INTERSPEECH 2021
30 August – 3 September, 2021, Brno, Czechia

Data Augmentation Methods for End-to-end Speech Recognition on Distant-Talk Scenarios

Emiru Tsunoo[1], *Kentaro Shibata*[1], *Chaitanya Narisetty*[2], *Yosuke Kashiwagi*[1], *Shinji Watanabe*[2]

[1]Sony Corporation, Japan
[2]Carnegie Mellon University, USA

Emiru.Tsunoo@sony.com

Abstract

Although end-to-end automatic speech recognition (E2E ASR) has achieved great performance in tasks that have numerous paired data, it is still challenging to make E2E ASR robust against noisy and low-resource conditions. In this study, we investigated data augmentation methods for E2E ASR in distant-talk scenarios. E2E ASR models are trained on the series of CHiME challenge datasets, which are suitable tasks for studying robustness against noisy and spontaneous speech. We propose to use three augmentation methods and thier combinations: 1) data augmentation using text-to-speech (TTS) data, 2) cycle-consistent generative adversarial network (Cycle-GAN) augmentation trained to map two different audio characteristics, the one of clean speech and of noisy recordings, to match the testing condition, and 3) pseudo-label augmentation provided by the pretrained ASR module for smoothing label distributions. Experimental results using the CHiME-6/CHiME-4 datasets show that each augmentation method individually improves the accuracy on top of the conventional SpecAugment; further improvements are obtained by combining these approaches. We achieved 4.3% word error rate (WER) reduction, which was more significant than that of the SpecAugment, when we combine all three augmentations for the CHiME-6 task.

Index Terms: speech recognition, data augmentation, RNN-transducer, text-to-speech, Cycle-GAN, label smoothing

1. Introduction

End-to-end automatic speech recognition (E2E ASR), which is an integrated neural network that directly estimates output text sequences from audio features, has attracted considerable attention. E2E ASR systems have many variations, such as connectionist temporal classification (CTC) [1], attention-based encoder–decoder models [2, 3], hybrid models [4], recurrent neural network transducers (RNN-Ts) [5], Transformers [6], and Conformers [7]. These E2E approaches achieve excellent performance, particularly when they exploit a large amount of training data.

Given such insights, data augmentation is known to be effective for E2E ASR. Conventionally, speed perturbation [8] and vocal tract length perturbation (VLTP) [9] are widely used. SpecAugmentation [10], which applies on-the-fly time warping and frequency/time masking, constantly achieves significant improvements in various tasks. Because the text-to-speech (TTS) generates various speech styles, it is suitable for augmentation, and the synthesized data is used in [11, 12, 13]. For semi-supervised learning, many label augmentation techniques have been proposed, where pseudo labels are generated by a model trained with limited supervised data [14, 15].

However, in the case of distant spontaneous talk in noisy environments with low resources, the situation becomes severe. Based on the solutions to the CHiME-6 Challenge, which is a competition to solve the aforementioned problem, only conventional hybrid systems have made a meaningful attempt in this regard [16, 17]. Follow-up studies have investigated E2E approaches on tasks but have failed to outperform the hybrid systems [18, 19]. Recently Andrusenko *et al.* explored model architectures; it was concluded that RNN-T achieved comparable performance to hybrid approaches [20]. Although many augmentation methods have been applied to the task, there is still scope for further exploration.

In general, there are not enough in-domain conversational data because conversational styles have many variations in nature. Therefore, augmentation using TTS-synthesized speeches is suitable to help this regard. Further, in many cases, simulating noise environments does not match the real testing environments. Cycle-consistent generative adversarial network (Cycle-GAN) [21] is a way to map two different domains; thus this can be used to transform from clean audio characteristics to the ones in noisy environments. Conversational distant-talk speech also has a difficulty in transcription such as label errors and recording failures. Therefore, the reference label distribution is biased and does not represent conversational speech properly. Pseudo labels generated by a statistical model, such as pretrained ASR, mitigate such label errors and biases.

In this study, we investigated data augmentation methods for E2E ASR in distant-talk scenarios. CHiME-6/CHiME-4 Challenge datasets are used to study its robustness against noisy and spontaneous speech. We propose to use aforementioned three augmentation methods and their combinations. Our contributions are as follows. 1) This study is the first to apply TTS augmentation to distant-talk problems like CHiME challenges. 2) It is also the first study to use Cycle-GAN to map training data from clean audio to noisy distant-talk for augmentation. 3) We investigate the usage of pseudo labels in supervised learning scenarios for data augmentation. 4) The combination of augmentation methods on top of the SpecAugment achieved 4.3% WER reduction from only the SpecAugment in the CHiME-6 task and 1.3/0.5% reduction in the CHiME-4 simulation/real tasks.

2. Distant-Talk Scenarios

In the distant-talk scenarios, where speech is heavily contaminated by reverberation and noises, it is more challenging for E2E ASR to acquire robustness with only limited audio resources. The series of CHiME Challenge is suitable for investigating in this regard; thus we use CHiME-6/CHiME-4 Challenges in this study.

http://dx.doi.org/10.21437/Interspeech.2021-958

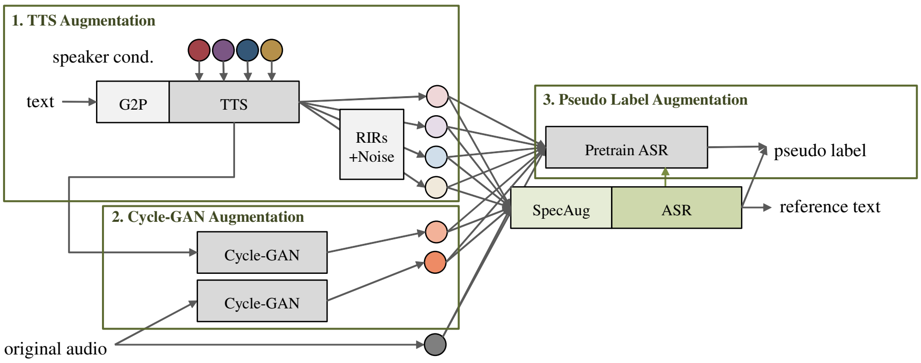

Figure 1: *Overview of augmentation methods for distant-talk ASR.*

2.1. CHiME-6 Challenge

The CHiME-6 Challenge [22] targets the problem of distant conversational ASR in everyday home environments. Conversations in twenty real dinner parties were recorded with multiple Microsoft Kinect having 4-channel microphone arrays and were fully transcribed. In addition, binaural microphones were worn by the participants. The close-talk binaural recordings play an important role for building an acoustic model as it is cleaner than the Kinect recordings. In the evaluation step, guided source separation (GSS) based speech enhancement preprocess is applied [23].

2.2. CHiME-4 Challenge

In the preceding CHiME-4 Challenge [24], WSJ prompts were recorded by 6 microphones embedded in a tablet device with 4 noisy locations, i.e., on the bus, cafe, pedestrian area and street junction. Along with the real recordings, simulation data is composed for 4 noisy environments by mixing with the original WSJ data or the booth recordings. We use the CHiME-4 Challenge data to confirm that our proposed augmentation methods are effective for different tasks and model architectures. Only isolated single-channel tracks are used for this purpose.

3. Augmentation Methods

For distant-talk low-resource scenarios, we propose to use three augmentation methods to train a robust E2E ASR model. An overview of the augmentation methods is shown in Fig. 1.

3.1. Text-to-speech data augmentation

TTS synthesized data augmentation has recently gained popularity [11, 12, 13]. It is a one-to-many mapping problem; thus it is suitable for augmenting data with a broad variation coverage. We use a Transformer TTS [25] conditioned on speaker-conditioning information, i.e., x-vector [26], and a global-style token (GST) [27]. Let the target label sequence be $\mathbf{y}$, the x-vector be $\mathbf{z}$, and the GST be $\mathbf{g}$, then the paired data of the synthesized audio signal sequence $\hat{\mathbf{s}}^{\text{tts}}$ and its label is obtained as

$$(\hat{\mathbf{s}}^{\text{tts}}, \mathbf{y}) \leftarrow \hat{\mathbf{s}}^{\text{tts}} \sim q_{\text{tts}}(\mathbf{s}|\mathbf{y}, \mathbf{z}, \mathbf{g}), \quad (1)$$

where $\mathbf{z}$ and $\mathbf{g}$ are added to the hidden output of the encoder in the Transformer.

Because the training dataset is noisy and small, it is unfeasible to train a TTS model. Therefore, we use a TTS model trained with external Librispeech dataset [28]. The reference text in the training datasets is tokenized into phonetic inputs; the waveform is directly synthesized with a given x-vector $\mathbf{z}$ and GST $\mathbf{g}$ computed using the original audio data. When the recordings are noisy, the x-vectors and GSTs also become noisy. Therefore, we use only the clean speech for the speaker/style conditions, i.e., worn binaural recordings for CHiME-6 and the original WSJ dataset for CHiME-4. Instead of shuffling speaker conditioning information, as in [22], we perturb synthesized speech with various RIRs generated by a room simulator and additive noise for the CHiME-6 setup.

3.2. Cycle-GAN data augmentation

Cycle-GAN learns two mapping functions between two domains, given the sample sets of each domain [21]. It is used in speech domain adaptation by mapping between male and female speeches [29], and is also used in other speech studies [30, 31]. According to [32], GSS-based speech enhancement applied to the training Kinect data improves the ASR performance significantly because it aligns training more to the testing environment. However, GSS-based speech enhancement can only be applied to multichannel data, which does not include TTS-synthesized data. Therefore, instead, we use Cycle-GAN to map the clean training data to the speech-enhanced noisy testing speeches. An advantage of using Cycle-GAN is that these two domain datasets do not need to be paired datasets. In real situations, the paired data does not always exist; paired binaural data is a peculiar case for CHiME-6 Challenge. In addition, the TTS synthesized speech is not strictly paired data to the target speech because the duration may vary or synthesis error may occur, which increases the difficulty of regression-based training.

Let the clean speech or TTS synthesized speech be S, and the target noisy observations be X. Thus, Cycle-GAN trains two mapping functions $G : S \to X$ and $F : X \to S$ with two adversarial discriminators D_S and D_X, where D_S distinguishes between clean speech S and translated speech $F(X)$ and so forth. We further extend it to the multi-discriminator Cycle-GAN by following [29]. The spectra $\mathbf{S}$ and $\mathbf{X}$ are computed from $\mathbf{s}$ and $\mathbf{x}$ with short-time Fourier transform (STFT) before divided into m and n frequency bands. Subsequently, $D_S^{f_i}$ and $D_X^{f_i}$ are applied to each subband f_i as the discriminators.

Generally, adversarial loss is computed as follows.

$$\begin{aligned}\mathcal{L}_{\text{gan}}(G, D_X^{f_i \in n}, S, X) =& \mathbb{E}\left[\sum_i^n \log D_X^{f_i}(\mathbf{X})\right] \\ &+ \mathbb{E}\left[\sum_i^n \log(1 - D_X^{f_i}(G(\mathbf{S})))\right] \quad (2)\end{aligned}$$

In addition, cycle consistency loss is defined as

$$\begin{aligned}\mathcal{L}_{\mathrm{cyc}}(G,F) =& \mathbb{E}\left[||F(G(\mathbf{S}))-\mathbf{S}||_1\right] \\ &+ \mathbb{E}\left[||G(F(\mathbf{X}))-\mathbf{X}||_1\right], \end{aligned} \tag{3}$$

where $||\cdot||_1$ is L1 loss. Then, the total loss is optimized in combination with a tunable parameter λ.

$$\begin{aligned}\mathcal{L}(G,F,D_S^{f_i\in m},D_X^{f_i\in n}) =& \mathcal{L}_{\mathrm{gan}}(G,D_X^{f_i\in n},S,X) \\ &+ \mathcal{L}_{\mathrm{gan}}(F,D_S^{f_i\in m},X,S) \\ &+ \lambda\mathcal{L}_{\mathrm{cyc}}(G,F) \end{aligned} \tag{4}$$

Thus, by using only G and the reference label $\mathbf{y}$, we obtain paired data by applying inverse-STFT as follows.

$$(\hat{\mathbf{x}}^{\mathsf{cgan}},\mathbf{y}) \leftarrow \hat{\mathbf{x}}^{\mathsf{cgan}} = \mathrm{ISTFT}(G(\mathbf{S})) \tag{5}$$

3.3. Pseudo-label augmentation

Although pseudo-label augmentation is widely used in semi-supervised learning [14, 15], we believe it is also effective for supervised setups where the labels are noisy because of conversational speech containing many ambiguities and disfluency. The framework can also be considered as knowledge distillation [33], except that we can use the same model size for both teachers and students. Beacuse of low resource data, the reference labels are biased; thus pseudo labels generated by statistical models such as neural networks would help interpolate its distribution.

Pseudo labels are estimated by a pretrained ASR model, as

$$(\mathbf{x},\hat{\mathbf{y}}^{\mathsf{pl}}) \leftarrow \hat{\mathbf{y}}^{\mathsf{pl}} = \arg\max_{\mathbf{y}} p_{\mathrm{pre}}(\mathbf{y}|\mathbf{x}). \tag{6}$$

Although a rigorous knowledge distillation for RNN-T was proposed in [34], we use the pseudo-label sequence directly as a reference sequence because it is efficient and sufficiently effective.

Because distant-talk transcription is a challenging task, the pseudo label may contain a noticeable number of errors. Therefore, the pseudo labels were filtered based on the character error rates (CERs). Given the CER threshold δ, pseudo labels not greater than δ are kept in $\hat{Y}_{\delta}^{\mathsf{pl}}$. We investigate the effectiveness of the filtering in Section 4.1.3.

3.4. Combination of the augmentation methods

The aforementioned augmentation methods can be used in combinations. With all the augmented paired data from (1), (5), and (6), the combined loss is computed as

$$\begin{aligned}\mathcal{L} =& -\sum_{\mathbf{x},\mathbf{y}\in(X,Y)}\log p(\mathbf{y}|\mathbf{x}) - \sum_{\hat{\mathbf{s}}^{\mathsf{tts}},\mathbf{y}\in(\hat{S}^{\mathsf{tts}},Y)}\log p(\mathbf{y}|\hat{\mathbf{s}}^{\mathsf{tts}}) \\ &- \sum_{\hat{\mathbf{x}}^{\mathsf{cgan}},\mathbf{y}\in(\hat{X}^{\mathsf{cgan}},Y)}\log p(\mathbf{y}|\hat{\mathbf{x}}^{\mathsf{cgan}}) - \sum_{\mathbf{x},\hat{\mathbf{y}}^{\mathsf{pl}}\in(X,\hat{Y}_{\delta}^{\mathsf{pl}})}\log p(\hat{\mathbf{y}}^{\mathsf{pl}}|\mathbf{x}). \end{aligned} \tag{7}$$

In the case of RNN-T, $p(\mathbf{y}|\mathbf{x})$ is a forward–backward probability [5]. We further investigate every possible combinations in the following section.

4. Experiments

4.1. CHiME-6 Challenge

4.1.1. Experimental setup

We carried out experiments on the CHiME-6 Challenge dataset. It consisted of 44 h of worn binaural data as well as multiple Kinect recordings. The dataset was prepared following the baseline system in [22], which included the perturbation using RIRs generated by a room simulator and the speed perturbation with a factor of $\{0.9, 1.0, 1.1\}$. The baseline training set was in total 1400 h. The development and evaluation sets were two dinner party sessions respectively. We applied GSS-based enhancement [23], following the baseline setup, as in [22].

We extracted log-Mel filterbanks using the Kaldi toolkit. The output units were characters (26 alphabets and 21 auxiliary symbols). As a baseline, RNN-T was trained following [20]. SpecAugment [10] was applied to the training set. All models were trained with 10 epochs, and last 5 models were averaged for regularization. External word-level LSTM language model (LM) was also trained with CHiME-6 training text corpus. The model consisted of a one-layer uni-directional LSTM with 500 units. When the LM was used, decoding was done with the shallow fusion with a weight of 0.1, and a beam size of 10. The training was carried out using ESPNet [35].

Transformer TTS was trained with the Librispeech corpus [28] for augmentation conditioned on an x-vector predicted by a TDNN trained with Voxceleb [36] and VoxCeleb2 [37]. The reference text was tokenized into pronunciation using the CMU dictionary and fed into the TTS module, which directly predicted the speech signal sequence. The synthesized speech was further perturbed with RIRs as in Section 3.1. Using both synthesized speech and original CHiME-6 training data, the GMM-HMM was trained, which was used for cleanup based on the decoded scores. The aforementioned speed perturbation was then applied, which ended up augmenting the data to 2070 h.

Two Cycle-GANs were trained with two pairs of domains: binaural worn recordings and GSS applied enhanced training Kinect set, and TTS synthesized speech and the GSS training set. Residual networks were trained following [21], except the fact that input and output features were spectrograms normalized with the mean and variance. We set $m = n = 3$ in (4). The Cycle-GANs followed by the speed perturbation added merely 123 h to the original set, but when it was combined with TTS augmentation the data increased up to 2366 h.

The baseline RNN-T model was also used for generating pseudo labels. All generated labels were evaluated once, and values that were not greater than $\delta = 50$ of the CER were used for augmentation. This nearly doubled the dataset from 1400 h to 2630 h. When all the proposed augmentation methods were combined, the dataset reached 3988 h in total. The data sizes are summarized in the last column in Table 1.

4.1.2. ASR results in the CHiME-6 task

We first evaluate the effectiveness of each proposed augmentation method. For comparison, other E2E ASR architectures are listed as well as the baseline RNN-T model in Table 1. Although, Andrusenko *et al.* reported in [20] that LM fusion degraded its performance, we observed that the external LM improved the WERs consistently.

With each proposed augmentation on top of the SpecAugment, the WER was respectively reduced comparing to only the SpecAugment (54.4% in eval set). Pseudo label augmentation significantly improved the WER (53.0%), which indicated that

Table 1: *WERs of RNN-T E2E models in the CHiME-6 task*

	w/o LM		w/ LM		
	dev	eval	dev	eval	hrs
TDNN-F hybrid [22]			51.8	51.3	1400
Joint CTC/Attention E2E [18]			82.1	71.8	
CNN Multi-ch. E2E [19]			80.7	-	
RNN-T E2E (reprod.) [20]	59.5	58.2	58.7	56.7	1400
+ SpecAugment (reprod.) [20]	55.8	55.7	55.4	54.4	
++ TTS Aug.	54.4	53.9	53.6	52.8	2070
++ Cycle-GAN Aug.	54.2	52.6	53.8	52.0	1523
++ Label Aug.	55.1	54.4	54.6	53.0	2630
++ TTS + Cycle-GAN Aug.	51.5	50.7	51.1	50.5	2366
++ Cycle-GAN + Label Aug.	53.7	52.1	53.0	51.7	2834
++ TTS + Label Aug.	53.1	53.3	52.1	51.7	3630
++ All combined	**50.9**	**50.5**	**50.4**	**50.1**	3988
+ SpecAugment + *gss_train* [20]	53.6	51.4	53.7	51.0	1437
++ All combined	**49.6**	**48.7**	**49.5**	**48.6**	4025

it is not only effective for unsupervised learning but also supervised learning with noisy label scenarios. Among three methods, Cycle-GAN augmentation archived the best result (52.0%).

Further, we investigated combinations of the three augmentation methods. The combination of TTS and Cycle-GAN augmentation significantly dropped the error rates, and when all three were combined, we achieved the best WER (50.1%). Our proposed augmentation methods reduced the WER on eval set by 4.3%, which was more significant than the WER reduction of the SpecAugment (2.3%). Improvements were also seen when we applied our proposed methods to TDNN-F hybrid model[1].

According to [32], train set enhanced by GSS improved the recognition accuracy. Therefore, we also included *train_gss* set in training and the word error rate (WER) of dev set was reduced by 2.2% absolute without LM, which almost matched the report in [20] (2.4% reduction). On top of that, the combination of our proposed augmentation methods successfully reduced the WERs further.

4.1.3. *Effectiveness of pseudo label filtering*

We also investigated the effectiveness of filtering for pseudo-label augmentation. The threshold δ was sampled from $\{20, 50, 70, \infty\}$, and the results were compared with the one without augmentation. The results are shown in Table 2. $\delta = 50$ achieves the best performance in the dev set; $\delta = 70$ was the best in the eval set.

4.2. CHiME-4 Challenge

4.2.1. *Experimental setup*

We conducted experiments on the CHiME-4 Challenge to confirm that our proposed augmentation methods were also effective for other datasets and model architectures. As the baseline, the Conformer [7] model was trained following [38]. SpecAugment was also applied to the task. Each training set was trained for 100 epochs and 10 models with the best validation accuracies were averaged for regularization. We also trained an external word-level one-layer LSTM LM with 1000 units using CHiME-4 training text corpus. The LM was fused with a weight of 1.0 and the beam size was fixed to 6.

For augmentation, we used the same Librispeech TTS model to generate synthesized speeches as in the CHiME-6 task.

Table 2: *Pseudo label augmentation with various filtering criteria.*

Filtering criteria	dev	eval	hrs
No pseudo label	55.4	54.4	1400
≤20 CER	55.0	53.5	2376
≤50 CER	**54.6**	53.0	2630
≤70 CER	54.9	**52.7**	2691
All pseudo label	55.2	53.5	2799

Table 3: *WERs of Conformer E2E models for the CHiME-4 task*

	dev		eval		
	sim	real	sim	real	hrs
Conformer [38]	13.6	12.0	21.5	22.1	190
+ SpecAugment [38]	11.0	9.3	17.4	16.2	
++ TTS Aug.	10.5	8.7	17.3	15.8	253
++ Cycle-GAN Aug.	**10.2**	8.5	17.1	**15.5**	271
++ Label Aug.	10.7	9.1	17.2	16.4	202
++ TTS + Cycle-GAN Aug.	10.4	**8.4**	16.3	15.7	334
++ All combined	10.4	8.7	**16.1**	15.7	346

Cycle-GAN for CHiME-4 Challenge was trained using the original WSJ dataset and the real isolated recording for the development set. Pseudo labels were generated using the baseline Conformer model. We found that approximately 90% of the pseudo labels did not contain errors, i.e., $\mathrm{CER} = 0$. Therefore, we excluded these labels and used only the remains. We applied filtering using $\delta = 2$, because our preliminary experiments provided the best results in the development set. The duration of training data is summarized in the last column in Table 3.

4.2.2. *ASR results in the CHiME-4 task*

First, we evaluated individual augmentation methods. The results are summarized in Table 3. All the proposed methods improved WERs in both simulation and real dataset. Pseudo-label augmentation was not as significant as the other methods, which indicated that the reference label contained few errors and was not as biased as the CHiME-6 dataset.

Subsequently, therefore, we combined TTS and Cycle-GAN augmentations, which provided the best WER in real development set. Finally, the combination of all augmentations provided a slight improvement in the simulation evaluation set.

5. Conclusion

We investigated three data augmentation methods for E2E ASR in the distant-talk scenarios. TTS-synthesized speech was used for data augmentation which was further perturbed by applying RIRs and additive noise. Cycle-GAN was also used to augment domain-matched speech by mapping from cleand recordings or TTS speech to the target noisy recordings. Finally, pseudo-label augmentation was proposed for the supervised scenarios to smoothen the label distribution, where the labels were also noisy. Experiments on the CHiME-6 and CHiME-4 tasks indicated that our proposed augmentation methods were effective in the RNN-T/Conformer E2E ASR models. Further, combinations of those improved its performance, particularly when all the methods were combined in the CHiME-6 setup.

Future work includes on-the-fly augmentation of TTS and pseudo-labels, by introducing consistency losses, and shuffling speaker information. Pseudo labels can also be used in a KL-divergence style as in knowledge distillation studies.

[1] We obtained marginal improvements in TTS/Label augmentation (1.0%/1.6% WER reductions) while Cycle-GAN degraded by 1.1%. Combinations were not always promising in TDNN-F. We leave it as our future investigation.

6. References

[1] A. Graves, S. Fernández, F. Gomez, and J. Schmidhuber, "Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks," in *Proc. of 23rd International Conference on Machine Learning*, 2006, pp. 369–376.

[2] J. K. Chorowski, D. Bahdanau, D. Serdyuk, K. Cho, and Y. Bengio, "Attention-based models for speech recognition," in *Proc. of NIPS*, 2015, pp. 577–585.

[3] W. Chan, N. Jaitly, Q. Le, and O. Vinyals, "Listen, attend and spell: A neural network for large vocabulary conversational speech recognition," in *Proc. of ICASSP*, 2016, pp. 4960–4964.

[4] S. Watanabe, T. Hori, S. Kim, J. R. Hershey, and T. Hayashi, "Hybrid CTC/attention architecture for end-to-end speech recognition," *Journal of Selected Topics in Signal Processing*, vol. 11, no. 8, pp. 1240–1253, 2017.

[5] A. Graves, A.-R. Mohamed, and G. Hinton, "Speech recognition with deep recurrent neural networks," in *Proc. of ICASSP*, 2013, pp. 6645–6649.

[6] S. Karita, N. Chen, T. Hayashi, T. Hori, H. Inaguma, Z. Jiang, M. Someki, N. E. Y. Soplin, R. Yamamoto, X. Wang *et al.*, "A comparative study on transformer vs RNN in speech applications," in *Proc. of ASRU Workshop*, 2019, pp. 449–456.

[7] A. Gulati, J. Qin, C.-C. Chiu, N. Parmar, Y. Zhang, J. Yu, W. Han, S. Wang, Z. Zhang, Y. Wu *et al.*, "Conformer: Convolution-augmented transformer for speech recognition," in *Proc. of Interspeech*, 2020, pp. 5036–5040.

[8] T. Ko, V. Peddinti, D. Povey, and S. Khudanpur, "Audio augmentation for speech recognition," in *Sixteenth Annual Conference of the International Speech Communication Association*, 2015.

[9] N. Jaitly and G. E. Hinton, "Vocal tract length perturbation (VTLP) improves speech recognition," in *Proc. ICML Workshop on Deep Learning for Audio, Speech and Language*, vol. 117, 2013.

[10] D. S. Park, W. Chan, Y. Zhang, C.-C. Chiu, B. Zoph, E. D. Cubuk, and Q. V. Le, "SpecAugment: A simple data augmentation method for automatic speech recognition," in *Proc. of Interspeech*, 2019.

[11] T. Hayashi, S. Watanabe, Y. Zhang, T. Toda, T. Hori, R. Astudillo, and K. Takeda, "Back-translation-style data augmentation for end-to-end ASR," in *2018 IEEE Spoken Language Technology Workshop (SLT)*, 2018, pp. 426–433.

[12] A. Tjandra, S. Sakti, and S. Nakamura, "End-to-end feedback loss in speech chain framework via straight-through estimator," in *Proc. of ICASSP*, 2019, pp. 6281–6285.

[13] G. Wang, A. Rosenberg, Z. Chen, Y. Zhang, B. Ramabhadran, Y. Wu, and P. Moreno, "Improving speech recognition using consistent predictions on synthesized speech," in *Proc. of ICASSP*, 2020, pp. 7029–7033.

[14] T. Kemp and A. Waibel, "Unsupervised training of a speech recognizer: Recent experiments," in *Sixth European Conference on Speech Communication and Technology*, 1999.

[15] Y. Chen, W. Wang, and C. Wang, "Semi-supervised ASR by end-to-end self-training," in *Proc. of Interspeech*, 2020, pp. 2787–2791.

[16] J. Du, Y.-H. Tu, L. Sun, L. Chai, X. Tang, M.-K. He, F. Ma, J. Pan, J.-Q. Gao, D. Liu *et al.*, "The USTC-NELSLIP systems for CHiME-6 challenge," in *Proc. of CHiME-6 Workshop*, 2020.

[17] H. Chen, P. Zhang, Q. Shi, and Z. Liu, "The IOA systems for CHiME-6 challenge," in *Proc. of CHiME-6 Workshop*, 2020.

[18] S. Dalmia, S. Kim, and F. Metze, "Situation informed end-to-end ASR for noisy environments," in *Proc. of CHiME-5 Workshop*, 2018.

[19] N. Yalta, S. Watanabe, T. Hori, K. Nakadai, and T. Ogata, "CNN-based multichannel end-to-end speech recognition for everyday home environments," in *2019 27th European Signal Processing Conference (EUSIPCO)*, 2019, pp. 1–5.

[20] A. Andrusenko, A. Laptev, and I. Medennikov, "Towards a competitive end-to-end speech recognition for CHiME-6 dinner party transcription," in *Proc. of Interspeech*, 2020, pp. 319–323.

[21] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 2223–2232.

[22] S. Watanabe, M. Mandel, J. Barker, and E. Vincent, "CHiME-6 challenge: Tackling multispeaker speech recognition for unsegmented recordings," in *Proc. of CHiME-6 Workshop*, 2020.

[23] C. Boeddeker, J. Heitkaemper, J. Schmalenstroeer, L. Drude, J. Heymann, and R. Haeb-Umbach, "Front-end processing for the CHiME-5 dinner party scenario," in *Proc. of CHiME-5 Workshop*, 2018.

[24] 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 & Language*, vol. 46, pp. 535–557, 2017.

[25] N. Li, S. Liu, Y. Liu, S. Zhao, and M. Liu, "Neural speech synthesis with transformer network," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 33, no. 01, 2019, pp. 6706–6713.

[26] D. Snyder, D. Garcia-Romero, G. Sell, D. Povey, and S. Khudanpur, "X-vectors: Robust dnn embeddings for speaker recognition," in *Proc. of ICASSP*. IEEE, 2018, pp. 5329–5333.

[27] Y. Wang, D. Stanton, Y. Zhang, R.-S. Ryan, E. Battenberg, J. Shor, Y. Xiao, Y. Jia, F. Ren, and R. A. Saurous, "Style tokens: Unsupervised style modeling, control and transfer in end-to-end speech synthesis," in *International Conference on Machine Learning*. PMLR, 2018, pp. 5180–5189.

[28] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "LibriSpeech: an ASR corpus based on public domain audio books," in *Proc. of ICASSP*, 2015, pp. 5206–5210.

[29] E. Hosseini-Asl, Y. Zhou, C. Xiong, and R. Socher, "A multi-discriminator CycleGAN for unsupervised non-parallel speech domain adaptation," in *Proc. Interspeech*, 2018, pp. 3758–3762.

[30] T. Kaneko and H. Kameoka, "Cyclegan-VC: Non-parallel voice conversion using cycle-consistent adversarial networks," in *2018 26th European Signal Processing Conference (EUSIPCO)*, 2018, pp. 2100–2104.

[31] P. S. Nidadavolu, J. Villalba, and N. Dehak, "Cycle-GANs for domain adaptation of acoustic features for speaker recognition," in *Proc. of ICASSP*, 2019, pp. 6206–6210.

[32] C. Zorila, M. Li, D. Hayakawa, M. Liu, N. Ding, and R. Doddipatla, "Toshiba's speech recognition system for the chime 2020 challenge," in *Proc. of CHiME-6 Workshop*, 2020.

[33] Y. Kim and A. M. Rush, "Sequence-level knowledge distillation," in *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, 2016, pp. 1317–1327.

[34] S. Panchapagesan, D. S. Park, C.-C. Chiu, Y. Shangguan, Q. Liang, and A. Gruenstein, "Efficient knowledge distillation for RNN-transducer models," *arXiv preprint arXiv:2011.06110*, 2020.

[35] S. Watanabe, T. Hori, S. Karita, T. Hayashi, J. Nishitoba, Y. Unno, N. E. Y. Soplin, J. Heymann, M. Wiesner, N. Chen *et al.*, "ESPnet: End-to-end speech processing toolkit," in *Proc. of Interspeech*, 2019, pp. 2207–2211.

[36] A. Nagrani, J. S. Chung, and A. Zisserman, "Voxceleb: a large-scale speaker identification dataset," in *Proc. of Interspeech*, 2017, pp. 2616–2620.

[37] J. S. Chung, A. Nagrani, and A. Zisserman, "Voxceleb2: Deep speaker recognition," in *Proc. of Interspeech*, 2018, pp. 1086–1090.

[38] P. Guo, F. Boyer, X. Chang, T. Hayashi, Y. Higuchi, H. Inaguma, N. Kamo, C. Li, D. Garcia-Romero, J. Shi *et al.*, "Recent developments on espnet toolkit boosted by conformer," *arXiv preprint arXiv:2010.13956*, 2020.