Multi-Task Based Mispronunciation Detection of Children Speech Using Multi-Lingual Information

Linxuan Wei*, Wenwei Dong*, Binghuai Lin†, and Jinsong Zhang*
* Beijing Advanced Innovation Center for Language Resources, Beijing Language and Culture University, Beijing, China
E-mail: willix.wei@gmail.com, dongwenwei_blcu@163.com, jinsong.zhang@blcu.edu.cn
† MIG, Tencent Science and Technology Ltd., Beijing, China
E-mail: binghuailin@tencent.com

***Abstract*— In developing a Computer-Aided Pronunciation Training (CAPT) system for Chinese ESL (English as a Second Language) children, we suffered from insufficient task-specific data. To address this issue, we propose to utilize first language (L1) and second language (L2) knowledge from both adult and children data through multitask-based transfer learning according to Speech Learning Model (SLM). Experimental set-up includes the TDNN acoustic modelling using the following training data: 70 hours of English speech by American Children (AC), 100 hours by American Adults (AA), 5 hours of Chinese speech by Chinese Children (CC), and 89 hours by Chinese Adults (CA). Testing data includes 2 hours of ESL speech by Chinese children. Experimental results showed that the inclusion of AA data brought about 13% relative Detection Error Rate (DER) reduction compared to AC only. Further inclusion of CC and CA data through L1 transfer learning brought about a total of 21% relative improvement in DER. These results suggested the proposed method is effective in mitigating insufficient data problem.**

I. INTRODUCTION

Computer-Aided Pronunciation Training (CAPT) systems with individualized feedback are becoming more and more popular. Automatic Mispronunciation Detection (AMD), a key component of CAPT, aims to precisely identify pronunciation errors from the practice recording of students. To identify mispronunciations, a statistic model was trained by L2 standard pronunciation data to represent the distribution of standard pronunciations. With this model, the deviation of the practice pronunciation of students from the distribution can be calculated, which leads us to find mispronunciations through setting a threshold. Therefore, it is important that whether the training data are enough to represent the whole distribution of standard pronunciations precisely. That is to say, lack of enough data makes AMD tasks more challenging.

When developing a CAPT system for Chinese ESL children, at the current stage, we only have 70 hours of native speech corpus of American children. With high variability of children speech and various English proficiency of learners, we suffered from the shortage of task-related data. Therefore, more techniques are worth exploring to tackle this issue.

To mitigate the shortage and the high variability of children data [1-4], it is common to leverage adult data of the same language to adapt acoustic model. Due to the significant acoustic differences between children and adult speech, simply mixing data usually do not work very well[5]. To bridge this gap, some studies are conducted from different perspectives [6-8]. A fMLLR-based stochastic feature mapping method [9] was used to reduce their differences in feature space. Owing to the shorter vocal tract length of children, Vocal Tract Length Normalization (VTLN) [10-12] was investigated on adult's data to compensate the difference between adult and children speech. With multi-task-based transfer learning mechanism, neural network based acoustic modelling [5, 13, 14],which is sharing hidden layers but separating output layer for two ASR (Automatic Speech Recognition) tasks, was explored to achieve better performance than the way that simply mixing data in a single ASR task. The above experiments show that adult data help in the condition of the shortage and the high variability of children data. Therefore, the native English corpus of adult speech is appropriate to be transferred to our system.

When all the L2 speech corpora are still not enough, L1 speech corpora are adopted by some research to further enhance the AMD acoustic model. According to the Speech Learning Model (SLM), when asked to learn L2 phonemes, learners tend to use the phonemes in L1 to substitute them firstly, because some of L2 phonemes are very phonetically similar to those in their L1. That is to say, some sounds in L1 corpora can be used to enhance the modelling of L2 standard pronunciations. Based on this idea, Multi-lingual acoustic modelling methods [15-17] is thus introduced to utilize the L1 native data for Chinese AMD tasks of adults. Accent scoring method of adult learners based on Multi-lingual ASR is also investigated [18, 19]. With above explorations, L1 native speech corpus of children is promising for promoting the performance of our system.

All the L1 children native data we have is only 5 hours, which indicate that it is not enough for a children task. To resolve this problem, given the fact that adult data from the same language will also help, we hold that L1 adult data are also supposed to be valuable in improving our system.

Based on above analyses, we propose to utilize four kinds of related databases including English corpus by American children and adult, and Mandarin corpus by Chinese children and adult to mitigate the data shortage on the AMD task for Chinese ESL children. To integrate the corpora of different languages in a single model, we use multi-task-based transfer

learning mechanism with a Time Delay Neural Network (TDNN) acoustic model. We also investigate the efficacy of different types of data in AMD task for Chinese ESL children.

The rest of the paper is organized as follows. In section II, an overview of Multi-Task and Multi-Lingual based AMD framework is provided. In section III, multi-lingual corpora used are introduced. The experiment set-up is listed in section IV. Results and discussion are given in section V followed by conclusions in section VI.

II. Multi-Task and Multi-Lingual based TDNN Acoustic Modelling

A. Mispronunciation detection framework

Figure 1 shows the adopted AMD framework. There are four components in this framework including feature extraction module, acoustic model, phone-level aligner module and GOP (Goodness of Pronunciation) calculator module. Input pronunciations are extracted into frame-level acoustic features by front-end feature extraction module. With these features, frame-level posteriors of different phones can be obtained by a trained acoustic model which simulates the role of native judgers. With phone-level alignment, phone-level posteriors can be calculated through summing all posteriors of frames in corresponding phone. By comparing the posterior of current phone and that of the most competing one, The GOP score can be obtained. Thus, mispronunciations can be identified through setting a threshold of GOP score.

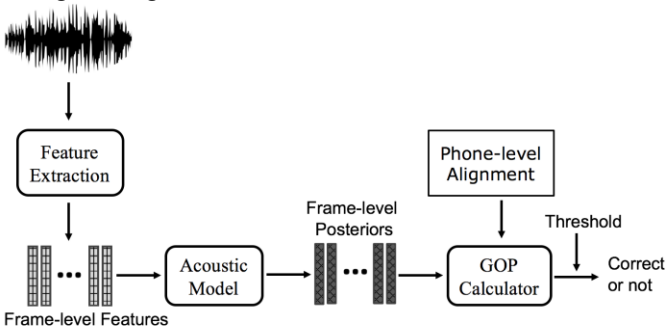

Fig. 1 Mispronunciation detection framework

The GOP measurement revised by Hu et al. is used, which is as following:

$$\mathrm{GOP}(p) \approx log \frac{p(p|\boldsymbol{o}; t_s, t_e)}{max_{\{q \in Q\}} p(q|\boldsymbol{o}; t_s, t_e)} \quad (1)$$

Where $\boldsymbol{o}$ is the whole observations; p is the canonical phoneme; t_s and t_e are the start and end frame indexes, respectively; $p(p)$ is the prior of phone; Q is whole phone set.

B. Acoustic modelling

The proposed multi-task and multi-lingual based TDNN acoustic modelling is illustrated in Figure 2. Multi-lingual and multi-age data are provided in the input layer. The multiple task s consists of Mandarin ASR task and English ASR task, which empower the model to utilize the data from both L2 and L1. The multi-task learning architecture is implemented by sharing the hidden layers of different tasks but separating their output layers providing the language dependent posteriors of senones. The loss functions of these output layers are combined by weights.

Adult data are also provided in the input layer. According to SLM, adult speech data can be treated as the phonological destination of the development of children speech. From this point of view, adult speech data are the standard and stable enough to mitigate the high variability of children speech, which give the model a better understanding of the linguistic similarity of children speech.

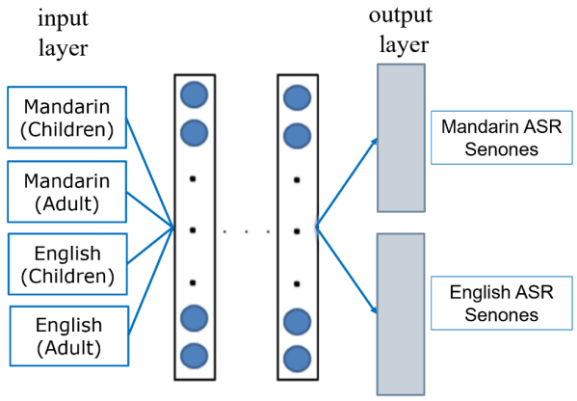

Fig. 2 Acoustic modelling diagram of TDNN based on multi-task and multi-lingual joint learning.

III. Speech Corpora

A. L1 Native Adult Corpus (cn-adult)

The native mandarin corpus of adult speech is from the Chinese National Hi-Tech Project 863 for Mandarin LVCSR system development [20]. A total of 94,000 utterances spoken by 166 speakers (114 hours) were used. 70% of it were used for acoustic modelling (cn-adult-train). 20% of it as a development set (cn-adult-dev), 10% as testing set (cn-adult-test).

B. L1 Native Children Corpus (cn-kids)

The native mandarin corpus of children speech (cn-kids-train). is an academic-free subset of King-ASR-409 multisource corpus which is balanced distributed in age (4-9), gender and regional accents. It is 5 hours in total and recorded in mainland China.

C. L2 Native Adult Corpus – adult (en-adult)

The native American English corpus of adult speech (en-adult-train) is a part of the training set of LibriSpeech corpus [21] which is a read speech dataset based on LibriVox's audio books from public domain recorded by American volunteers. The selected training set is 100 hours in total recorded by 251 speakers balanced in gender. Dev-clean, a subset of LibriSpeech corpus is used as development set (en-adult-dev). Test-other, a subset of LibriSpeech corpus is used as testing set (en-adult-test).

D. L2 Native Children Corpus (en-kids)

The native American English corpus of children speech (en-kids-train) is a subset of the CSLU corpus [22] including American English recordings by 1116 students from Portland, Oregon. There are 70 hours in total of 71999 utterances. The age of speakers range from 5 to 15 years old.

E. L2 Non-Native Children Corpus (en-kids-non)

The non-native English corpus is used as English test set. It is collected from Chinese children by a mobile CAPT app in a real environment which means there are full of various noises, emotions, channels, and volume. The signal channel varies with the difference of users' devices. The speakers range from 6 to 12 years old. The whole corpus is still under construction. The subset used in this study containing 2105 utterances.

This corpus is annotated by three trained Chinese annotators for three reasons. First, annotating a corpus by English native speakers seems expensive and impractical, especially when we aim to build a large dataset, for it is not easy to find that much English native speakers, who will live in Beijing for a long time and are willing to do this kind of hard job. Second, it is more reasonable for students to imitate the pronunciation from skilled Chinese speakers of English who can be fully understand by native speakers. Third, the mispronunciations identified by Chinese annotators may mush easier for Chinese learners to perceive them.

TABLE I
SUMMARY OF TRAINING DATA.

Data	no. spks	no. utts	hrs. w/sil
cn-adult-train	130	73565	89
cn-kids-train	20	4225	5
en-adult-train	251	28539	100
en-kids-train	1118	71999	70

IV. EXPERIMENT

A. Evaluation Metric

In this study, four metrics are used, containing Detection Accuracy (DA), False Rejection Rate (FRR), False Acceptance Rate (FAR), Detection Error Rate (DER). The sum of DA and DER is 1.

FAR is the percentage that system treats mispronunciation as correct one and FRR is the percentage that system rejects the correct pronunciation from learners wrongly. The formula of them are as following:

$$\mathrm{FAR} = \frac{\mathrm{FA}}{\mathrm{FA} + \mathrm{TR}} \tag{2}$$

$$\mathrm{FRR} = \frac{\mathrm{FR}}{\mathrm{FR} + \mathrm{TA}} \tag{3}$$

FA (False Acceptance) is the number of mispronunciations detected as correct, and TR (True Rejection) is the number of mispronunciations detected as incorrect. FR (False Rejection) is the number of correct pronunciations detected as incorrect, and TA (True Acceptance) is the number of correct pronunciation detected as correct.

FAR and FRR are trade-off indexes. However, from the teaching perspective, minimizing FRR is more important than FAR, for courage is a key factor when they learning a language. When their correct pronunciation is rejected by CAPT systems, they will be discouraged and feel hopeless and do not know what to do next. DA is a more comprehensive index which balances the influence between FAR and FRR. Its formula is as following:

$$\mathrm{DA} = \frac{\mathrm{TA} + \mathrm{TR}}{\mathrm{TA} + \mathrm{TR} + \mathrm{FA} + \mathrm{FR}} \tag{4}$$

B. Experiment Set-up

Phoneme-level AMD experiment is under DNN-HMM framework with GOP algorithm. All neural networks used in this study have 7 hidden layers. The 27-dimensional input feature is used, containing a 23-dimensional fbank feature, a 3-dimensional pitch feature and a 1-dimensional energy feature. The summation weight of objective functions in multi-task learning is 0.5 for each task. All training sets and testing set have no speaker overlap.

For English AMD task of Chinese ESL children's speech, TDNN modelling is used in the **baseline** acoustic model which is trained by English children's data *en-kids-train* only. Then, English adult data *en-adult-train* is used to adapt the baseline by mixed data approach (**en-adult_adapt model**). Next, native Chinese data of children speech *cn-kids-train* is further included for adapt *en-adult_adapt* model through a multi-task way (**en-all_cn-child_adapt model**). Finally, Chinese adult's data is further included to utilize L1 phonology information, for the constant phonological feature is easier to be extracted in adult's speech than children's (**en-all_cn-all_adapt model**). all models mentioned above is evaluated in English data of Chinese children *en-kids-non.*

V. RESULTS AND DISCUSSIONS

Results on English AMD task of Chinese children's speech *en-kids-non* are shown in Table II. *en-adult_adapt* model outperforms baseline with 13% DER reduction due to the adaptation of L2 adult speech. By further including native Chinese data of children speech *cn-kids-train*, the performance of *en-all_cn-child_adapt* model is barely changed, which is not consistent with our expectation. This is may be caused by insufficient Chinese children's speech and high variability of children's speech, which make the model hard to extract valuable feature to adapt the original model. A few Chinese children's speech here seems to act like noise due to its variability. When we further include Chinese adult's data, the model achieves the best performance.

TABLE II
RESULTS OF ENGLISH AMD OF CHINESE CHILDREN'S SPEECH.

	DA	FRR	FAR
Baseline	49.24	52.56	**35.28**
en-adult_adapt	56.01	44.03	43.55
en-all_cn-child_adapt	55.57	44.8	41.22

en-all_cn-all_adapt	**60.03**	**39.32**	45.51

VI. Conclusions and Future Work

In this paper, we propose to utilize multi-perspective information in English data (L2) by American children and adults, Chinese speech data (L1) by Chinese children and adults, through multi-task-based transfer learning mechanism to improve AMD performance for Chinese ESL children. Both L1 and L2 help in AMD task. When we include all data of L1 and L2 in acoustic modelling, the proposed system achieves the best performance in English AMD task. When we only have a few L1 children data, a wiser adaptation method needed to be explored.

We will continue investigating more efficient adaptation methods against the low-resource condition. Further analysis will be conducted of the relationship between SLM prediction and results of the AMD model.

Acknowledgment

This study was supported by Advanced Innovation Center for Language Resource and Intelligence (KYR17005), Discipline Team Support Program of Beijing Language and Culture University (GF201906), National social Science foundation of China (18BYY124), BLCU support project for young researchers program (19YCX110) (the Fundamental Research Funds for the Central Universities), and, "Intelligent Speech technology International Exchange" Introduced Intelligence Project. Jinsong Zhang is the corresponding author.

References

[1] N. F. Chen, R. Tong, D. Wee, P. X. Lee, B. Ma, and H. Li, "SingaKids-Mandarin: Speech Corpus of Singaporean Children Speaking Mandarin Chinese," 2016.

[2] M. Gerosa, D. Giuliani, and F. J. S. C. Brugnara, "Acoustic variability and automatic recognition of children's speech," vol. 49, no. 10-11, pp. 847-860, 2007.

[3] H. Liao *et al.*, "Large vocabulary automatic speech recognition for children," in *Sixteenth Annual Conference of the International Speech Communication Association*, 2015.

[4] A. Potamianos, S. Narayanan, and S. Lee, "Automatic speech recognition for children," in *Fifth European Conference on Speech Communication and Technology*, 1997.

[5] K. Evanini and X. Wang, "Automated speech scoring for non-native middle school students with multiple task types," 2013.

[6] K.-n. Hassanali, S.-Y. Yoon, and L. Chen, "Automatic scoring of non-native children's spoken language proficiency."

[7] M. Russell, S. D'Arcy, and L. J. I. S. P. L. Qun, "The effects of bandwidth reduction on human and computer recognition of children's speech," vol. 14, no. 12, pp. 1044-1046, 2007.

[8] A. Potamianos, S. J. S. Narayanan, and A. P. I. T. on, "Robust recognition of children's speech," vol. 11, no. 6, pp. 603-616, 2003.

[9] P. G. Shivakumar, A. Potamianos, S. Lee, and S. Narayanan, "Improving speech recognition for children using acoustic adaptation and pronunciation modeling."

[10] J. Fainberg, P. Bell, M. Lincoln, and S. Renals, "Improving Children's Speech Recognition Through Out-of-Domain Data Augmentation," in *Interspeech*, 2016.

[11] S. S. Gray, D. Willett, J. Lu, J. Pinto, P. Maergner, and N. Bodenstab, "Child Automatic Speech Recognition for US English: Child Interaction with Living-Room-Electronic-Devices."

[12] R. Serizel and D. Giuliani, "Vocal tract length normalisation approaches to DNN-based children's and adults' speech recognition," in *Spoken Language Technology Workshop*, 2015.

[13] R. Serizel and D. J. N. L. E. Giuliani, "Deep-neural network approaches for speech recognition with heterogeneous groups of speakers including children," vol. 23, no. 3, pp. 325-350, 2016.

[14] R. Duan, T. Kawahara, M. Dantsujii, and J. Zhang, "Pronunciation error detection using DNN articulatory model based on multi-lingual and multi-task learning," in *International Symposium on Chinese Spoken Language Processing*, 2017.

[15] R. Tong, N. F. Chen, and B. Ma, "Multi-Task Learning for Mispronunciation Detection on Singapore Children's Mandarin Speech," 2017.

[16] R. Duan, T. Kawahara, M. Dantsuji, and J. Zhang, "Multi-lingual and multi-task DNN learning for articulatory error detection," in *Signal & Information Processing Association Summit & Conference*, 2017.

[17] R. Duan, T. Kawahara, M. Dantsuji, J. J. I. T. o. I. Zhang, and Systems, "Articulatory Modeling for Pronunciation Error Detection without Non-Native Training Data Based on DNN Transfer Learning," vol. E100.D, no. 9, pp. 2174-2182, 2017.

[18] J. T. Huang, J. Li, Y. Dong, D. Li, and Y. Gong, "Cross-language knowledge transfer using multilingual deep neural network with shared hidden layers," in *IEEE International Conference on Acoustics*, 2013.

[19] N. Moustroufas, V. J. C. S. Digalakis, and Language, "Automatic pronunciation evaluation of foreign speakers using unknown text," vol. 21, no. 1, pp. 219-230, 2007.

[20] M. Tu, A. Grabek, J. Liss, and V. J. a. p. a. Berisha, "Investigating the role of L1 in automatic pronunciation evaluation of L2 speech," 2018.

[21] R. J. M. l. Caruana, "Multitask learning," vol. 28, no. 1, pp. 41-75, 1997.

[22] S. Gao, B. Xu, H. Zhang, B. Zhao, C. Li, and T. Huang, "Update progress of Sinohear: advanced Mandarin LVCSR system at NLPR," in *Sixth International Conference on Spoken Language Processing*, 2000.

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

[24] K. Shobaki, J. P. Hosom, and R. A. Cole, "The OGI kids² speech corpus and recognizers," in *The Proceedings of the*, 2000.

[25] C. Wen, D. Wang, J. Zhang, and Z. Xiong, "Developing a Chinese L2 speech database of Japanese learners with narrow-phonetic labels for computer assisted pronunciation training," in *Interspeech, Conference of the International Speech Communication Association, Makuhari, Chiba, Japan, September*, 2010.

[26] W. Hu, Q. Yao, F. K. Soong, and W. J. S. C. Yong, "Improved mispronunciation detection with deep neural network trained acoustic models and transfer learning based logistic regression classifiers," vol. 67, pp. 154-166, 2015.