Data Augmentation Using CycleGAN for End-to-End Children ASR

Dipesh K. Singh[1], Preet P. Amin[1], Hardik B. Sailor[2], and Hemant A. Patil[1]

[1]Dhirubhai Ambani Institute of Information and Communication Technology (DA-IICT), Gandhinagar, Gujarat, India.
[2]Samsung R&D Institute, Bangalore (SRI-B), India
[1]{*dipesh_singh, preet_amin, hemant_patil*}*@daiict.ac.in*
[2]*h.sailor@samsung.com*

***Abstract*—Recent deep learning algorithms are known to perform better for Automatic Speech Recognition (ASR) of adult speakers, however, yet remains a challenge to recognize children's speech with the similar accuracy. Due to less availability of children's speech data to train the deep neural network, data augmentation is one of the key research areas for children ASR. In this paper, voice conversion-based data augmentation using CycleGAN is explored and performance comparison with and without data augmentation is presented. ASR experiments were performed using TLT school corpus. In our experiment, CycleGAN-based 200 hours of converted adult speech showed good performance improvement with the reduction of 5.58% WER compared to the baseline system. In addition, the combination of SpecAugment, speed perturbed, and CycleGAN converted adult speech showed the highest reduction of 7.44% WER compared to baseline system[1].**

***Index Terms*—Hybrid HMM/DNN architecture, CTC, attention, Transformer, speed perturbation, SpecAugment, speech recognition, and ASR**

I. Introduction

Language is the engine of civilization, and speech is its most natural and powerful form, resulting in various successful speech technologies, such as speech and speaker recognition, text-to-speech synthesis, voice conversion, etc. Voice assistant or Intelligent Personal Assistant (IPAs) is a device that uses these technologies to provide various services through voice. Voice assistants can be used in day-to-day life, e.g., in education, retails, healthcare, inside our car or home. The voice assistant needs to overcome many challenges to give consistent services to the user irrespective of age and gender. Automatic Speech Recognition (ASR) for children is one of the key challenges in the literature. The voice assistants can be used by the children for remote learning, playing games, in-vehicle, entertainment, etc. However, there are several challenges in children's ASR.

Due to shorter vocal tract and smaller (i.e., smaller mass of) vocal folds, children have higher fundamental frequency (F_O) and formant frequencies than those of adults. There are more repetitive segments, stammering, and false start in children's speech as compared to the adult speech [1]. Children use to stretch the utterance primarily because of hesitation and lack of language information [2], [3], [4]. End-to-end (E2E) deep neural network (DNN) architecture showed a tremendous growth for ASR task in the past few years [5]. As E2E framework consists of a neural architecture, during training the number of training samples plays a vital role [6], [7]. Generally, these ASR systems are trained on adult speech and thereby does not perform well when tested on children's speech due to various acoustic differences between adult and children's speech [8]. There are different challenges associated with training deep neural network, such as less amount of children's speech available for training. Even if the amount of data used for children ASR will be the same as that of adult ASR, the performance of children ASR will be poorer than adult ASR (because of the acoustic variability and poor spectral resolution due to sampling of vocal tract spectrum by high pitch source harmonics) [9], [10]).

In this context, data augmentation is one of the well established regularization techniques in the literature [11]. There are various conventional data augmentation techniques for ASR system, such as SpecAugment, speed, and tempo perturbation. SpecAugment operates on the principle of time warping, frequency masking, and time masking [12], [13]. The children database has more variations in different acoustic parameters, such as pitch (F_0), speaking rate. However, SpecAugment does not introduces acoustic variability which is one of the limitations of SpecAugment technique. Speed perturbation is a technique which is used to generate the audio data by re-sampling the audio signal by a factor α [14]. This method introduces prosody variation in the database resulting in the elimination of the limitation for SpecAugment technique. It also introduces new speaker information into the database which helps in making the E2E system robust. Tempo perturbation also called speech rate perturbation is a technique in which speech rate or tempo of the audio data is altered while keeping pitch and spectral envelope the same [15]. Tempo perturbation does not include any new speaker information into the database, however, it modifies the speaking rate of utterances present in the dataset. Data augmentation strategy utilizing various GAN models explored in past including WGAN-GP (Wasserstein GAN with gradient penalty) [16]

[1]The converted speech samples are provided at https://preetx.github.io/Children-ASR-Samples/

and CycleGAN [17]. CycleGAN-based data augmentation has shown significant performance improvement for children ASR [16].

In this paper, data augmentation using CycleGAN is explored for end-to-end children ASR. To advance the research on CycleGAN-based data augmentation, we propose novel CycleGAN network inspired from its application in non-parallel voice conversion [18]. It is an improved version of previous CycleGAN network [16] incorporating three new techniques:

- An improved objective (two-step adversarial loss)
- Improved generator (2-1-2D CNN and ConvTranspose layer)
- Improved discriminator (Patch GAN).

Furthermore, techniques for improving CycleGAN training for voice conversion task is also proposed. It is aimed to convert large amount of adult speech corpus into children's speech so that the converted speech samples can be used for training children ASR.

The rest of the paper is organized as follows. Section 2 briefly introduces the voice conversion using CycleGAN and the novelty of paper for stable training of CycleGAN. Section 3 introduces the experimental setup of CycleGAN architecture and ASR system. In Section 4, experimental results and spectrographic analysis are presented for baseline and proposed system. Section 5 summarizes the paper along with future research directions.

II. Proposed Approach

A. Voice Conversion using CycleGAN

For voice conversion, let $a \in \mathbb{R}^{D\times N}$, and $c \in \mathbb{R}^{D\times N}$ be the features belonging to adult class *A* and children class *C*, respectively. Here, *D* is the feature dimension and *N* is the number of frames. Then, the aim of CycleGAN is to learn the mapping from $G_{A\to C}$, and $G_{C\to A}$. CycleGAN uses four different loss functions in order to perform the voice conversion task, which are as follows [18]:

Adversarial loss: It's goal is to make feature $G_{A\to C}(a)$ indistinguishable from the children's speech feature, c. In particular,

$$\mathcal{L}_{adv}(G_{A\to C}, D_C) = \mathbb{E}_{c\sim P_C(c)}[logD_C(c)] + \mathbb{E}_{a\sim P_A(a)}[log(1 - D_C(G_{A\to C}(a)))]. \quad (1)$$

Cycle-consistency loss: For voice conversion, it is necessary to preserve the context during conversion. The cycle-consistency loss is used to meet that condition, which ensures the network can learn the forward-inverse and inverse-forward mapping simultaneously, i.e.,

$$\mathcal{L}_{cyc}(G_{A\to C}, G_{C\to A}) = \mathbb{E}_{a\sim P_A(a)}[\|G_{C\to A}(G_{A\to C}(a)) - a\|_1] + \mathbb{E}_{c\sim P_C(c)}[\|G_{A\to C}(G_{C\to A}(c)) - c\|_1], \quad (2)$$

where $\|.\|_1$ is L_1-norm and $\mathbb{E}[.]$ is expectation operator.

Identity-mapping loss: It is used to preserve the identity of target class in the network. It is generally employed to train the network for initial learning, i.e.,

$$\mathcal{L}_{id}(G_{A\to C}, G_{C\to A}) = \mathbb{E}_{a\sim P_A(a)}[\|G_{C\to A}(a) - a\|_1] + \mathbb{E}_{c\sim P_C(c)}[\|G_{A\to C}(c) - c\|_1]. \quad (3)$$

Two-step adversarial loss: We impose an additional adversarial loss on circular feature to further improve adversarial training of our GAN network. In particular,

$$\mathcal{L}_{2-step}(G_{A\to C}, G_{C\to A}, D_C) = \mathbb{E}_{c\sim P_C(c)}[logD_C(c)] + \mathbb{E}_{a\sim P_A(a)}[log(1 - D_C(G_{A\to C}(G_{C\to A}(c))))]. \quad (4)$$

Full objective: Overall objective of the CycleGAN network is given as:

$$\begin{aligned}\mathcal{L}_{full} = {} & \lambda_{adv}\mathcal{L}_{adv}(G_{A\to C}, D_C) + \lambda_{adv}\mathcal{L}_{adv}(G_{C\to A}, D_A) \\ & + \lambda_{adv}\mathcal{L}_{2-step}(G_{A\to C}, G_{C\to A}, D_C) \\ & + \lambda_{adv}\mathcal{L}_{2-step}(G_{C\to A}, G_{A\to C}, D_A) \\ & + \lambda_{cyc}\mathcal{L}_{cyc}(G_{A\to C}, G_{C\to A}) \\ & + \lambda_{id}\mathcal{L}_{id}(G_{A\to C}, G_{C\to A}), \end{aligned} \quad (5)$$

where λ_{adv}, λ_{cyc}, and λ_{id} are the weights associated with adversarial, cycle-consistency, and identity-mapping loss, respectively. These values are used as hyperparameters in the network during training.

B. Stable training of CycleGAN

The high variations in feature space, making the learning of the CycleGAN generator difficult and thereby causes instability in training. In our previous experiments, we observed that the generator learning becomes too slow compared to the discriminator due to the this reason. It results in collapse of the entire CycleGAN network. To improve the CycleGAN training for the voice conversion, we kept a lower limit or threshold on discriminator loss, after which the training of discriminator gets stopped. If the overall discriminator loss goes below 0.05, the weight update will not take place for the discriminator, allowing the generator to learn the mapping. Once the generator learns the mapping, it automatically makes the discriminator loss above 0.05, thereby including the discriminator for weight update.

Based on synthetic speech, we observed during CycleGAN training that due to the low weight of adversarial losses ($\lambda_{adv} = 1$) compared to the other two losses, voice conversion was not taking place as we desired. Since the weights of cycle loss and identity loss are high ($\lambda_{cyc} = 10$, and $\lambda_{id} = 5$), generator focuses more on reducing these losses than its adversarial loss. This setup is desirable in the initial phase of the training since it forces our generators to learn to reconstruct original input feature. However, since adversarial loss has main role in voice conversion, it is desirable to increase its weight after generator had learned to reconstruct original features. Hence, we kept on reducing λ_{cyc} and increasing λ_{adv} by a factor of 1.2 after every two epochs.

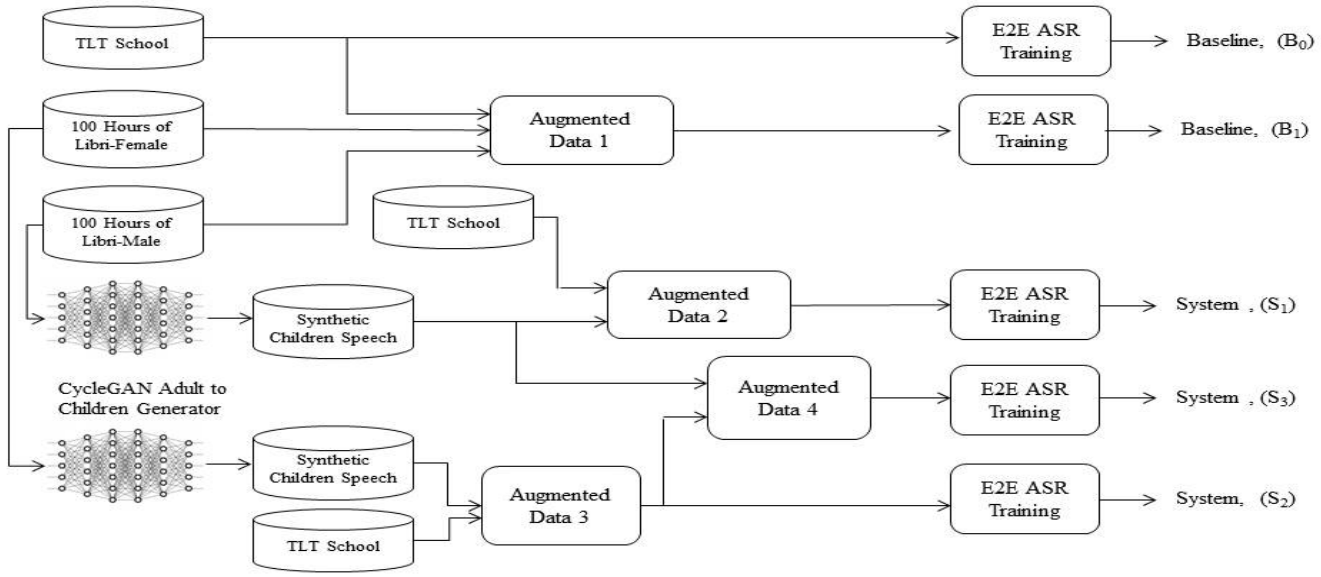

Fig. 1. Functional block diagram of proposed data augmentation approach for children ASR.

III. Experimental Setup

A. Database

In this paper, TLT School corpus is used for ASR experiments. The TLT-School corpus is obtained from INTERSPEECH TLT2020 shared task on ASR for non-native children's speech [19]. The corpus contains 3518 speakers of age group between 9 to 16 years with three different English proficiency stages. The duration of training, development, and test corpora are 49 hours, 2 hours, and 2 hours, respectively. The transcription consists of symbols of laughter, mispronounced words, whispered speech, and special characters, such as <unk-it>, for Italian, and German words.

To generate synthetic voices, My Science Tutor (MyST) children database is used along with Libri clean 360 adult speech database [20]. The MyST corpus contains unsubscribed utterances from 1372 speakers with an age range between 8 to 11 years, and the duration of the corpus is 499 hours [21]. This corpus is used for training CycleGAN because the noise is very low compared to the TLT School corpus. In addition, MyST corpus contains utterances of native English speakers, Our initial experiments with TLT School corpus for CycleGAN produced very noisy synthetic speech.

The Libri-clean-360 corpus is a subset of the Librispeech corpus [22]. It consists of 363 hours of speech data collected from 921 adult speakers. This corpus is used to train the CycleGAN network for voice conversion. In addition, this corpus is converted into children's speech and those converted utterances further utilized in training ASR.

B. CycleGAN Architecture

Generator:

For generator, we have used generator from CycleGAN-VC2 network [18]. Input feature are first downsampled using two strided 2-D CNN. For voice conversion, six 1-D CNN residual layers are used. For upsampling layers, we have used the two strided 2-D convolution transpose layer as opposed to interpolation used in original CycleGAN-VC2 study in [18]. Convolution transposed layer are primarily used so that network can learn upsampling on its own. 2-D CNN and 1-D CNN layers are bridged together using 1x1 convolution which is applied before and after reshaping of the features.

Discriminator:

In previous GAN-based models, FullGAN (i.e., discriminator with fully-connected layers) were used [23], [24], however, studies have shown that assigning only a single penalty to entire features using FullGAN can sometimes lead to blurriness of features [25]. To alleviate this, we used PatchGAN that assigns individual penalties to NxN patches of features using CNN (i.e., discriminator with CNN as output layer). PatchGAN increases difficulty of discriminator that helps in generating more crispier (fine) features as observed in computer vision literature [26].

C. CycleGAN training

In CycleGAN, two generators and two discriminators are trained simultaneously in an adversarial manner [27]. For voice conversion, we have extracted *24*-dimensional MCEPs, pitch (F_0) contour, and aperiodicity using WORLD vocoder [28]. *24*-dimensional MCEPs are transformed using the CycleGAN architecture whereas pitch contour is converted by using logarithmic Gaussian normalized transformation. Aperiodicity was used without any modification.

MCEPs are normalized with global zero-mean and unit variance in the pre-processing stage. We used Adam optimizer with betas $\beta_1 = 0.5$, and $\beta_2 = 0.999$. Values of 0.0004 and 0.0002 were taken as initial learning rates for generators and

TABLE I
WER (%) COMPARISON FOR CONVENTIONAL DATA AUGMENTATION SCENARIOS

Model Id	System	WER (%)
B_0	Baseline - 1	30.94
B_1	Baseline - 2	28.45
S_6	B_1 + SpecAugment	27.61
S_7	B_1 + SpecAugment + Speed Perturb	**27.11**

TABLE II
WER (%) COMPARISON FOR PROPOSED DATA AUGMENTATION

Model Id	System	WER (%)
S_1	B_0 + Converted male	27.55
S_2	B_0 + Converted female	27.12
S_3	B_0 + Converted male + Converted female	25.42
S_4	S_3 + SpecAugment	25.20
S_5	S_3 + SpecAugment + Speed Perturb.	**23.50**

discriminators, respectively. We used 10 seconds segments (2001 frames) for our training of CycleGAN. Identity loss ($\mathcal{L}_{id}$) was used for first 10^4 iterations only. We trained our model for 10^5 iterations with the batch size of 2. Learning rate was decreased using exponential decay with the decay of 1.2 every 2x10^4 steps. Initially, we set $\lambda_{cyc} = 10$, and $\lambda_{id} = 5$.

D. ASR training

We have used ESPNET-based E2E ASR system [29]. We adopted librispeech transformer model recipe in ESPNET [30]. *80*-dimensional Mel filterbank feature along with pitch are used as feature set to train transformer model. LSTM-based language model (LM) is also used with 5000 BPE throughout the experiments [31]. We kept 0.6 as LM shallow fusion weight.

IV. EXPERIMENTAL RESULTS

A. Baseline Systems

We have two baseline systems in this paper. For baseline 1, we have trained the E2E ASR system with only TLT school corpus. For baseline 2, we have augmented 100 hours of male and 100 hours of female Librispeech subset into the TLT school corpus and then trained the ASR system.

Table 1 shows the word error rate (WER) comparison for conventional data augmentation techniques. Augmenting 200 hours of adult speech shows the relative WER reduction by 2.49%. However, only SpecAugment over children training data reduces the performance WER by 3.33%. SpecAugment and speed perturbation together shows the relative WER reduction of 3.83%. Hence, using convention data augmentation and adding 200 hours of adult speech, system S_7 shows absolute reduction of 3.83% compared to the B_0.

B. Data Augmentation results using CycleGAN

Table 2 shows the WER comparison for data augmentation using CycleGAN. Augmenting 100 hours of converted male speech shows the relative WER reduction of 3.39%. However, augmenting 100 hours of converted female speech shows the relative WER reduction of 3.82%. In addition,

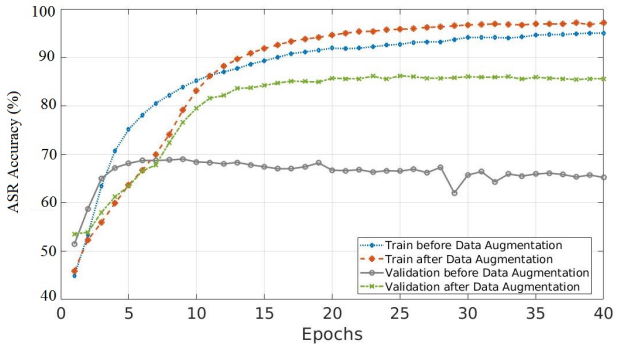

Fig. 2. Learning curves obtained during the training of transformer E2E architecture with and without data augmentation.

augmentation of 100 hours of converted male and 100 hours of converted female together (S_3) shows the relative WER reduction of 5.52% compared to the B_0. We also explored the significance of proposed method with conventional data augmentation. Model Id S_4 denotes SpecAugment on top of proposed method. System S_4 shows the relative WER reduction of 5.74% compared to the B_0. However, Model Id S_5 denotes SpecAugment and speed perturbation together on top of the proposed method. System S_5 shows the relative WER reduction of 7.74% compared to the B_0. Using conventional data augmentation on top of proposed method (S_5) shows absolute reduction of 3.61% compared to the S_7. Hence, using synthetic children speech is more beneficial than using adult speech for end-to-end children ASR.

We can observe from Table I that conventional data augmentation methods, such as augmentation of adult speech with children's speech, SpecAugment and speed perturbation contributes performance improvement of children ASR. However, the contribution is not significant in comparison with the proposed approach.

C. Spectrographic Analysis

The spectrogram and pitch contour are shown in Fig.3 for adult and synthetic children speech. From Fig.3, it can be observed that children's speech converted from adult speech is showing characteristics that are similar to the children's speech signal.

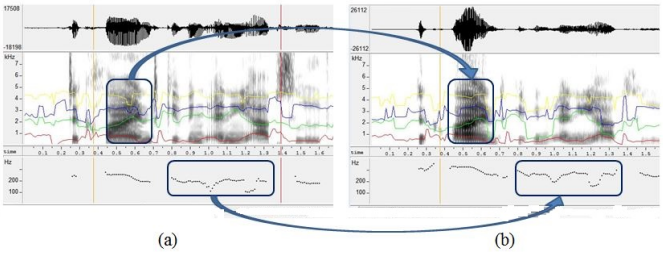

Fig. 3. Spectrogram, pitch, and formant contours of (a) adult speech signal, and (b) synthetic children's speech signal generated from the same utterance shown in Fig.3.(a). Highlighted box shows the formant and pitch contour shift before and after passing through CycleGAN generator, $G_{A\to C}$.

Thus, formant contours and pitch of the adult speech signal are shifted upwards after passing the utterance through trained Cycle-GAN generator,$G_{A\rightarrow C}$. In addition, the shape of the pitch contour is mostly unchanged due to the ability of the Cycle-GAN to reconstruct fine objects as observed in computer vision literature.

V. Summary and Conclusions

In this work, we evaluated the performance of data augmentation using CycleGAN for children ASR task. The ASR system is built using E2E transformer model of ESPNET and ASR experiments are performed on TLT school corpus. In this paper, novel stratergies are discussed for stable training of CycleGAN for voice conversion of adult speech into children's speech. The experimental results shows that the voice conversion of adult female is giving better WER improvement (i.e., reduction in WER) for children ASR. However, voice conversion of male adult speech into children's speech contains some outliers. Our future work will involve exploration of CycleGAN to enhance the quality of voice conversion of male adult speech into children's speech.

References

[1] M. Gerosa, D. Giuliani, S. Narayanan, and A. Potamianos, "A review of ASR technologies for children's speech," in *Proceedings of the* 2^{nd} *Workshop on Child, Computer and Interaction*, New York, United States, 2009, pp. 1–8.

[2] Y. Gao, B. M. L. Srivastava, and J. Salsman, "Spoken English Intelligibility Remediation with PocketSphinx Alignment and Feature Extraction improves Substantially over the State-of-the-Art," in 2^{nd} *IEEE Advanced Information Management, Communicates, Electronic and Automation Control Conference (IMCEC)*, Xi'an, China, 2018, pp. 924–927.

[3] L. Chen, J. Tao, S. Ghaffarzadegan, and Y. Qian, "End-to-end neural network based automated speech scoring," in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Calgary, AB, Canada, 2018, pp. 6234–6238.

[4] N. Jose, B. R. Chakravarthi, S. Suryawanshi, E. Sherly, and J. P. McCrae, "A survey of current datasets for code-switching research," in *IEEE* 6^{th} *International Conference on Advanced Computing and Communication Systems (ICACCS)*, Coimbatore, Tamil Nadu, India, 2020, pp. 136–141.

[5] B. Li, S.-y. Chang, T. N. Sainath, R. Pang, Y. He, T. Strohman, and Y. Wu, "Towards fast and accurate streaming end-to-end ASR," in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2020, pp. 6069–6073.

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

[7] 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 (SLT) Workshop*, Athens, Greece, 2018, pp. 426–433.

[8] A. Potamianos and S. Narayanan, "Robust recognition of children's speech," *IEEE Transactions on Speech and Audio Processing*, vol. 11, no. 6, pp. 603–616, 2003.

[9] M. Russell and S. D'Arcy, "Challenges for computer recognition of children's speech," in *Workshop on Speech and Language Technology in Education*, Farmington, PA, USA, 2007.

[10] J. G. Wilpon and C. N. Jacobsen, "A study of speech recognition for children and the elderly," in *IEEE International Conference on Acoustics, Speech, and Signal Processing Conference (ICASSP)*, vol. 1, Atlanta, Georgia, 1996, pp. 349–352.

[11] X. Song, Z. Wu, Y. Huang, D. Su, and H. Meng, "Specswap: A simple data augmentation method for end-to-end speech recognition," *INTERSPEECH 2020* Shanghai, China, pp. 581–585, 2020.

[12] 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," *arXiv preprint arXiv:1904.08779*, 2019, {Last accessed : 15-01-2021}.

[13] C. Li and Y. Qian, "Prosody usage optimization for children speech recognition with zero resource children speech." in *INTERSPEECH*, Graz, Austria, 2019, pp. 3446–3450.

[14] T. Ko, V. Peddinti, D. Povey, and S. Khudanpur, "Audio augmentation for speech recognition," in *INTERSPEECH* pp. 3586-3589, Dresden, Germany, 2015.

[15] E. D. Scheirer, "Tempo and beat analysis of acoustic musical signals," *The Journal of the Acoustical Society of America (JASA)*, vol. 103, no. 1, pp. 588–601, 1998.

[16] P. Sheng, Z. Yang, and Y. Qian, "Gans for children: A generative data augmentation strategy for children speech recognition," in *2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, 2019, pp. 129–135.

[17] S. Shahnawazuddin, N. Adiga, K. Kumar, A. Poddar, and W. Ahmad, "Voice conversion based data augmentation to improve children's speech recognition in limited data scenario," *INTERSPEECH 2020* Sanghai, China, pp. 4382–4386, 2020.

[18] T. Kaneko, H. Kameoka, K. Tanaka, and N. Hojo, "Cyclegan-vc2: Improved cyclegan-based non-parallel voice conversion," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, 2019.

[19] R. Gretter, M. Matassoni, D. Falavigna, K. Evanini, and C. W. Leong, "Overview of the INTERSPEECH TLT2020 Shared Task on ASR for Non-Native Children's Speech," *INTERSPEECH* Shanghai, China, pp. 245–249, 2020.

[20] R. Gretter, M. Matassoni, S. Bannò, and D. Falavigna, "Tlt-school: A corpus of non native children speech," *arXiv preprint arXiv:2001.08051*, 2020, {Last accessed : 15-01-2021}.

[21] R. Cole, W. Ward, and S. Pradhan, "My Science Tutor and the MyST Corpus," 2019.

[22] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: an asr corpus based on public domain audio books," in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, South Brisbane, Queensland, Australia, pp. 5206–5210.

[23] J. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," in *IEEE International Conference on Computer Vision (ICCV)* Venice, Italy, 2017, pp. 2242–2251.

[24] T. Kaneko and H. Kameoka, "Parallel-Data-Free Voice Conversion using Cycle-Consistent Adversarial Networks," 2017, {Last accessed : 15-01-2021}.

[25] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros, "Image-to-image translation with conditional adversarial networks," 2018.

[26] I. Goodfellow, "NIPS 2016 tutorial: Generative adversarial networks," *arXiv preprint arXiv:1701.00160*, 2017, {Last Accessed: 15-01-2021}.

[27] T. Kaneko, H. Kameoka, K. Tanaka, and N. Hojo, "Cyclegan-vc3: Examining and improving cyclegan-vcs for mel-spectrogram conversion," *arXiv preprint arXiv:2010.11672*, 2020, {Last accessed : 15-01-2021}.

[28] M. Morise, F. Yokomori, and K. Ozawa, "World: a vocoder-based high-quality speech synthesis system for real-time applications," *IEICE Trans. on Information and Systems*, vol. 99, no. 7, pp. 1877–1884, 2016.

[29] 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," *arXiv preprint arXiv:1804.00015*, 2018, {Last accessed : 15-01-2021}.

[30] 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 *2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, Singapore, 2019, pp. 449–456.

[31] X. Liu, D. Cao, and K. Yu, "Binarized lstm language model," in *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, 2018, pp. 2113–2121.