Using Convolution and Sequence-discriminative Training to Improving Children Speech Recognition

Fanchang Meng, Shouye Peng, Guohui Zhang
Beijing Century TAL Education Technology Co.,Ltd, Beijing 100190, China
E-mail: {mengfanchang, pengshouye, zhangguohui}@100tal.com

***Abstract*—The conclusion that ASR for children's speech is especially difficult compared to adult was given by the robotics community from recent works. Challenges on Children's speech recognition mainly due to the increased variability in acoustic and linguistic correlates depending on a young age. This work focused on the recognition of oral English spoken by Chinese children aging six to twelve. Experiments were conducted on: (1) Speaker Normalization algorithms, including Cepstral Mean and Variance Normalization (CMVN) and Vocal Tract Length Normalization (VTLN) techniques; (2) Acoustic models adapting techniques, such as Maximum Likelihood Linear Transform (MLLT) and Speaker Adaptive Training (SAT) based on Constrained MLLR; (3) Different acoustic models, GMM-HMM, DNN-HMM, CNN-DNN; (4) Training criterion, with frame-level training such as Cross entropy (CE), and sequence-discriminative training (SDT) such as MMI, MPE and sMBR were conducted in this paper. The results included: (1) with the increase of age, the variability of children's pronunciation decreased significantly; (2) the convolution on the frequency axis has a great performance contribution (34.72%) to the variability of children over the baseline system.**

I. INTRODUCTION

Speech recognition has become an indispensable part of our life. ASR system has been applied in a range of fields, such as education, communication, human-computer interaction and translation. Comparing with the dramatically improved ASR for adults which have been studied for decades [1][2][3][4][5], children's speech recognition is facing huge challenges due to the rapid physical development of children and the lack of language skills, especially at a very young age [6][7][8][9][10][11]. Therefore, it is necessary to address the challenges brought by variability in children's speech.

The variability in acoustic mainly comes from the developmental changes associated with vocal tract growth. On the linguistic side, it is related to children's proficiency in language skills, for example, limited vocabulary, lack of clarity in pronunciation and grammatical structure.

Much has been done in the past to analyze the acoustic characteristics of children's speech. Acoustic variability in the children's speech can be attributed to three major factors (i) the shift of overall spectral content and formant frequencies for children [12], (ii) high within-subject variability in the spectral content, which affects formant locations [8], (iii) high inter-speaker variability observed across age groups, due to developmental changes, especially vocal tract [13]. S. Lee, A. Potamianos, and S. Narayanan found that temporal and spectrum parameters of children speech were analyzed and studied in detail [8].

In recent years, several techniques for dealing with acoustic variability have been proposed. Different front-end features, such as Perceptual Linear Prediction (PLP) cepstral coefficients, Mel-Frequency Cepstral Coefficients (MFCC), and spectrum-based filter bank features were studied [12]. In addition, [14][15][16][17][18] also investigate several slight changes in front-end characteristics.

Speaker normalization algorithms have been investigated, for example, Cepstral Mean and Variance Normalization (CMVN) and Vocal Tract Length Normalization (VTLN) technique were explored in ASR. The front-end frequency warping like VTLN have been proved useful to deal with the aforementioned speech variability in children speakers [19][20].

In this paper, we concentrate on two aspects of speech recognition: front-end processing and acoustic modeling for building robust ASR for children. The rest of the paper is organized as follows. In Section Ⅱ, we give an overview of the databases used to conduct the experiments. Section Ⅲ describes our experi-

mental setup. Section Ⅳ presents the recognition experiments and their results. Finally, we conclude our views in Section Ⅴ.

II. DATABASES

The dataset used in this study was collected from *Xueersi Online School*, an APP which contains a function to collect students' oral English practice demos. The users are mainly from 6 to 12 year's old including boys and girls. There were 3993 children speakers included in the 166.06 hours of corpus through grade 1 to 6. The average number of words per sentence increased as the grade increased. For example, the average number of words is 6 in grade 1 and 14 in grade 6. Table Ⅰ shows the grade distribution of training and testing databases.

TABLE Ⅰ

GRADE DISTRIBUTION OF TRAINING AND TESTING DATA

Dataset	Grade	1	2	3	4	5	6	Total
Training-Set	utterances	13165	12989	14850	14768	12942	13051	81765
	speakers	647	643	605	753	720	625	3993
	hours	20.47	22.33	31.19	30.87	30.12	31.08	166.06
Testing-Set	utterances	1000	1000	1000	1000	1000	1000	6000
	speakers	291	275	323	286	312	222	1709
	hours	1.50	1.61	1.80	2.12	2.20	2.32	11.55

III. SPEECH RECOGNITION SETUP

We use the open-source speech recognition toolkit Kaldi [21] to run the experiments. The sampling frequency of 16 KHz was adopted in the work. The standard MFCC features with 13 mel-cepstrum coefficients with their first and second order derivatives were used as the front-end features. The MFCCs were extracted using 23-channel filter banks using frame-length of 25ms and frame-shift of 10ms.

GMM-HMM system: The HMMs were modeled using 3 states for non-silence phones and 5 for silence phones. A total of 1000 Gaussian densities are shared among HMMs.

Dictionary: The CMU Pronunciation dictionary [22] was employed which corresponds to American-English pronunciations and was made compatible with our available children data. To account for the out-of-vocabulary (OOV) words during training, a grapheme to phoneme converter was used to generate phoneme transcripts for OOV words.

Language Model: The language model was training using the transcriptions from children's training data sets. Reference [15] conducted the experiment using unigram, bigram and trigram models, and the trigram was proved be the best. Trigram model was training in this work.

IV. RECOGNITION EXPERIMENTS AND RESULTS

A. Baseline System

Except for the experimental configuration mentioned in section 3 above, Cepstral Mean and Variance Normalization (CMVN) was explored in the baseline experiments as a standard practice. We modeled and evaluated the monophone and triphone models.

TABLE Ⅱ

WERS (%) OF MONOPHONE AND TRIPHONE MODEL

Grade	Model	
	Monophone	Triphone
1	52.58	35.84
2	49.13	32.77
3	49.10	31.98
4	48.51	31.63
5	44.77	29.20
6	44.93	28.93
Average	48.17	31.73

Table Ⅱ shows that the Triphone model performs much better and reduces the WER to 31.73% an absolute improvement of 16.44% in average over the monophone models. Therefore, we used the triphone model for subsequent experiments.

Acoustic model adaptation techniques like Maximum Linear Likelihood Transform (MLLT), Speaker Adaptive Training (SAT) have shown improvements with children speech in the past [19][20]. Because of the increase of variability in children speech and the LDA working by transforming features, the Linear Discriminant Analysis (LDA) was employed to reduce the intra-class variability and increase the inter-class

variability.

Table Ⅲ shows the performance of children speech before and after adapting the acoustic model adaptation techniques. The model transformed provides a significant reduction in WER of about 3.13% in average absolute compared to the original triphone model. Thus we use the adapating triphone model (LDA+MLLT+SAT) as the baseline system for subsequent experiments.

TABLE Ⅲ

WERS (%) OF ACOUSTIC MODEL AND ADAPTATION

Grade	Model	
	Triphone	LDA+MLLT+SAT
1	35.84	32.85
2	32.77	30.15
3	31.98	28.65
4	31.63	28.87
5	29.20	25.42
6	28.93	25.67
Average	31.73	28.60

B. Speaker Normalization Algorithms

There have been a number of papers that describe implementations of Vocal Tract Length Normalization (VTLN) that work out a linear feature transform corresponding to each VTLN warp factor in recent years [23].

TABLE Ⅳ

WERS (%) OF BASELINE MODEL AND VTLN MODEL

Grade	Model	
	baseline	VTLN
1	32.85	30.1
2	30.15	29.83
3	28.65	28.11
4	28.87	27.72
5	25.42	24.84
6	25.67	24.66
Average	28.60	27.54

In the Table Ⅳ, the VTLN technique achieved a word error rate (WER) of 27.54% in average which is a 1.06% gain over the raw MFCC features.

C. Sequence-discriminative training (SDT)

Speech recognition is inherently a sequence classification problem. As such, speech recognizers using Gaussian mixture model (GMM) as the emission density of an HMM achieve state-of-the-art performance when trained using sequence-discriminative criteria like maximum mutual information (MMI) [24][25], boosted MMI (BMMI) [26]and minimum phone error (MPE) [27]. In this part, we evaluated the effectiveness of SDT in children speech corpus.

Sequence-discrimination training begins with a set of alignments and lattices generated by decoding training data using unigram LM.

TABLE Ⅴ

WERS (%) OF BASELINE, MMI, BMMI AND MPE

Grade	Model			
	baseline	MMI	MMI-boost0.05	MPE
1	32.85	30.78	29.9	29.59
2	30.15	28.23	27.28	26.55
3	28.65	25.87	25.16	25.38
4	28.87	26.07	25.42	25.25
5	25.42	23.07	23.59	22.96
6	25.67	22.9	22.6	22.32
Average	28.60	26.15	25.66	25.34

Table Ⅴ shows the effectiveness of SDT techniques on children datasets with significant variability. Similar to the adults speech recognition [28][29], the SDT achieves an improvement of 2.45% absolute on MMI, 2.94% absolute on BMMI and 3.26% absolute on MPE.

D. DNN-HMM and DNN-sMBR

A hybrid DNN-HMM system was employed, where the DNN was used to replace the posterior probabilities of a traditional GMM system. The 40-dimensional features are MFCC-LDA-MLLT-fMLLR with CMVN, extracted with window 25 msec and frame rate 10 msec. The DNN consumes high resolution MFCC features with a context of 5 left and 5 right frames. The DNN has 7 hidden layers, each of dimension 1024. The output Softmax layer consists of 1470 units trained to predict the posterior.

In addition, sMBR sequence-discriminative training [28][29] was explored to train the deep neural network to jointly optimize for whole children speech sentences, which is closer to the general ASR objective than frame-level training.

Table Ⅵ shows the huge advantage of the CD-DNN-HMM over the baseline system trained from GMM-HMM with an absolute improvement of 4.81%. Further, we have seen the significant contribution of CD-DNN-sMBR that has reduced the WER to 21.95% on children's datasets.

TABLE Ⅵ

WERs (%) OF BASELINE, DNN-HMM AND DNN-SMBR

Grade	Model		
	baseline	DNN-HMM	DNN-sMBR
1	32.85	26.8	24.22
2	30.15	25.28	23.08
3	28.65	23.58	21.35
4	28.87	24.34	22.78
5	25.42	21.54	20.18
6	25.67	21.22	20.07
Average	28.60	23.79	21.95

E. CNN-DNN-HMM and CNN-DNN-sMBR

Convolutional layers provide robustness to vocal tract length variation similar to VTLN, however do so by normalizing small shifts in frequency rather than feature warping [30][31][32][33]. Thus, to tackle the large inter-speaker variability and the intra-speaker variability, Convolutional layers were employed in the deep neural networks.

The acoustic model is comprised of 2 CNN layers, where the first CNN layer has 4224 units and the second has 2048 units. Both the CNN layers are followed by Maxpooling layers. The output of CNN model is sent to a 6 layers DNN and the last layer is a softmax with 1470 context-dependent triphone states.

TABLE Ⅶ

WERS (%) OF DNN-SMBR, CNN-HMM AND CNN-DNN-SMBR

Grade	Model		
	DNN-sMBR	CNN-HMM	CNN-DNN-sMBR
1	24.22	25.14	22.27
2	23.08	23.52	20.07
3	21.35	21.39	18.79
4	22.78	21.08	18.05
5	20.18	18.27	16.48
6	20.07	18.14	16.37
Average	21.95	21.26	18.67

Table Ⅶ shows an exhilarating result of the CNN-HMM and CNN-DNN-sMBR. Achieving an absolute improvement of 3.28% over the DNN-sMBR model, the CNN-DNN-sMBR has reduced the WER to 18.67% in average which is a relative improvement of 34.72% over the baseline models.

F. Overall performances of the experiments

In this part, we combine the results of all the aforementioned trials for a more intuitive comparison in Fig. 1. The Word Error Rate decreases over the level from grade 1 to grade 6. The GMM-HMM model using SDT outperformers the baseline system. Both DNN and DNN-sMBR performer better than the GMM-HMM model, and the CNN-DNN-sMBR model achieves the best effectiveness of tackling with the variability in children speech recognition.

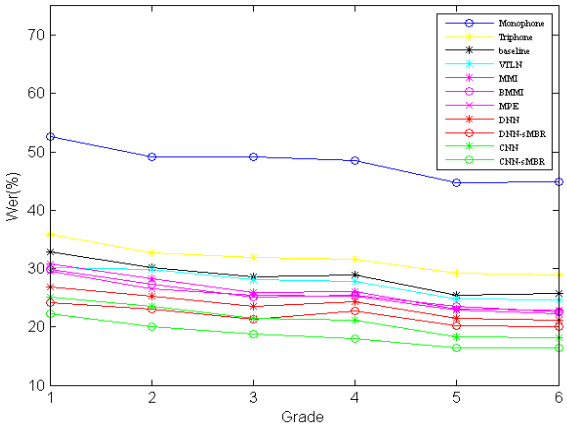

Fig. 1 Performances of all the experiments.

V. CONCLUSION

In this work, we have conducted the experiments with children dataset on acoustic model adapting techniques, speaker normalization algorithms, different acoustic models and training criterions. We found that (1) with the increase of age, the variability of children's pronunciation decreased significantly, (2) the CNN-DNN-sMBR are proven to be more effective (WER, 18.67%) in tackling with the large inter-speaker variability and the intra-speaker variability. The convolution on the frequency axis has a great performance contribution (34.72%) to the frequency variability of children over the baseline system.

REFERENCES

[1]. Xiong W , Droppo J , Huang X , et al. "Achieving Human Parity in Conversational Speech Recognition"*[J]. IEEE/ACM Transactions on Audio Speech & Language Processing*, 2016.

[2]. W. Xiong, L. Wu, F. Alleva, J. Droppo, X. Huang, and A. Stolcke, "The microsoft 2017 conversational speech recognition system," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Apr. 2018, pp. 5934–5938.

[3]. Chiu C C , Sainath T N , Wu Y , et al. "State-of-the-Art Speech Recognition with Sequence-to-Sequence Models"*[C]// ICASSP* 2018 - 2018 *IEEE International*

Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2018.

[4]. S. Zhang, M. Lei, Z. Yan, and L. Dai, "Deep-fsmn for large vocabulary continuous speech recognition," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2018.

[5]. X. Yang, J. Li, and X. Zhou, "A novel pyramidal-fsmn architecture with lattice-free MMI for speech recognition," in *Proc. ICASSP*, 2018.

[6]. 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

[7]. L. L. Koenig and J. C. Lucero, "Stop Consonant Voicing and Intraoral Pressure Contours in Women and Children," *The Journal of the Acoustical Society of Americ*a, vol. 123, no. 2, pp. 1077–1088, 2008.

[8]. 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.

[9]. ——, "Analysis of Children's Speech: Duration, Pitch and Formants" in *Proc. of EUROSPEECH*, 1997, pp. 473–476.

[10]. H. K. Vorperian and R. D. Kent, "Vowel Acoustic Space Development in Children: A Synthesis of Acoustic and Anatomic Data," *Journal of Speech, Language, and Hearing Research*, vol. 50, no. 6, pp. 1510–1545, 2007.

[11]. B. L. Smith, "Relationships Between Duration and Temporal Variability in Children's Speech," *The Journal of the Acoustical Society of America*, *v*ol. 91, no. 4, pp. 2165–2174, 1992

[12]. 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.

[13]. M. Gerosa, D. Giuliani, and S. Narayanan, "Acoustic analysis and automatic recognition of spontaneous children's speech," in *Ninth International Conference on Spoken Language Processing*, 2006.

[14]. Q. L. M. J. Russell, "Why is automatic recognition of children's speech difficult?" 2001.

[15]. P. G. Shivakumar, A. Potamianos, S. Lee, and S. Narayanan, "Improving speech recognition for children using acoustic adaptation and pronunciation modeling," in *Proc. Workshop on Child, Computer and Interaction* (WOCCI), 2014.

[16]. S. Umesh and R. Sinha, "A study of filter bank smoothing in mfcc features for recognition of children's speech," *IEEE Transactions on audio, speech, and language processing*, vol. 15, no. 8, pp. 2418–2430, 2007.

[17]. S. Ghai and R. Sinha, "Pitch adaptive mfcc features for improving childrens mismatched asr," *International Journal of Speech Technology* vol. 18, no. 3, pp. 489–503, 2015.

[18]. S. Shahnawazuddin, A. Dey, and R. Sinha, "Pitch-adaptive front-end features for robust children's asr." in *INTERSPEECH*, 2016, pp. 3459–3463

[19]. D Elenius and M Blomberg. "Adaptation and normalization experiments in speech recognition for 4 to 8 year old children." In: *INTERSPEECH*. 2005, pp. 2749–2752

[20]. A Potamianos, S Narayanan, and S Lee. "Automatic speech recognition for children." In: *Eurospeech*. Vol. 97. 1997, pp. 2371–2374.

[21]. Daniel Povey et al. "The Kaldi Speech Recognition Toolkit". In: *IEEE 2011 Workshop on Automatic Speech Recognition and Understanding*. *IEEE Catalog No.: CFP11SRW-USB. Hilton Waikoloa Village, Big Island, Hawaii, US: IEEE Signal Processing Society*, Dec. 2011.

[22]. R. Weide, "The cmu pronunciation dictionary, release 0.6," 1998.

[23]. D Y Kim et al. "Using VTLN for broadcast news transcription". In: *Proc. ICSLP*. Vol. 4. 2004.

[24]. L. R. Bahl, P. F. Brown, P. V. de Souza, and R. L. Mercer, "Maximum mutual information estimation of hidden Markov model parameters for speech recognition," in *Proc. IEEE ICASSP*, vol. 1, April 1986, pp. 49–52.

[25]. V. Valtchev, J. Odell, P. Woodland, and S. Young, "MMIE training of large vocabulary recognition systems," *Speech Communication*, vol. 22, no. 4, pp. 303–314, September 1997.

[26]. D. Povey, D. Kanevsky, B. Kingsbury, B. Ramabhadran, G. Saon, and K. Visweswariah, "Boosted MMI for model and feature-space discriminative training," in *Proc. IEEE ICASSP*, 2008, pp. 4057–4060.

[27]. D. Povey, "Discriminative training for large vocabulary speech recognition," Ph.D. dissertation, University of Cambridge, Cambridge, UK, 2003.

[28]. Veselý K, Ghoshal A, Burget L, et al. "Sequence discriminative training of deep neural networks"*[C] Proc. INTERSPEECH*. 2013: 2345-2349.

[29]. Naing, H. M. S. , et al. "A Myanmar large vocabulary continuous speech recognition system." *Asia-pacific Signal & Information Processing Association Summit & Conference (APSIPA ASC) IEEE*, 2015.

[30]. O. Abdel-Hamid, A. Mohamed, H. Jiang, and G. Penn, "Applying convolutional neural networks concepts to hybrid NN-HMM model for speech recognition," in *Proc. ICASSP*, 2012.

[31]. Hu, X. , M. Saiko , and C. Hori . "Incorporating tone features to convolutional neural network to improve Mandarin/Thai speech recognition." *Asia-pacific Signal & Information Processing Association, Summit & Conference (APSIPA ASC) IEEE*, 2015.

[32]. Lim, Hyungjun , et al. "CNN-based bottleneck feature for noise robust query-by-example spoken term detection." *2017 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)* 2017.

[33]. Fu, Szu Wei , et al. "Raw waveform-based speech enhancement by fully convolutional networks." *Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)* (2017):006-012.