Mun and Ko, 2017). More recently, some work use GAN structure to improve speech recognition (Hu et al., 2018; Mimura et al., 2017; Wang et al., 2018), and they have shown some promising results.

In this work, we propose a new data augmentation strategy by utilizing generative adversarial networks to improve the performance of noise robust speech recognition systems. The basic acoustic model we use is an advanced very deep convolutional neural network (VDCNN) (Qian et al., 2016). Based on the input feature map of VDCNN, the generated speech samples are based on spectrum feature level and produced frame by frame. Due to there are no corresponding labels for these generated data, an unsupervised learning framework is designed for acoustic modeling. Furthermore, in order to better guide the data generation, we implement conditional GAN and introduce conditional information in the data generation. The method with conditional GAN can obtain transcribed data, thus makes it possible to use the true label directly in the acoustic modeling. Two different conditions are introduced, including the acoustic state for each speech frame and the original paired clean speech for each speech frame. Our proposed data augmentation approaches are evaluated on both Aurora4 (Pearce and Picone, 2002) and AMI-SDM (Hain et al., 2012b), and the experimental results show that the system performance can be improved significantly by the proposed strategy.

The remainder of the paper is organized as follows. Section 2 briefly introduces the basic generative adversarial networks and conditional generative adversarial networks. In Section 3, the proposed data augmentation strategies with basic GAN or conditional GAN is described in detail. The acoustic modeling process with the augmented data is presented in Section 4. Section 5 shows the experimental results and analysis. Section 6 concludes the paper.

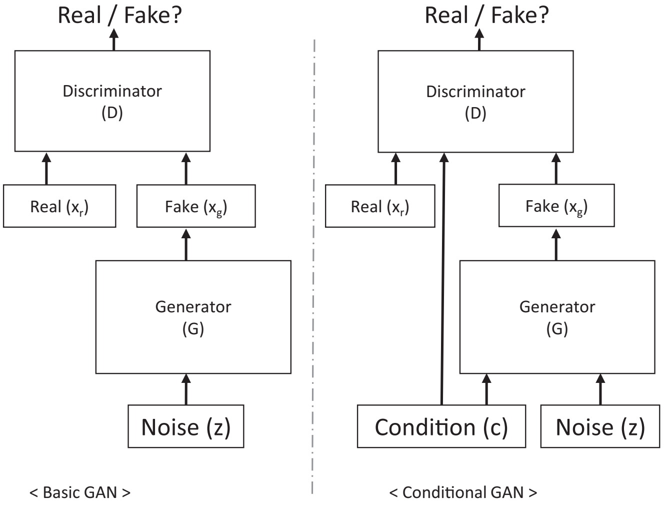


Fig. 1. The architecture of the basic GAN and conditional GAN.

2. Generative adversarial networks

2.1. GAN

Generative adversarial network (GAN) was firstly introduced in Goodfellow et al. (2014) as a powerful generative model for a wide range of applications. A basic GAN consists of two neural neural networks, as the left part in Fig. 1: a discriminator D performs classification between the real samples and fake samples; a generator G produces samples from a data distribution, which is usually a low dimensional random noise. The generator is optimized to fool the discriminator while the discriminator is trained to distinguish the fake samples from the real samples. More specifically, the game between the generator G and the discriminator D is formulated as a two-player minimax game as:

$$\min_{G}\max_{D}\mathop{\mathbf{E}}_{x\sim P_r}[log(D(x))]+\mathop{\mathbf{E}}_{z\sim P_f}[log(1-D(G(z)))] \tag{1}$$

where P_r and P_f are the real and generated fake data distributions respectively. $D(x)$ represents the probability that x comes from the real data, and z is the random noise as the input to G.

More recently, researchers proposed Wasserstein generative adversarial network (WGAN), which uses Wasserstein distance to measure the difference between real and fake distributions (Arjovsky et al., 2017; Arjovsky and Bottou, 2017). D and G are trained by the following expression:

$$\min_{G}\max_{D\in L}\mathop{\mathbf{E}}_{x\sim P_r}[D(x)]-\mathop{\mathbf{E}}_{z\sim P_f}[D(G(z))] \tag{2}$$

where L is the set of 1-Lipschitz functions introduced by WGAN to restrict the discriminator. The Wasserstein distance has the desirable property of being continuous and differentiable almost everywhere under mild assumptions. Thus, WGAN is more stable to be applied in many scenarios.

2.2. Conditional GAN

There is no condition information in original GAN structures, i.e. there is no control to guide the data generation process. Thus, conditional GAN (Mirza and Osindero, 2014; Odena et al., 2017; Isola et al., 2016) is introduced by integrating additional conditional information, which is shown as the right part in Fig. 1. With the integrated condition, conditional GAN can generate data in a desired type. The objective

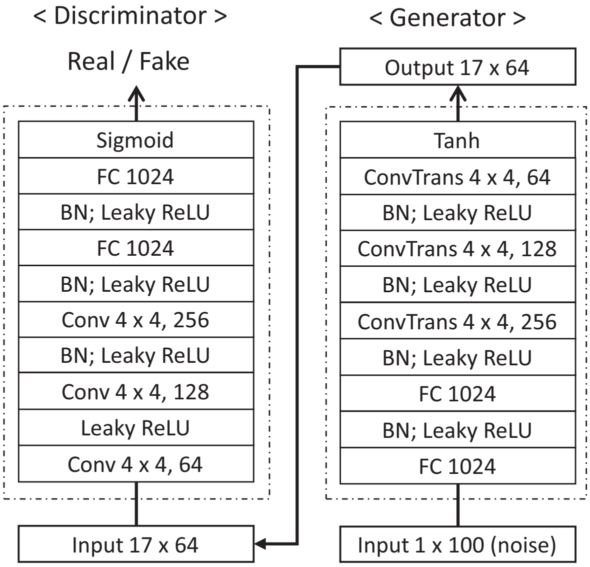


Fig. 2. The architecture of the proposed generator and discriminator with basic GAN structure. **Conv** means convolutional layer, **ConvTrans** means transposed convolutional layer, **FC** means fully connected layer, and **BN** means batch normalization. The model configuration, such as [4 × 4, 64] indicates that the layer uses a 4 × 4 filter and the output contains 64 feature maps.

function in Expression 1 is changed to:

$$\min_G \max_D \mathop{\mathbf{E}}_{x\sim P_r}[log(D(x,c))] + \mathop{\mathbf{E}}_{z\sim P_f}[log(1-D(G(z,c),c))] \tag{3}$$

where c is the condition.

3. GAN for data augmentation in speech recognition

In this work, the proposed GAN-based framework is implemented on the speech feature frame level. The basic unit we choose to generate data is the feature map on speech spectrum, such as FBANK (log Mel filter banks) feature. In our framework, when given a K-dimension FBANK feature, the context expansion is applied with N frames on each side at first. Then we can get a $(2N+1)\times K$-dimension feature map in the time-frequency domain, which is finally used as the real data input for the discriminator. In our experiments. The frame size is 25ms and the step size is 10ms. And we set $K=64$ and $N=8$ to form the 17×64 feature map. The output of generator is also a feature map with the same size, which will be utilized for acoustic modeling.

3.1. Augmentation with basic GAN

At first we implement our generative model with basic GAN structure, just as the left part of Fig. 1 shows. The input of the generator is a vector randomly sampled from a normal distribution and the output is the feature map of FBANK feature. The discriminator is trained to distinguish generated feature maps produced by the generator and real feature maps from original noisy data set. In our experiments, we adopted the loss function and configuration of WGAN framework, which is described in Expression 2.

According to some previous works on GAN, the structure configuration and training setting of GAN are very important for the model optimization. For our structure using basic GAN, the configuration is illustrated in Fig. 2. For the discriminator, there are three convolutional layers, followed with two fully connected layers to classify the real and fake data. For the generator, similar to the discriminator, there are two fully connected layers to transfer the input random noise, and then the generator uses three transposed convolutional layers to generate feature maps. After each convolutional, transposed convolutional and fully connected layer, batch normalization is adopted. The Leaky ReLU is applied in both discriminator and generator, and the negative slope is set to 0.2 in our experiments.

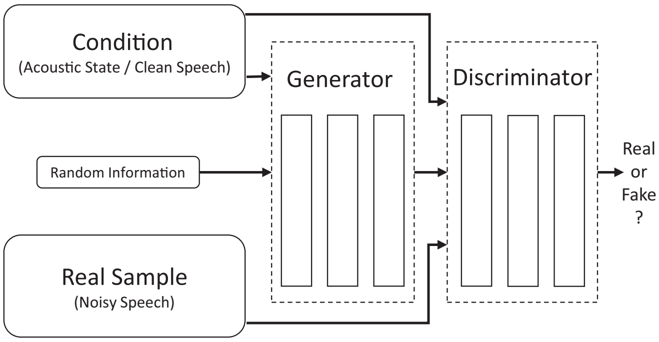


Fig. 3. The architecture of the proposed Conditional GAN.

It is noted that due to the randomness of the noise input for the generator and our frame-level data generation strategy, the labels are unknown for the generated samples (feature maps), and all the generated samples are independent from each other.

3.2. Augmentation with conditional GAN

Furthermore, in order to obtain the augmented data with labels, we introduce conditional information to basic GAN structure, in which the generated data is more specific. As the Fig. 3 illustrates, conditional information is integrated into both generator and discriminator. Random information is also added to the generator by a random vector or dropout operation in the generation process. In this work, two different conditions are introduced respectively: acoustic state and clean speech.

3.2.1. Conditioned on acoustic state

The first condition is the acoustic state for each input frame, i.e. the senone label for each frame in acoustic modeling. As illustrated in Fig. 3, conditional information is applied in both generator and discriminator. As for generator, the state information is firstly prepared with a one-hot vector, and then combined with the input noise vector to be fed into the generator. For discriminator, each dimension of this one-hot vector needs to be enlarged to the same size as feature maps (padding with 0 or 1). The extra condition maps are added on the channel dimension, i.e. extending the input feature map $17\times 64\times 1$ into $17\times 64\times(1+2787)$. All the extra maps are set zero except the label layer. To be specific, if it's the N_{th} senone, the N_{th} feature map is set with one, and the others are set with zero. And then it is stacked with the real noisy speech feature map to be fed into the discriminator. Noted that the real noisy speech feature used here belongs to the corresponding acoustic state condition, and the state information can be obtained by the training data alignment in advance. This acoustic state based conditional GAN learns to simulate the state-related real data as similar as possible. After the model training, the generator is used to generate new data by varying the state condition, and this state condition can be used as the label for that generated data.

The network structure is based on the basic GAN shown in Fig. 2. Both the generator and the discriminator have the same configuration. And the WGAN training criterion is used here to optimize the conditional GAN. With the introduction of acoustic state as conditional information, the loss function in Expression 2 is rewritten as follows:

$$\min_G \max_{D\in L} \mathop{\mathbf{E}}_{x\sim P_r}[D(x,c)] - \mathop{\mathbf{E}}_{z\sim P_f}[D(G(z,c),c)] \tag{4}$$

where c is the condition.

3.2.2. Conditioned on clean speech

The second condition is the paired clean speech feature in the training data. From the work in Isola et al. (2016) for image style transformation, instead of adding a Gaussian noise as an input to G, we add the random information in the form of dropout. In the conditional GAN model training, the parallel paired data is prepared, such as the original clean vs. manually added noisy speech or close-talk vs. far-field recorded speech. The generator takes the clean speech feature map as the input and generate corresponding noisy one. Then the generated noisy speech and real noisy speech is stacked with the original clean speech individually to be fed into the discriminator. The discriminator learns whether it is a real or fake speech pairs. Two feature maps in a certain pair share the same underlying speech pattern. Above these patterns various noise styles are presented for different pairs in the training set, corresponding to the different noisy conditions. In addition, we add the L1 term to generator's loss to encourage the respect of the input pattern:

$$\mathbf{L}_{L1} = \mathbf{E}_{x,c\sim P_r, z\sim P_f}[||x - G(z,c)||_1] \tag{5}$$

where x is the real sample from original noisy data set, and c is the condition, which is the paired sample form clean data set. The combination of extra L1 loss is designed to ensure that the generated samples retain more original acoustic information. Thus, the final training objective is

$$\min_{G} \max_{D\in L} \mathbf{L}_{cGAN} + \beta\, \mathbf{L}_{L1} \tag{6}$$

where $\mathbf{L}_{cGAN}$ is the objective function in Expression 4, and β is set to 100 in our setup.

The generator can learn the different noise styles from training noisy speech and transfer them to other clean speech. In this way, we are able to obtain new types of noisy speech which cannot be collected in the real world. After the model training, we can use a large quantity of existing transcribed clean speech as the condition, and the generated noisy speech has the same labels as the original paired clean speech.

The networks structure of conditional GAN conditioned on clean speech is based on the configuration in Isola et al. (2016). Both generator and discriminator use modules of the form convolution-BatchNorm (Ioffe and Szegedy, 2015), and we use ReLU for the generator and leakyReLU for the discriminator. As for generator, skip connections are added between each layer i and layer $n - i$, where n is the total number of layers. Each skip connection simply concatenates all channels at layer i with those at layer $n - i$. For discriminator, each input pair is transformed with a convolutional layer with different strides on each side to a regular form. Then it is forwarded to convolutional layers with batch normarization and leakyReLU operation. The final output of the discriminator will be regularized to [0,1] by a Sigmoid operation.

4. Acoustic modeling with augmented data

As for generated data with basic GAN, each output feature map of GAN is generated from a random noise vector, so it is hard for us to obtain the true labels for the generated feature maps. Thus an unsupervised learning strategy is developed to utilize these augmented data, whose idea is similar to the teacher-student learning (Li et al., 2014b; Hinton et al., 2015). Assuming that the distributions between the original data and generated data are similar from the well-trained GAN model, the augmented data from GAN can be firstly processed by the original acoustic model to collect the soft labels (the corresponding posterior probabilities). Once the soft label of each feature map is obtained, the original data can be mixed with the new generated data to train a new acoustic model. The learning framework is illustrated in Fig. 4. More specific, it consists of 4 steps.

- *Step* 1: Use the original dataset D_{orig} to train an original acoustic model A for ASR and a GAN model N for data augmentation.
- *Step* 2: Use GAN model N to generate extra dataset D_{gen}.
- *Step* 3: Use original acoustic model A to get the soft label for each augmented feature map frame by frame.

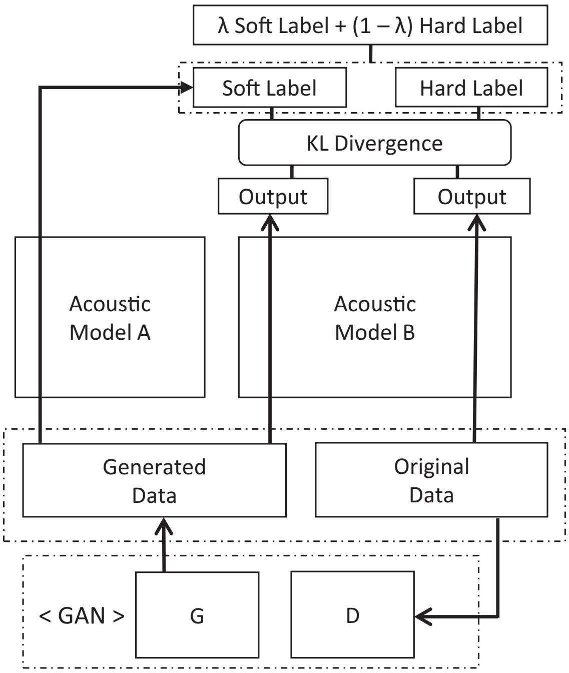


Fig. 4. The proposed unsupervised learning architecture with GAN-based data augmentation for acoustic modeling. The top part represents two acoustic modeling strategies, where inside the dashed box is for the basic GAN, outside the dashed box is for the conditional GAN.

- *Step* 4: Pool the original dataset D_{orig} with hard labels and augmented dataset D_{gen} with soft labels to train a new acoustic model B.

The KullbackLeibler (KL) divergence between the acoustic model output distribution and the related labels is used as the training criteria. In our experiments, minimizing the KL divergence is equivalent to maximizing the following expression:

$$J = \sum_{\mathbf{o}_t\in Dgen} \sum_{s} p_{gen} \log p_B(s|\mathbf{o}_t) + \sum_{\mathbf{o}_t\in Dorig} \sum_{s} p_{ref} \log p_B(s|\mathbf{o}_t) \tag{7}$$

$$p_{gen} = p_A(y|\mathbf{o}_t) \tag{8}$$

where $\mathbf{o}_t$ is the input feature and and s is the acoustic state. D_{orig} and D_{gen} are the original dataset and generated dataset respectively. p_{ref} is the reference alignment for the original transcribed data, which is the true label (also can be regarded as the hard label). The posterior distributions of the acoustic model A and B are denoted as p_{gen} and $p_B(y|\mathbf{o}_t)$, where $p_{gen} = p_A(y|\mathbf{o}_t)$ is the soft label, i.e. the posterior generated by the original acoustic model A. This approach allows us to utilize the large quantity of untranscribed augmented data more effectively.

In contrast, for the generated data with conditional GAN, true labels can be obtained directly during the generation process. Therefore, in addition to the soft labels, the hard labels also can be used for acoustic modeling. In our experiments, we find that the combination of soft labels and true labels leads to a better result, which is illustrated in the top part of Fig. 4. Thus, the Expression 8 is changed to

$$p_{gen} = \lambda\ p_A(s|\mathbf{o}_t) + (1-\lambda)\ p_{ref} \tag{9}$$

where p_{ref} is the hard label from conditional GAN directly and $p_A(s|\mathbf{o}_t)$ is the soft label generated by the original acoustic model A. λ is the hyper-parameter which is between 0 to 1. In our experiments, λ is set to 0.5 to get a consistent better position.

5. Experiments

In this section the proposed approaches are evaluated on two tasks: Aurora4 and AMI meeting transcription, which have different noisy scenarios. The reasons we choose these two data set are listed as follows.

- Aurora4 has been a benchmark task for noise-robust speech recognition for a long time (more than a decade). All the noisy data are generated artificially with additive noise and channel distortion.
- Aurora4 is based on WSJ0, which has a small vocabulary and with reading speech. To make the evaluation more realistic, AMI meeting transcription is also used, with large vocabulary and more spontaneous real speech, and the reverberation is the main challenge in this noisy task.

The evaluation on these two tasks cover not only the different noisy conditions, but also includes both the simulated and real noisy data.

5.1. Experimental setup and baseline systems

5.1.1. Data sets

- Aurora4 (Pearce and Picone, 2002) is a medium vocabulary task. Transcriptions are based on the Wall Street Journal corpus (WSJ0) (Paul and Baker, 1992). It contains 16 kHz speech data in the presence of additive noises and linear convolutional channel distortions, which were introduced synthetically to clean speech. The multi-condition training set contains 7138 utterances from 83 speakers, including clean speech and speech corrupted by one of six different noises at 10–20 dB SNR. Some utterances in the training set are from the primary Sennheiser microphone and others are from the secondary microphone. Similar to the training data, the same types of noise and microphones are used to generate the test set, grouped into 4 subsets: clean, noisy, clean with channel distortion, and noisy with channel distortion, which are referred to as A, B, C, and D, respectively.
- AMI (Hain et al., 2012b) contains around 100 hours of meetings recorded in specifically equipped instrumented meeting rooms at three sites in Europe (Edinburgh, IDIAP, TNO). The acoustic signal is captured and synchronized by multiple microphones including individual head microphones (IHM, close-talk), lapel microphones and one or more microphone arrays. For the distant speech recognition in this work, we use the single distant microphone (AMI-SDM) data paired with the individual head microphone (AMI-IHM) data to evaluate. Our experiments adopted the suggested AMI corpus partition that contains about 80 h and 8 h in the training and evaluation sets, respectively (Swietojanski et al., 2013).

5.1.2. Baseline systems and experimental setup

Gaussian mixture model based hidden Markov models (GMM-HMMs) are first built with Kaldi (Povey et al., 2011) using the standard recipes. After the GMM-HMM training, a forced-alignment is performed to get the state level labels. The standard testing pipelines in the Kaldi recipes are used for decoding and scoring. The number of states in the Aurora4 and AMI models were 2787 and 3916 respectively. The task-standard WSJ0 bi-gram language model and WSJ0 5K-word closed vocabulary are used for decoding on Aurora4. A tri-gram language model and 50K-word dictionary interpolated on the training transcripts and Fisher English transcripts were used for AMI decoding. Very deep convolutional neural network (VDCNN) is used as the acoustic model for all the experiments (Qian et al., 2016), which is a very strong acoustic model for noise robust ASR. It consists of 10 convolutional layers and 4 fully connected layers. The neural networks based acoustic models are built using CNTK (Yu et al., 2014). They were trained using cross entropy (CE) criterion with stochastic gradient descent (SGD) based back propagation (BP) algorithm. The learning rate starts at 0.1, with a momentum of 0.9. The learning rate halves when validation loss stops decreasing. More details about the acoustic models and experimental setup on acoustic modeling can be referred to Qian et al. (2016).

All the GAN / conditional GAN models for data augmentation used in this paper are implemented with PyTorch (Paszke et al., 2017). Batch normalization is used after the convolutional or transposed convolutional layers in both generator and discriminator. During the training process, the mini-batch size for all GAN structures is set to 64. As for the proposed basic GAN and conditional GAN conditioned on the acoustic state, the discriminator is updated five times then followed one time update of the generator. The networks are trained using RMSprop with the learning rate set to 0.00005 in the model optimization for both discriminator and generator. As for the conditional GAN with the clean speech, we adopt the recommended configuration in Isola et al. (2016). We alternate between one gradient descent step on discriminator and then one step on generator. Mini-batch based SGD and Adam optimizer (Kingma and Ba, 2014) are used, in which the learning rate is set to 0.0002 for both discriminator and generator. The maximum training epoch is set to 20 for the model optimization.

Table 1
WER (%) comparison of acoustic modeling with different training data on Aurora4. `original` means only using original Aurora4 multi-condition training data, `Manual` means data directly adding noise to the original speech waveform manually, `GAN` means data generated by basic GAN structure, and `cGAN-state` and `cGAN-clean` means the conditional GAN based data augmentation with acoustic state or clean speech as conditions respectively.

Systems	Data	A	B	C	D	AVG
(1)	`original`	3.62	5.81	5.12	13.77	9.02
(2)	`+Manual`	3.94	6.16	5.60	12.49	8.67
(3)	`+GAN`	3.31	5.53	4.93	13.04	8.55
(4)	`+cGAN-state`	3.05	5.57	5.19	12.44	8.30
(5)	`+cGAN-clean`	3.49	5.59	5.10	12.46	8.35
(6)	`+(4)&(5)`	3.23	5.32	4.78	12.39	8.16
(7)	`+(2)&(4)&(5)`	3.38	5.24	4.88	11.53	7.78

5.2. Evaluation on Aurora4

The baseline system using original Aurora4 data is shown as the first line of Table 1, and VDCNN based acoustic model is optimized with multi-condition training. Noted that this performance is slightly worse than our previous number in Qian et al. (2016) (9.02 vs. 8.81), since the different CNTK versions are used here.

5.2.1. Evaluation on the different data augmentation approaches

For data augmentation using basic GAN / conditional GAN, the generated data is pooled with the original Aurora4 data to build the acoustic model. For the better comparison, we also perform the data generation using the normal mode: directly adding six noise types from Aurora4 on the original clean speech waveform manually. In detail, extra noise from six noise types in training set is added to real data with channel distortion. Each utterance is added with noise with different SNR ranging from 5db to 30db with 5db stride. Thus we can obtain new noisy speech data with both additive noise and channel distortion for acoustic modeling, which has the same noise type with subset D data in testing set. For the different data augmentation methods, total 15-hour speech data is newly generated for each, which is the same size as that of the original data. The results are shown in Table 1.

The second line of results shows that the traditional data augmentation approach with a manual model can indeed get a gain, but it is more easier to obtain the biased performance on some conditions and gets degradations on the others. For the data generation using GAN, it is observed that all the newly proposed methods using GAN can obtain a significant improvement upon the strong acoustic model VDCNN, and they also outperform the manual model obviously. There is relative 6.0% ~ 8.0% WER reduction compared to the baseline system only using the original noisy training data, and the cGAN approaches are slightly better than the basic GAN. Another interesting finding is that although no noise type is assigned at the generation stage, most of the gain is from the subset D with both additive noise and channel distortion, which may due to the randomness of the generation from GAN. The generated data is more likely to be data with both additive noise and channel distortion.

Table 2
WER (%) comparison of different training data sizes generated by basic GAN on Aurora4.

Data size	A	B	C	D	AVG
`original`	3.62	5.81	5.12	13.77	9.02
15h	3.31	5.53	4.93	13.04	8.55
30h	3.36	5.60	4.99	12.91	8.53
60h	3.34	5.70	4.93	12.79	8.51
90h	3.36	5.68	5.01	12.68	8.47

This observation further demonstrates the effectiveness of the proposed GAN-based data augmentation for noise robust speech recognition.

In addition, the augmented data from different approaches is pooled together to train acoustic models, and the results are illustrated in the last two lines of Table 1. It shows that the generated data from two different conditional GAN models seems to be complementary. The combination of augmented data from two strategies can make the training data with more diversity, and achieve another additional improvement. Moreover, when further adding the data with manual mode, another obvious gain can be obtained. The best system can achieve a relative ∼ 14.0% WER reduction compared to the baseline system only using the original noisy training data.

5.2.2. *Evaluation on different augmented data sizes*

Then, the different data sizes of augmented data from the same GAN model are compared, and the results using basic GAN are illustrated in Table 2. It is observed that increasing the augmented data size from GAN indeed can gradually improve the system, but the performance difference is not very large. The performance of system using 15 h augmented data can even approaches that using 90 h augmented data (less than absolute 0.1% difference on averaged WER on Aurora4). Therefore, in our following experiments for Aurora4, the data size we generate is 15 h, which is as the same as the size of the original data.[1]

5.2.3. *Investigation on training labels*

Considering that output feature maps from basic GAN are only generated from a random noise vector, it is impossible for us to get the true labels for the generated feature maps. Thus, an unsupervised learning frame work is proposed for generated data from basic GAN, in which we utilize the purely soft labels of the generated data in acoustic models. In contrast, when using conditional GAN as generator to generate extra training data, we can obtain the true labels (hard labels) for each generated frame at the same time. Hence, these true labels can also be utilized for acoustic modeling, and the training criterion is changed from Expression 7 to 9 when using conditional GAN based data augmentation. In the experiments here, we investigate the hyper-parameter λ in Expression 9 for different kinds of training labels.

The related experimental results is shown in Table 3. The first three lines is the baseline system that only using original data for training. For a better comparison, in addition to the normal hard label, we also construct the baselines using soft labels and half soft labels for original data. From the results we can see that use purely soft labels can not lead to a better performance, and the results in all the four subsets are almost the same as using hard labels. When we combine the soft labels and hard labels, the system can get an improvement. The other lines in Table 3 shows the experiments using conditional GAN based data augmentation. It is observed that for the cGAN-based newly generated data, using purely soft labels is obviously better than using purely hard labels.

Table 3
WER (%) comparison of different training labels of generated data using conditional GAN conditioned on Aurora4. `original` means only using original Aurora4 multi-condition training data, `cGAN-state` and `cGAN-clean` means the conditional GAN based data augmentation with acoustic state or clean speech as conditions respectively. λ in Expression 9 is varied: `soft` means $\lambda = 1$, `hard` means $\lambda = 0$, and `half` means $\lambda = 0.5$.

Systems	Labels	A	B	C	D	AVG
`original`	`soft`	3.27	5.80	5.17	13.84	9.02
	`half`	3.44	5.50	5.36	13.36	8.71
	`hard`	3.62	5.81	5.12	13.77	9.02
`+cGAN-state`	`soft`	3.36	5.72	4.88	12.79	8.52
	`half`	3.05	5.57	5.19	12.44	8.30
	`hard`	3.68	5.73	5.32	14.03	9.11
`+cGAN-clean`	`soft`	3.31	5.66	5.01	12.93	8.56
	`half`	3.49	5.59	5.10	12.46	8.35
	`hard`	3.36	5.53	5.29	13.67	8.85

Table 4
WER (%) comparison of acoustic modeling with different training data on AMI-SDM. `original` means only using original AMI-SDM training data, `Manual` means generating reverberant data manually, `GAN` means data generated by basic GAN structure, and `cGAN-state` and `cGAN-clean` means the new conditional GAN based data augmentation with acoustic state or clean speech as conditions, respectively.

Data	dev	eval
`original`	49.3	54.2
`+Manual`	48.1	53.9
`+GAN`	47.7	52.6
`+cGAN-state`	46.5	51.8
`+cGAN-clean`	46.5	51.1

The combination of soft labels and hard labels will contribute an additional gain, and obtain the best system performance. This observation demonstrates that the introduction of true labels can indeed lead to an effective improvement with data augmentation using conditional GAN.

5.3. *Evaluation on AMI-SDM*

The proposed basic GAN / conditional GAN based data augmentation strategy is also evaluated on AMI-SDM, in which the reverberation is the main challenge. The baseline system is illustrated as the first line of Table 4, which only uses the original training data (AMI-SDM) for the VDCNN acoustic modeling. The experiments using different data augmentation methods are also shown and compared in Table 4.

Similar as the experiment in Aurora4, the data augmentation with manual model is also evaluated and compared in this reverberant scenario. It is noted that we only evaluated and compared to the manual mode with the reverberations due to the main challenge is the room reverberation in this meeting transcription task: at first, a shoebox model based on the image source method (Allen and Berkley, 1979) was used to simulate RIRs (Room impulse response) via RIR-generator (Habets, 2010). For every configuration, a room with random dimensions was chosen, and the reverberation time was picked randomly from 0.1 to 0.8 s. Then a microphone was randomly placed inside the room. The distance between a speaker and the microphone was picked randomly from a reasonable range as AMI corpus. After that, a speech signal was convolved with the simulated RIRs of a single room to obtain the far-field signal. In our experiments, AMI-IHM data is used as original clean data to generate the reverberation data. The data size of newly generated data is the same as the AMI-IHM data, i.e. 80 hours. The results using manually generated data is shown as the second line of Table 4. It shows that the manually generated data only with reverberations can

[1] Actually we also investigated the different sizes of manually generated data in our experiments, and the results also showed that increasing manual data size can not reduce the WER with a significant scale, which is the same as that using the GAN generated data in Table 2.

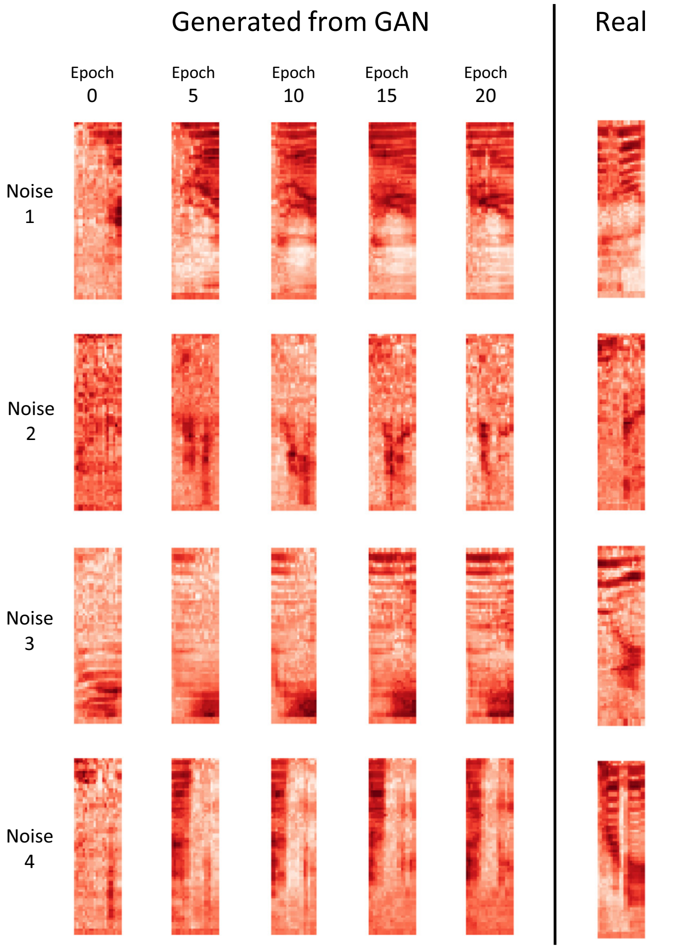
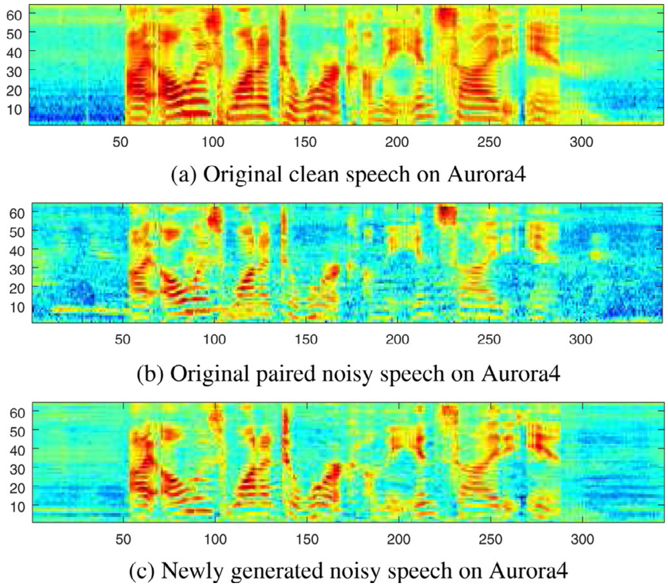


Fig. 5. Spectrum illustrations of individual frames from the basic GAN model on Aurora4.

also get a gain when compared with the baseline, but the improvement is not very large for this challenging meeting transcription task.[2]

The results using proposed GAN-based data augmentations are listed as the bottom part of Table 4, including both basic GAN and conditional GAN, and also 80 hours data is generated. For cGAN with clean speech condition, the speech pair from IHM and SDM is formed for cGAN training, and the IHM data (close-talk) is regarded as the clean speech condition. It is observed that all the GAN-based methods can improve the system performance significantly in this reverberant scenario, and the improvement is larger than that from the manual method only with reverberations. For the different GAN models, conditional GAN is consistent better than the basic GAN, as the same conclusion on Aurora4, and it demonstrates the superiority of the introduction of condition information once more.

The best system using cGAN with clean speech condition obtains a relative ~6% WER reduction on both testing sets compared to the baseline using the original AMI-SDM training data. Besides the additive noise and channel distortion scenarios as in Aurora4, the proposed GAN-based data augmentation approach can also improve the speech recognition system under the reverberant scenario.



(a) Original clean speech on Aurora4

(b) Original paired noisy speech on Aurora4

(c) Newly generated noisy speech on Aurora4

Fig. 6. Spectrum illustrations of an utterance example from Aurora4, including original clean speech, original paired noisy speech and newly generated noisy speech by cGAN with clean speech condition. The vertical axis is the FBANK channel index and the horizontal axis is the time.

5.4. *Visualization of generated data*

In order to better understand generated training samples from the proposed basic GAN or conditional GAN model, the spectrum of the augmented data is plotted and compared.

5.4.1. *Visualization of data by basic GAN*

The new data from the basic GAN is firstly illustrated. Considering that there is no dependence between the different generated samples, so we plot the spectrum for each feature map (17×64) individually. The generated data for Aurora4 is shown in Fig. 5. It shows the comparison of different feature maps generated by the GAN model on different training stages. Four random noise vectors are used for data generation and each row corresponds the feature maps generated from the same random noise vector input but on different training epochs. For a better comparison between the generated data and real speech, several real feature maps are selected from the original noisy corpus, and they are illustrated in the right part of Fig. 5.

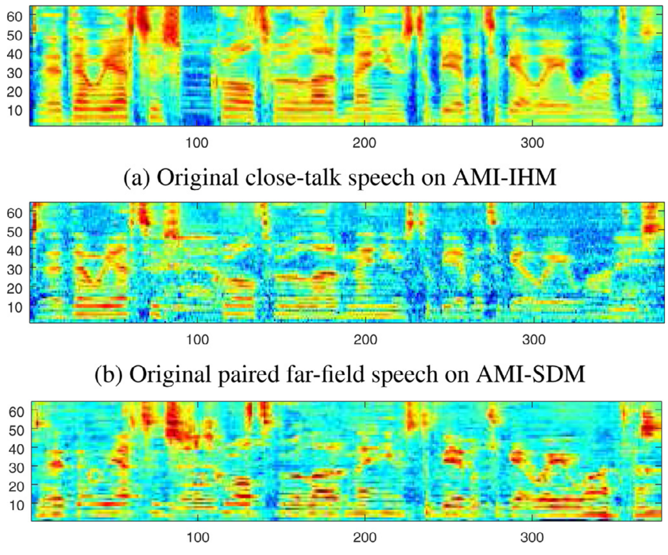
Fig. 5 shows that all the generated feature maps indeed look like the real speech spectrum very much. As the training process proceeds, the speech patterns can be observed more obviously, which means that the quality of the generated data is gradually improved with more training epochs. Doing the comparison within the samples from different noise inputs for generator, the randomness and difference is obvious significant. This property can enable us to produce noisy data with more random patterns with GAN for robust speech recognition.

5.4.2. *Visualization of data by conditional GAN*

The examples from conditional GAN are then compared. Due to the generated samples can be controlled by the condition, so we can plot the spectrum of the whole sequence. The comparisons between the original and generated speech are shown in Fig. 6 and Fig. 7 for Aurora4 and AMI-SDM, respectively, and conditional GAN with clean speech condition is used.

We can observe: (1) Most of the speech patterns related to the content are retained within the generated speech. (2) The whole spectrum of the generated utterance looks like to the corrupted one of the original noisy data, so regarding the augmented data as noisy speech is reasonable. (3) There are still many differences between the generated speech and

[2] It will be fairer if adding both reverberation and additive noise with the manual mode, and that could get a better performance as the conclusion in Ko et al. (2017) for the far-field conditions. Considering the main challenge is the reverberant condition in this task, so the manual mode only with reverberation still can show some useful conclusions.



(a) Original close-talk speech on AMI-IHM

(b) Original paired far-field speech on AMI-SDM

(c) Newly generated noisy speech on AMI-SDM

Fig. 7. Spectrum illustrations of an utterance example from AMI, including original close-talk speech, original paired far-field speech and newly generated noisy speech by cGAN with clean speech condition. The vertical axis is the FBANK channel index and the horizontal axis is the time.

the original noisy speech. It indicates that it is possible to obtain noisy speech with more diverse noises using the proposed conditional GAN, and some noise types are unseen in the original training data. All these properties of the cGAN-based augmented data can improve the system robustness.

6. Conclusion

In this paper we propose a new framework on data augmentation for noise robust speech recognition. Different from most traditional data augmentation approaches by directly adding noise to the original waveform, the generative adversarial network (GAN) is utilized. The augmented data from GAN is based on spectrum feature level and generated frame by frame (one frame corresponds one feature map). As for data augmentation with basic GAN, there is no true labels existing for them. Thus an unsupervised learning framework is designed to utilize these untranscribed data for acoustic modeling. Furthermore, conditional information is introduced to GAN to better guide the data generation. Two different conditional GANs are implemented respectively, including the acoustic state condition and the original paired clean speech condition. The combination of soft labels and hard labels can get the best performance when using conditional GAN based data augmentation.

The proposed framework is evaluated on both Aurora4 and AMI-SDM. The results show that, the proposed GAN / conditional GAN based data augmentation strategy can significantly improve the results of ASR systems, and it is also better than the traditional data augmentation with a manual mode. The conditional GAN based data augmentation approach is consistently better than the basic GAN based one. With these extra generated data by basic GAN / conditional GAN, the system robustness can be enhanced substantially and the speech recognition accuracy can be improved 6% to 14% relatively under the noisy scenarios.

Although obtaining promising results with the feed-forward neural networks based acoustic models in this paper, the current framework is still not suitable for the recurrent models or sequence training due to the independence between the frame-level generated data. The research on the RNN application and sequence training with the augmented data using GAN is another very interesting topic, and we will leave it to our future work.

Declaration of Competing Interest

The authors declared that they have no conflicts of interest to this work.

Acknowledgment

This work has been supported by the National Key Research and Development Program of China (Grant No. 2017YFB1302402) and the China NSFC projects (No. 61603252 and No. U1736202). Experiments have been carried out on the PI supercomputer at Shanghai Jiao Tong University.

References

Abdel-Hamid, O., Deng, L., Yu, D., 2013. Exploring convolutional neural network structures and optimization techniques for speech recognition. In: Proceedings of the INTERSPEECH. ISCA.

Abdel-Hamid, O., r. Mohamed, A., Jiang, H., Deng, L., Penn, G., Yu, D., 2014. Convolutional neural networks for speech recognition. IEEE/ACM Trans. Audio Speech Lang. Process. 22 (10), 1533–1545. doi:10.1109/TASLP.2014.2339736.

Abdel-Hamid, O., r. Mohamed, A., Jiang, H., Penn, G., 2012. Applying convolutional neural networks concepts to hybrid NN-HMM model for speech recognition. In: Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 4277–4280. doi:10.1109/ICASSP.2012.6288864.

Allen, J.B., Berkley, D.A., 1979. Image method for efficiently simulating small-room acoustics. Acoust. Soc. Am. J. 65, 943–950.

Arjovsky, M., Bottou, L., 2017. Towards principled methods for training generative adversarial networks. In: International Conference on Learning Representations.

Arjovsky, M., Chintala, S., Bottou, L., 2017. Wasserstein GAN. arXiv preprint arXiv:1701.07875.

Cui, X., Goel, V., Kingsbury, B., 2015. Data augmentation for deep neural network acoustic modeling. IEEE/ACM Trans. Audio Speech Lang. Process. 23 (9), 1469–1477.

Dahl, G.E., Yu, D., Deng, L., Acero, A., 2012. Context-dependent pre-trained deep neural networks for large-vocabulary speech recognition. IEEE Trans. Audio Speech Lang. Process. 20, 30–42.

Donahue, C., Li, B., Prabhavalkar, R., 2017. Exploring speech enhancement with generative adversarial networks for robust speech recognition. arXiv preprint arXiv:1711.05747.

Gong, Y., 1995. Speech recognition in noisy environments: a survey. Speech Commun. 16 (3), 261–291.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y., 2014. Generative adversarial nets. In: Advances in Neural Information Processing Systems 27, pp. 2672–2680.

Habets, E.A.P., 2010. Internal Report, 01, 2006, 1–17.

Hain, T., Burget, L., Dines, J., Garner, P.N., Grezl, F., Hannani, A.E., Huijbregts, M., Karafiat, M., Lincoln, M., Wan, V., 2012a. Transcribing meetings with the amida systems. IEEE Trans. Audio Speech Lang. Process. 20 (2), 486–498.

Hain, T., Burget, L., Dines, J., Garner, P.N., Grezl, F., Hannani, A.E., Huijbregts, M., Karafiat, M., Lincoln, M., Wan, V., 2012b. Transcribing meetings with the amida systems. IEEE Trans. Audio Speech Lang. Process. 20 (2), 486–498.

Higuchi, T., Kinoshita, K., Delcroix, M., Nakatani, T., 2017. Adversarial training for data–driven speech enhancement without parallel corpus. In: Proceedings of the ASRU, pp. 40–47.

Hinton, G., Deng, L., Yu, D., Dahl, G.E., r. Mohamed, A., Jaitly, N., Senior, A., Vanhoucke, V., Nguyen, P., Sainath, T.N., Kingsbury, B., 2012. Deep neural networks for acoustic modeling in speech recognition: the shared views of four research groups. IEEE Signal Process. Mag. 29 (6), 82–97.

Hinton, G.E., Vinyals, O., Dean, J., 2015. Distilling the knowledge in a neural network. arXiv preprint arXiv:1503.02531.

Hsu, C., Hwang, H., Wu, Y., Tsao, Y., Wang, H., 2017. Voice conversion from unaligned corpora using variational autoencoding wasserstein generative adversarial networks. arXiv preprint arXiv:1704.00849.

Hu, H., Tan, T., Qian, Y., 2018. Generative adversarial networks based data augmentation for noise robust speech recognition. In: Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP.

Ioffe, S., Szegedy, C., 2015. Batch normalization: accelerating deep network training by reducing internal covariate shift. In: Proceedings of the ICML.

Isola, P., Zhu, J.-Y., Zhou, T., Efros, A.A., 2016. Image-to-image translation with conditional adversarial networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017.

Kaneko, T., Kameoka, H., Hojo, N., Ijima, Y., Hiramatsu, K., Kashino, K., 2017a. Generative adversarial network-based postfilter for statistical parametric speech synthesis. In: Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP, pp. 4910–4914.

Kaneko, T., Kameoka, H., Hojo, N., Ijima, Y., Hiramatsu, K., Kashino, K., 2017b. Generative adversarial network-based postfilter for statistical parametric speech synthesis. In: IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP, pp. 4910–4914.

Kim, C., Misra, A., Chin, K., Hughes, T., Narayanan, A., Sainath, T.N., Bacchiani, M., 2017. Generation of large-scale simulated utterances in virtual rooms to train deep-neural networks for far-field speech recognition in Google home. In: Proceedings of the 18th Annual Conference of the International Speech Communication Association, pp. 379–383.

Kingma, D.P., Ba, J., 2014. Adam: A method for stochastic optimization. CoRR abs/1412.6980.

Ko, T., Peddinti, V., Povey, D., Khudanpur, S., 2015. Audio augmentation for speech recognition. In: Proceedings of the INTERSPEECH.

Ko, T., Peddinti, V., Povey, D., Seltzer, M.L., Khudanpur, S., 2017. A study on data augmentation of reverberant speech for robust speech recognition. In: Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP, pp. 5220–5224.

Lang, K.J., Waibel, A.H., Hinton, G.E., 1990. A time-delay neural network architecture for isolated word recognition. Neural Netw. 3 (1), 23–43. doi:10.1016/0893-6080(90)90044-L.

Li, J., Deng, L., Gong, Y., Haeb-Umbach, R., 2014a. An overview of noise-robust automatic speech recognition. IEEE/ACM Trans. Audio Speech Lang. Process. 22 (4), 745–777.

Li, J., Zhao, R., Huang, J.-T., Gong, Y., 2014b. Learning small-size DNN with output-distribution-based criteria. In: Proceedings of the INTERSPEECH.

Liao, H., 2013. Speaker adaptation of context dependent deep neural networks. In: Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, pp. 7947–7951. doi:10.1109/ICASSP.2013.6639212.

Lippmann, R., Martin, E., Paul, D., 1987. Multi-style training for robust isolated-word speech recognition. In: Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing, 12, pp. 705–708.

Mimura, M., Sakai, S., Kawahara, T., 2017. Cross-domain speech recognition using nonparallel corpora with cycle-consistent adversarial networks. In: Proceedings of the IEEE Automatic Speech Recognition and Understanding Workshop (ASRU), pp. 134–140.

Mirza, M., Osindero, S., 2014. Conditional generative adversarial nets. CoRR abs/1411.1784.

Narayanan, A., Wang, D., 2013. Ideal ratio mask estimation using deep neural networks for robust speech recognition. In: Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, pp. 7092–7096. doi:10.1109/ICASSP.2013.6639038.

Odena, A., Olah, C., Shlens, J., 2017. Conditional image synthesis with auxiliary classifier gans. In: Proceedings of the 34th International Conference on Machine Learning, ICML, pp. 2642–2651. Sydney, NSW, Australia, 6–11 August 2017.

Pascual, S., Bonafonte, A., Serrà, J., 2017. SEGAN: speech enhancement generative adversarial network. arXiv preprint arXiv:1703.09452.

Paszke, A., Gross, S., Chintala, S., 2017. Pytorch.

Paul, D.B., Baker, J.M., 1992. The design for the wall street journal-based CSR corpus. In: Proceedings of the Workshop on Speech and Natural Language. Association for Computational Linguistics, Stroudsburg, PA, USA, pp. 357–362.

Pearce, D., 2002. Aurora working group: DSR front-end LVCSR evaluation AU/384/02. Mississippi State Univ. Ph.D. dissertation.

Pearce, D., Picone, J., 2002. Aurora working group: DSR front end lvcsr evaluation AU/384/02. Tech. Rep. Inst. for Signal & Inform. Process., Mississippi State Univ..

Shen, P., Lu, X., Li, S., Kawai, H., 2017. Conditional generative adversarial nets classifier for spoken language identification. In: Proceedings of the 18th Annual Conference of the International Speech Communication Association.

Povey, D., Ghoshal, A., Boulianne, G., Burget, L., Glembek, O., Goel, N., Hannemann, M., Motlicek, P., Qian, Y., Schwarz, P., Silovsky, J., Stemmer, G., Vesely, K., 2011. The kaldi speech recognition toolkit. In: Proceedings of the IEEE Workshop on Automatic Speech Recognition and Understanding.

Qian, Y., Bi, M., Tan, T., Yu, K., 2016. Very deep convolutional neural networks for noise robust speech recognition. IEEE Trans. Audio Speech Lang. Process. 24 (12), 2263–2276.

Sainath, T.N., Vinyals, O., Senior, A.W., Sak, H., 2015a. Convolutional, long short--term memory, fully connected deep neural networks. In: Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP, pp. 4580–4584. South Brisbane, Queensland, Australia, April 19–24, 2015.

Sainath, T.N., Weiss, R.J., Senior, A.W., Wilson, K.W., Vinyals, O., 2015b. Learning the speech front-end with raw waveform CLDNNs. In: Proceedings of the INTERSPEECH.

Saito, Y., Takamichi, S., Saruwatari, H., 2017. Statistical parametric speech synthesis incorporating generative adversarial networks. IEEE Trans. Audio Speech Lang. Process. 26 (1), 84–96.

Salimans, T., Goodfellow, I.J., Zaremba, W., Cheung, V., Radford, A., Chen, X., 2016. Improved techniques for training gans. arXiv preprint arXiv:1606.03498.

Seide, F., Li, G., Chen, X., Yu, D., 2011a. Feature engineering in context-dependent deep neural networks for conversational speech transcription. In: Proceedings of the IEEE Workshop on Automatic Speech Recognition Understanding, pp. 24–29. doi:10.1109/ASRU.2011.6163899.

Seide, F., Li, G., Yu, D., 2011b. Conversational speech transcription using context-dependent deep neural networks. In: Proceedings of the INTERSPEECH.

Mun, S., Park, S., Han, D.K, Ko, H., 2017. Generative adversarial network based acoustic scene training set augmentation and selection using SVM hyper-plane. Detection and Classification of Acoustic Scenes and Events.

Shrivastava, A., Pfister, T., Tuzel, O., Susskind, J., Wang, W., Webb, R., 2016. Learning from simulated and unsupervised images through adversarial training. arXiv preprint arXiv:1612.07828.

Swietojanski, P., Ghoshal, A., Renals, S., 2013. Hybrid acoustic models for distant and multichannel large vocabulary speech recognition. In: Proceedings of the IEEE Workshop on Automatic Speech Recognition and Understanding, pp. 285–290.

Tan, T., Qian, Y., Hu, H., Zhou, Y., Ding, W., Yu, K., 2018. Adaptive very deep convolutional residual network for noise robust speech recognition. IEEE/ACM Trans. Audio Speech Lang. Process. 26 (8), 1393–1405. doi:10.1109/TASLP.2018.2825432.

Wang, Q., Rao, W., Sun, S., Xie, L., Chng, E.S., Li, H., 2018. Unsupervised domain adaptation via domain adversarial training for speaker recognition. In: Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP, pp. 4889–4893. Calgary, AB, Canada, April 15–20, 2018.

Wang, Y., Gales, M.J.F., 2012. Speaker and noise factorization for robust speech recognition. IEEE Trans. Audio Speech Lang. Process. 20 (7), 2149–2158.

Yoshioka, T., Gales, M., 2015. Environmentally robust ASR front-end for deep neural network acoustic models. Comput. Speech Lang. 31 (1), 65–86.

Yu, D., Deng, L., Droppo, J., Wu, J., Gong, Y., Acero, A., 2008. A minimum-mean-square-error noise reduction algorithm on Mel-frequency cepstra for robust speech recognition. In: Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, pp. 4041–4044.

Yu, D., Eversole, A., Seltzer, M., Yao, K., Kuchaiev, O., Zhang, Y., Seide, F., Huang, Z., Guenter, B., Wang, H., Droppo, J., Zweig, G., Rossbach, C., Gao, J., Stolcke, A., Currey, J., Slaney, M., Chen, G., Agarwal, A., Basoglu, C., Padmilac, M., Kamenev, A., Ivanov, V., Cypher, S., Parthasarathi, H., Mitra, B., Peng, B., Huang, X., 2014. An Introduction to Computational Networks and the Computational Network Toolkit. Technical report.