INTERSPEECH 2019
September 15–19, 2019, Graz, Austria

Prosody Usage Optimization for Children Speech Recognition with Zero Resource Children Speech

Chenda Li, Yanmin Qian

MoE Key Lab of Artificial Intelligence
SpeechLab, Department of Computer Science and Engineering
Shanghai Jiao Tong University, Shanghai, China
lichenda1996@sjtu.edu.cn, yanminqian@sjtu.edu.cn

Abstract

Children's speech recognition remains a big challenge for automatic speech recognition. Due to the more difficult process and higher cost on data collection, most current ASR systems are optimized only using lots of adult speech with limited or even none children's speech. Accordingly, the acoustic mismatch between children's and adult speech is the primary reason for the ASR performance degradation when facing children's speech. To overcome this problem, we proposed several approaches to improve children's speech recognition without using any children's speech data. A better utilization strategy on prosody-based features is developed. First, pitch and prosody modification is explored in both training and testing respectively, which can significantly reduce the mismatch between two types of speech. Furthermore, joint-decoding with both the prosody modified speech and the original speech is designed to get a more robust performance on both children's and adult speech. Experiments are evaluated on a Mandarin speech recognition task, with only 400-hour adult speech in the training. The results show that our proposed method can obtain a large gain on children's speech, with relative ∼20% WER reduction compared to the baseline, and also no obvious degradation is observed on the adult speech for the proposed system.

Index Terms: children's speech recognition, pitch feature, prosody feature, joint decoding

1. Introduction

In recent decades, a great number of methods have been proposed for improving the performance of automatic speech recognition (ASR) system[1, 2, 3, 4]. With a large amount of training data and advanced model structure, significant progress has been made in ASR system developing. However, one challenge that still remains for modern ASR systems is children's speech recognition. As far as we know, compared to adult ASR system, fewer efforts have been taken on children's ASR system in previous researches.

One way to improve the ASR systems' performance for children is introducing more children's corpus in training [5]. Since the DNN based ASR systems [6] are driven by data, it is commonly recognized that with a larger amount of training data, the performance of ASR systems is even better. However, most of the public corpora are collected with adult speakers. Children's corpora for ASR training is difficult to be collected, usually smaller than that of adults [7]. Another way is by reducing the acoustic mismatch between children's and adult voice by algorithms. There are some forms of these acoustic mismatches [8, 9, 10]. The acoustic mismatch is mainly because of that children's vocal tract is shorter than that of adults [11, 12, 13, 14]. The mismatch between children's and adult voice leads to the performance degradation when applying the ASR system trained with an adult corpus to the children's speech.

In most practical applications, the costs on time and resources are very large to obtain well-labeled children corpus, especially for some low-resource languages. Accordingly, in this paper, we tried to solve the challenge through the second approach mentioned above, reducing the acoustic mismatch between children's and adult speech using algorithms.

One of the major acoustic mismatches is that children's fundamental frequency is usually higher than that of adults [14]. Focusing on the fundamental frequency mismatch, a prosody modification method is proposed to reduce the mismatch. We perform the prosody modification method in two ways. The first approach is making prosody modification on adult training corpus, make the acoustic features closer to children's. The second approach, on the contrary, is performing prosody modification on children's speech directly when testing.

In practice, the above prosody modification method works well on improving children's speech recognition performance, and a significant improvement in children's speech recognition is obtained. However, this method also leads to performance degradation in adult speech. To overcome this shortcoming and make the system more robust, a joint decoding method is then introduced, which can further improve the system performance. The joint decoding method do not need to retrain the built system, which is flexible and low cost.

This paper is organized as follows. In section 2, the prosody features are introduced for children's speech recognition, including the prosody modification and pitch feature. In section 3, the joint decoding architecture with different prosody modifications is designed. The detailed experimental results and analysis are described in section 4, and conclusions are finally given in section 5.

2. Prosody Feature for Children's Speech

2.1. Motivation

As mentioned in section 1 , the fundamental frequency of children's speech is higher than that of adults. The range of adult fundamental frequency, for male is usually from 85 Hz to 180 Hz, and for female is usually from 165 Hz to 255 Hz [15, 16]. The range of children's fundamental frequency is from about 200 Hz to 350 Hz[17]. Thus, adding prosody-related features into the system may improve system performance. In this paper, two types are explored, including prosody modification by tuning the fundamental frequency, and explicitly using the pitch

Yanmin Qian is the corresponding author.

http://dx.doi.org/10.21437/Interspeech.2019-2659

feature.

2.2. Prosody Modification

The prosody modification[1] procedure that we use can be described as the following steps:

Firstly, by resampling the original audio signal $f(t)$ with factor λ, we get a new signal $f(\lambda t)$. Denote Fourier transform of $f(t)$ as $\hat{f}(\omega)$. Then, the Fourier transform of $f(\lambda t)$ can be presented as $\lambda^{-1}\hat{f}(\lambda^{-1}\omega)$. This resampling procedure shifts frequency components and changes speech duration at the same time. For example, the fundamental frequency of adult speech can be tuned up through downsampling the original speech, while the speech duration will become shorter. Secondly, since we assume that the speech duration of children and adults are the same, we then perform the WSOLA[18] procedure on the frequency-tuned signals. WSOLA is a high-quality time-scale modification algorithm based on waveform similarity, which keeps the fundamental frequency of the original signal unchanged.

In order to enhance the performance of ASR system which is trained with adult corpus to recognize children's speech, we make the prosody modification to reduce the acoustic mismatch between adult and children's speech. We propose two prosody modification methods to eliminate the mismatch between the adult speech in the training set and the children's speech in the evaluation set. One is to tune up the prosody of the adult speech in the training corpus and to retrain the acoustic model with the prosody-modified corpus. The other way is to tune down the prosody of the children's speech in the evaluation directly.

SoX [19] is an audio manipulation tool, and we used it to make prosody modification. For example, to tune up prosody of an utterance, we can downsample the original audio with the factor λ by using *speed* command of *SoX*. This procedure changes the length of the original signal at the same time, in other words, the speaking rate becomes higher. For the second procedure we mentioned early in this section, we use the *tempo* command provided by *SoX* which is implemented based on WSOLA[18], to modify the tempo of the audio signal while keeping the original pitch and spectral unchanged. Combining these procedures, we can finish prosody modification without changing the speaking rate.

Fig.1 shows the comparison of the spectrograms of the original adult speech and the related prosody-tuned-up speech. This utterance is randomly picked from the adult training corpus. The original speech is downsampled with $\lambda = 1.1$. WSOLA algorithm is then performed to make the duration the same as the original signal. From these two spectrogram illustrations, it can be observed that the pitch and formant frequencies in figure (b) are higher than those in figure (a).

2.3. Pitch Feature

The motivation of adding extra pitch features in optimizing children's speech recognition can be expressed in two ways. (1) On the one hand, the pitch feature is a presentation for the auditory perception of tone [20]. In the process of prosody modification proposed in section 2.2, the prosody of speech is modified. Intuitively, extracting pitch features helps the DNN acoustic model explicitly focus more on the prosody. (2) On the other hand, in [21], an effective pitch extraction algorithm is proposed. In that algorithm, apart from pitch features, the probability of voicing feature and the delta-pitch feature will also be extracted. Previous works [21, 22] have shown that adding extra pitch features can improve the performance on tonal languages, such as Mandarin and Cantonese ASR. In this paper, children's speech recognition with Mandarin is explored and evaluated.

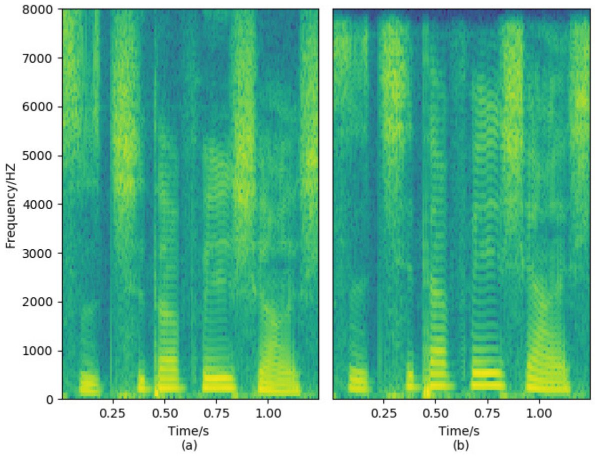

Figure 1: *The comparison of the spectrograms of the original adult speech (a) and the prosody-tuned-up speech (b).*

3. Joint Decoding with Prosody Modification

3.1. Shortcoming of prosody modification

Prosody modification method can get a significant performance improvement on children's speech. However, it is usually observed that there may be a performance degradation on adult evaluation set if we train the acoustic model with prosody tuned training data or modifying prosody on children's testing speech directly. The reason may be that the prosody modification simply applied to the training set or testing set, can reduce the acoustic mismatch for children's speech, but in contrast, it may increase the mismatch for adult speech.

3.2. Joint Decoding

To overcome this shortcoming, we propose a joint decoding architecture, which is much easier to apply to the already trained ASR system. Inspired by the previous work in acoustic system combination [3, 23], during evaluation, both the original speech and the prosody-modified speech are forwarded through the acoustic model as Fig.2 shows. The acoustic model generates two acoustic likelihood at the same time, then the two likelihood is combined by the weight of α. Denoting $\mathbf{o}$ and $\hat{\mathbf{o}}$ as the original and the prosody tuned acoustic features, the new likelihood of DNN output can be expressed as:

$$p_{joint}(x|\mathbf{o}) = \alpha p(x|\mathbf{o}) + (1-\alpha)p(x|\hat{\mathbf{o}}) \qquad (1)$$

The joint acoustic likelihood $p_{joint}(x|\mathbf{o})$ is then passed through the standard decoding pipeline to obtain the final results. This joint decoding framework with different prosody modification can take the advantages from both the original and new speech, which can further enhance the system robustness and improve the performance for both adult and children's speech.

[1] The proposed method does not affect some factors of the prosody, however in this paper, we still call it prosody modification.

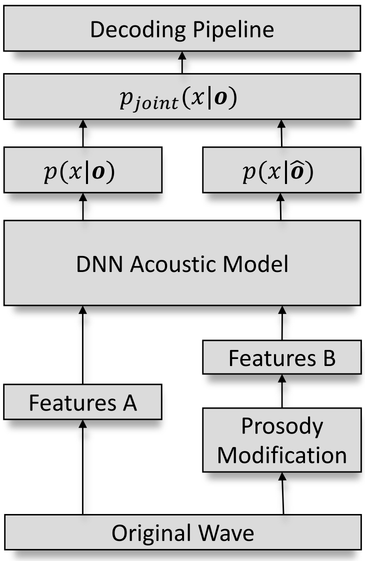

Figure 2: *Joint decoding with the original and prosody modified speech*

4. Experiments

4.1. Experimental setup and baseline system

A 400-hour hand-transcribed Mandarin adult corpus is used to train our baseline system. There are 481K utterances with an average duration of 3 seconds in the corpus, 95% of which are used as training set and the rest 5% are used as the validation set. There are two testing sets to evaluate our proposed methods. The first testing set containing 15626 utterances of children's speech is used to evaluate the system performance on children's speech recognition task. The other testing set containing 8272 utterances of adult speech is used to evaluate the performance on adult speech recognition task.

Gaussian mixture model based hidden Markov models (GMM-HMM) is first trained, which consists of 9663 clustered states. Then, a forced-alignment is performed over the 400-hour corpus using the GMM-HMM model to get state level labels. The Kaldi toolkit [24] is used to build the deep neural network (DNN) acoustic models. The DNN contains 5 hidden layers with 2048 units in each layer, and the ReLU activation function is used after each layer; The input layer has 1320 units since we use 40-dimension filter bank features with delta order 2 and ±5 frame expansion; The output layer consisted of 9663 units corresponding to GMM-HMM clustered states.

Word error rate (WER) on children's and adult testing set is listed in Table 1 as the first line. It is observed that the children's speech is much more difficult to be recognized than adult speech, and the performance gap is large when using the traditional acoustic modeling method only with adult speech.

4.2. Pitch feature

3-dimension pitch features, including probability of voicing feature, pitch feature and pitch-delta feature, are extracted following the recipe in [21] with Kaldi toolkit. The pitch features are combined with the 40-dimension filter bank features. Experiment setup is similar to that we mentioned in section 4.1, 5 hidden layers with 2048 units per layer is used in DNN. The activation function is ReLU. For the input layer, 43-dimension features consisting of filter bank and pitch with delta order 2 and ±5 frame expansion is used. So the input layer in this setup contains 1419 units considering the addition of 3-dimension pitch features, which is different from the setup in section 4.1.

As Table 1 shows, assisted by the pitch feature, there is a consistent improvement on both the adult and children's speech.

Table 1: *WER (%) comparison of baseline systems with/without pitch feature on adult/child testing set*

	Adult	Child
baseline	16.26	29.23
+ pitch feature	15.65	28.66

4.3. Prosody modification on training

The prosody modification procedure that mentioned in section 2.2 is performed on the 400-hour adult corpus with factor $\lambda_{train} = \{1.05, 1.1, 1.15\}$. Then the acoustic model trained with the prosody-modified adult corpus is evaluated on children's speech and adult speech. The model configuration and training procedure are exactly the same as the baseline, and the performance comparison of the proposed approach using prosody modification in training is listed in Table 2.

Table 2: *WER (%) comparison of the system trained with the prosody-modified training set using different λ_{train} parameters*

λ_{train}	1.0	1.05	1.1	1.15
Adult	**16.26**	17.04	20.10	25.34
+ pitch	**15.65**	16.54	19.26	24.56
Child	29.23	26.47	**25.81**	26.13
+ pitch	28.66	26.21	**25.28**	25.72

From Table 2, it can be seen that when $\lambda_{train} = 1.1$, the performance of children's speech recognition achieves the best position. However, the system trained only with prosody modified corpus suffers performance degradation while it is evaluated on adult speech. On the one hand, this phenomenon shows that the prosody modification on adult training corpus indeed works for improving children's speech recognition. On the other hand, this simple prosody modification on training corpus leads to acoustic mismatches between real adult speech and prosody modified adult speech, which causes performance degradation on adult speech.

To reduce this degradation, the prosody modified training corpus is combined with the original training corpus, getting an 800-hour training corpus. The new system trained with the 800-hour corpus significantly reduces the impact on adult speech recognition shown as Table 3. It shows that by combining the original training corpus with the prosody modified corpus, 15% relative WER reduction can be obtained for children's speech, and without an obvious performance degradation on adult speech.

4.4. Prosody modification on testing

Prosody modification on testing corpus is more flexible in practice. The model is not re-trained and it can be performed on

Table 3: *WER (%) comparison of the systems trained with both the prosody-modified training set and the original training set using* $\lambda_{train} = 1.1$

λ_{train}/hours	1.0/400hr	1.1/400hr	1.0&1.1/800hr
Adult	**16.26**	20.10	16.29
+ pitch	**15.65**	19.26	15.77
Child	29.23	25.81	**25.51**
+ pitch	28.66	25.28	**24.93**

the testing directly with the original adult model. The proposed prosody modification on testing is evaluated on the original 400-hour adult trained systems. The modification factors $\lambda_{test} = \{0.86, 0.88, 0.9, 0.92, 0.94\}$ have been compared. As Table 4 shows, the similar conclusion as that in section 4.3 is observed. The performance is significantly improved on children's speech with prosody modification on testing speech directly, and it can achieve the best position when $\lambda_{test} = 0.9$. In contrast, the accuracy on adult speech degrades gradually with the reduced prosody modification factors.

Table 4: *WER (%) comparison of the prosody modification on testing speech directly. The ASR system is built on the original 400-hour adult corpus.*

λ_{test}	0.86	0.88	0.9	0.92	0.94	1.0
Adult	26.77	23.32	20.47	18.50	17.56	**16.26**
+ pitch	25.19	21.98	19.42	17.83	16.88	**15.65**
Child	27.48	27.07	**26.89**	27.00	27.55	29.23
+ pitch	26.70	26.29	**26.06**	26.20	26.76	28.66

4.5. Joint decoding with prosody modified speech

In this subsection, the proposed joint decoding method for children's speech recognition is evaluated. The DNN acoustic model is trained with 400-hour adult corpus. In the evaluation, the prosody modification method is first performed with $\lambda_{test} = 0.9$, and both the modified speech and the original speech are fed into the acoustic model. The two streams of likelihood generated from the DNN acoustic model are then combined following the method described in section 3. The decoding pipeline is the same as the baseline setup.

Table 5: *WER (%) comparison of the proposed joint decoding with the original and prosody modified testing speech. The ASR system is built on the original 400-hour adult corpus.*

λ_{test}	1.0	0.9	joint
Adult	**16.26**	20.47	16.53
+ pitch	**15.65**	19.42	15.85
Child	29.23	26.89	**25.73**
+ pitch	28.66	26.06	**25.02**

The experimental results are illustrated in Table 5, and the acoustic model is built on the original 400-hour adult corpus. It shows that the proposed joint decoding can further improve the system performance for the children's speech when compared to the direct prosody modification on testing speech in Table 4. On the other hand, the accuracy on adult speech is also boosted, and the performance degradation compared to the baseline adult speech is very small when performing joint decoding.

4.6. Evaluation summary of the proposed approaches

Finally, we tried to combine the different methods proposed in this paper to construct our best children's speech recognition system, and the performance comparison is summarized in Table 6.

Table 6: *WER (%) comparison of the newly proposed methods for children's speech recognition.*

System	WER	
	Adult	Child
Baseline	16.26	29.23
+ pitch	15.65	28.66
++ prosody modified on test	19.42	26.06
+++ joint decoding	15.85	25.02
++ prosody modified on train	15.77	24.93
+++ joint decoding	15.79	23.71

The results show that all the newly proposed approaches can improve children's speech recognition significantly. Different methods utilize the prosody knowledge on the different levels, and these individual techniques can be combined to get a further improved system. Our final system can obtain a large gain on children's speech, with relative ~20% WER reduction, and still keeps the same high-performance on adult speech compared to the baseline.

5. Conclusions

In this paper, we explored several ways on the prosody usage to improve the speech recognition system on children's speech, only using the adult data in training. A better utilization strategy on prosody-based features is developed. First, pitch and prosody modification is explored in both training and testing respectively, which can significantly reduce the mismatch between two types of speech, and an obvious improvement in children's speech can be obtained. Furthermore, joint-decoding with both the prosody modified speech and the original speech is designed to get a more robust performance on both children's and adult speech. The final system, built with all the proposed technologies, can obtain a large improvement on Mandarin children's speech recognition, and also no obvious degradation on WER is observed for adult speech.

6. Acknowledgement

This work was supported by the China NSFC projects (No. 61603252 and No. U1736202). Experiments have been carried out on the PI supercomputer at Shanghai Jiao Tong University.

7. References

[1] L. Deng, J. Li, J.-T. Huang, K. Yao, D. Yu, F. Seide, M. Seltzer, G. Zweig, X. He, J. Williams *et al.*, "Recent advances in deep learning for speech research at Microsoft," in *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, 2013, pp. 8604–8608.

[2] M. Bi, Y. Qian, and K. Yu, "Very deep convolutional neural networks for LVCSR," in *Sixteenth Annual Conference of the International Speech Communication Association*, 2015.

[3] Y. Qian, M. Bi, T. Tan, and K. Yu, "Very deep convolutional neural networks for noise robust speech recognition," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 24, no. 12, pp. 2263–2276, 2016.

[4] V. Peddinti, D. Povey, and S. Khudanpur, “A time delay neural network architecture for efficient modeling of long temporal contexts,” in *Sixteenth Annual Conference of the International Speech Communication Association*, 2015.

[5] H. Liao, G. Pundak, O. Siohan, M. K. Carroll, N. Coccaro, Q.-M. Jiang, T. N. Sainath, A. Senior, F. Beaufays, and M. Bacchiani, “Large vocabulary automatic speech recognition for children,” in *Sixteenth Annual Conference of the International Speech Communication Association*, 2015.

[6] G. Hinton, L. Deng, D. Yu, G. Dahl, A.-r. Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, B. Kingsbury *et al.*, “Deep neural networks for acoustic modeling in speech recognition,” *IEEE Signal processing magazine*, vol. 29, 2012.

[7] A. Batliner, M. Blomberg, S. D’Arcy, D. Elenius, D. Giuliani, M. Gerosa, C. Hacker, M. Russell, S. Steidl, and M. Wong, “The PF_STAR children’s speech corpus,” in *Ninth European Conference on Speech Communication and Technology*, 2005.

[8] J. Fainberg, P. Bell, M. Lincoln, and S. Renals, “Improving children’s speech recognition through out-of-domain data augmentation.” in *INTERSPEECH*, 2016, pp. 1598–1602.

[9] S. Shahnawazuddin, A. Dey, and R. Sinha, “Pitch-adaptive front-end features for robust children’s ASR.” in *INTERSPEECH*, 2016, pp. 3459–3463.

[10] R. Sinha and S. Shahnawazuddin, “Assessment of pitch-adaptive front-end signal processing for childrens speech recognition,” *Computer Speech & Language*, vol. 48, pp. 103–121, 2018.

[11] S. Ghai, “Addressing pitch mismatch for children’s automatic speech recognition,” Ph.D. dissertation, 2011.

[12] M. Russell and S. DArcy, “Challenges for computer recognition of childrens speech,” in *Workshop on Speech and Language Technology in Education*, 2007.

[13] M. Gerosa, D. Giuliani, S. Narayanan, and A. Potamianos, “A review of ASR technologies for children’s speech,” in *Proceedings of the 2nd Workshop on Child, Computer and Interaction*. ACM, 2009, p. 7.

[14] S. Lee, A. Potamianos, and S. Narayanan, “Acoustics of childrens speech: Developmental changes of temporal and spectral parameters,” *The Journal of the Acoustical Society of America*, vol. 105, no. 3, pp. 1455–1468, 1999.

[15] I. Titze, “Principles of voice production. prentice-hall,” *Englewood Cliffs, NJ*, 1994.

[16] R. J. Baken and R. F. Orlikoff, “Clinical measurement of speech and voice. london,” *Cengage Learning*, pp. 561–570, 2000.

[17] I. Chandra Yadav, A. Kumar, S. Shahnawazuddin, and G. Pradhan, “Non-uniform spectral smoothing for robust children’s speech recognition,” 09 2018, pp. 1601–1605.

[18] W. Verhelst and M. Roelands, “An overlap-add technique based on waveform similarity (WSOLA) for high quality time-scale modification of speech,” in *Acoustics, Speech, and Signal Processing, 1993. ICASSP-93., 1993 IEEE International Conference on*, vol. 2. IEEE, 1993, pp. 554–557.

[19] “Sox, audio manipulation tool,” saf, // http://sox.sourceforge.net/.

[20] D. Talkin, “A robust algorithm for pitch tracking (RAPT),” *Speech coding and synthesis*, vol. 495, p. 518, 1995.

[21] P. Ghahremani, B. BabaAli, D. Povey, K. Riedhammer, J. Trmal, and S. Khudanpur, “A pitch extraction algorithm tuned for automatic speech recognition,” in *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2014, pp. 2494–2498.

[22] H. C.-H. Huang and F. Seide, “Pitch tracking and tone features for Mandarin speech recognition,” in *2000 IEEE International Conference on Acoustics, Speech, and Signal Processing. Proceedings (Cat. No. 00CH37100)*, vol. 3. IEEE, 2000, pp. 1523–1526.

[23] S. Kundu, G. Mantena, Y. Qian, T. Tan, M. Delcroix, and K. C. Sim, “Joint acoustic factor learning for robust deep neural network based automatic speech recognition,” in *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2016, pp. 5025–5029.

[24] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz *et al.*, “The Kaldi speech recognition toolkit,” in *IEEE 2011 workshop on automatic speech recognition and understanding*, no. EPFL-CONF-192584. IEEE Signal Processing Society, 2011.