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

UNSW System Description for the Shared Task on Automatic Speech Recognition for Non-Native Children's Speech

Mostafa Shahin[1,2], *Renée Lu*[1], *Julien Epps*[1], *Beena Ahmed*[1]

[1]UNSW, School of Electrical Engineering and Telecommunications, Sydney, Australia
[2]Data61, CSIRO, Tasmania, Australia

Abstract

In this paper we describe our children's Automatic Speech Recognition (ASR) system for the first shared task on ASR for English non-native children's speech. The acoustic model comprises 6 Convolutional Neural Network (CNN) layers and 12 Factored Time-Delay Neural Network (TDNN-F) layers, trained by data from 5 different children's speech corpora. Speed perturbation, Room Impulse Response (RIR), babble noise and non-speech noise data augmentation methods were utilized to enhance the model robustness. Three Language Models (LMs) were employed: an in-domain LM trained on written data and speech transcriptions of non-native children, a LM trained on non-native written data and transcription of both native and non-native children's speech and a TEDLIUM LM trained on adult TED talks transcriptions. Lattices produced from the different ASR systems were combined and decoded using the Minimum Bayes-Risk (MBR) decoding algorithm to get the final output. Our system achieved a final Word Error Rate (WER) of 17.55% and 16.59% for both developing and testing sets respectively and ranked second among the 10 teams participating in the task.

Index Terms: children's speech recognition, non-native children's speech, data augmentation

1. Introduction

Despite the remarkable leap in the performance of Automatic Speech Recognition (ASR) of adult speech in recent years, the accuracy of children's ASR still lags significantly. The shape and size of children's vocal tract alter rapidly as they grow causing a high variation in their acoustic characteristics [1]. Moreover, children tend to use non-standard language with imaginative words, grammatically incorrect sentences, and pronunciation errors.

To model these variations a considerable amount of supervised speech data is needed, however, limited transcribed children's speech corpora are available. Furthermore, non-native speech introduces more acoustic and language challenges to ASR systems due to differences between the speaker's mother language and the spoken foreign language.

To cope with the scarcity of children's speech data, most existing research focuses on leveraging the abundant adult speech corpora. Vocal Tract Length Normalization (VTLN) is one of the early and commonly used methods to alleviate the acoustic mismatch between child and adult speech [2-4]. Other feature-based methods such as Stochastic Feature Mapping (SFM) [5] and Pitch Adaptive Mel Frequency Cepstral Coefficients (PAMFCCs) [6] have also been proposed to effectively combine adult and children's speech for acoustic model training. Recently, deep learning-based domain adaptation techniques were investigated to adapt the adult domain to children's domain [7-9]. However, the most effective way to build a reliable children's acoustic model is still to use substantial amounts of children's speech training data [10]. Language Model (LM) interpolation has also been used to combine adult and children LMs [11]. Here the authors also used confusion matrix-based pronunciation modeling to handle the non-standard pronunciation in children's speech.

In this paper, we detail our design of a children's ASR system. We used large amounts of publicly available children's speech data from 5 different corpora with ~380 hours of speech collected from ~5700 speakers to train our system. We further increased the training data by manipulating the original speech files to simulate different speaking styles and environmental noise effects. Three LMs were employed in this work, two of which were trained using pure children's data, and one was produced by interpolating adult and children's LMs. The system was tested against non-native children English speech from Italian students as part of the first Interspeech shared task on ASR for non-native children's speech.

The rest of the paper is organized as follow, Section 2 contains a brief description of the challenge. Our proposed system is detailed in Section 3, experimental results summarized in Section 4 and conclusions given in Section 5.

2. Challenge Description

The goal of the challenge was to achieve the lowest WER in a test set of non-native English spontaneous speech produced by 3618 Italian students in the context of English proficiency assessment exam. Three datasets were released, 50 hours of transcribed speech for training (TLT-train) and 2 hours each for developing (TLT-dev) and testing (TLT-test).

The data was collected from children aged between 9 to 16 years with three levels of English proficiency. The transcription contained symbols of background speech noise, unrecognized words, mispronounced words, whispered speech, laughter, cough, and other non-speech noises. The transcription also contained special symbols for Italian and German speech.

Two text materials were provided for the training of LM, written data produced by non-native children and manual transcriptions of their speech.

3. System Description

3.1. Speech Corpora

In this work 5 different speech corpora were utilized, the TLT-school speech corpus distributed by the challenge organizers [12], Oregon Graduate Institute (OGI) kids' speech corpus [13], Carnegie Mellon University (CMU) kids' speech corpus [14], Colorado University (CU) Kid's prompted, read and

http://dx.doi.org/10.21437/Interspeech.2020-3111

summarized speech corpus [15, 16] and My Science Tutor (MyST) Children's speech corpus [17].

Other than the TLT corpus, all other datasets were collected from native English speakers with different age ranges that contained both read and spontaneous speech. Manual word-level transcriptions of each dataset were provided with symbols representing speech and non-speech noise events such as laugh, cough, line noise, background speech, etc.

Our acoustic model was trained using the TLT-train along with all the other four datasets. Table 1 summarizes the details of each dataset. As shown, MyST is the largest dataset with around 200 hours of transcribed speech recorded from 677 children through interaction with a virtual science tutor.

Table 1 *The distribution of the speech corpora*

Speech Corpus	Age range	N# of speakers	N# of hours	N# of segments
OGI	5 – 15	794	59.4	42316
MyST	8 – 11	677	208.4	90902
CU	6 – 11	716	49	13999
CMU	6 – 11	54	6.3	3699
TLT-train	9 – 16	3450	59.1	66694
All-train	5 – 16	5691	382.2	217610
TLT-dev	9 – 16	84	2	562
TLT-test	9 – 16	84	2	578

3.2. Data Augmentation

To increase the amount of training data and enhance the robustness of the model against different speaking styles, reverberation, and background noises, we employed four types of data augmentation methods, namely speed perturbation, real Room Impulse Response (RIR) addition, babble noise addition and non-speech noise addition.

Two versions of the speech data with speed factors 0.9 and 1.1 were generated to form the speed perturbed data. The BUT ReverbDB [18] dataset was utilized to simulate the reverberation effect caused by different room environments and generate reverberated speech while babble noise and non-speech noises were simulated using MUSAN corpus [19]. RIR as well as babble and non-speech noises were added to the original and speed perturbed speech data leading to a 9-fold increase in the training data (~ 3500 hours). To keep the training time of the acoustic model reasonable, we randomly sampled 2000 total hours from the manipulated speech data.

3.3. Acoustic Model

Figure 1 presents the process used to state-align the training data. We first trained a GMM-HMM acoustic model using the original version of the training dataset, i.e. without augmentation, to obtain the tied states' alignment (senones) of the training speech corpora for use as outputs of the deep-learning acoustic model. The speech data was segmented into frames of 25 msec with an overlap of 10 msec. A feature vector of size 39 was extracted from each frame containing the 13 MFCC coefficients along with their first and second derivatives. Seven frames were spliced to form one feature vector then projected to a size of 40 by applying the Linear Discriminant Analysis (LDA) algorithm. The projected features were then transformed using the global Semi-Tied Covariance (STC) [20] transform and the resultant features used to train a Speaker-Adaptive Training (SAT) model. We shared the GMMs of all the speech and non-speech noise models as well as the silence model to improve the system performance.

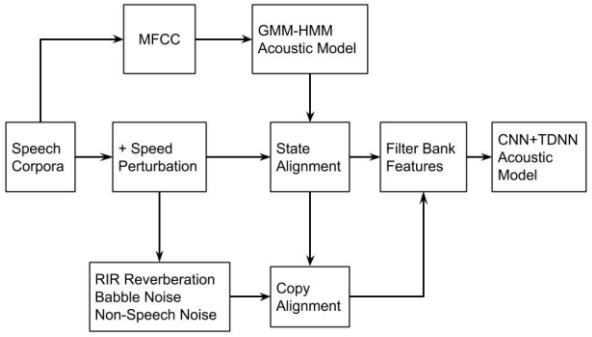

Figure 1 *The acoustic model building flow diagram*

This trained model was also used to align the speed perturbed version of the speech as speed perturbation modified the phoneme duration while for the other augmented data, we used the state alignment corresponding to their original speech.

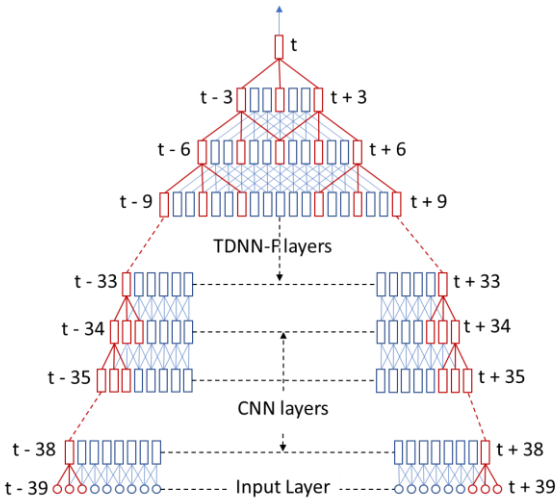

Figure 2 *The deep learning acoustic model architecture. Each block represents a layer in a specific time step. Each layer receives input from multiple time steps of the previous layer. The model consists of 6 lower Convolutional Neural Network (CNN) layers followed by 12 factored Time Delay Neural Network (TDNN-F) layers.*

We then trained a deep learning-based acoustic model from the state-aligned training data. The input features to the model were 40 Mel-frequency filter banks per frame with an additional 100-dimensional per speaker iVector appended to facilitate speaker adaptation [21]. Figure 2 shows the complete architecture of the model. As shown the model consists of 6 CNN layers followed by 12 TDNN-F layers [22]. The effective temporal context of the model is 39 frames before and after the underlying frame (-39, +39).

We adopted the lattice-free Maximum Mutual Information (MMI) training criteria as proposed in [23]. Dropout and L2 regularization techniques were utilized to alleviate the impact of overfitting. Furthermore, batch normalization [24] was applied to the activation outputs of all layers. The acoustic model was trained using the Kaldi ASR toolkit [25].

3.4. Language Model

Three Language Models (LMs) were used in this work. First, the TLT-school, a 4-gram LM trained using the training text provided with the challenge dataset, CHILD, a 4-gram LM trained by combining the challenge text and the transcriptions of all training speech corpora listed in Table 1 and TEDLIUM, a pre-trained 4-gram LM trained on captions of TED talks corpus released by LIUM University [26]. The TLT-school and CHILD LMs were used mainly for first pass decoding while the TEDLIUM LM was used for lattice rescoring.

Table 2 shows the vocabulary size of each LM along with the perplexity (ppl) of development and test sets. As shown, TEDLIUM suffers from high perplexity for both test and development sets. This is due to variation between adults' and children's use of language in addition to the non-native nature of the TLT test and development sets. We thus interpolated between CHILD LM, which includes training text from both native and non-native children, and TEDLIUM LM, to create an output model, TEDLIUM+CHILD.

The standard CMU pronunciation dictionary [27] was utilized to obtain the phoneme sequence of each word, while the language models were trained using SRILM toolkit [28].

Table 2 *The vocabulary size and the test and development perplexity (ppl) for each language model (LM)*

LM	Vocab Size	Dev. ppl	Test ppl
TLT-school	3930	54	56
CHILD	17445	71	72
TEDLIUM	152215	238	221
TEDLIUM + CHILD	154392	77	76

4. Results

Table 3 shows the details and evaluation results of different models starting from the baseline model to the model that achieved the best performance. The table contains details of only those experiments that led to a significant improvement in the performance.

The challenge baseline model consists of 13 TDNN-F layers trained with 9 hours of speech, augmented to 27 hours with two speed perturbated versions. The baseline accuracies of development and test sets were 38% and 35% respectively. A 30% decrease in the WER was achieved by increasing the training data to around 150 hours using an additional 40 hours of speech data provided by the challenge organizers and augmentation using speed perturbation method (Model 1).

Model 2 shows the improvement gained by sharing the states of all silence phones, which included silence, hesitation, cough, laugh, background speech, unrecognized words, and impulsive noises. By analyzing the results from Models 1 and 2, we noticed a dramatic decrease in the insertion errors when using state sharing (~24% decrease). A possible reason is that sharing the states of all silence phones increased the amount of data used to train their models and make them more robust. This therefore improved the alignment of the training data. Given this improvement, state sharing of silence phones was included in all subsequent models.

In Model 3, a slight drop in the WER was observed when adding CNN layers. We then further increased the training data to ~570 hours by utilizing three more data augmentation techniques, the RIR, babble noise, and non-speech noise as explained in section 3.2. As the training data increased, we also boosted the model capacity from 7M parameters (Model 3) to ~17M parameters (Model 4) by adding three more TDNN-F layers. This led to an extension in the effective input temporal context from (-30, +30) frames to (-39, +39) frames. Consequently, the WER dropped from 21.3% and 20%, in Model 3, to 20.5% and 19.2%, in Model 4, for both development and test sets, respectively.

Model 5 was trained using all children's speech corpora listed in Table 1 and the four types of data augmentation to produce a training data of 2200 hours while the architecture of the model was the same as Model 4. The model achieved a WER of 19.6% and 18.8% for the test and development sets respectively when decoding against the 4-gram CHILD LM.

Table 3 *Word Error Rates (WER) of the TLT development (Dev.) and test sets for different model architectures with the Language Models (LM) used for decoding, number of trainable parameters (Num. Param), data augmentation method (Aug), effective temporal context and amount of training data after augmentation in hours. The data augmentation methods included Speed perturbation (SP), Room Impulse Response (RIR), babble noise (Babble) and non-speech noise (Noise).*

	Architecture	LM	Num. Param	Training Corpus	Aug.	Temp Context	Training Hours	Dev WER (%)	Test WER (%)
Baseline	13 TDNN-F	TLT-school	7.8M	TLT-train	SP	(-29, +29)	27.1	37	35
Model 1	13 TDNN-F	TLT-school	8.2M	TLT-train	SP	(-29, +29)	148.5	23.9	22.4
Model 2	13 TDNN-F + ShareSil	TLT-school	8.4M	TLT-train	SP	(-29, +29)	148.5	22	20.4
Model 3	6 CNN + 9 TDNNF	TLT-school	7M	TLT-train	SP	(-30, +30)	148.5	21.3	20
Model 4	6 CNN + 12 TDNNF	TLT-school	17.6M	TLT-train	SP+ RIR+ Babble + Noise	(-39, +39)	573.7	20.5	19.2
Model 5	6 CNN + 12 TDNNF	CHILD	19M	All-train	SP+ RIR+ Babble+ Noise	(-39, +39)	2200	19.6	18.8
LM-rescore of the decoded lattice of model5 with TEDLIUM+CHILD LM								18.11	17.99
Minimum Bayes-Risk (MBR) decoding of lattices from different model configurations								**17.55**	**16.59**

The output lattice of Model 5 was further rescored using TEDLIUM+CHILD LM which was obtained by interpolating the adult TEDLIUM LM and the children's CHILD LM and gave ~7% reduction in WER.

As a final step, we combined rescored lattices from different model configurations and decoded the resultant lattice using the Minimum Bayes-Risk (MBR) decoding algorithm as proposed in [29]. The MBR works to estimate the word sequence that directly minimizes the WER rather than maximize the posterior probability. The best WER of 17.55% and 16.59% were attained by combining lattices from 4 models for development and testing sets, respectively.

Table 4 shows the number of insertion, deletion, and substitution errors in both test and development sets.

Table 4 Breakdown of the system evaluation of development and testing sets in terms of Insertion errors (I), Deletion errors (D), and Substitution errors (S).

	Dev.	**Test**
N# Words	5087	6038
I	212	216
D	360	377
S	321	409
WER (%)	17.55	16.59

5. Conclusions

In this paper, we present a detailed description of our ASR system for children's speech developed for the Interspeech shared task on ASR for non-native children's speech. The acoustic model was based on a deep learning model consisting of 6 CNN layers followed by 12 TDNN-F layers. Each layer received input from proceeding and succeeding time steps of the previous layer leading to an effective temporal context of (-39, +39) time frames. The acoustic model was fed by ~2000 hours of speech generated by data augmentation of ~380 hours of original speech dataset from 5 different native and non-native children speech corpora.

A 4-gram language model estimated from children's transcribed speech was used for the first decoding step. The resultant lattice was rescored using an interpolated language model combining pretrained children's and adult language models. The final decoding step was performed using MBR decoding of a combination of 4 different model configurations lattices. The proposed system achieved a WER of 17.55% and 16.59% when tested against the challenge development and test sets respectively and placed second among the 10 participants.

In this work we used a native English pronunciation dictionary, however, further improvement could be gained by using a knowledge-based or data-driven pronunciation model designed for non-native speech [30].

6. References

[1] R. Mugitani and S. Hiroya, "Development of vocal tract and acoustic features in children," *Acoustical Science Technology and Health Care,* vol. 33, no. 4, pp. 215-220, 2012.

[2] A. Potamianos, S. Narayanan, and S. Lee, "Automatic speech recognition for children," in *Eurospeech'97*, 1997, pp. 2371-2374.

[3] A. Potamianos and S. Narayanan, "Robust recognition of children's speech," *IEEE Trans Audio Speech Lang Process.,* vol. 11, no. 6, pp. 603-616, 2003.

[4] R. Serizel and D. Giuliani, "Vocal tract length normalisation approaches to dnn-based children's and adults' speech recognition," in *SLT*, 2014, pp. 135-140: IEEE.

[5] 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.

[6] S. Shahnawazuddin, A. Dey, and R. Sinha, "Pitch-Adaptive Front-End Features for Robust Children's ASR," in *INTERSPEECH*, 2016, pp. 3459-3463.

[7] P. G. Shivakumar and P. Georgiou, "Transfer Learning from Adult to Children for Speech Recognition: Evaluation, Analysis and Recommendations," *arXiv:1805.03322,* 2018.

[8] J. Li, M. L. Seltzer, X. Wang, R. Zhao, and Y. Gong, "Large-scale domain adaptation via teacher-student learning," *Interspeech,* pp. 2386-2390, 2017.

[9] R. Serizel and D. Giuliani, "Deep neural network adaptation for children's and adults' speech recognition," in *Italian Computational Linguistics Conference (CLiC-it)*, 2014.

[10] H. Liao *et al.*, "Large vocabulary automatic speech recognition for children," in *Interspeech*, 2015, pp. 1611-1615.

[11] P. G. Shivakumar, A. Potamianos, S. Lee, and S. Narayanan, "Improving speech recognition for children using acoustic adaptation and pronunciation modeling," in *WOCCI*, 2014.

[12] R. Gretter, M. Matassoni, S. Bannò, and D. Falavigna, "TLT-school: a Corpus of Non Native Children Speech," *arXiv,* 2020.

[13] K. Shobaki, J.-P. Hosom, and R. A. Cole, "The OGI kids' speech corpus and recognizers," in *Sixth International Conference on Spoken Language Processing*, 2000.

[14] M. Eskenazi, J. Mostow, and D. Graff, "The CMU Kids Corpus LDC97S63," *LDC database,* 1997.

[15] R. Cole, P. Hosom, and B. Pellom, "University of colorado prompted and read childrens speech corpus," in "Technical Report TR-CSLR-2006-02," Center for Spoken Language Research, University of Colorado, Boulder2006.

[16] R. Cole and B. Pellom, "University of colorado read and summarized story corpus," in "Technical Report TR-CSLR-2006-03," Center for Spoken Language Research, University of Colorado, Boulder2006.

[17] W. Ward, R. Cole, and S. Pradhan, "My Science Tutor and the MyST Corpus," 2019.

[18] I. Szöke, M. Skácel, L. Mošner, J. Paliesek, and J. H. Černocký, "Building and evaluation of a real room impulse response dataset," *IEEE JSTSP,* vol. 13, no. 4, pp. 863-876, 2019.

[19] D. Snyder, G. Chen, and D. Povey, "Musan: A music, speech, and noise corpus," *arXiv:.08484,* 2015.

[20] M. J. Gales, "Semi-tied covariance matrices for hidden Markov models," *IEEE Trans Audio Speech Lang Process.,* vol. 7, 1999.

[21] G. Saon, H. Soltau, D. Nahamoo, and M. Picheny, "Speaker adaptation of neural network acoustic models using i-vectors," in *ASRU*, 2013, pp. 55-59: IEEE.

[22] D. Povey, G. Cheng, Y. Wang, K. Li, H. Xu, and S. Khudanpur, "Semi-Orthogonal Low-Rank Matrix Factorization for Deep Neural Networks," in *Interspeech*, 2018.

[23] D. Povey *et al.*, "Purely sequence-trained neural networks for ASR based on lattice-free MMI," in *Interspeech*, 2016.

[24] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," *arXiv:.03167,* 2015.

[25] D. Povey *et al.*, "The Kaldi speech recognition toolkit," in *ASRU*, 2011, no. CONF: IEEE Signal Processing Society.

[26] A. Rousseau, P. Deléglise, and Y. Esteve, "Enhancing the TED-LIUM corpus with selected data for language modeling and more TED talks," in *LREC*, 2014, pp. 3935-3939.

[27] R. L. Weide. (1998). *The CMU pronouncing dictionary*. Available: URL: http://www. speech. cs. cmu. edu/cgibin/cmudict

[28] A. Stolcke, "SRILM-an extensible language modeling toolkit," in *Interspeech*, 2002.

[29] H. Xu, D. Povey, L. Mangu, and J. Zhu, "Minimum bayes risk decoding and system combination based on a recursion for edit distance," *Computer Speech & Language,* vol. 25, 2011.

[30] R. E. Gruhn, W. Minker, and S. Nakamura, *Statistical pronunciation modeling for non-native speech processing*. Springer Science & Business Media, 2011.