Proceedings Paper:
Do, C.-T., Imai, S., Doddipatla, R. et al. (2024) Improving accented speech recognition using data augmentation based on unsupervised text-to-speech synthesis. In: 2024 32nd European Signal Processing Conference (EUSIPCO). 2024 32nd European Signal Processing Conference (EUSIPCO), 26-30 Aug 2024, Lyon, France. Institute of Electrical and Electronics Engineers (IEEE), pp. 136-140. ISBN: 9798331519773. ISSN: 2219-5491. EISSN: 2076-1465.

https://doi.org/10.23919/eusipco63174.2024.10715166

Improving Accented Speech Recognition using Data Augmentation based on Unsupervised Text-to-Speech Synthesis

Cong-Thanh Do
Toshiba Research Europe
Cambridge, UK
cong-thanh.do@toshiba.eu

Shuhei Imai
Tohoku University
Sendai, Japan
shuhei.imai@tohoku.ac.jp

Rama Doddipatla
Toshiba Research Europe
Cambridge, UK
rama.doddipatla@toshiba.eu

Thomas Hain
University of Sheffield
Sheffield, UK
t.hain@sheffield.ac.uk

***Abstract*—This paper investigates the use of unsupervised text-to-speech synthesis (TTS) as a data augmentation method to improve accented speech recognition. TTS systems are trained with a small amount of accented speech training data and their pseudo-labels rather than manual transcriptions, and hence unsupervised. This approach enables the use of accented speech data without manual transcriptions to perform data augmentation for accented speech recognition. Synthetic accented speech data, generated from text prompts by using the TTS systems, are then combined with available non-accented speech data to train automatic speech recognition (ASR) systems. ASR experiments are performed in a self-supervised learning framework using a Wav2vec2.0 model which was pre-trained on large amount of unsupervised accented speech data. The accented speech data for training the unsupervised TTS are read speech, selected from L2-ARCTIC and British Isles corpora, while spontaneous conversational speech from the Edinburgh international accents of English corpus are used as the evaluation data. Experimental results show that Wav2vec2.0 models which are fine-tuned to downstream ASR task with synthetic accented speech data, generated by the unsupervised TTS, yield up to 6.1% relative word error rate reductions compared to a Wav2vec2.0 baseline which is fine-tuned with the non-accented speech data from Librispeech corpus.**

***Index Terms*—Accented speech recognition, text-to-speech synthesis, data augmentation, self-supervised learning, Wav2vec2.0**

I. Introduction

Accented speech recognition is an important research topic of automatic speech recognition (ASR). Because of its importance, this research topic has been receiving attention and being addressed with various research approaches. In general, these approaches can be classified as accent-agnostic approaches, in which the modeling of accents inside the ASR systems is not made specific, and accent-aware approaches in which additional information about the accents of the input speech are used [1]. Among accent-agnostic approaches, adversarial learning was used to establish accent classifier and accent relabeling which led to performance improvement [2], [3], [4]. In addition, similarity losses such as cosine losses or contrastive losses were used to build accent neutral models [5]. In accent-aware approaches, multi-domain training [6], accent embeddings [7], or accent information fusion [8] are among the approaches which have been investigated.

Text-to-speech synthesis (TTS) is an useful technology which can be used to improve ASR in a number of ways, for instance to improve the pre-training of self-supervised learning (SSL) models [9] or to improve the recognition of out-of-vocabulary words in end-to-end ASR [10]. TTS was also used as a data augmentation method to improve speech recognition for Librispeech task [11] and low-resource speech recognition [12], [13], [14]. More specifically, synthetic data were used for data augmentation in the context of low-resource ASR using conventional hybrid structure [13] and to augment the training of RNN-T (recurrent neural network - transducer) ASR model [15]. In [14], cross-lingual multi-speaker speech synthesis and cross-lingual voice conversion were applied to data augmentation for ASR. The authors showed that it is possible to achieve promising results for ASR model training with just a single speaker dataset in a target language, making it viable for low-resource scenarios [14].

While having been widely used in various ASR tasks, TTS, especially unsupervised TTS which is trained with unsupervised audio data [16], has not been extensively studied as a data augmentation method in accented speech recognition. In a recent study on using TTS as data augmentation for accented speech recognition [17], accented speech were generated by passing English text prompts through TTS system for a language corresponding to the target accent. For example, English text prompts passing through Spanish TTS will approximate Spanish-accented English. The study in [17] used commercial TTS systems whose training data were not accessible by users.

In this paper, we investigate the use of unsupervised TTS as a data augmentation method to improve accented speech recognition. In our approach, we make use of a small amount of accented speech data which do not have manual transcriptions to train TTS systems. This approach enables the use of accented speech data without manual transcriptions to perform data augmentation for accented speech recognition. Indeed, from a small amount of unsupervised accented speech data used to train the TTS systems, we can generate larger amount of synthetic accented speech data once the TTS systems are trained. In this paper, from 58 hours of accented speech data, selected from two speech corpora of read speech: L2-ARCTIC [18] and British Isles [19], we train unsupervised TTS and generate 250 hours more of synthetic accented speech data which help to achieve better gains on the evaluation data of spontaneous conversational speech from the Edinburgh international accents of English corpus (EdAcc) [20].

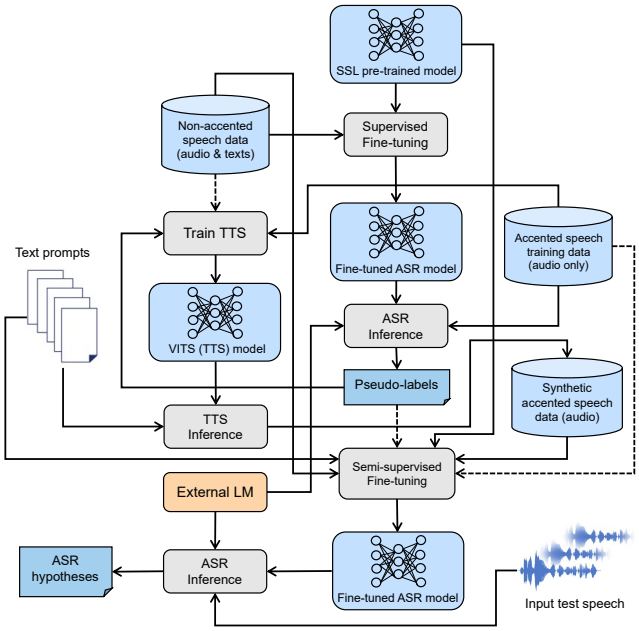

Fig. 1: Unsupervised accented speech training data and theirs pseudo-labels are used to train unsupervised TTS. The pseudo-labels are generated by decoding the unsupervised accented speech training data using the baseline ASR model obtained from the supervised fine-tuning of the SSL pre-trained model with the supervised non-accented speech data. The unsupervised accented speech training data may be included in the semi-supervised fine-tuning for ASR, and the non-accented speech data may be used to train a TTS system.

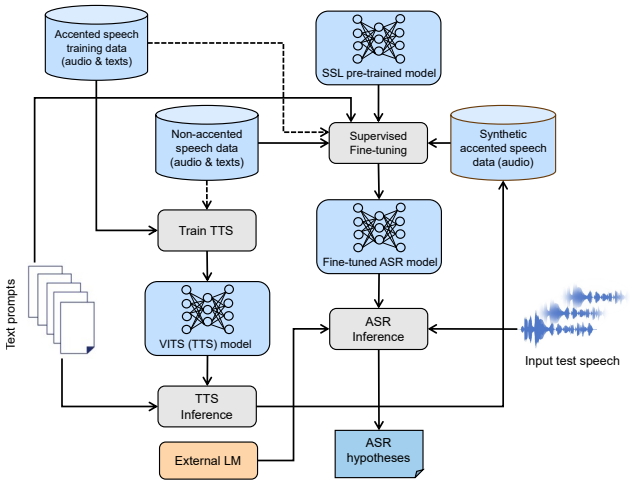

Fig. 2: Supervised accented speech training data are used to train supervised TTS. These data may be included in the supervised fine-tuning for ASR, and the non-accented speech training data may be used to train a TTS system. The fine-tuned ASR model is used in the ASR inference, with an external language model (LM), to decode test speech.

The paper is organized as follows. In section II, details of the data augmentation for accented speech recognition based on unsupervised TTS are introduced. The training and inference of TTS systems are presented in section III. Section IV introduces the data used in the experiments, experimental results, and discussion. Finally, section V concludes the paper.

II. Data Augmentation for Accented Speech Recognition based on Unsupervised TTS

We use Wav2vec2.0 SSL framework [21] for our experiments with accented speech recognition. Assume that a Wav2vec2.0 model was pre-trained via SSL on large amount of unsupervised speech data to cover various English accents and speakers, we can fine-tune this pre-trained model to downstream ASR task using available non-accented speech data. The non-accented speech data could be any available data which can be used to train ASR systems, for instance Librispeech training data [22]. When using publicly available Wav2vec2.0 pre-trained models, we assume that only the models are available and their training data are not available.

In addition to the non-accented speech data, we assume that a small amount of accented speech training data, named AccD, is available. These accented speech training data will be used to train the TTS systems. In accented speech recognition, it is not practical to find accented speech training data spoken in the same speaking styles and by the same speakers in the evaluation data. It is actually more viable to find accented speech training data which are spoken by speakers whose first languages are similar to those of the speakers in the evaluation data. Using these speech data to train TTS systems and generate more accented speech data for ASR training should create more accent variability, and hence, improve accented speech recognition performance.

A. Unsupervised scenario

Fig. 1 shows unsupervised scenario where the manual transcriptions of the accented speech training data AccD are not available. Hence, pseudo-labels for the unsupervised accented speech training data are generated by decoding these data using the baseline ASR model obtained by fine-tuning the SSL pre-trained model with the supervised non-accented speech data. The unsupervised accented speech training data AccD and their pseudo-labels are then used to train TTS model, which is a Variational Inference with adversarial learning for end-to-end Text-to-Speech (VITS) model [23], to generate synthetic accented speech data for data augmentation. The unsupervised accented speech training data and their pseudo-labels may be included in the semi-supervised fine-tuning of the SSL pre-trained model. The ASR model obtained after the semi-supervised fine-tuning is used in the ASR inference to decode input test speech. The ASR inferences use an external language model (LM) when decoding audio data.

A word-level 4-gram LM is used as external LM during ASR inferences. This 4-gram LM is trained on the manual transcriptions of Librispeech training data. The pre-training and fine-tuning of the Wav2vec2.0 models as well as the inference follow the same settings used for the LARGE Wav2vec2.0 models in [21]. These large models consist of 6 convolutional neural network (CNN) and 24 transformer layers, and have 350 millions parameters.

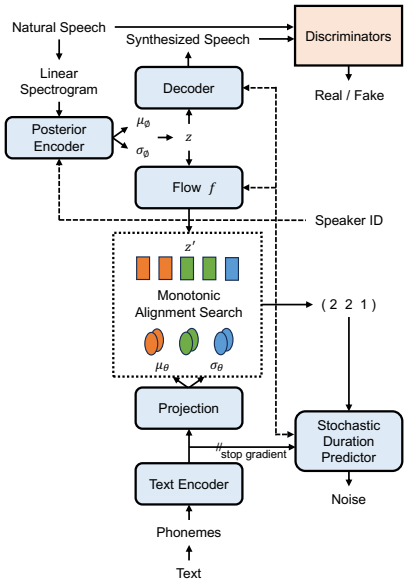

Fig. 3: Training of VITS parallel end-to-end TTS system. The input text can be either manual transcriptions or pseudo-labels.

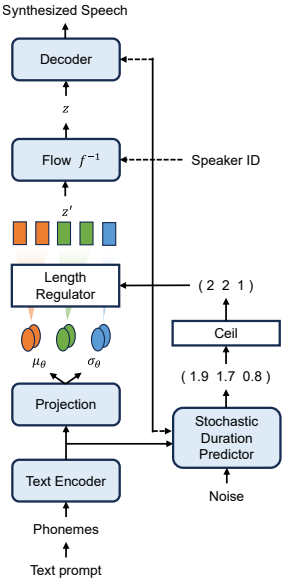

Fig. 4: TTS inference using VITS model where text prompt and speaker ID are used as input.

B. Supervised scenario

For comparison, we also examine supervised scenario where the manual transcriptions of the accented speech training data AccD are available. In Fig. 2, the accented speech training data AccD and their manual transcriptions can be directly used to train supervised TTS model. They may also be included in the training data which are used for the supervised fine-tuning of the SSL pre-trained model. Once the TTS model is trained, it can be used during the TTS inference to generate synthetic accented speech data using independent text prompts. Both the synthetic accented speech data and the text prompts can then be used for data augmentation in the supervised fine-tuning of the SSL pre-trained model.

III. Text-to-speech Synthesis

VITS is an end-to-end multi-speaker TTS system which can generate high-quality waveforms [23]. During the training of VITS (see Fig. 3), a Posterior Encoder encodes linear spectrogram from natural speech into a latent variable z [24] which is then used in a Decoder to restore waveform. HiFi-GAN (Generative Adversarial Network) [25], a GAN-based neural vocoder [26], is used in the decoder to synthesize high-fidelity speech. The latent variable z is also fed into the Flow f which computes the Kullback-Leibler divergence with the Text Encoder outputs. The Flow f is trained to remove speaker information and reduce posterior complexity [27]. During training, speakers identities (IDs) are used to extract speaker embeddings for training multi-speaker TTS.

During TTS inference (see Fig. 4), an inverse transform f^{-1} of the Flow f is used to synthesize speech. The output of the Text Encoder is stretched by the Length Regulator based on the predicted duration, and then the sampled latent variable $z' \sim \mathcal{N}(z'; \mu_\theta(text), \sigma_\theta(text))$ is transformed by the inverse Flow f^{-1} together with speaker information. Speech is subsequently generated by the decoder. All the VITS systems in this paper use the same architecture and are trained with the same number of training iterations, i.e. 300K. We observe that after 300K training iterations, the quality of the synthesized waveforms saturates. Greater details of the VITS models and their implementation can be found in [23].

IV. Experiments

The Wav2vec2.0 model is pre-trained with more than 60K hours of unsupervised speech data from Libri-Light, Common-Voice, Switchboard, and Fisher corpora. These speech training data were spoken in various English accents. The non-accented speech data consist of 960 hours of training data from Librispeech corpus [22] which include 2200 speakers. Although these speakers spoke US-English, we consider Librispeech as non-accented in the context of this study because US-English is only one of the English accents in the evaluation data. Subsequent experimental results confirm our assumption.

A. Data

1) Accented speech training data (AccD): We combine data from the L2-ARCTIC corpus [18] and the British Isles corpus [19] as accented speech training data. These are corpora of read speech which were recorded in controlled environments. The L2-ARCTIC corpus is a speech corpus of non-native English which contains 26,867 utterances from 24 non-native English speakers with equally distributed number of speakers per accent. The total duration of the corpus is 27.1 hours, with an average of 67.7 minutes of speech per speaker. On average, each utterance is 3.6 seconds in duration. The utterances in L2-ARCTIC are spoken in 6 non-native accents: Arabic, Chinese, Hindi, Korean, Spanish, and Vietnamese.

The British Isles corpus includes speech utterances recorded by volunteers speaking with different accents of the British

TABLE I: Word error rates (WERs) on the development and test sets of the Edinburgh international accents of English corpus (EdAcc), and on the test-clean and test-other sets of Librispeech (LS) corpus.

Fine-tuning data \ Test data	EdAcc dev-set	EdAcc test-set	LS test-clean	LS test-other
LS 960h (M) (baseline in [20])	33.4	36.1	2.9	5.6
LS 960h (M) (our baseline)	32.8	35.1	2.2	4.2
LS 960h (M) + AccD (P)	32.4	34.6	2.1	4.1
LS 960h (M) + AccD (M)	31.1	33.4	2.1	4.0
LS 960h (M) + TTS-LS 960h (M)	31.4	33.8	2.1	4.0
LS 960h (M) + TTS-AccD (P)	31.0	33.2	2.1	4.1
LS 960h (M) + TTS-AccD (M)	30.8	33.0	2.1	4.1
LS 960h (M) + AccD (P) + TTS-AccD (P) + TTS-LS 960h (M)	30.8	33.2	2.1	4.2
LS 960h (M) + AccD (M) + TTS-AccD (M) + TTS-LS 960h (M)	30.4	32.7	2.1	4.1

Isles, namely Ireland, Scotland, Wales, the Midlands, Northern, and Southern of England. The corpus consists of 17,877 utterances spoken by 120 speakers of which 49 are female and 71 are male. The total duration of the corpus is 31 hours. When being decoded in the unsupervised scenario, the WERs of the pseudo-labels obtained on the L2-ARCTIC and British Isles training data are 10.7% and 10.2%, respectively.

2) Evaluation data: We use the development and test sets from the Edinburgh international accents of English corpus (EdAcc) [20], which consist of spontaneous conversational speech, as evaluation data. The corpus includes a wide range of first- and second-language varieties of English in the form of dyadic video call conversations between friends. The conversations range in durations from 20 to 60 minutes. These conversations are segmented into shorter utterances based on manual annotations and are then separated into development and test sets which consist of 9079 and 8494 utterances, respectively. In total, the development set contains 14 hours and the test set contains 15 hours of speech. There are more than 40 self-reported English accents from 51 different first languages. The statistics and analyses show that EdAcc is linguistically diverse and challenging for current English ASR systems [20]. With more than 40 English accents, the EdAcc corpus covers English accents from four continents, including Africa, America, Asia, and Europe. The conversations were manually transcribed by professional transcribers to obtain manual transcriptions which are used in the evaluation.

3) Synthetic speech data: Synthetic speech data are generated using the TTS systems and English text prompts. The text prompts used in the TTS inference are selected from the manual transcriptions of the training data in three speech corpora: LJSpeech [28], TED-LIUM [29], and VCTK [30]. The objective of selecting text prompts from independent TTS and ASR corpora is to ensure that these prompts are not related to the evaluation data and are phonetically balanced, since they were designed for TTS and ASR applications. In total, there are 120K text prompts resulting in 250 hours of synthetic speech data which are spoken by the speakers presented in the training data of the TTS systems.

B. Results & Discussion

Experimental results, in terms of WERs, are shown in Table I. In Table I, the WERs computed on the EdAcc development & test sets and the Librispeech (LS) test-clean & test-other sets are shown. The ASR models in Table I are fine-tuned from one Wav2vec2.0 pre-trained model, which was pre-trained on the unsupervised training data of Libri-Light, Common Voice, Switchboard, and Fisher, using different fine-tuning data. The abbreviations used in Table I have the meaning as follows:

- LS 960h (M): 960 hours of training speech from Librispeech, manual (M) transcriptions are used as labels.
- AccD (P), AccD (M): 58 hours of accented speech training data, using either pseudo-labels (P) or manual (M) transcriptions as labels.
- TTS-LS 960h (M): 250 hours of synthetic non-accented speech data generated by TTS system trained on LS 960h (M) data. The speakers are from the LS 960h data.
- TTS-AccD (P), TTS-AccD (M): 250 hours of synthetic accented speech data generated by TTS systems trained on either AccD (P) or AccD (M) data, with speakers from the AccD data.

We build a baseline model by fine-tuning the Wav2vec2.0 pre-trained model with the LS 960h (M) data. The Wav2vec2.0 pre-trained model and the fine-tuning data that we use are the same as those used to train the baseline model in [20]. We will compare the results with our baseline model which has lower WERs, compared to those of the Wav2vec2.0 model reported in [20], on the development and test data of both EdAcc and Librispeech (see Table I). Combining the unsupervised accented speech training data AccD (P) with the non-accented speech data LS 960h (M) to fine-tune the pre-trained model yields 1.2% and 1.4% relative WER reductions on EdAcc dev and test sets, respectively, while the respective relative WER reductions on these sets are 5.2% and 4.8% when the supervised accented speech training data AccD (M) are used.

When the synthetic non-accented speech data TTS-LS 960h (M) which are generated by the supervised TTS system, trained on the non-accented speech data LS 960h (M) with manual transcriptions, are included in the fine-tuning, 4.3% and 3.7% relative WER reductions are obtained on the EdAcc dev and test sets, respectively. Since the synthetic non-accented speech data TTS-LS 960h (M) are spoken by the same speakers in the LS 960h (M) data, the relative WER reductions are made mainly thanks to more acoustic realizations, based on the independent text prompts, are added

to the fine-tuning data from the synthetic non-accented speech data. Larger gains are obtained when the synthetic accented speech data TTS-AccD (P) and TTS-AccD (M) are used, even though the amount of data and the number of speakers in the AccD data used to train TTS systems are much smaller compared to those of the non-accented speech data LS 960h (M): 58 hours compared to 960 hours, and 144 speakers compared to 2200 speakers. More specifically, the TTS-AccD (P) data generated by unsupervised TTS help to achieve 5.5% and 5.4% relative WER reductions on the EdAcc dev and test sets, respectively, while the TTS-AccD (M) data generated by supervised TTS help to achieve 6.1% and 6.0% relative WER reductions on the EdAcc dev and test sets, respectively.

When the accented speech training data AccD and all the synthetic speech data are combined with the non-accented speech data to fine-tune the pre-trained model, further gains are obtained. In the unsupervised scenarios, 6.1% and 5.4% relative WER reductions are obtained on the EdAcc dev and test sets, respectively, while the respective relative WER reductions obtained on these sets in the supervised scenario are 7.3% and 6.8%, respectively. Actually, using natural accented speech training data and synthetic accented speech data improves the performance on EdAcc dev and test sets but does not harm or improve the ASR performance on Librispeech test sets. This confirms that considering Librispeech training data as non-accented speech data in our experiments is relevant.

V. Conclusion

Unsupervised TTS, trained on unsupervised accented speech training data, was used to generate synthetic accented speech data for data augmentation in accented speech recognition. Experiments showed that the Wav2vec2.0 models which used the synthetic accented speech data yielded up to 6.1% relative WER reductions compared to a large Wav2vec2.0 baseline. These gains are close to those obtained in the supervised scenario. The results demonstrate that unsupervised accented speech data, even when available in limited quantities and are spoken in different styles by speakers who differ from those in the evaluation data, can be effectively used to train TTS systems for data augmentation. This approach improves accented speech recognition, particularly when the speakers in the unsupervised accented speech data and those in the evaluation data have some overlaps on speakers' first languages.

References

[1] D. Prabhu, P. Jyothi, S. Ganapathy, and V. Unni, "Accented speech recognition with accent-specific codebooks," in *Proc. 2023 Conference on Empirical Methods in Natual Language Processing (EMNLP)*.

[2] Y.-C. Chen, Z. Yang, C.-F. Yeh, M. Jain, and M. Seltzer, "Aipnet: generative adversarial pre-training of accent-invariant networks for end-to-end speech recognition," in *Proc. 2020 IEEE ICASSP*, pp. 6979–6983.

[3] N. Das, S. Bodapati, M. Sunkara, S. Srinivasan, and D. Horng Chau, "Best of both worlds: robust accented speech recognition with adversarial transfer learning," in *Proc. INTERSPEECH 2021*, pp. 1314–1318.

[4] H. Hu et al., "REDAT: accent-invariant representation for end-to-end ASR by domain adversarial training with relabeling," in *Proc. 2021 IEEE ICASSP*, pp. 6408–6412.

[5] K. Deng, S. Cao, and L. Ma, "Improving accent identification and accented speech recognition under a framework of self-supervised learning," in *Proc. INTERSPEECH 2021*, pp. 1504–1508.

[6] M. Lucas and Y. Estève, "Improving accented speech recognition with multi-domain training," in *Proc. 2023 IEEE ICASSP*, pp. 1–5.

[7] A. Jain, M. Upreti, and P. Jyothi, "Improved accented speech recognition using accent embeddings and multi-task learning," in *Proc. INTERSPEECH 2018*, pp. 2454–2458.

[8] X. Wang, Y. Long, Y. Li, and H. Wei, "Multi-pass training and cross-information fusion for low-resource end-to-end accented speech recognition," in *Proc. INTERSPEECH 2023*, pp. 2923–2927.

[9] Z. Chen, Y. Zhang, A. Rosenberg, B. Ramabhadran, G. Wang, and P. Moreno, "Injecting text in self-supervised speech pretraining," in *Proc. 2021 IEEE ASRU Workshop*, pp. 251–258.

[10] X. Zheng, Y. Liu, D. Gunceler, and D. Willett, "Using synthetic audio to improve the recognition of out-of-vocabulary words in end-to-end ASR systems," in *Proc. 2021 IEEE ICASSP*, pp. 5674–5678.

[11] A. Rosenberg, Y. Zhang, B. Ramabhadran, Y. Jia, P. Moreno, Y. Wu, and Z. Wu, "Speech recognition with augmented synthesized speech," in *Proc. 2019 IEEE ASRU*, pp. 996–1002.

[12] S. Ueno, M. Mimura, S. Sakai, and T. Kawahara, "Data augmentation for ASR using TTS via discrete representation," in *Proc. 2021 IEEE Automatic Speech Recognition and Understanding Workshop*, pp. 68–75.

[13] G. Zhong, H. Song, R. Wang, L. Sun, D. Liu, J. Pan, X. Fang, J. Du, J. Zhang, and L. Dai, "External text based data augmentation for low-resource speech recognition in the constrained condition of OpenASR21 challenge," in *Proc. INTERSPEECH 2022*, pp. 4860–4864.

[14] E. Casanova, C. Shulby, A. Korolev, A.C. Junior, A.d.S. Soares, S. Aluísio, and M.A. Ponti, "ASR data augmentation in low-resource settings using cross-lingual multi-speaker TTS and cross-lingual voice conversion," in *Proc. INTERSPEECH 2023*, pp. 1244–1248.

[15] A. Fazel, W. Yang, Y. Liu, R. Barra-Chicote, Y. Meng, R. Maas, and J. Droppo, "SynthASR: unlocking synthetic data for speech recognition," in *Proc. INTERSPEECH 2021*, pp. 896–900.

[16] J. Ni, L. Wang, H. Gao, K. Qian, Y. Zhang, S. Chang, and M. Hasegawa-Johnson, "Unsupervised text-to-speech synthesis by unsupervised automatic speech recognition," in *Proc. INTERSPEECH 2022*, pp. 461–465.

[17] G. Karakasidis, N. Robinson, Y. Getman, A. Ogayo, R. Al-Ghezi, A. Ayasi, S. Watanabe, Mortensen D. R., and M. Kurimo, "Multilingual TTS accent impressions for accented ASR," in *Proc. 2023 International Conference on Text, Speech, and Dialogue (TSD)*, pp. 317–327.

[18] G. Zhao, S. Sonsaat, A. Silpachai, I. Lucic, E. Chukharev-Hudilainen, J. Levis, and R. Gutierrez-Osuna, "L2-ARCTIC: a non-native English speech corpus," in *Proc. INTERSPEECH 2018*, pp. 2783–2787.

[19] I. Demirsahin, O. Kjartansson, A. Gutkin, and C. Rivera, "Open-source multi-speaker corpora of the English accents in the British Isles," in *Proc. 2020 Conference on Language Resources and Evaluation (LREC)*.

[20] R. Sanabria, N. Bogoychev, N. Markl, A. Carmantini, O. Klejch, and P. Bell, "The Edinburgh international accents of English corpus: towards the democratization of English ASR," in *Proc. 2023 IEEE ICASSP*.

[21] A. Baevski, Y. Zhou, A. Mohamed, and M. Auli, "Wav2vec 2.0: a framework for self-supervised learning of speech representations," in *Proc. 2020 Advances in Neural Information Processing Systems*.

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

[23] J. Kim, J. Kong, and J. Son, "Conditional variational autoencoder with adversarial learning for end-to-end text-to-speech," in *2021 International Conference on Machine Learning*.

[24] D. P. Kingma and M. Welling, "Auto-encoding variational bayes," in *2014 International Conference on Learning Representations (ICLR)*.

[25] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in *Proc. 2014 Advances in Neural Information Processing Systems (NIPS)*.

[26] J. Kong, J. Kim, and J. Bae, "Hifi-gan: Generative adversarial networks for efficient and high fidelity speech synthesis," *Proc. 2020 Advances in Neural Information Processing Systems*.

[27] D. Rezende and S. Mohamed, "Variational inference with normalizing flows," in *Proc. 2015 International Conference on Machine Learning*.

[28] K. Ito and L. Johnson, "The LJ Speech dataset," https://keithito.com/LJ-Speech-Dataset, 2017.

[29] A. Rousseau, P. Deléglise, and Y. Estève, "TED-LIUM: an automatic speech recognition dedicated corpus," in *Proc. 2012 Conference on Language Resources and Evaluation (LREC)*, pp. 125–129.

[30] J. Yamagishi, C. Veaux, and K. MacDonald, "CSTR VCTK corpus: English multi-speaker corpus for CSTR voice cloning toolkit (version 0.92)," https://doi.org/10.7488/ds/2645, 2017.