INTERSPEECH 2019
September 15–19, 2019, Graz, Austria

Jointly Adversarial Enhancement Training for Robust End-to-End Speech Recognition

Bin Liu[1,2], *Shuai Nie*[1], *Shan Liang*[1], *Wenju Liu*[1], *Meng Yu*[3], *Lianwu Chen*[4], *Shouye Peng*[5], *Changliang Li*[6]

[1] National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, China
[2] School of Artificial Intelligence, University of Chinese Academy of Sciences, China
[3] Tencent AI Lab, Bellevue, WA, USA
[4] Tencent AI Lab, Shenzhen, China
[5] Xueersi Online School, China
[6] kingsoft AI lab, China

{bin.liu2015,shuai.nie,sliang,lwj}@nlpr.ia.ac.cn, {raymondmyu,lianwuchen}@tencent.com
pengshouye@100tal.com, lichangliang@kingsoft.com

Abstract

Recently, the end-to-end system has made significant breakthroughs in the field of speech recognition. However, this single end-to-end architecture is not especially robust to the input variations interfered of noises and reverberations, resulting in performance degradation dramatically in reality. To alleviate this issue, the mainstream approach is to use a well-designed speech enhancement module as the front-end of ASR. However, enhancement modules would result in speech distortions and mismatches to training, which sometimes degrades the ASR performance. In this paper, we propose a jointly adversarial enhancement training to boost robustness of end-to-end systems. Specifically, we use a jointly compositional scheme of mask-based enhancement network, attention-based encoder-decoder network and discriminant network during training. The discriminator is used to distinguish between the enhanced features from enhancement network and clean features, which could guide enhancement network to output towards the realistic distribution. With the joint optimization of the recognition, enhancement and adversarial loss, the compositional scheme is expected to learn more robust representations for the recognition task automatically. Systematic experiments on AISHELL-1 show that the proposed method improves the noise robustness of end-to-end systems and achieves the relative error rate reduction of 4.6% over the multi-condition training.

Index Terms: end-to-end speech recognition, robust speech recognition, speech enhancement, generative adversarial networks

1. Introduction

Recently, end-to-end neural networks have made significant breakthroughs in the field of speech recognition [1, 2, 3], challenging the dominance of DNN-HMM hybrid architectures [4]. Attention-based encoder-decoder network integrates the acoustic and language modeling components with a single neural architecture. However, speech inputs for ASR systems are generally interfered by various background noises and reverberations in realistic environments. The single end-to-end architecture is not especially robust to the input variations and the performance drops dramatically in reality, which remains a challenge to improve the robustness of end-to-end ASR systems.

The mainstream approach to boost noise robustness is adding a speech enhancement component during the front-end of ASR, including traditional statistical methods like Wiener filter [5] and DNN-based speech enhancement methods, such as the time-frequency (T-F) masking [6, 7, 8], signal approximation [9, 10] and spectral mapping [11, 12]. However, the speech enhancement part is usually distinct from the recognition part and therefore enhancement method fails to optimize towards the final objective, which leads to a suboptimal solution [13]. Moreover, the enhancement method generally uses hand-engineering loss functions such as mean squared error, which tends to generate over-smoothed spectra that lack the fine structures that are near to those of the true speech. The speech distortions and mismatches to training sometimes degrade the end-to-end ASR performance [14].

In order to obtain an optimal performance and alleviate the speech distortions, integrating the speech enhancement and end-to-end recognition network via jointly training is proposed for robust speech recognition [14, 15]. A key concept of the joint end-to-end framework is to optimize the entire inference procedure based on the final ASR objectives, such as word/character error rate (WER/CER). In addition, generative adversarial nets (GANs) [16] have been applied to speech enhancement [17, 18] and robust ASR [19, 20], where the generator synthesizes increasingly more realistic data in attempt to fool a competing discriminator.

The end-to-end system predicts the next output symbol conditioned on the full sequence of previous predictions. If a mistake occurred in one estimation step due to the noise interference, the next prediction steps will be disrupted, which would lead to a series of mistakes. Therefore, it is critical to improve the robustness of end-to-end ASR system for the practical application.

In this paper, we propose a jointly adversarial enhancement training to boost noise robustness of end-to-end ASR systems. Specifically, we use a jointly compositional scheme of mask-based enhancement network (for the enhancement component), attention-based encoder-decoder network (for the recognition component) and discriminant network in the training phase. The discriminant network is used to distinguish between the enhanced features from enhancement network and clean features, which could guide enhancement network to output towards the

http://dx.doi.org/10.21437/Interspeech.2019-1242

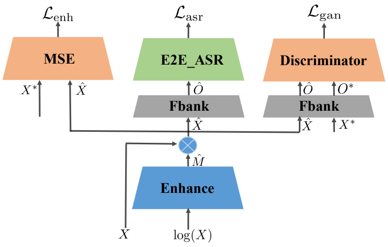

Figure 1: *Overview of the robust end-to-end ASR system architecture: Enhance module is the mask-based enhancement network; E2E_ASR module is the attention-based recognition network; Fbank module is used to extract fbank features; Discriminator is the discriminant network.*

realistic clean distribution. The use of adversarial training circumvents the limitation of hand-engineering loss functions and captures the underlying structural characteristics from the noisy signals. With the joint optimization of the recognition, enhancement and adversarial loss, the compositional scheme is expected to learn more robust representations suitable for the recognition task automatically.

2. Related Work

Generative adversarial nets (GANs) have been applied speech enhancement in the time domain [17] and frequency domain [18]. They have also been employed to improve the robustness of traditional hybrid [20] and end-to-end ASR models [19]. However, these methods don't investigate the joint training with the speech enhancement component.

The joint training framework is proposed to integrate the components of speech enhancement and recognition into a single neural-network-based architecture [14, 15]. And we propose a jointly adversarial enhancement training to boost noise robustness of end-to-end ASR systems.

3. Robust End-to-end ASR

3.1. Overview

Fig. 1 illustrates an overview of our proposed jointly adversarial enhancement training framework for robust end-to-end speech recognition (JAE E2E_ASR). The system consists of a mask-based enhancement network, an attention-based encoder-decoder network, a fbank feature extraction network and a discriminant network. Given the noisy speech input X and clean input X^*, which consists of a short-time Fourier transform (STFT) feature sequence, we represent the entire procedure of the JAE E2E_ASR system in the following forms:

$$\hat{X} = \text{Enhance}(X), \tag{1}$$

$$\hat{O} = \text{Fbank}(\hat{X}), \tag{2}$$

$$O^* = \text{Fbank}(X^*), \tag{3}$$

$$P(Y|\hat{O}) = \text{E2E\_ASR}(\hat{O}), \tag{4}$$

$$P(D|\hat{O}, O^*) = \text{Discriminate}(\hat{O}, O^*). \tag{5}$$

Here, Enhance$(\cdot)$ is a speech enhancement function realized by the mask-based enhancement network, which transforms the noisy STFT features X to the enhanced STFT features $\hat{X}$. Fbank$(\cdot)$ is a function to extract the normalized log fbank features, which converts $\hat{X}$ to $\hat{O}$. Subsequently, E2E_ASR$(\cdot)$ is an end-to-end ASR function realized by the attention-based encoder-decoder network, which estimates the posteriori probabilities for output labels Y. Moreover, Discriminate$(\cdot)$ is a discriminant network to distinguish between the enhanced features and clean ones, which gets the clean and enhanced fbank features as inputs.

3.2. Mask-based enhancement network

The mask-based enhancement method estimate a masking function to multiply by the frequency-domain feature of the noisy speech, in order to form an estimate of the clean speech.

We consider the complex short-time spectrum of the noisy speech $x_{f,t}$, the noise $n_{f,t}$, and the clean speech $x^*_{f,t}$, where t and f index time and frequency respectively. Given an estimated masking function $\hat{m}_{f,t}$, the estimated clean speech is $\hat{x}_{f,t} = \hat{m}_{f,t} x_{f,t}$. In the rest of this section, we drop f, t and consider a single time-frequency bin for simplicity.

In parallel training, the clean and noisy speech signals are provided and the signal approximation objective measures the error between the enhanced signal and the target clean speech: $\mathcal{L}_{\text{sa}}(\hat{m}) = \mathcal{L}(x^*|\hat{m}x) = |\hat{m}x - x^*|^2$. And the ideal mask is the complex filter $m^{\text{icf}} = x^*/x$. [10] proposed the phase-sensitive filter which keeps the noisy phase under the constraint $m \in \mathbb{R}$. The formulation is

$$m^{\text{psf}} = \text{Re}(\frac{x^*}{x}) = \frac{|x^*|}{|x|}\text{Re}\left(e^{i(\theta^{x^*} - \theta^{x})}\right) = \frac{|x^*|}{|x|}\cos(\theta) \tag{6}$$

where $\theta = \theta^{x^*} - \theta^x$. For training, the proposed phase-sensitive spectrum approximation (PSA) objective is $\mathcal{L}_{\text{psa}}(\hat{m}) = (\hat{m}\,|x| - |x^*|\cos(\theta))^2$. And the enhancement loss function is defined as:

$$\mathcal{L}_{\text{enh}} = \frac{1}{T}\sum_{t,f}\mathcal{L}_{\text{psa}}(\hat{m}) = \frac{1}{T}\sum_{t,f}\left(\hat{m}\,|x| - |x^*|\cos(\theta)\right)^2. \tag{7}$$

where $\hat{m}$ is the estimated mask at time t and frequency f, and T is the number of total frames in the dataset.

At the test stage, after obtaining the estimated mask from the noisy speech using the trained network, we multiply it pointwisely with the spectrogram of the noisy speech to get the enhanced spectrogram, i.e., $\hat{X} = \hat{M} \otimes X$, where $\hat{X}$ is the enhanced STFT features, $\hat{M}$ is the estimated mask, X is the noisy STFT features and $\otimes$ denotes point-wise matrix multiplication.

3.3. Fbank extraction network

We extract the normalized log Mel filterbank feature $\hat{\mathbf{o}}_t \in \mathbb{R}^{D_{\text{O}}}$ as an input of attention-based encoder-decoder, which is computed from the enhanced STFT feature $\hat{\mathbf{x}}_t \in \mathbb{R}^{D_{\text{F}}}$:

$$\hat{\mathbf{o}}_t = \text{Fbank}(\hat{\mathbf{x}}_t) = \text{Norm}(\log(\text{Mel}(\hat{\mathbf{x}}_t))) \tag{8}$$

where Mel$(\cdot)$ is the operation of $D_{\text{O}} \times F$ Mel matrix multiplication, and Norm$(\cdot)$ is the operation of global mean and variance normalization so that its mean and variance become 0 and 1. Therefore, the fbank feature extraction procedure used as a layer of network is differentiable.

3.4. Attention-based encoder-decoder network

Fig. 2 illustrates an overview of the attention-based encoder-decoder network, which consists of an encoder network that maps the input feature sequence into a higher-level representation and an attention-based decoder that predicts the next output conditioned on the full sequence of previous predictions.

Given feature sequence $O = \{\mathbf{o}_t \in \mathbb{R}^{D_{\text{O}}} \mid t = 1, \cdots, T\}$, where $\mathbf{o}_t$ is a D_{O}-dimensional fbank feature at input time step

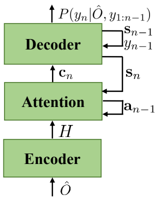

Figure 2: *Overview of the attention-based e2e system.*

t and T is the input sequence length, the network estimates the posteriori probabilities for output label sequence $Y = \{y_n \in \mathcal{V} \mid n = 1, \cdots, N\}$, where y_n is a label symbol (e.g., character) at output time step n, N is the output sequence length and $\mathcal{V}$ is a set of labels as follows:

$$P(Y|O) = \prod_n P(y_n|O, y_{1:n-1}), \tag{9}$$

$$H = \text{Encoder}(O), \tag{10}$$

$$\mathbf{c}_n = \text{Attention}(\mathbf{a}_{n-1}, \mathbf{s}_n, H), \tag{11}$$

$$y_n = \text{Decoder}(\mathbf{c}_n, y_{1:n-1}), \tag{12}$$

where $y_{1:n-1}$ is a label sequence from y_1 to y_{n-1}. Eqs. (9) to (12) correspond to E2E_ASR in Eq. (4).

For input sequence O, the encoder in Eq. (10) first transforms it to the L-length representation $H = \{\mathbf{h}_l \in \mathbb{R}^{D_\text{H}} \mid l = 1, \cdots, L\}$, where $\mathbf{h}_l$ is a D_H-dimensional state vector at time step l. Next the location-based attention mechanism [2] in Eq. (11) computes L-dimensional attention weight vector $\mathbf{a}_n \in [0, 1]^L$ to integrate all encoder outputs H into a context vector $\mathbf{c}_n \in \mathbb{R}^{D_\text{H}}$. Then the decoder in Eq. (12) estimates the posteriori probability for output label y_n at output time step n conditioned on the previous predictions $y_{1:n-1}$ and context vector $\mathbf{c}_n$. If a mistake occurred in one estimation step due to the noise interference, the next prediction steps will be disrupted, which would lead to a series of mistakes.

Based on the cross-entropy criterion, the loss function is defined using Eq. (9) as follows:

$$\mathcal{L}_\text{asr} = -\ln P(Y^*|O) = -\sum_n \ln P(y_n^*|O, y_{1:n-1}^*) \tag{13}$$

where Y^* is the ground truth of a whole sequence of output labels and $y_{1:n-1}^*$ is the ground truth from output step 1 to $n-1$.

3.5. Discriminant network

The discriminant network aims to distinguish between the enhanced features and clean ones, which could guide the enhancement network to output towards the realistic distribution. Given clean features O^* from the dataset and enhanced features $\hat{O}$ from the enhancement network, with the LSGANs [21] approach, the formulation is

$$\begin{aligned}\mathcal{L}_\text{d} = &\frac{1}{2}\mathbb{E}_{X^* \sim p_\text{data}(O^*)}\left[(\text{Discriminate}(O^*) - 1)^2\right] \\ &+ \frac{1}{2}\mathbb{E}_{X \sim p_\text{data}(\hat{O})}\left[(\text{Discriminate}(\hat{O}))^2\right],\end{aligned} \tag{14}$$

where $O^* = \text{Fbank}(X^*)$ is the clean features and $\hat{O} = \text{Fbank}(\text{Enhance}(X))$ is the enhanced features.

3.6. Training

We build a robust end-to-end speech recognition system, which converts noisy speech signals to texts with a single network. Note that all procedures, such as enhancement, feature extraction, attention-based encoder-decoder and discriminant network are implemented with neural networks and the parameters are updated by stochastic gradient descent.

Besides the phase sensitive spectrum approximation objective in Section 3.2, the enhancement network is also trained to produce outputs that cannot be distinguished from clean samples by the discriminator. In this way, the discriminator is in charge of transmitting information to enhancement network of what is real and what is fake, such that enhancement network can correct its output towards the realistic distribution. The adversarial loss, with the LSGANs approach, becomes

$$\mathcal{L}_\text{gan} = \frac{1}{2}\mathbb{E}_{X \sim p_\text{data}(\hat{O})}\left[(\text{Discriminate}(\hat{O}) - 1)^2\right]. \tag{15}$$

We alternatively train the parameters of the enhancement, recognition and discriminant network based on the jointly adversarial enhancement training. For the enhancement and recognition network, we combine three losses $\mathcal{L}_\text{asr}$, $\mathcal{L}_\text{enh}$ and $\mathcal{L}_\text{gan}$ based on Eqs. (7), (13) and (15),

$$\mathcal{L} = \mathcal{L}_\text{asr} + \alpha\mathcal{L}_\text{enh} + \beta\mathcal{L}_\text{gan}. \tag{16}$$

The magnitude of the enhancement loss and adversarial loss is controlled by hyperparameters α and β. And we train the discriminant network according to Eq. (14).

4. Experiments

4.1. Data

We systemically evaluate the proposed jointly adversarial enhancement training on an open-source Mandarin speech corpus called AISHELL-1 [22], which contains 400 speakers and over 170 hours of Mandarin speech data. Training set contains 120,098 utterances from 340 speakers; development set contains 14,326 utterance from 40 speakers and test set contains 7,176 utterances from 20 speakers.

For multi-condition training (MCT), we artificially corrupt each utterance of AISHELL-1 with background noises at SNRs randomly sampled between [0dB, +20dB]. And the background noises are from CHiME-4 corpus [23]. Apart from the “matched” noisy AISHELL-1 test set corrupted with CHiME-4 noises, we also create the “unmatched” noisy test set corrupted with NOISE-92 corpus noises [24].

4.2. Configurations

For the enhancement network, the input is the 257-dimensional logarithmic STFT features and all input vectors are normalized to have zero mean and unit variance using the training data statistics. We use three LSTM layers with 128 nodes. And a linear layer with the sigmoid activation function is connected to the last LSTM layer, whose output size is equal to the input size. The network outputs the masking estimate to multiply by the STFT of the noisy speech and forms the estimate of the clean speech.

The fbank extraction network is a linear layer to transform the STFT features to fbank features, which operates 257 × 80 matrix multiplication. After the matrix multiplication, we also do the logarithmic operation and global mean and variance normalization based on Eq. (8).

For the attention-based encoder-decoder network, we use 80-dimensional normalized log Mel filterbank features from the fbank extraction network as an input feature. We use 4-layer bidirectional LSTM with 320 cells in the encoder and 1-layer unidirectional LSTM with 320 cells in the decoder. After every BLSTM layer, a linear projection layer with 320 units is used to combine the forward and backward LSTM outputs. In the encoder, we subsample the hidden states of the first and second layers and use every second of hidden states for the subsequent layer's inputs [25]. For the location-based attention mechanism, 10 centered convolution filters of width 100 are used to extract the convolutional features and the attention inner product dimension is set as 320. We adopt a joint CTC-attention multi-task loss function [26] and set the CTC loss weight as 0.1.

The discriminant network is the 4-layer convolution network, where the num of channel is $32, 64, 128, 256$, the kernel size is 3×3 and stride is 2×2. All convolutions are followed by rectified linear unit (Relu) activation function [27]. We use PatchGANs formulation [28], which can be applied to arbitrarily-sized inputs in a fully convolutional fashion and averaging all responses to provide the final output.

For decoding, we use a beam search algorithm with the beam size 12. CTC scores are also used to re-score the hypotheses with 0.1 weight [26]. An end detection technique [26] is used to stop the beam search. We also integrate external RNN language model with 0.2 weight during decoding and the language model is trained with the training transcripts.

All the parameters are initialized with the range $[-0.1, 0.1]$ of a uniform distribution. We use the AdaDelta algorithm [29] with gradient clipping [30] for optimization and the AdaDelta hyper-parameters are initialized $\rho = 0.95$ and $\epsilon = 1\text{e}-8$. Once the performance of the validation set is degraded, we decrease the hyper-parameter ϵ by multiplying it by 0.01 at each subsequent epoch. The training procedure is stopped after 15 epochs.

4.3. Results

In the following results, we use character error rate (CER) to quantify the system performance. We report CER of the AISHELL-1 test set with three conditions. "clean" refers to the original clean AISHELL-1 test set. "matched" denotes the noisy test set corrupted with CHiME-4 background noises that are matched to the training. "unmatched" refers to the noisy test set corrupted with NOISE-92 corpus noises.

Table 1: *CER results of E2E_ASR system trained by clean data and multi-condition training (MCT) without the enhancement.*

E2E_ASR	CER Results(%)		
	clean	matched	unmatched
E2E_ASR-Clean	12.3	83.1	85.7
E2E_ASR-MCT	12.9	51.8	59.7

Firstly, we train the E2E_ASR network using the clean speech data (E2E_ASR-Clean) and multi-condition training strategy (E2E_ASR-MCT), i.e., optimization with both the clean and noisy speech. The result is shown in Table 1. The E2E_ASR-Clean network performs very poorly in the noisy test set, which demonstrates the necessity of the robust E2E_ASR investigation. The E2E_ASR-MCT significantly improves the system robustness, which outperforms E2E_ASR-Clean by 37.7% in "matched" test set and 30.3% in "unmatched".

Table 2: *CER results of the E2E_ASR system trained by the clean data and multi-condition training with the enhancement.*

E2E_ASR	CER Results(%)		
	clean	matched	unmatched
E2E_ASR-Clean	13.1	63.2	65.5
E2E_ASR-MCT	13.6	55.5	64.5

Secondly, we train the mask-based enhancement network according to Section 3.2, which converts the noisy speech into the enhanced version. Then the enhanced features are fed into the well-trained E2E_ASR-Clean and E2E_ASR-MCT network to generate the final label sequence. The result is shown in Table 2. For the E2E_ASR-Clean network, the speech enhancement component significantly improves the system robustness and outperforms E2E_ASR-Clean system without the enhancement module by 23.9% in "matched" test set and 23.6% in "unmatched", which confirms the effectiveness of combining the speech enhancement with E2E_ASR framework. However, for the E2E_ASR-MCT network, the speech enhancement degrades the system performance, consistent with observations in [18]. The enhancement and E2E_ASR network are separately trained by the different objectives and the enhanced process may introduce unseen distortions that degrades the performance.

Table 3: *CER results of the E2E_ASR system using the retraining and joint training with and without the GAN procedure.*

Model	CER Results(%)		
	clean	matched	unmatched
E2E_ASR-Retraining	12.3	51.5	59.2
Joint-Enhance-E2E_ASR	12.3	50.2	58.5
Joint-Enhance-E2E_ASR-GAN	**12.2**	**49.1**	**57.3**

Next, we retrain E2E_ASR-MCT network using enhanced features hoping to alleviate the speech distortion problem. We feed all the training data into the well-trained enhancement network and the enhanced features are used to retrain E2E_ASR network, which is referred to E2E_ASR-Retraining. The result is shown in the first row of Table 3. Compared to the E2E_ASR-MCT network, the performance improvement of the E2E_ASR-Retraining is limited.

Finally, we jointly train the enhancement and E2E_ASR network without and with the adversarial training according to Eq. (16). The joint model is initialized from the existing enhancement and E2E_ASR-MCT network checkpoint. Joint-Enhance-E2E_ASR denotes the system trained by the joint optimization of the recognition and enhancement loss. We set magnitude of the enhancement loss $\alpha = 5.0$ and adversarial loss $\beta = 0$. Joint-Enhance-E2E_ASR-GAN refers to the system with the adversarial training and magnitude of the adversarial loss is set $\beta = 2.0$. The result is shown in the last two rows of Table 3. Compared to the E2E_ASR-MCT network, Joint-Enhance-E2E_ASR improves the system performance. And Joint-Enhance-E2E_ASR-GAN improves performance further, exceeding the performance of E2E_ASR-MCT by 5.2% in "matched" test set and 4.0% in "unmatched", suggesting the potential of the adversarial training.

5. Conclusions

In this paper, we propose a jointly adversarial enhancement training to imporove robustness of end-to-end systems. We use a jointly compositional scheme of enhancement, recognition and discriminant network. The discriminator is used to distinguish between the clean and enhanced features. Experiments on AISHELL-1 demonstrate effectiveness of the proposed method. In future, we will investigate different network architectures and training strategies to obtain greater performance improvement.

6. Acknowledgements

This work was supported in part by the National Natural Science Foundation of China (no. 61573357, no. 61503382, no. 61403370, no. 61273267, no. 91120303) and Project of Xueersi Online School(no. Y9D2M10101) and sponsored by CCF-Tencent Open Research Fund.

7. References

[1] A. Graves and N. Jaitly, "Towards end-to-end speech recognition with recurrent neural networks," in *International Conference on Machine Learning*, 2014, pp. 1764–1772.

[2] J. K. Chorowski, D. Bahdanau, D. Serdyuk, K. Cho, and Y. Bengio, "Attention-based models for speech recognition," in *Advances in neural information processing systems*, 2015, pp. 577–585.

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

[4] G. Hinton, L. Deng, D. Yu, G. Dahl, A.-r. Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, B. Kingsbury *et al.*, "Deep neural networks for acoustic modeling in speech recognition," *IEEE Signal processing magazine*, vol. 29, 2012.

[5] J. S. Lim and A. V. Oppenheim, "All-pole modeling of degraded speech," *Acoustics Speech and Signal Processing IEEE Transactions on*, vol. 26, no. 3, pp. 197–210, 1978.

[6] A. Narayanan and D. Wang, "Ideal ratio mask estimation using deep neural networks for robust speech recognition," in *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, 2013, pp. 7092–7096.

[7] Y. Wang, A. Narayanan, and D. Wang, "On training targets for supervised speech separation," *IEEE/ACM transactions on audio, speech, and language processing*, vol. 22, no. 12, pp. 1849–1858, 2014.

[8] S. Nie, S. Liang, W. Xue, X. Zhang, W. Liu, L. Dong, and H. Yang, "Two-stage multi-target joint learning for monaural speech separation," in *Sixteenth Annual Conference of the International Speech Communication Association*, 2015.

[9] F. Weninger, J. R. Hershey, J. Le Roux, and B. Schuller, "Discriminatively trained recurrent neural networks for single-channel speech separation," in *2014 IEEE Global Conference on Signal and Information Processing (GlobalSIP)*. IEEE, 2014, pp. 577–581.

[10] H. Erdogan, J. R. Hershey, S. Watanabe, and J. Le Roux, "Phase-sensitive and recognition-boosted speech separation using deep recurrent neural networks," in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2015, pp. 708–712.

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

[12] S. Nie, S. Liang, W. Liu, X. Zhang, and J. Tao, "Deep learning based speech separation via nmf-style reconstructions," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 26, no. 11, pp. 2043–2055, 2018.

[13] M. L. Seltzer, "Bridging the gap: Towards a unified framework for hands-free speech recognition using microphone arrays," in *Hands-Free Speech Communication and Microphone Arrays*, 2008, pp. 104–107.

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

[15] T. Ochiai, S. Watanabe, T. Hori, and J. R. Hershey, "Multichannel end-to-end speech recognition," 2017.

[16] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in *International Conference on Neural Information Processing Systems*, 2014, pp. 2672–2680.

[17] S. Pascual, A. Bonafonte, and J. Serr, "Segan: Speech enhancement generative adversarial network," 2017.

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

[19] A. Sriram, H. Jun, Y. Gaur, and S. Satheesh, "Robust speech recognition using generative adversarial networks," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 5639–5643.

[20] B. Liu, S. Nie, Y. Zhang, D. Ke, S. Liang, and W. Liu, "Boosting noise robustness of acoustic model via deep adversarial training," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 5034–5038.

[21] X. Mao, Q. Li, H. Xie, R. Y. K. Lau, Z. Wang, and S. P. Smolley, "Least squares generative adversarial networks," 2016.

[22] H. Bu, J. Du, X. Na, B. Wu, and H. Zheng, "Aishell-1: An open-source mandarin speech corpus and a speech recognition baseline," in *2017 20th Conference of the Oriental Chapter of the International Coordinating Committee on Speech Databases and Speech I/O Systems and Assessment (O-COCOSDA)*. IEEE, 2017, pp. 1–5.

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

[24] M. T. A. P. Varga, H. J. M. Steeneken and D. Jones, "The noise -92 study on the effect of ad- ditive noise on automatic speech recognition," http://spib.rice.edu/spib/select, 1992.

[25] D. Bahdanau, J. Chorowski, D. Serdyuk, P. Brakel, and Y. Bengio, "End-to-end attention-based large vocabulary speech recognition," in *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2016, pp. 4945–4949.

[26] S. Kim, T. Hori, and S. Watanabe, "Joint ctc-attention based end-to-end speech recognition using multi-task learning," in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2017, pp. 4835–4839.

[27] V. Nair and G. E. Hinton, "Rectified linear units improve restricted boltzmann machines," in *Proceedings of the 27th international conference on machine learning (ICML-10)*, 2010, pp. 807–814.

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

[29] M. D. Zeiler, "Adadelta: an adaptive learning rate method," *arXiv preprint arXiv:1212.5701*, 2012.

[30] R. Pascanu, T. Mikolov, and Y. Bengio, "On the difficulty of training recurrent neural networks," in *International conference on machine learning*, 2013, pp. 1310–1318.