INTERSPEECH 2020
October 25–29, 2020, Shanghai, China

Improving Speech Recognition using GAN-based Speech Synthesis and Contrastive Unspoken Text Selection

Zhehuai Chen[1], *Andrew Rosenberg*[1], *Yu Zhang*[1], *Gary Wang*[2,*]
Bhuvana Ramabhadran[1], *Pedro J. Moreno*[1]

[1]Google
[2]Simon Fraser University
{zhehuai,rosenberg,ngyuzh,bhuv,pedro}@google.com, ywa289@sfu.ca

Abstract

Text-to-Speech synthesis (TTS) based data augmentation is a relatively new mechanism for utilizing text-only data to improve automatic speech recognition (ASR) training without parameter or inference architecture changes. However, efforts to train speech recognition systems on synthesized utterances suffer from limited acoustic diversity of TTS outputs. Additionally, the text-only corpus is always much larger than the transcribed speech corpus by several orders of magnitude, which makes speech synthesis of all the text data impractical. In this work, we propose to combine generative adversarial network (GAN) and multi-style training (MTR) to increase acoustic diversity in the synthesized data. We also present a contrastive language model-based data selection technique to improve the efficiency of learning from unspoken text. We demonstrate that our proposed method allows ASR models to learn from synthesis of large-scale unspoken text sources and achieves a 35% relative WER reduction on a voice-search task.
Index Terms: Speech Synthesis, Speech Recognition, Generative Adversarial Network, Contrastive Data Selection

1. Introduction

End-to-End (E2E) automatic speech recognition (ASR) has become popular as a result of recent advances in neural modeling of context and history in sequences [1, 2]. Nevertheless, it needs a larger amount of transcribed speech data to perform well [3]. Several fusion-based approaches to incorporating a traditional language model (LM) into E2E ASR frameworks were introduced in [4, 5]. However, incorporating these traditional LMs and the corresponding decoding techniques into E2E systems [6] can be difficult, while complicating the inference framework [5]. The recently introduced Hybrid Autoregressive Transducer (HAT) model offers one approach to evaluate the value of an external, traditional language model in the E2E framework but concludes that further in-depth analysis is needed before a complete E2E model can be built [7].

Speech synthesis (TTS) based data augmentation paves the way for utilizing text-only data in a novel manner to improve the ASR model. As state-of-the-art speech synthesis can be indistinguishable from human speech [8, 9], it can be utilized to synthesize text-only data which can subsequently serve as additional training data for ASR [10–13]. Such approaches do not require ASR parameter or inference architecture changes [13, 14]. However, efforts to train speech recognition systems on synthesized utterances suffer from limited acoustic diversity of TTS data. Synthesized speech exhibits much less variation than real speech, and almost no speech disfluencies. Previous work shows that ASR models trained mostly on synthesized data are hard to generalize to real speech utterances [15]. Additionally, prior work has limited the use of synthetic corpora to sizes similar to the transcribed speech corpus [12–14]. Encoding speech signals needs more memory than encoding text, while the conversion between speech and text modalities in TTS inference and ASR training adds to the computational load. These factors make the synthesis of all available text impractical. Moreover, the large differences between spoken and written material requires intelligent strategies to balance their contributions [15].

In this work, we propose to combine generative adversarial networks (GANs) [16] and multistyle training (MTR) [17] to increase the acoustic diversity of synthesized data. Previous work has applied GANs in TTS to achieve high quality and fidelity [18]. The motivation behind this work is to drive TTS audio closer to the acoustics seen in ASR training corpora under more challenging acoustic environments. Therefore, to train robust ASR models, we inject noise in a tiled fashion and apply SpecAugment [19] to the synthesized audio. Next, to realize the potential of large-scale text corpora, we introduce contrastive language models to select data from unspoken text for synthesis. Lastly, we include a diverse, on-the-fly realization of each selected sentence during ASR training for efficiency. We demonstrate how this significantly improves the efficiency of unspoken text learning through TTS.

2. Speech Synthesis for Speech Recognition

TTS based data augmentation is a successful, novel approach to utilizing large-scale, text-only data to improve ASR. Such approaches add a TTS module during training but do not require changes to the ASR model architecture. State-of-the-art synthesized speech can be indistinguishable from human speech in quality. Nevertheless, synthesized speech exhibits much less variation than real speech. This prevents ASR models trained on it from generalizing well to real scenarios.

2.1. Speech synthesis Model

To synthesize a diverse set of speaker and noise characteristics, we base our TTS model on Tacotron 2D [8, 20], which takes text sequences as input, conditioned on speaker embeddings and outputs a sequence of Mel spectrogram frames. The autoregressive decoder network uses the phoneme encoding of the input sequence and combines it with a speaker embedding obtained from a separately trained speaker encoder [20]. We directly generate Mel-filter bank features as input for training ASR models. We never synthesize waveforms, thereby eliminating the need for any vocoder. To model prosody and increase its vari-

[*] Work performed during an internship at Google.

http://dx.doi.org/10.21437/Interspeech.2020-1475

ability during inference, we further augment the model with a variational auto encoder (VAE) as in [21]. We modify its global VAE to a hierarchical version [12]. This helps in capturing local and global speaking styles separately and make TTS more stable. The hierarchical VAE includes a local encoder which encodes fixed two-second chunks with a one-second overlap and a global encoder encodes the whole utterance.

2.2. Consistent predictions on synthesized speech

A consistency loss term [13] to encourage the ASR model to generate consistent predictions on both real and synthesized presentations of the same utterance is included in the training objective. This is done by minimizing divergence between predictions based on real speech and TTS speech, $\mathcal{J}_{\text{cons}}(\theta)$. By behaving similarly in response to real and synthetic input, the model learns from optimization on synthetic training data with greater impact on real evaluation data. In the case of transcribed material, $\mathbb{I}$, we use two cross-entropy supervised loss terms from real speech, $\mathcal{J}_{real}(\theta)$ and TTS speech, $\mathcal{J}_{tts}(\theta)$, along with the consistency loss for transcribed data. For unspoken text, $\mathbb{N}$, we use cross-entropy based supervised loss terms from TTS speech only. Our overall training objective is, thus:

$$\min_{\theta} \mathcal{J}_{ASR}(\theta) = \lambda_r \mathcal{J}^{(\mathbb{I})}_{real}(\theta) + \lambda_{t_{\mathbb{I}}} \mathcal{J}^{(\mathbb{I})}_{tts}(\theta) + \lambda_c \mathcal{J}^{(\mathbb{I})}_{\text{cons}}(\theta) + \lambda_{t_{\mathbb{N}}} \mathcal{J}^{(\mathbb{N})}_{tts}(\theta) \quad (1)$$

3. Proposed method

3.1. Generative adversarial network (GAN) based TTS

It is challenging to train a TTS model using ASR training corpora [12] because of speaker biasing and adverse acoustic environments. However, if the TTS model is trained on clean TTS corpora, the distinct acoustics inferred by it is not diverse. Therefore, we finetune a well-trained TTS model from Section 2.1 to synthesize features resembling that of an ASR training corpus and use this model to synthesize unspoken text. The motivation of applying a GAN to finetune the TTS model is to synthesize audio with similar acoustics as the ASR training corpus under adverse acoustic environments. On the contrary, previous work have applied GANs in TTS to achieve high quality and fidelity [18, 22].

The framework of the GAN based TTS model jointly trained with the E2E ASR model is shown in Figure 1. As the post-processing network (PostNet) of the Tacotron model [8] can see the full decoded sequence, it is reasonable to inject background noises at this stage to meet our motivation of matching acoustics of the ASR corpus. Therefore, we update only the Postnet. The generated TTS feature $G(\mathbf{y})$ is fed into the discriminator $D(\cdot)$. The discriminator distinguishes between ASR features $\mathbf{x}$ and the TTS variants. However, for unspoken text, only synthesized features are available. Therefore, we randomly pick a real feature $\hat{\mathbf{x}}$ derived from the speech corpus and feed it to the discriminator.

As mentioned earlier, a consistency loss term helps with the intelligibility of the synthesized speech. We also synthesize the same unspoken text using a reference E2E TTS model $G_{ref}(\cdot)$ with the same architecture. We calculate the mean squared error (MSE) loss between the spectra from this model and the GAN-based TTS model, while back-propagating gradients only through the GAN-based TTS model. This loss serves as a regularizer and ensures that the TTS model is not driven too far by GAN training. The final criterion using unspoken text $\mathbf{y}$

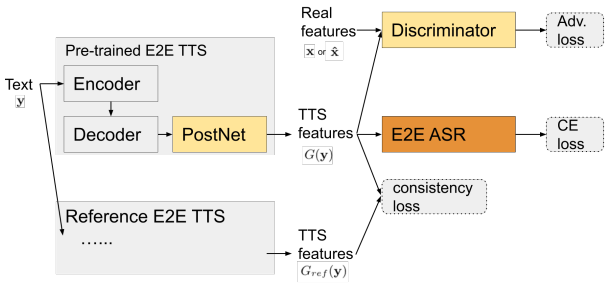

Figure 1: *Generative adversarial network (GAN) based TTS joint training framework. The colored boxes are trained while other parts are fixed.*

and randomly selected feature $\hat{\mathbf{x}}$ is as follows.

$$\min_{\theta} \mathcal{J}_{TTS}(\theta) = E_{(\hat{\mathbf{x}},\mathbf{y})} \min_{G} \max_{D} [\log(1 - D(G(\mathbf{y}))) + \log D(\hat{\mathbf{x}})] + \lambda_{\text{cons}} \min_{G} \|G(\mathbf{y}) - G_{ref}(\mathbf{y})\| \quad (2)$$

where the first two terms represent GAN training and the last term is the consistency loss. We use $\lambda_{\text{cons}} = 0.1$ here.

3.2. Contrastive unspoken text selection

Although the text-only corpus is always much larger than the transcribed speech corpus, most of the previous work limit the size of synthetic corpora be similar to the transcribed speech corpora. There are several challenges in synthesizing large-scale text corpora: i) Encoding speech requires a lot more than text [1], thereby rendering converting all text to speech impractical. ii) The computational costs associated with the joint training of ASR and TTS are quite high. State-of-the-art ASR models are generally trained using several thousand hours of speech [3, 23], which in iteself is a relatively small fraction of synthesized speech from all available text corpora. iii) Intelligent strategies are needed to balance the contributions of written and spoken material [15].

In this section, we propose techniques for text selection that will help realize the potential of large-scale text corpora as a source for augmenting ASR training corpora. The selection aims to improve the match between the selected subset and the desired ASR application. This, in turn, reduces the computational resources needed to benefit from the availability of a large amount of non-domain-specific data. Similar data selection methods were first proposed in language modeling [24, 25] with the intent of improving performance rather than reducing the computational overhead.

The proposed algorithm is as follows. We build two language models [25, 26] to contrastively select from unspoken text: a background model $\mathbb{B}$, trained on the entire unspoken text corpus, and an in-domain model $\mathbb{D}$, obtained by interpolating the background model with an LM trained on the transcriptions from the ASR corpora. We assume that the transcribed ASR material matches the domain of interest. We evaluate each sentence in the unspoken text corpus using the following equation:

$$\mathcal{S} = \frac{\log P(\mathbf{w}|\mathbb{D}) - \log P(\mathbf{w}|\mathbb{B})}{\#(\mathbf{w})} \quad (3)$$

where, the probabilities from the two language models are compared, normalized by the number of words to eliminate any

[1] for example, 16KHz 960-hour Librispeech corpus takes 100GB while its ASCII transcription only takes 100MB.

length bias. We select the sentences with the top $\mathcal{S}$ scores, thereby, selecting sentences that are relatively close to the domain of interest $\mathbb{D}$. We use N-Gram LMs for both the background and adapted LMs [2].

An alternative approach to the above unspoken text selection is the sampling of data from a well-trained language model [12]. Here, we train a large maximum-entropy language model using all the available text corpora and adapt it towards the spoken domain [27]. Assuming that the model learns the distribution of the entire data, we sample a fixed number of sentences from this language model.

3.3. Noise Injection

To regularize ASR training on synthetic speech and prevent over-fitting to the synthesized spectra, we apply SpecAugment [19] and multistyle training (MTR) [17] on the synthesized data. Notably, this is different from [14] in that SpecAugment is applied not only on real speech but also on the TTS audios. We further inject different varieties of environmental noises to TTS audios with random signal-to-noise ratios as a form of MTR [28]. The noise snippets used here, originate from a collection of several real-life noises detailed in [28]. The snippet is additively tiled along the time dimension of the synthesized audio. MTR aims to make the acoustics of synthetic speech closer to the adverse characteristics commonly seen in usage of ASR systems [28], while SpecAugment has proven to be a strong regularization approach to train robust E2E ASR models.

3.4. Putting it all together

This section presents the framework to generate on-the-fly, acoustically diverse realizations of the sentences selected using the methods proposed in Section 3.2. Within each batch, the framework randomly mixes transcribed speech, selected unspoken text, speaker embedding to use in the Tacotron 2D TTS model (Section 2.1) and the MTR noise sources. For each utterance in the transcribed speech corpora, we infer the VAE latent variable given the acoustic feature. We randomly assign a VAE latent variable for each unspoken text sentence from within this set. Next, we randomly shuffle speaker embeddings, VAE latent variables and MTR noise sources within each training batch. The shuffling provides a diverse combination of speaker characteristics, prosody and background noise with which to synthesize each sentence. The GAN TTS and ASR model are trained jointly with the training batches constructed on-the-fly in the manner described above.

The size of text-only corpora are several orders of magnitude larger than the size of transcribed speech corpora. Even with contrastive data selection, balancing these two sources is critical to train ASR Models. We use a curriculum training [29] strategy to balance the contribution of written and spoken source material. The proposed curriculum training begins with a large portion of transcribed speech data in each batch and gradually reduces it during training. While decreasing the amount of transcribed speech in each batch, we increase the weights λ_r and λ_c in Equation (1) to keep their loss contributions the same. We find that this strategy has the added benefit of speeding up model convergence, a key factor to consider when incorporating large quantities of synthesized material.

[2] We experimented with training N-Gram language models and neural network-based language models (NNLM) for both the background and adapted LMs. Both types of LMs yielded similar results.

4. Related Work

Incorporating text through the use of decoding or rescoring methodologies in E2E ASR [5,6,30,31] can result in a complex inference framework. TTS based data augmentation allows E2E models to utilize text-only data efficiently, without any changes to the decoding or inference architecture. SpeechChain [10, 11] while utilizing TTS as an data augmentation method for ASR, allows for the joint training of both models. Cycle consistent loss was proposed to bridge the representation mismatch between TTS and ASR systems [32,33]. In [34,35], an alternative approach to address this mismatch involves sharing encoders and decoders between TTS and ASR.

The following highlights the differences between prior work and the proposed method in this paper. Our motivation for applying GANs in TTS is to synthesize audio representing adverse acoustic environments similar to what is seen during training by ASR models. Our method can also be viewed as using GANs to augment the synthesized data similar to [36]. With this motivation, we only update the PostNet of the TTS model and jointly train it with the ASR model. We also introduce a consistency loss to maintain intelligibility of the synthesized data. Additionally, we combine GAN TTS training with traditional data augmentation methods like multistyle training (MTR) [17] and SpecAugment [19]. Notably, this is different from [14] in that SpecAugment is applied on not only real speech but also on the synthesized features to regularize the ASR training on synthetic speech.

To the best of our knowledge, this is the first work that introduces text data selection for TTS based data augmentation. As explained in Section 3.2, while the data selection method itself is similar to what has been previously proposed in language modeling, its application to reduce memory and computational load in the joint training of TTS and ASR is novel.

5. Experiments

5.1. Experimental Setup

The ASR model presented in this paper is a listen-attend-spell (LAS) [37] model which includes two convolutional layers of 32 filters with shape 3 x 1 and a 2 x 2 stride, followed by four bidirectional LSTM layers of 1024 units for each direction [13, 37]. The decoder contains two unidirectional LSTM layers with 1024 units and graphemic targets. The architecture of the TTS model is similar to the model described in [8, 13] with the addition of hierarchical VAE [21] discussed in Section 2.1. The decoder is followed by a PostNet with five convolutional layers of 512 filters with shape 5 x 1. The discriminator in GAN training also comprises of five convolutional layers followed by a softmax.

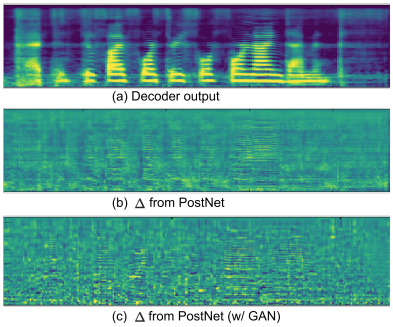

Figure 2: *Sample TTS output w/ and w/o GAN training. Comparing (b) and (c) shows the injected noise from PostNet.*

The TTS model is described in [12, 13, 38] and was trained on the segmented Librispeech corpus. The speaker embeddings are learnt from that. The ASR model is trained on an internal corpus of isolated sentences, presented in [12], containing 76 hours of voice search queries. The external text corpus for data selection comprises of anonymized and aggregated typed search query data. The background LM for data selection was trained on this external text corpus, while the adapted LM was trained on the transcribed speech used for training ASR models.

5.2. Results

The baseline model is trained only on transcribed speech with SpecAugment resulting in a WER of 18.5% (Row 1 in Table 1). When data augmentation with consistency loss [13] is applied between two SpecAugment copies of transcribed speech, a stronger baseline with a WER of 18.2% is obtained (Row 2 in Table 1). Adding synthesized data derived from the transcribed speech corpus in conjunction with SpecAugment applied to the synthesized data results in further improvement, resulting in a WER of 17.1% (Row 3 in Table 1). Next, we randomly select 30M text sentences and integrate the proposed curriculum training routine (Section 3.4). The result is shown in the fourth row with a reduction in WER to 16.2%, consistent with the work presented in [12]. We attribute this relative improvement of 5% relative only to the fact that synthesized speech exhibits much less variation than real speech, causing the ASR model to overfit to mismatched acoustics. To obtain further improvements from external text and overcome the overfitting issue, we first integrate SpecAugment & MTR based data augmentation schemes described in Section 3.4. As can be seen in the fifth row of Table 1, we see a reduction in WER to 13.3%, a reduction of 22% relative in WER. The next three rows present the additional improvements obtained using the proposed GAN TTS and contrastive data selection methods. Collectively, the proposed methods reduce the WER to 11.0%. Overall, we achieve a relative 35% improvement in performance over a strong baseline. Figure 2 illustrates an example of synthesized spectra after GAN training. We can see that TTS introduces different noise patterns compared to traditional augmentation schemes such as SpecAugment & MTR. We believe that this diversity is one major source of improvement.

Table 1: *Performance of the Proposed Method*

Method	Unspoken Text	WER
SpecAug	×	18.5
+ Consistency loss	×	18.2
+ TTS (w/ SpecAug)	×	17.1
+ TTS (w/o SpecAug)	√	16.2
+ SpecAug & MTR	√	13.3
+ GAN TTS	√	12.2
+ Data selection	√	11.5
+ GAN TTS	√	11.0

5.3. Ablation Study

Table 2 summarizes our efforts to increase acoustic diversity of TTS. Both SpecAugment and MTR obtain additive improvements, consistent to previous data augmentation research [19], while GAN TTS obtains further improvement of relative 8.3%. The difference between data augmentation methods and the proposed GAN TTS is that the former focuses on increasing acoustic diversity by noise injection while the latter tries to match acoustics to that of the transcribed speech corpus under adverse environments. We hypothesize that GAN TTS can help in cases where target acoustics cannot be simulated by SpecAugment and MTR.

Table 2: *Noise Injection on the TTS Output*

Method on Unspoken Text	WER
TTS (w/o SpecAug)	16.2
+ SpecAug	14.0
+ MTR	13.3
+ GAN TTS	12.2

Table 3 shows the effect of the amount of unspoken text based TTS data on ASR performance, where more data always yields better results. We believe this improved result, compared to the findings in [12], can be attributed to the increased acoustic diversity and on-the-fly learning framework proposed in Section 3.4

Table 3: *The Amount of Unspoken Text v.s. WER*

# of Unpaired Text Utt.	WER
0	17.1
1M	14.7
5M	13.7
30M	13.3

To understand the impact of data selection, Table 4 compares the proposed contrastive selection method with the data sampling method [12] discussed in Section 3.2. It can be seen that data selection obtains larger gains, which not only indicates that the selection metric is effective but also shows the advantage of using real data versus data generated from sampling methods.

Table 4: *Unspoken Text Selection (30M utterances) v.s. WER*

# of Unpaired Text Utt.	WER
Random	13.3
Sampling from LM	12.5
Contrastive selection	11.5

6. Conclusions

We have shown that generative adversarial network (GAN) and multistyle training (MTR) are complimentary in increasing acoustic diversity of synthesized data. Contrastive language model based data selection combined with curriculum training is key to learning from synthesis of large volumes of text. Together, these two proposed strategies yield a large reduction in WER of 35% relative over a state-of-the-art baseline on the described train and test sets.

7. References

[1] H. Sak, A. Senior, and F. Beaufays, "Long short-term memory recurrent neural network architectures for large scale acoustic

modeling," in *Fifteenth Annual Conference of the International Speech Communication Association*, 2014.

[2] 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," *arXiv preprint arXiv:1909.06317*, 2019.

[3] H. Soltau, H. Liao, and H. Sak, "Neural speech recognizer: Acoustic-to-word lstm model for large vocabulary speech recognition," *arXiv preprint arXiv:1610.09975*, 2016.

[4] 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.

[5] S. Toshniwal, A. Kannan, C.-C. Chiu, Y. Wu, T. N. Sainath, and K. Livescu, "A comparison of techniques for language model integration in encoder-decoder speech recognition," in *2018 IEEE Spoken Language Technology Workshop (SLT)*. IEEE, 2018, pp. 369–375.

[6] T. Hori, S. Watanabe, and J. R. Hershey, "Multi-level language modeling and decoding for open vocabulary end-to-end speech recognition," in *Automatic Speech Recognition and Understanding Workshop (ASRU), 2017 IEEE*. IEEE, 2017, pp. 287–293.

[7] E. Variani, D. Rybach, C. Allauzen, and M. Riley, "Hybrid autoregressive transducer (hat)," *arXiv preprint arXiv:2003.07705*, 2020.

[8] Y. Wang, R. Skerry-Ryan, D. Stanton, Y. Wu, R. J. Weiss, N. Jaitly, Z. Yang, Y. Xiao, Z. Chen, S. Bengio *et al.*, "Tacotron: Towards end-to-end speech synthesis," *arXiv preprint arXiv:1703.10135*, 2017.

[9] S. Ö. Arik, M. Chrzanowski, A. Coates, G. Diamos, A. Gibiansky, Y. Kang, X. Li, J. Miller, A. Ng, J. Raiman *et al.*, "Deep voice: Real-time neural text-to-speech," in *ICML-Volume 70*. JMLR. org, 2017, pp. 195–204.

[10] A. Tjandra, S. Sakti, and S. Nakamura, "Listening while speaking: Speech chain by deep learning," in *ASRU*. IEEE, 2017, pp. 301–308.

[11] S. Nakayama, A. Tjandra, S. Sakti, and S. Nakamura, "Speech chain for semi-supervised learning of japanese-english code-switching asr and tts," in *SLT*. IEEE, 2018, pp. 182–189.

[12] A. Rosenberg, Y. Zhang, B. Ramabhadran, Y. Jia, P. Moreno, Y. Wu, and Z. Wu, "Speech recognition with augmented synthesized speech," in *IEEE ASRU*, 2019.

[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 *ICASSP*. IEEE, 2020.

[14] N. Rossenbach, A. Zeyer, R. Schlüter, and H. Ney, "Generating synthetic audio data for attention-based speech recognition systems," *arXiv preprint arXiv:1912.09257*, 2019.

[15] J. Li, R. Gadde, B. Ginsburg, and V. Lavrukhin, "Training neural speech recognition systems with synthetic speech augmentation," *arXiv preprint arXiv:1811.00707*, 2018.

[16] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in *Advances in neural information processing systems*, 2014, pp. 2672–2680.

[17] L. Deng, A. Acero, M. Plumpe, and X. Huang, "Large-vocabulary speech recognition under adverse acoustic environments," in *Sixth International Conference on Spoken Language Processing*, 2000.

[18] C. Donahue, J. McAuley, and M. Puckette, "Adversarial audio synthesis," *arXiv preprint arXiv:1802.04208*, 2018.

[19] D. S. Park, Y. Zhang, C.-C. Chiu, Y. Chen, B. Li, W. Chan, Q. V. Le, and Y. Wu, "Specaugment on large scale datasets," *arXiv preprint arXiv:1912.05533*, 2019.

[20] Y. Jia, Y. Zhang, R. Weiss, Q. Wang, J. Shen, F. Ren, P. Nguyen, R. Pang, I. L. Moreno, Y. Wu *et al.*, "Transfer learning from speaker verification to multispeaker text-to-speech synthesis," in *Advances in neural information processing systems*, 2018, pp. 4480–4490.

[21] W.-N. Hsu, Y. Zhang, R. J. Weiss, H. Zen, Y. Wu, Y. Wang, Y. Cao, Y. Jia, Z. Chen, J. Shen *et al.*, "Hierarchical generative modeling for controllable speech synthesis," *arXiv preprint arXiv:1810.07217*, 2018.

[22] H. Guo, F. K. Soong, L. He, and L. Xie, "A new gan-based end-to-end tts training algorithm," *arXiv preprint arXiv:1904.04775*, 2019.

[23] T. N. Sainath, Y. He, B. Li, A. Narayanan, R. Pang, A. Bruguier, S.-y. Chang, W. Li, R. Alvarez, Z. Chen *et al.*, "A streaming on-device end-to-end model surpassing server-side conventional model quality and latency," *arXiv preprint arXiv:2003.12710*, 2020.

[24] K. Yasuda, R. Zhang, H. Yamamoto, and E. Sumita, "Method of selecting training data to build a compact and efficient translation model," in *Proceedings of the Third International Joint Conference on Natural Language Processing: Volume-II*, 2008.

[25] R. C. Moore and W. Lewis, "Intelligent selection of language model training data," in *Proceedings of the ACL 2010 conference short papers*. Association for Computational Linguistics, 2010, pp. 220–224.

[26] W. Wang, B. Liang, M. Hughes, T. Watanabe, T. Nakagawa, and A. Rudnick, "Contrastive sequence-to-sequence data selector," Nov. 14 2019, uS Patent App. 16/376,254.

[27] F. Biadsy, M. Ghodsi, and D. Caseiro, "Effectively building tera scale maxent language models incorporating non-linguistic signals," in *Interspeech*, 2017.

[28] A. Narayanan, A. Misra, K. C. Sim, G. Pundak, A. Tripathi, M. Elfeky, P. Haghani, T. Strohman, and M. Bacchiani, "Toward domain-invariant speech recognition via large scale training," in *2018 IEEE Spoken Language Technology Workshop (SLT)*. IEEE, 2018, pp. 441–447.

[29] Y. Bengio, J. Louradour, R. Collobert, and J. Weston, "Curriculum learning," in *Proceedings of the 26th annual international conference on machine learning*, 2009, pp. 41–48.

[30] K. Hu, T. N. Sainath, R. Pang, and R. Prabhavalkar, "Deliberation model based two-pass end-to-end speech recognition," *arXiv preprint arXiv:2003.07962*, 2020.

[31] Z. Chen, M. Jain, Y. Wang, M. L. Seltzer, and C. Fuegen, "End-to-end contextual speech recognition using class language models and a token passing decoder," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 6186–6190.

[32] T. Hori, R. Astudillo, T. Hayashi, Y. Zhang, S. Watanabe, and J. Le Roux, "Cycle-consistency training for end-to-end speech recognition," in *ICASSP*. IEEE, 2019, pp. 6271–6275.

[33] M. K. Baskar, S. Watanabe, R. Astudillo, T. Hori, L. Burget, and J. Černockỳ, "Self-supervised sequence-to-sequence asr using unpaired speech and text," in *Interspeech*, 2019.

[34] Y. Ren, X. Tan, T. Qin, S. Zhao, Z. Zhao, and T.-Y. Liu, "Almost unsupervised text to speech and automatic speech recognition," *arXiv preprint arXiv:1905.06791*, 2019.

[35] S. Karita, S. Watanabe, T. Iwata, M. Delcroix, A. Ogawa, and T. Nakatani, "Semi-supervised end-to-end speech recognition using text-to-speech and autoencoders," in *ICASSP*. IEEE, 2019, pp. 6166–6170.

[36] A. Antoniou, A. Storkey, and H. Edwards, "Data augmentation generative adversarial networks," *arXiv preprint arXiv:1711.04340*, 2017.

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

[38] H. Zen, V. Dang, R. Clark, Y. Zhang, R. J. Weiss, Y. Jia, Z. Chen, and Y. Wu, "Libritts: A corpus derived from librispeech for text-to-speech," *ArXiv*, vol. abs/1904.02882, 2019.