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

Voice Conversion Based Data Augmentation to Improve Children's Speech Recognition in Limited Data Scenario

*S. Shahnawazuddin†, Nagaraj Adiga‡, Kunal Kumar†, Aayushi Poddar† and Waquar Ahmad**

†Department of Electronics and Communication Engineering, NIT Patna, India
‡Department of Computer Science, University of Crete, Greece
*Department of Electronics and Communication Engineering, NIT Calicut, India
s.syed@nitp.ac.in, nagaraj@csd.uoc.gr, waquar@nitc.ac.in

Abstract

Automatic recognition of children's speech is a challenging research problem due to several reasons. One among those is unavailability of large amounts of speech data from child speakers to develop automatic speech recognition (ASR) systems employing deep learning architectures.Using a limited amount of training data limits the power of the learned system. To overcome this issue, we have explored means to effectively make use of adults' speech data for training an ASR system. For that purpose, generative adversarial network (GAN) based voice conversion (VC) is exploited to modify the acoustic attributes of adults' speech making it perceptually similar to that of children's speech. The original and converted speech samples from adult speakers are then pooled together to learn the statistical model parameters. Significantly improved recognition rate for children's speech is noted due to VC-based data augmentation. To further enhance the recognition rate, a limited amount of children's speech data is also pooled into training. Large reduction in error rate is observed in this case as well. It is worth mentioning that GAN-based VC does not change the speaking-rate. To demonstrate the need to deal with speaking-rate differences we report the results of time-scale modification of childrens speech test data.

Index Terms: Children's ASR, out-of-domain data augmentation, voice conversion, generative adversarial network.

1. Introduction

The demand for speech-based smart devices is increasing day by day. There are many user applications that employ speech-based interface to take commands [1, 2]. Since such devices are expected to be used by anyone, the embedded automatic speech recognition (ASR) system should be highly robust towards speaker-dependent acoustic variations. The speaker-dependent acoustic attributes vary with age and gender of the speaker. Collecting representative speech data that capture all kinds of speaker-dependent acoustic variations is a very challenging task. When the training data insufficiently represents the testing scenario, the recognition performance is known to degrade significantly. Therefore, in such cases, researchers generally resort to synthetically creating more data by simulation and augment the original data in order to supply the missing acoustic attributes. For example, data augmentation based on vocal tract length perturbation was studied in [3] to robustly model the variations in acoustic attributes resulting from differences in the geometry of vocal organs.

Motivated by earlier studies, we have explored out-of-domain data augmentation in this work to enhance the recognition performance of children's ASR system when the domain-specific data is limited. In this regard, two different cases are studied. In CASE-I, sufficient amount of children's speech data for training an ASR system is assumed to be unavailable. The reason behind studying this scenario is that, most of the children's speech corpora are expensive and limited in terms of hours of data available. On the other hand, large amount of adults' speech data are freely available (e.g., TED-LIUM [4] and Librispeech [5]). Furthermore, there are a large number of low resource languages where the amount of speech data from adult as well as child speakers is limited. To deal with unavailability of training data from children, voice-conversion-based out-of-domain data augmentation is explored wherein the acoustic attributes adults' speech are modified using a cycle-consistent generative adversarial network (GAN) [6]. The GAN is employed to learn a mapping from adults' speech to the children's speech. Consequently, the modified adults' speech samples become perceptually similar to those of children's speech as observed during our listening tests. The modified data is then merged with the original one from adult speakers. Next, an ASR system is trained on the augmented data using deep learning architectures [7]. By pooling modified adults' speech into training, the ASR system is able to learn some of the dominant acoustic attributes of children's speech. Subsequently, enhanced recognition performances are noted on transcribing speech data from children. At the same time, there is no significant change in the recognition performance when adults' speech is transcribed.

In CASE-II of out-of-domain data augmentation studied in this work, only a limited amount of children's speech is assumed to be available for training. Using a limited amount of training data to train an ASR system employing deep learning architecture limits the power of the learned system. Therefore, children's speech is mixed with original and voice converted adults' data so that the ASR system may generalize well for both groups of speakers. On transcribing children's speech, improved speech recognition accuracy is noted even in this case. It is worth mentioning here that, voice conversion through cycle GAN does not change the speaking-rate (SR). Differences in SR between adults' and children's speech is an acoustic mismatch factor that adversely affects the recognition rate. To overcome this shortcoming, we have employed optimal time scaling of children's test speech prior to decoding. Further reduction in error rates are noted due to speaking-rate adaptation of test data. The experimental evaluations presented is this work demonstrate the effectiveness of VC-based data augmentation as well as the combination of speaking-rate adaptation with data augmentation.

The rest of this paper is organized as follows: In Section 2, a brief discussion on cycle GAN is presented along with the

http://dx.doi.org/10.21437/Interspeech.2020-1112

description of the explored data augmentation schemes. In Section 3, the experimental evaluations are presented. Finally, the paper is concluded in Section 4.

2. Voice-conversion-based data augmentation

Voice conversion (VC) has been applied to various tasks such as text-to-speech synthesis, speaking assistance, speech enhancement, and pronunciation conversion [6]. VC is basically intended towards modifying non- or para-linguistic information of speech. At the same time, the linguistic information is preserved. In this paper, we have exploited VC for improving children's ASR when the available domain specific data is limited. For that purpose, VC is applied on adults' speech to synthetically generate speech data with acoustic attributes similar to those of child speakers. The synthetically generated data is then pooled into training in order to improve the recognition performance. In the following subsection, a discussion on VC through cycle consistent GAN is presented. This is followed by the description of proposed approaches for effectively exploiting VC for out-of-domain data augmentation in order to improve children's speech recognition.

2.1. Voice conversion through cycle GAN

GANs are popular generative models that were initially used in image processing. GANs consist of a pair neural networks: the generator (G) and the discriminator (D). The G network learns the forward mapping function ($G_{S->T}$) from source ($s \in S$) to target ($t \in T$) data. On the other hand, the D network classifies whether the generated output is real or fake. To learn the mappings, adversarial loss is used, which is given by:

$$L_{Adv}(G_{S->T}, D_T) = \mathbb{E}_{t\sim p_{data}(t)}[\log(D_T(t))] + \mathbb{E}_{s\sim p_{data}(s)}[\log(1 - D(G_{S->T}(s)))]. \quad (1)$$

GANs have been reported to produce impressive results in image processing. Therefore, they have been subsequently applied to other fields such as speech and video processing. Kaneko *et al.* [6] proposed a voice conversion method using cycle consistent GAN (Cycle GAN). In this method, apart from learning a source to target data forward mapping function, an inverse mapping function ($G_{T->S}$) from target to source data is also learned. For this purpose, two additional loss functions were defined, namely, cycle consistency and identity mapping loss, respectively. The cycle consistency loss is given as:

$$L_{Cyc}(G_{S->T}, G_{T->S}) = \mathbb{E}_{s\sim p_{data}(s)}[\| G_{T->S}(G_{S->T}(s)) - s \|_1] + \mathbb{E}_{t\sim p_{data}(t)}[\| G_{S->T}(G_{T->S}(t)) - t \|_1] \quad (2)$$

L_{Cyc} provides consistency between the contextual information of data s and t. Further, to preserve the linguistic information, additional identity mapping loss is defined as follows:

$$L_{Id}(G_{S->T}, G_{T->S}) = \mathbb{E}_{t\sim p_{data}(t)}[\| G_{S->T}(t) - t \|_1] + \mathbb{E}_{s\sim p_{data}(s)}[\| G_{T->S}(s)) - s \|_1]. \quad (3)$$

The overall objective function is written with penalty factors λ_{Cyc} and λ_{Id}:

$$L_{Full} = L_{Adv}(G_{S->T}, D_T) + L_{Adv}(G_{T->S}, D_S) + \lambda_{Cyc}L_{Cyc}(G_{S->T}, G_{T->S}) + \lambda_{Id}L_{Id}(G_{S->T}, G_{T->S}). \quad (4)$$

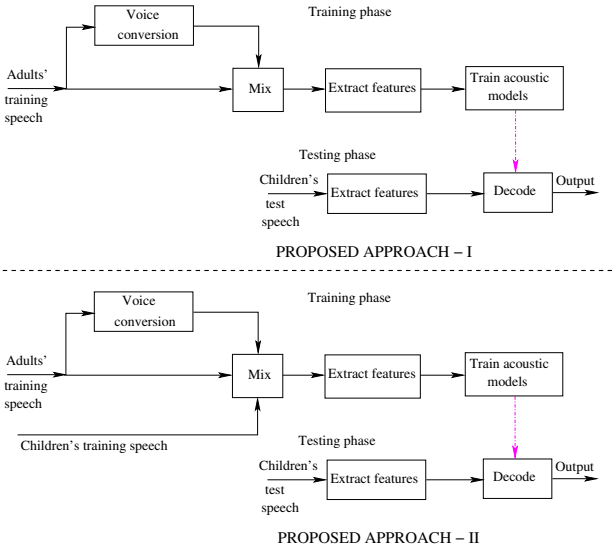

Figure 1: *Proposed schemes for improving children's ASR exploiting voice-conversion-based out-of-domain data augmentation.*

Further, strided convolutional neural networks (CNN) are used to construct both the discriminator and generator networks. After each layer, gated linear units (GLUs) activation function is used as non-linearity [8]. It is reported in literature that Cycle GAN yields better results with GLUs while training the generator networks [6]. In this work, Cycle GAN has been used for voice conversion using the original python code [6] available online.

2.2. Approaches for out-of-domain data augmentation

Generally, a large amount of speech data is used to learn the statistical model parameters in order to enhance robustness towards speaker-dependent acoustic variations. However, in the context of children's speech, there is a paucity of publicly available database. Unlike hundreds of hours of speech data available from adult speakers, most of the publicly available children's speech corpora contain only a few tens of hours of data [9, 10]. At the same time, those are expensive propriety items. In addition to that, there are a large number of low resource languages having limited amount of data from both adult as well as children. As a result, training a robust ASR system for children using state-of-the-art approaches is very challenging. Deep learning architectures that are used in ASR involve huge number of parameters [7]. Using a limited amount of training data limits the power of the learned system. Therefore, out-of-domain data has been employed in earlier works for improving children's speech recognition [11, 12, 13, 14].

As mentioned earlier, when the training data does not sufficiently represent the targeted acoustic conditions, synthetically generated data is augmented with the original one. As a consequence of that, the ASR system is able to robustly learn those missing targeted acoustic attributes. Similarly, in the absence of children's speech or when the available data is limited, one can resort to out-of-domain data augmentation. The out-of-domain-data used in this study is from adult speakers. Two different data augmentation scenarios are explored and those are summarized in Fig. 1. In CASE-I, it is assumed that only adults' speech is available for training the ASR system. However, a very small amount children's speech data is available so that a GAN can

Table 1: *Specifications of the speech corpora used in this work.*

Corpus	WSJCAM0		PF-STAR	
Language	British English		British English	
Name of data set	AD_TR	AD_TS	CH_TR	CH_TS
Purpose	Training	Testing	Training	Testing
Speaker group	Adult	Adult	Child	Child
No. of speakers	92	20	122	60
Age group	>18 years	>18 years	4-14 years	4-14 years
No. of words	132,778	5,608	46,974	5,067
Duration (hrs.)	15.5	0.6	8.3	1.1

Table 2: *Baseline WERs for AD_TS and CH_TS with respect to TDNN systems trained using either CH_TR or AD_TR or a mix of both CH_TR and AD_TR.*

Data used for training	WER (in %)	
	CH_TS	AD_TS
CH_TR	9.40	32.14
AD_TR	19.26	5.63
AD_TR + CH_TR	7.41	5.53

be trained. It is worth mentioning here that, acoustic attributes of children's and adults' speech are starkly different [15, 14]. Therefore, poor speech recognition accuracy is obtained when children's speech is transcribed using an adult data trained ASR system. To overcome this issue, the acoustic attributes of adults' speech are modified by learning a mapping from adults to children domain using a cycle consistent GAN. The original and voice converted adults' speech samples are then pooled together and statistical model parameters are trained. As a consequence the speech recognition accuracy is improved. In CASE-II, on the other hand, a limited amount of children's speech data is assumed to be available. As mentioned earlier, learning a large number of network parameters using a limited amount of children's speech data limits the power of the developed ASR system. In order to address this shortcoming, both adults' and children's speech data along with the voice converted adults' data are pooled together before learning the network parameters as shown in Fig. 1. Consequently, the error rates for children's speech recognition decrease significantly. The experimental evaluations presented in paper substantiate these claims.

3. Experimental evaluation

In this section, experimental evaluations demonstrating the effectiveness of VC-based data augmentation are presented.

3.1. Experimental setup and baseline performances

For the experimental evaluations presented in this paper, Kaldi toolkit [16] was used. Two different *British English* corpora were employed for experimental evaluations. The chosen databases are very similar in terms of accent of speakers and recording conditions. More details of the two corpora and the training and test sets derived from them are given in Table 1. Wide-band speech sampled at 16kHz rate has been employed for system development and evaluation in this study.

For front-end speech analysis, overlapping Hamming windows of length 20 ms and frame-rate of 100 Hz were employed. For MFCC feature extraction, a 40-channel Mel-filterbank was used for spectral warping before extracting 13-dimensional base MFFC features. The base features were time-spliced using a context size of $\pm$ 4. This was followed by dimensionality reduction and de-correlation using linear discriminant analysis and maximum likelihood linear transformation to derive 40-dimensional feature vectors. Cepstral mean and variance normalization was applied to all the features. Robustness towards speaker-dependent variations was further enhanced through feature-space maximum-likelihood linear-regression (fMLLR).

Three-states hidden Markov models (HMM) were used for statistically modeling context-dependent cross-word triphones with the maximum number of senones being fixed at 2000. The observation densities for the HMM states were then generated using time-delay neural networks (TDNN) [17, 18]. The initial effective learning-rate was chosen as 0.015 while the final effective learning-rate was 0.002. The TDNN architecture consisted of 5 hidden layers with splicing indices being "0", "$-2, -2$", "0", "$-4, -4$" and "0". The ReLU dimension was chosen as 250. Prior to learning the TDNN parameters, the fMMLR-normalized feature vectors were spliced again considering a context size of $\pm$ 4. Minibatch size of 512 was used. While decoding CH_TS, a domain-specific 1.5k bi-gram language model (LM) was employed. The out-of-vocabulary (OOV) rate and perplexity of the employed LM with respect to CH_TS were 1.20% and 95.8, respectively. Further, a lexicon consisting of $1,969$ words including the pronunciation variations was employed. For decoding AD_TS, the standard MIT-Lincoln 5k Wall Street Journal bi-gram LM was used. The lexicon employed in this case consisted of $5,850$ words including the pronunciation variations. For evaluating the recognition performances, the word error rate (WER) metric is used.

Initially, three different ASR systems were trained using CH_TR, AD_TR and a mix of both, respectively. The WERs for the two test sets with respect to those ASR systems are given in Table 2. Due of aforementioned problems of data scarcity and acoustic mismatch, the WER for AD_TS is extremely poor when only children's speech is used for system development. On the other hand, when AD_TR set is used for training, WER for CH_TS is still comparatively degraded. On pooling AD_TR and CH_TR, the WERs are reasonably good for both the test sets. These results show that out-of-domain-data augmentation significantly improves children's ASR.

3.2. Evaluating CASE-I of VC-based data augmentation

To implement CASE-I of data augmentation, the acoustic attributes of AD_TR were modified using Cycle GAN in order to render the speech samples perceptually similar to those of children's speech. For learning the model parameters of the Cycle GAN, 10 minutes of data from each group was used. A mapping from adult to child domain as well as from child to adult domain was learned. The number of epochs employed in training the GAN parameters was equal to 5000. The modified adult data training set is referred to as AD_TR-VC in this work. Next the TDNN network parameters were trained after pooling AD_TR and AD_TR-VC. The effect of pooling the two training sets is demonstrated by the WERs enlisted in Table 3. On comparing the baseline WER for CH_TS with that obtained using voice conversion, a relative reduction of 31.4% is noted. At the same time, there is no degradation in the WER for AD_TS.

Table 3: *WERs for AD_TS and CH_TS demonstrating the effect of CASE-I of VC-based out-of-domain data augmentation.*

Data used for training	WER in (%)	
	AD_TS	CH_TS
AD_TR	5.63	19.26
AD_TR + AD_TR-VC	5.63	**13.22**
Percentage relative reduction	0	31.4

Table 4: *WERs for AD_TS and CH_TS comparing VTLP- and SP-based data augmentation with CASE-I of proposed approach.*

Technique employed	WER in (%)	
	AD_TS	CH_TS
Baseline	5.63	19.26
VTLP	5.66	15.17
SP	5.39	18.47
VC	5.63	**13.22**

3.2.1. Comparison with VTLP and speed perturbation

Vocal tract length perturbation (VTLP) [3, 19] and speed perturbation (SP) [20] are two of the dominant approaches for data augmentation reported in literature. In order to further substantiate the effectiveness of the proposed approach, it was compared with VTLP and SP. In this study, we have used the the same warping factors for all the utterances in AD_TR. It is well known that, the formant frequencies are higher in the case of children's speech. Hence, warping factor was varied from 1.1 to 1.14 in steps of .01 to create 3-fold training set. The best case performance was obtained when the 3-fold training set was obtained by using warping factors 1.12 and 1.14. In the case of SP, two copies of AD_TR were derived by modifying the speed by a factor of 0.9 and 1.1 of the original rate leading to the creation of 3-fold training set. To implement SP, the default pipeline in Kaldi was used. The WERs for those experiments are summarized in Table 4. On comparing the WERs obtained through VTLP- and SP-based data augmentation, the proposed approach is noted to be better for children's speech test set.

3.2.2. Speaking-rate adaptation

Even though voice-converted adults' speech utterances sound very similar to children's speech, the speaking-rate remains unchanged. Earlier works on children's speech have noted that the speaking-rate for children is much slower than that for the adults' [15]. This mismatch in speaking-rate leads to degraded speech recognition accuracy. To overcome this shortcoming, we resorted to explicit time-scale modification (TSM) of children's speech test set. In this work, we have employed a recently proposed TSM approach based on fuzzy classification of spectral bins [21]. For each of the test utterances, the optimal scaling factor was selected using a two-pass maximum likelihood grid search. The modified test data was then decoded using the trained acoustic models. The WER obtained by performing speaking-rate adaptation of the test data is given in Table 5. As evident from the tabulated WERs, a relative improvement of 21.1% is obtained by combining VC-based data augmentation with time-scaling of the test utterances.

Table 5: *WERs for CH_TS demonstrating the effect of time scaling (TS) of test utterances in CASE-I of data augmentation.*

Technique employed	WER in (%)
VC	13.22
VC + TS	**10.43**
Percentage relative reduction	21.1

Table 6: *WERs for AD_TS and CH_TS demonstrating the effect of CASE-II of VC-based out-of-domain data augmentation.*

Data used for training	WER in (%)	
	AD_TS	CH_TS
AD_TR + CH_TR	5.53	7.41
AD_TR + CH_TR + AD_TR-VC	**5.14**	**6.49**
Percentage relative reduction	7.1	12.5

Table 7: *WERs for CH_TS demonstrating the effect of combining time scaling (TS) with CASE-II of VC-based data augmentation.*

Technique employed	WER in (%)
VC	6.49
VC + TS	**5.99**
Percentage relative reduction	7.7

3.3. Evaluating CASE-II of VC-based data augmentation

In CASE-II of VC-based out-of-domain data augmentation, as illustrated by Fig. 1, AD_TR, CH_TR and AD_TR-VC datasets were pooled together and TDNN-based acoustic models were trained. On pooling the three training sets, the WERs are observed to decrease significantly not only of CH_TS test set but also for AD_TS set as evident from Table 6. Further reduction in WER for CH_TS is noted on combining proposed CASE-II of VC-based out-of-domain data augmentation with time-scaling of the test utterances. The WER for that study is enlisted in Table 7.

4. Conclusion

Voice-conversion-based out-of-domain data augmentation has been explored in this study in order to improve children's speech recognition when the available domain-specific data is limited. For that purpose, the acoustic attributes of adults' speech samples are modified using cycle consistent GAN so that they become perceptually similar to children's speech. Significantly reduced WERs are obtained by VC-based out-of-domain data augmentation. Since cycle-GAN-based voice conversion does not affect the speaking-rate, the acoustic mismatch induced by differences in speaking-rate is overcome by explicit time-scaling of the children's speech test set. Added improvements are noted by combining data augmentation with speaking-rate adaptation of test data.

5. References

[1] Amazon:, "Amazon lex is a service for building conversational interfaces," https://aws.amazon.com.

[2] Google:, "Google assistant," https://assistant.google.com/.

[3] N. Jaitly and G. E. Hinton, "Vocal tract length perturbation (VTLP) improves speech recognition," in *Proc. ICML*, vol. 117, 2013.

[4] A. Rousseau, P. Delglise, and Y. Estve, "TED-LIUM: an automatic speech recognition dedicated corpus," in *Proc. LREC*, 2012, pp. 125–129.

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

[6] T. Kaneko and H. Kameoka, "Parallel-data-free voice conversion using cycle-consistent adversarial networks," *arXiv preprint arXiv:1711.11293*, 2017.

[7] D. Yu and L. Deng, *Automatic Speech Recognition: A Deep Learning Approach*. Springer, 2016.

[8] Y. N. Dauphin, A. Fan, M. Auli, and D. Grangier, "Language modeling with gated convolutional networks," in *Proc. 34th International Conference on Machine Learning*, 2017, pp. 933–941.

[9] K. Shobaki, J.-P. Hosom, and R. A. Cole, "The ogi kids^2 speech corpus and recognizers," in *INTERSPEECH*, 2000.

[10] A. Batliner, M. Blomberg, S. D'Arcy, D. Elenius, D. Giuliani, M. Gerosa, C. Hacker, M. Russell, and M. Wong, "The PF_STAR children's speech corpus," in *Proc. INTERSPEECH*, 2005, pp. 2761–2764.

[11] H. Liao, G. Pundak, O. Siohan, M. K. Carroll, N. Coccaro, Q. Jiang, T. N. Sainath, A. W. Senior, F. Beaufays, and M. Bacchiani, "Large vocabulary automatic speech recognition for children," in *Proc. INTERSPEECH*, 2015, pp. 1611–1615.

[12] J. Fainberg, P. Bell, M. Lincoln, and S. Renals, "Improving children's speech recognition through out-of-domain data augmentation," in *Proc. INTERSPEECH*, 2016.

[13] S. Shahnawazuddin, A. Dey, and R. Sinha, "Pitch-adaptive front-end features for robust children's ASR," in *Proc. INTERSPEECH*, 2016.

[14] R. Sinha and S. Shahnawazuddin, "Assessment of pitch-adaptive front-end signal processing for children's speech recognition," *Computer Speech & Language*, vol. 48, no. Supplement C, pp. 103 – 121, 2018.

[15] M. Gerosa, D. Giuliani, S. Narayanan, and A. Potamianos, "A review of ASR technologies for children's speech," in *Proc. Workshop on Child, Computer and Interaction*, 2009, pp. 7:1–7:8.

[16] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz, J. Silovsky, G. Stemmer, and K. Vesely, "The Kaldi Speech recognition toolkit," in *Proc. ASRU*, December 2011.

[17] A. H. Waibel, T. Hanazawa, G. E. Hinton, K. Shikano, and K. J. Lang, "Phoneme recognition using time-delay neural networks," *IEEE Trans. Acoustics, Speech, and Signal Processing*, vol. 37, pp. 328–339, 1989.

[18] V. Peddinti, D. Povey, and S. Khudanpur, "A time delay neural network architecture for efficient modeling of long temporal contexts," in *Proc. INTERSPEECH*, 2015.

[19] X. Cui, V. Goel, and B. Kingsbury, "Data augmentation for deep neural network acoustic modeling," in *Proc. ICASSP*, May 2014, pp. 5582–5586.

[20] T. Ko, V. Peddinti, D. Povey, and S. Khudanpur, "Audio augmentation for speech recognition," in *Proc. INTERSPEECH*, 2015, pp. 100–104.

[21] E.-P. Damskägg and V. Välimäki, "Audio time stretching using fuzzy classification of spectral bins," *Applied Sciences*, vol. 7, no. 12, 2017.