USING MODIFIED ADULT SPEECH AS DATA AUGMENTATION FOR CHILD SPEECH RECOGNITION

Zijian Fan, Xinwei Cao, Giampiero Salvi, Torbjørn Svendsen

Norwegian University of Science and Technology
Department of Electronic Systems

ABSTRACT

Data augmentation is a technique which enhances the size and quality of training data such that deep learning or machine learning models can achieve better performance. This paper proposes a novel way of applying data augmentation for child speech recognition in the low data resource scenario. Data augmentation is achieved by modifying existing adult speech signals. The procedure consists of two main parts, resampling, and time scaling. The experiment involves both speech from children aged from kindergarten to grade 10, and adults' speech. We test the proposed method using both a TDNN-HMM and a GMM-HMM acoustic model. The results show that the proposed data augmentation scheme achieves a relative 7.95% reduction of WERs compared with 4.56% relative reduction when using a traditional bilinear frequency warping approach.

***Index Terms*—** Data augmentation, children's speech recognition, TDNN-HMM, GMM-HMM

1. INTRODUCTION

The need for automatic child speech recognition (ASR) systems has grown rapidly because child ASR plays an important role in the interaction between children and modern digital devices. Even though remarkable progress has been achieved in the area of adult ASR systems, the performance of child ASR systems degrades considerably [1]. A reasonable explanation is that there exist several mismatches between children's speech and adults' speech with respect to acoustic and linguistic characteristics [1, 2, 3]. These mismatches are caused by physiological and developmental differences [4], and are augmented by insufficient linguistic knowledge applicable to remedy these differences.

Previous studies have investigated several methods to improve the performance of child ASR systems. Vocal Tract Length Normalization (VTLN) [5] attempts to reduce the inter-speaker variability and normalize the speech spectra space by frequency warping, using a maximum log-likelihood method. It has been applied to child ASR systems [1] showing some improvements. Serizal et al [6] applied VTLN into a hybrid DNN - HMM ASR system and achieved 20% relative improvement of phone error rate (PER). Azizi et al [7] applied VTLN into an isolated word recognition system, and found out the improvement of adult model was more than child model after using VTLN.

Speaker adaptive training [8] is another useful technique which has been applied to child ASR [1]. It aims to reduce the inter-speaker variability by modeling speaker characteristics as linear transformations of the speaker independent acoustic parameters. Sanand et al [9] found out that there was a significant performance improvement when applying SAT together with VLTN to a child ASR model.

Recently, more and more researchers try to build child ASR systems based on DNN. Liao et al [10] built a large vocabulary child ASR system which was used in the YouTube Kids mobile application. They found that many techniques, including VTLN, were not effective for their task. The major contribution towards better results was the new collected child speech corpus which has the same size as adult speech corpus. It also revealed that lack of publicly available child speech databases holds back the development of child ASR.

Instead of collecting large scale child speech databases, an effective alternative is to use a technique called data augmentation [11, 12, 13, 14]. Many methods have been proposed for data augmentation. Some methods seek to increase the variability of spectrum space by generating artificial spectra such as vocal tract length perturbation (VTLP) [15] and SpecAugment [16], while others seek to modify formant information based on linear prediction coding (LPC) [13, 14, 17]. However, there is still a need for better data augmentation methods.

In this study, a novel data augmentation method is proposed. It involves temporal and spectral modifications when linearly shifting the spectra. Then it involves local perturbations at the phase matching stage. A key difference between the proposed method and past methods is its strategy to discard the high frequency part information. We compared our method with other data augmentation methods, and analyzed the results with respect to age and gender. The proposed method consistently performed best.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No.857897.

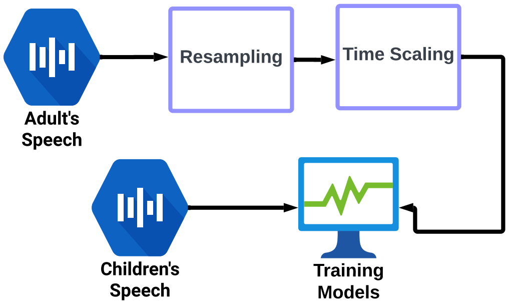

Fig. 1. A block diagram of proposed method

2. METHODS

Traditional spectral modification based methods seek to find an optimal mapping between original frequencies and warped frequencies. In e.g., VTLN this mapping is piece-wise linear while the mapping is continuous when using a bilinear transform. Sometimes, the time domain modification, for example speed perturbation [18], can also improve the recognition performance. With this in mind, we proposed a method to linearly shift the spectra and discard the high frequency information while performing speed perturbation. Fig. 2 summarizes the proposed data augmentation method.

2.1. Resampling

A time discrete sequence, $x[nT_s]$, sampled with a sampling period T_s, can be resampled to a new sampling period T_d through digital interpolation and/or decimation. Downsampling $x[nT_s]$ when $T_d > T_s$, yields $x[nT_d]$.

We propose to process the resampled signal $x[nT_d]$ at the original sampling frequency f_s. This results in a linear frequency warp by a ratio of f_s/f_d, including loss of the high frequency region ($f > f_d/2$) and a speaking rate speed-up by a ratio of f_s/f_d, where f_d is the downsampling rate. Whereas the frequency warp and high-frequency loss is desired, the change in speaking rate is not and needs to be addressed.

2.2. Time-scaling

The speaking style of children varies considerably. One significant difference between children's speech and adult's speech is that children's speaking rate is slower and they have more pauses while speaking. The operations in section 2.1 speed up the speech signals by a ratio of f_s/f_d. We, therefore, slow down the signals and include the variability of speaking rate in the augmented data. We use time scaling to achieve this because it tends to change the speaking rate without alerting the fundamental frequency.

The basic idea is to modify the spectral information and resynthesize the signals using DFT and the overlap-add method. A sequence of signals $x_m[nT_s]$ is divided into M overlapping frames where each frame contains N concatenated digital samples shifted L_a samples from the previous frame. The time duration can be easily extended by changing L_a into a larger integer L_s. This is equivalent to playing the audio $r = L_a/L_s$ times slower. In order to reduce speech discontinuities and glitches, a phase vocoder [19] and a Hamming window is be applied to each frame.

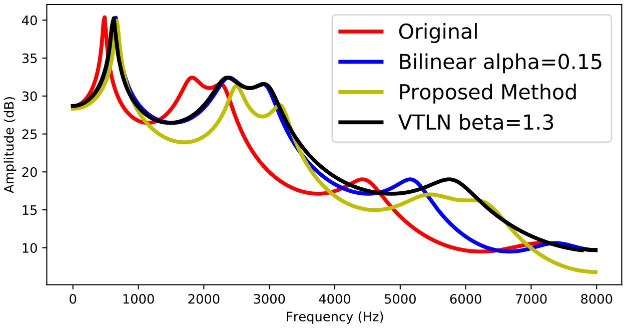

Fig. 2. Spectral envelopes of a segment of LibriSpeech audio with (a) no modification, (b) bilinear warping, $\alpha = 0.15$, (c) the proposed method, $f_s = 16$kHz, $f_d = 12$kHz, $r = 0.75$ (d) VTLN, $\beta = 1.3$

The advantage of the phase vocoder is phase matching. After computing the DFT of i-th frame $(X_a[k])_i$, we kept its amplitude, and estimated the new phase $(\phi_s[k])_i$ based on the information of the phase of previous frame $(\phi_s[k])_{i-1}$:

$$(\Delta\phi_a[k])_i = (\phi_a[k])_i - (\phi_a[k])_{i-1} - L_a\omega_k \tag{1}$$

$$(\Delta_p\phi_a[k])_i = \mod[((\Delta\phi_a[k])_i + \pi), 2\pi] - \pi \tag{2}$$

$$(\phi_s[k])_i = (\phi_s[k])_{i-1} + L_s \cdot (\Delta_p\phi_a[k])_i \tag{3}$$

After phase modification, we perform an inverse DFT of $X_s = |X_a|e^{\phi_s}$, and resynthesize $x_t[nT_s]$ using the overlap-add method. If $L_a/L_s = f_d/f_s$, the original speaking rate is restored, with the exception of local perturbations caused by the phase matching. In order to implement time scaling, we used a speech processing library called Rubberband [20].

2.3. Frequency Warping

In order to compare with spectrum based data augmentation method, we also implemented other frequency warping methods, namely bilinear frequency warping and VTLN. The bilinear frequency warping is defined by

$$\omega^{warp} = \omega + 2\arctan\left(\frac{\alpha\sin(\omega)}{1-\alpha\cos(\omega)}\right) \tag{4}$$

where α is called the warping factor. For adult-to-child warping we limit α to be in the range $[0.05, 0.15]$.

For piece-wise linear warping implementation, we used the Kaldi VTLN tools, and we also limited the VTLN warping factor β to the range $[1.1, 1.3]$. Fig.2 illustrates differences introduced by the frequency warping methods.

Age	Male		Female	
	#Sub	#Utt	#Sub	#Utt
6	2	85	7	327
7	5	152	16	1003
8	11	943	18	1166
≥ 9	6	566	11	938

Table 1. Speaker and utterance information of CMU kids speech corpus

3. EXPERIMENT SETUPS

3.1. Databases

Three databases containing both adult and child speech corpus were used. All of them are American English. The LibriSpeech ASR corpus [21] was used as the adult speech part. It contains speech from 1210 male and 1128 female adults reading fiction books. The speech was sampled at 16 kHz. We specifically used the training set of female speakers because female voice has a higher pitch than male voice in general. The female training part contains roughly 300 hours of clean speech.

The first child speech database was the CMU Kids Corpus [22]. It contains read speech from 76 children aged from six to eleven. There were 24 male and 52 female speakers. There were 5180 utterances in all sampled at 16 kHz. Table 1 summarizes the age and utterance information. 3670 (70%) utterances of the CMU kids corpus were used for training. and 1510 (30%) utterances were used for testing.

The second child speech database was the CSLU Kids Speech Corpus [23]. It contains read and spontaneous speech from approximately 1100 children from kindergarten to grade 10. Only a small portion of read speech from this corpus was used for training. There were 605 male and 512 female speakers.

Test data were solely from the CMU kids corpus data and consisted of the 1510 (30%) utterances not used for training.

3.2. Set-ups

All the experiments were performed using the Kaldi speech toolkit-based recipes for ”cmu_cslu_kids”. The baseline features for the following experiments were Mel-frequency cepstral coefficients (MFCCs). They were extracted using default Kaldi setting except for one difference, we used the log energy instead of C_0. For normalization, cepstral feature-space maximum likelihood linear regression (fMLLR) was used and the fMLLR transformations for the training and test data were generated using the SAT. There were two types of acoustic models explored in this study, GMM-HMM and TDNN-HMM. The LibriSpeech 3-gram model was used as the language model.

Six different combinations of databases were set up for training, shown in Table 2. They are (a) CMU data; (b) CMU data augmented with unmodified LibriSpeech data; (c) CMU data augmented with CSLU data; (d) CMU data augmented with warped LirbiSpeech data using VTLN; (e) CMU data augmented with warped LibriSpeech data using bilinear warping method; (f) CMU data augmented with warped LibriSpeech data using proposed method. The recognition performance was evaluated on the CMU test data. For data augmentation, the number of utterances was fixed as 2000 for each combination mentioned before. All warping methods were only applied on LibriSpeech data. For traditional methods, the choice of warping factor are mentioned in section 2.4, while for the proposed method, the re-sampling frequency f_d was randomly selected from a list $[10500, 12000, 13500, 14500, 16000]$. Utterances from same speakers shared the same warping factor or f_d. Besides, speed perturbation factor r is randomly selected in a range [0.55, 0.85].

Training set	ASR System	
	GMM	TDNN
CMU	32.45	21.51
CMU + LibriSpeech	34.75	21.88
CMU + CSLU	34.17	21.76
CMU + VTLN warping	36.00	21.99
CMU + Bilinear warping	35.07	20.53
CMU + Proposed method	34.31	19.80

Table 2. Word error rates (WERs) of GMM-HMM, TDNN-HMM ASR systems training on different combinations of databases. Each training combination contains same utterances of CMU kids.

4. RESULTS AND DISCUSSION

4.1. Overall Results

Table 2 shows the effect of different data augmentation methods on the word error rates (WERs) for both GMM-HMM and TDNN-HMM ASR systems. The first row shows the WERs of the baseline ASR system trained without augmentation. The next two rows show the WERs of the systems augmented with unmodified adult and child speech training data, respectively. The last three rows show the WERs of the systems augmented with modified adult training speech.

As all the systems were trained with limited child speech data, it was expected that the WERs should be improved with more training data. However, for the GMM-HMM systems, none of them outperformed the baseline model. One possible explanation is that these methods increased the feature variations which will cause GMM based systems to learn non-necessary variations.

As for TDNN-HMM ASR systems, the performance was improved in the last two rows. VTLN was not effective be-

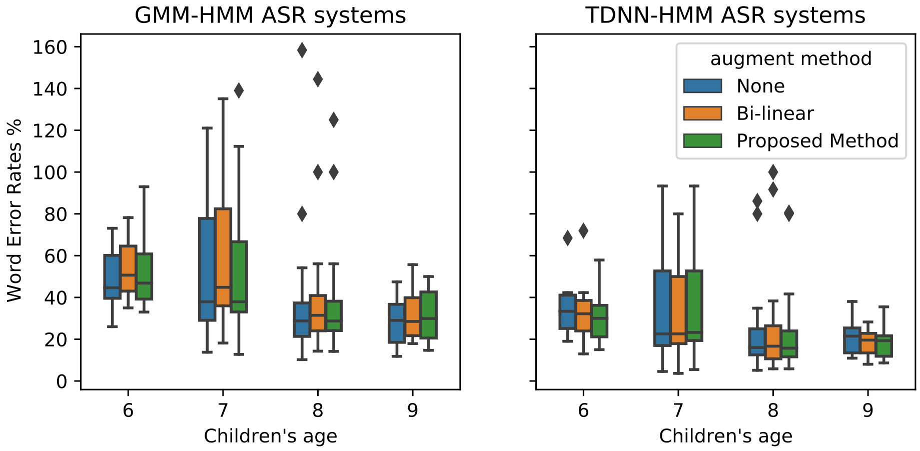

Fig. 3. Word error rates (WERs) of GMM-HMM, TDNN-HMM ASR systems grouped by children's age. Diamonds indicate outliers.

cause the warp factor was estimated using poor hypotheses. Our method achieved a relative 7.95% reduction of WERs than baseline model, which is better than the 4.56% relative reduction when using a traditional bilinear method.

4.2. Analysis by age

To better compare our method with the traditional bilinear warping method, we investigated how the WERs were distributed with respect to age. The age groups were set up according to Table. 1. As shown in Fig. 3, the performance of GMM-HMM ASR systems degraded in all age groups. It is consistent with the results in Table 2. Our method produced more outliers, but it is more reliable in the 6 and 7 year-old groups. For most age groups, both data augmentation methods had a smaller median and a smaller 3rd quartile. It means if the number of test data is equally distributed with respect to speakers, both methods could still help to improve the performance of TDNN-HMM child ASR systems. To our surprise, the most significant improvement happened in the 6-year-old age group, however the number of the utterances in this age group is the smallest. As a consequence, this age group contributed very little to the overall performance improvement. The second significant improvement happened in the age group who is at least 9 years old. This was expected because we already know that the mismatch between children's speech and adults' speech will decrease with age in general. One big advantage of the bilinear method is that all whiskers were smaller than with our proposed method, however our method had less outliers. It implies that the high frequency information loss is not always a benefit.

Training Set	Gender	
	Male	Female
CMU	22.57	20.97
CMU + Bilinear warping	21.44	20.06
CMU + Proposed method	20.92	19.28

Table 3. Word error rates (WERs) of TDNN-HMM ASR systems with respect to gender.

4.3. Analysis by gender

We further investigated the WERs with respect to gender, because the CMU speech database is not a gender-balanced database. As shown in Table 3, our method outperformed the bilinear method for both genders and the improvement of males is more significant than females. One possible explanation is that the percentage of utterance in the 7-8 age group is less for male speakers compared with female speakers.

5. CONCLUSION AND FUTURE WORK

The proposed data augmentation method provides a significant WER reduction for TDNN-HMM based child ASR systems. We showed comparable or better performance than the traditional bilinear warping method. We also analyzed the WER distribution with respect to age, and found out our method is better for most age groups. Future work includes optimizing the choice of down-sampling frequency and shift ratio. Future work also includes evaluating the proposed method using different languages.

6. REFERENCES

[1] V. Bhardwaj, M. T. B. Othman, V. Kukreja, Y. Belkhier, M. Bajaj, B. S. Goud, A. Rehman, M. Shafiq, and H. Hamam, "Automatic Speech Recognition (ASR) Systems for Children: A Systematic Literature Review," *Applied Sciences*, vol. 12, no. 9, pp. 4419, 2022.

[2] L. L. Koenig, J. C. Lucero, and E. Perlman, "Speech production variability in fricatives of children and adults: Results of functional data analysis," *The Journal of the Acoustical Society of America*, vol. 124, no. 5, pp. 3158–3170, 2008.

[3] L. Rumberg, H. Ehlert, U. Lüdtke, and J. Ostermann, "Age-Invariant Training for End-to-End Child Speech Recognition Using Adversarial Multi-Task Learning," *Interspeech 2021*, pp. 3850–3854, 2021.

[4] S. Lee, A. Potamianos, and S. Narayanan, "Acoustics of children's speech: Developmental changes of temporal and spectral parameters," *The Journal of the Acoustical Society of America*, vol. 105, no. 3, pp. 1455–1468, 1999.

[5] L. Lee and R. Rose, "A frequency warping approach to speaker normalization," *IEEE Transactions on Speech and Audio Processing*, vol. 6, no. 1, pp. 49–60, 1998.

[6] R. Serizel and D. Giuliani, "Vocal Tract Length Normalisation Approaches to DNN-Based Children's and Adults' Speech Recognition," *2014 IEEE Spoken Language Technology Workshop (SLT)*, pp. 135–140, 2014.

[7] S. Azizi, F. Towhidkhah, and F. Almasganj, "Study of VTLN Method to Recognize Common Speech Disorders in Speech Therapy of Persian Children," *2012 19th Iranian Conference of Biomedical Engineering (ICBME)*, pp. 246–249, 2012.

[8] T. Anastasakos, J. McDonough, and J. Makhoul, "Speaker adaptive training: a maximum likelihood approach to speaker normalization," *ICASSP*, vol. 2, pp. 1043–1046 vol.2, 1997.

[9] D. R. Sanand and T. Svendsen, "Synthetic speaker models using VTLN to improve the performance of children in mismatched speaker conditions for ASR," *Interspeech 2013*, pp. 3361–3365, 2013.

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

[11] A. Antoniou, A. Storkey, and H. Edwards, "Data Augmentation Generative Adversarial Networks," *arXiv*, 2017.

[12] X. Cui, V. Goel, and B. Kingsbury, "Data Augmentation for Deep Neural Network Acoustic Modeling," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 23, no. 9, pp. 1469–1477, 2015.

[13] A. Johnson, R. Fan, R. Morris, and A. Alwan, "LPC Augment: an LPC-based ASR Data Augmentation Algorithm for Low and Zero-Resource Children's Dialects," *ICASSP*, vol. 00, pp. 8577–8581, 2022.

[14] G. Yeung, R. Fan, and A. Alwan, "Fundamental frequency feature warping for frequency normalization and data augmentation in child automatic speech recognition," *Speech Communication*, vol. 135, pp. 1–10, 2021.

[15] N. Jaitly and G. E. Hinton, "Vocal tract length perturbation (VTLP) improves speech recognition," in *Proc. ICML Workshop on Deep Learning for Audio, Speech and Language*, 2013, vol. 117, p. 21.

[16] V. P. Singh, H. Sailor, S. Bhattacharya, and A. Pandey, "Spectral Modification Based Data Augmentation For Improving End-to-End ASR For Children's Speech," *arXiv*, 2022.

[17] H. K. Kathania, S. R. Kadiri, P. Alku, and M. Kurimo, "A formant modification method for improved ASR of children's speech," *Speech Communication*, vol. 136, pp. 98–106, 2022.

[18] V. Peddinti, D. Povey, and S. Khudanpur, "A time delay neural network architecture for efficient modeling of long temporal contexts," *Interspeech 2015*, pp. 3214–3218, 2015.

[19] J. Laroche and M. Dolson, "Improved phase vocoder time-scale modification of audio," *IEEE Transactions on Speech and Audio Processing*, vol. 7, no. 3, pp. 323–332, 1999.

[20] "Rubber band audio time stretcher library," `https://breakfastquay.com/rubberband/`, Accessed: 2022-09-30.

[21] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: An ASR Corpus Based on Public Domain Audio Books," *ICASSP*, 2015.

[22] M. Eskenazi, J. Mostow, and D. Graff, "The CMU Kids Corpus LDC97S63. Web Download. Philadelphia: Linguistic Data Consortium," 1997.

[23] K. Shobaki, J. Hosom, and R. A. Cole, "The OGI kids^2 speech corpus and recognizers," *ICSLP*, pp. vol. 4, 258–261–0, 2000.