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

Advances in Automatic Speech Recognition for Child Speech Using Factored Time Delay Neural Network

Fei Wu[1], *Leibny Paola Garcia*[1], *Daniel Povey*[1,2], *Sanjeev Khudanpur*[1,2]

[1]Center for Language and Speech Processing,
[2]Human Language Technology Center of Excellence,
Johns Hopkins University, Baltimore, MD, USA

fwu24@jhu.edu, lgarci27@jhu.edu, dpovey@gmail.com, khudanpur@jhu.edu

Abstract

Automatic speech recognition (ASR) has shown huge advances in adult speech; however, when the models are tested on child speech, the performance does not achieve satisfactory word error rates (WER). This is mainly due to the high variance in acoustic features of child speech and the lack of clean, labeled corpora. We apply the factored time delay neural network (TDNN-F) to the child speech domain, finding that it yields better performance. To enable our models to handle the different noise conditions and extremely small corpora, we augment the original training data by adding noise and reverberation. Compared with conventional GMM-HMM and TDNN systems, TDNN-F does better on two widely accessible corpora: CMU_Kids and CSLU_Kids, and on the combination of these two. Our system achieves a 26% relative improvement in WER.

Index Terms: speech recognition, child speech, deep neural network

1. Introduction

Automatic speech recognition (ASR) technologies enable automatic conversion of recorded spoken language into written text by computers. Techniques for ASR have evolved from filter bank analysis to Gaussian mixture-based hidden Markov models (GMM-HMM) to, most recently, deep neural networks (DNN) [1, 2]. These advances have lead to application of ASR in various fields, such as education, entertainment, medical assistance, and home automation [3].

Many of these applications can benefit children. For example, interactive reading tutors [4] and automatic reading assessment systems can help school-age and preschool children in learning both native and foreign languages [5, 6]. However, challenges in ASR for child speech have hindered its adoption for such applications. A study in [7] shows that the word error rate (WER) of child speech recognition can be 2 to 5 times higher than ASR for adults. Difficulties in acoustic and language modeling for child speech contribute to this inaccuracy: higher inter-speaker variance due to the development of vocal tract, different formant locations and spectral distribution [8], and the inaccuracy in pronunciation and grammar due to acquisition of languages [9]. More importantly, insufficient data also limits the performance of ASR for children. Adult speech corpora normally contain hundreds of hours of data, while most publicly available child speech corpora have less than 100 hours of data [10, 11].

Though the inaccuracy in ASR for child speech comes from multiple challenges, the major one is the last we have mentioned: a lack of available data. A neural network can learn complicated and highly varied features, leveraging deeper structure or a bigger model, as long as we have enough data to train it. In [12], the performance of convolutional long short term memory (CLDNN) systems and traditional LSTM systems in the scenario of massive training data (607M frames for adults and 459M for children) is studied. With this private child speech corpus of size comparable to the adults' ones, the CLDNN system achieves a state-of-the-art performance (WER of 9.4%). Such performance is competitive with some of the best ASR system for adults, which shows that a DNN is capable of modeling the complicated acoustic features for child speech.

However, most systems do not have the luxury of training their acoustic model with this large of a corpus, as the publicly accessible corpora are quite limited. Several strategies have been taken to tackle this problem. Earlier efforts include refining existing data with a more detailed annotation [13] and capturing the characteristics of speakers in the feature level. Vocal tract length normalization (VTLN), for example, aims to normalize the spectral distribution of different speakers by a transformation on the feature-level. Speaker-adaptive training (SAT), on the other hand, modified the objective function for parameter estimation [14].

In the context of DNNs, recent research focuses more on how to better utilize the available data by augmentation and transfer learning. A fixed-size, context-dependent ($\pm$5 frames) feed-forward DNN model trained on a small child speech corpus and an adult speech corpus is presented in [15]. By applying data augmentation and language model adaptation (a trigram model trained exclusively on transcripts of the training data), the proposed model achieved a WER of 16.9%. A transfer learning technique is presented in [9] to fine tune the top and bottom layers of a DNN acoustic model trained on adults' speech with a child speech corpus, reaching a competitive WER (17.8%).

Another approach to improve ASR for child speech with limited data is to find a more data-efficient network. We demonstrate that factored time delay (TDNN-F), a recently proposed modification of the traditional TDNN network [16], suits this need. Though TDNN-F has never been applied to ASR for children, it has shown impressive data efficiency in different experiment settings for adult ASR. With sufficient data, e.g. 1000 hours, TDNN-F achieves comparable or slightly better performance with about half of the size of a TDNN-LSTM or a BLSTM model. When tested in ASR for low-resource languages (each of which has about 80 hours of available data), TDNN-F outperforms the alternative models using a similar number of parameters. Both cases indicate that TDNN-F is more data-efficient, which is desired for modeling child speech. We found the low-resource scenario particularly encouraging,

http://dx.doi.org/10.21437/Interspeech.2019-2980

as it tackles similar problems to ours in ASR for children.

In this paper we extend previous TDNN-F training techniques to our child speech scenario. We present the progress of our research from the simplest GMM-HMM based system up to the TDNN-F. The effectiveness of the method is validated considering that only 80 hours of data is available for the complete setup. Furthermore, we include a complementary study on VTLN with TDNN-F. Although VTLN decreases WER in GMM-HMM systems, as expected, it shows minimal improvement with TDNN-F. Lastly, we explore the effect of data augmentation on extremely small data sets. By introducing data augmentation, the performance of TDNN-F is further improved on on a matched testing set, yet such improvement cannot be extended into the scenario where there is a mismatch between training and testing sets.

The rest of this paper is organized as follows. Section 2 gives a detailed description of our model. Section 3 shows the setup of our experiments. Section 4 presents and discusses the obtained results, followed by conclusions and future work in Section 5.

2. System Description

2.1. TDNN-F in Acoustic Modeling

Our system uses a hybrid DNN-HMM structure for acoustic modeling, where the HMM models each triphone with 3 states, and the emission probabilities of all states are estimated by one TDNN-F. As the TDNN-F is a refinement of the traditional TDNN, it is worthwhile to discuss the model with TDNN structure, and then escalate it into a TDNN-F model.

A traditional time delay neural network (TDNN) can be regarded as a fully-connected 1-dimensional convolutional neural network. In each layer, the input at time step t is fed forward to the next layer with its neighbors within a certain context window [17]. As the network gets deeper, it has the ability to capture features in a wider context window, but the overlap among frames makes this network computationally inefficient. As shown in Figure 1, instead of passing every frame to the next layer, we can sub-sample the input sequence and pass only the sampled frames (shown in blue). The window size at each layer is tuned as a hyper-parameter [18].

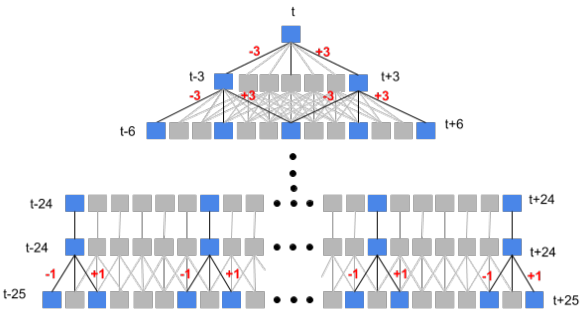

Figure 1: *TDNN with sub-sampling*

TDNN-F further improves the computation efficiency of the network by using singular value decomposition (SVD), decomposing the weight matrix of each layer into an approximation as the product of two lower rank matrices [16]:

$$\mathbf{W} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^{\mathbf{T}} = \mathbf{M}\mathbf{N}$$

where $\mathbf{\Sigma} \in \mathbb{R}^{m \times n}$ is a non-negative, rectangular diagonal matrix, $\mathbf{M} \in \mathbb{R}^{m \times k}$, and $\mathbf{N} \in \mathbb{R}^{k \times n}$. By choosing a suitable value of $k \leq \min\{m, n\}$, we can easily reduce the total number of parameters for this transformation. For this approach, we need to ensure that one of the two sub matrices is close to a semi-orthogonal matrix, as it is the equivalent of $\mathbf{U}\mathbf{\Sigma}$ (or $\mathbf{\Sigma}\mathbf{V}^{\mathbf{T}}$). When training the network, after every a few updates of the whole network, we specifically update $\mathbf{N}$ with SGD using an additional objective function to guarantee that $\mathbf{N}$ is not too far from being semi-orthogonal [16]. Since k is much smaller than m and n, TDNN-F can also be considered as introducing an extra bottleneck layer into the traditional TDNN. Figure 2 shows the factored version of the top layer from Figure 1. [16] also also illustrates that a 3-stage splicing structure can further improve the performance of TDNN-F by inserting two (instead of one) bottleneck layers, and allows convolution in between the bottleneck layers. This structure is similar to deep residual network [19], a deep CNN structure widely adopted in the field of computer vision.

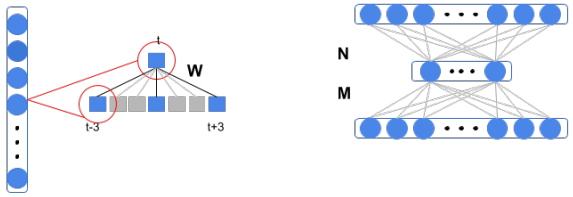

Figure 2: *Factored TDNN Layer*

To train a TDNN-F acoustic model, a GMM-HMM model is trained to provide labeled data by aligning the frames to phone states. The hybrid model is then trained using lattice-free maximum mutual information (LF-MMI). MMI criterion maximizes the posterior probability of the transcript given the audio signal, but it involves summing the joint probability over all possible word sequences allowed by the acoustic and language models (LM) in the system. MMI can be approximated by summing over the lattice instead of all possible sentences, but it is still computationally expensive. LF-MMI uses a phone-level language model instead of a word-level language model (backed by a lexicon), which allows a smaller frame rate, a faster decoding speed, and improvement in accuracy [20]. We interpolate the LF-MMI loss with cross-entropy loss during training. An additional output layer was built to calculate the cross-entropy loss. Only the LF-MMI output layer is used during decoding.

3. Experiment Setup

All of the experiments were carried out using the Kaldi toolkit [21]. In this section, we provide details of the corpora, data augmentation process, and the two models we used in our experiment.

3.1. Corpora and Data Preparation

We trained and tested our models using 2 child speech corpora, CMU_Kids [22] and CSLU_Kids [23], both of which are publicly available [24, 25]. The CMU_Kids corpus contains 5180 utterances, recording 78 speakers reading 1 sentence per utterance. The (scripted part of) the CSLU_Kids corpus contains

around 100 speakers saying a single word or reading a short phrase in each utterance. As shown in Table 1, we randomly reserved 30% of the utterances from each corpus for development testing. All of our models were trained on both of the corpora separately, as well as on the combined training set.

Table 1: *Child Speech Corpora and Training/Test Splits*

	CMU_Kids		**CSLU_Kids**		**Combined**	
	Train	Dev	Train	Dev	Train	Dev
# of Utterances	3621	1559	50026	21354	53647	2213
Duration (hours)	6.34	2.76	48.56	20.75	54.90	23.51

Since CMU_Kids is particularly small in size, and it is noisier than CSLU_Kids, we trained a TDNN-F model with augmented training set from CMU_Kids by adding babble noise and reverberation. Babble noise was created by combining 3 to 5 speakers from the Mixer-6 corpus [26], and then added to the original training set. On top of additive babble noise, we simulated room reverberation using the RIRS_NOISES database [27]. In total, we generated 2 augmented copies (babble noise and babble noise with reverberation) of the data and used them alongside the original, clean training data as the training set of the TDNN-F. Alignments used for the new augmented data come from their corresponding clean copies.

Mel-frequency cepstral coefficients (MFCC) were used as the front-end feature. We extracted MFCC features from a 25 ms window, and a frame rate of 10 ms. For the GMM-HMM system, 13 MFCCs and their corresponding Δ and Δ-Δ features were used. For TDNN-F system, high-resolution MFCCs were used. The input vector is the concatenation of 40-dimension cepstral coefficients of both the current and neighbor frames, and a 100-dimension i-vector of the current frame.

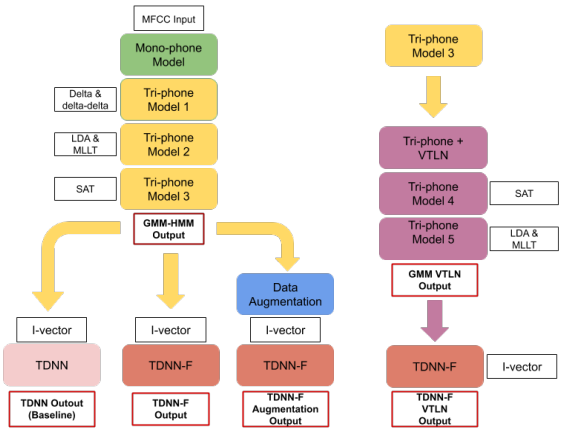

Figure 3: *Feature Processing and Model Training Recipes*

3.2. TDNN-F System

For the TDNN-F network, we followed the three-stage splicing structure [1]. As shown in Figure 3, we started by building a traditional 3-state left-to-right GMM-HMM triphone model to provide aligned (frame to phone state) training data for the DNN system. For better alignment, linear discriminant analysis (LDA), feature-space maximum likelihood linear regression (fMLLR), and speaker adaptive training (SAT) are included.

Our model is then trained using the aligned data. High-resolution MFCC of the current (t) and neighboring ($t-1, t+1$) frames were concatenated with an i-vector of the current frame, and used as the input feature. After transferring the input feature into the hidden dimension (1024), 12 TDNN-F layers were used as the hidden layers. Each TDNN-F layer can be regarded as a large TDNN layer (with a dimension of 1024), followed by two bottleneck layers (equivalent of two small TDNN layers with a dimension of 256). All (large and small) hidden layers concatenate the current frame with either the left or right neighbor before forwarding it to the next layer.

On top of the hidden layers, two separate output layers were built: one for the LF-MMI objective function and another for the cross-entropy objective function. When training the model, the losses with respect to the two objective functions were interpolated, while in decoding, only the LF-MMI output was used.

3.3. Baseline System

To illustrate the effectiveness of TDNN-F, we also built a traditional TDNN as the baseline system[2]. The baseline network was constructed to have a comparable size (in number of parameters) as the TDNN-F. The baseline system has the same input as the TDNN-F system, and the same number of hidden layers (12). Each hidden layer is a TDNN layer with a dimension of 768, and a window size of 1 (on both sides). Since the input and output layer of the two networks are approximately the same, we compare the size of two networks by comparing the number of parameters in each hidden layer, and the baseline system is about 1.5 times in size as the TDNN-F system.

3.4. Language Model

Both the baseline model and TDNN-F model use the same lexicon and language model (LM). We uses the CMU Pronunciation Dictionary as the lexicon, and the LM is a 3-gram model trained on LibriSpeech. An additional phone-level 4-gram LM was trained from the lexicon and used in the LF-MMI training for the TDNN-F.

4. Results and Discussion

All of our models described in Section 3 were tested with the datasets described in Section 3.1. Table 2 shows a comparison among all of our models. *Dev* denotes the development set from the same corpus as the training set, while *Test* denotes the development set from the combined set.

Table 2: *WER Comparison of Different Acoustic Models*

	CMU_Kids		**CSLU_Kids**		**Combined**
	Dev	Test	Dev	Test	Dev/Test
GMM	29.6%	75.9%	32.5%	38.4%	36.5%
GMM + VTLN	29.5%	**74.9%**	31.6%	37.4%	35.8%
TDNN *(baseline)*	20.1%	77.9%	13.7%	24.9%	15.8%
TDNN-F	**17.3%**	77.0%	**10.8%**	22.4%	**11.7%**
TDNN-F + VTLN	17.6%	78.2%	13.3%	**22.3%**	11.9%

[1] Adapted from the Kaldi-Material recipe.

[2] Adapted from the Kaldi MiniLibriSpeech recipe

4.1. Word Error Rate Analysis

The first row shows the WER result from the GMM-HMM system, which is used to align the training data for the baseline TDNN and TDNN-F system; the second and third row show a comparison between the TDNN and the TDNN-F systems. By comparing these two rows, it can be seen that TDNN-F outperforms the baseline in almost all combinations of training and testing data. On the combined data set, the TDNN-F achieves a relative improvement of 26%.

Though TDNN-F shows great strength in the *Dev* sets, it is worth noticing that when the training set is extremely small, i.e., on CMU_Kids, TDNN-F has no such advantage against the GMM-HMM model on the *Test* set. Part of the reason is that TDNN-F, like all other neural models, suffers more from the insufficiency of training data due to the massive number of parameters. Another reason is that, comparing to CSLU_Kids, CMU_Kids has more noise. Data augmentation is introduced with the expectation that it would compensate for the insufficiency and mismatch of data. More discussion on augmentation follows in Section 4.2.

Since previous research [8] has shown the effectiveness of VTLN on GMM-HMM models for recognizing child speech, we also explored applying VTLN along with TDNN-F. We find however, by comparing the first and last two rows in Table 2, that while VTLN does improve the WER performance of GMM-HMM model as reported in [8], it doesn't have a significant effect on TDNN-F.

4.2. Performance on Small Dataset

To further study how models are effected when the training set is extremely small, i.e., around 6 hours in our case, we compare the performance of different models (GMM, TDNN, TDNN-F) trained on only CMU_Kids. We examine WER separately on the CMU_Kids and CSLU_Kids development sets to study the effect of matched versus mismatched noise conditions in training and test, showing how well the models generalize.

Table 3: *WER for Very Small Training and Mismatched Test Sets*

	CMU_Kids	**CSLU_Kids***(Mismatched)*
GMM	29.5%	**85.4**%
TDNN *(baseline)*	20.1%	89.8%
TDNN-F	17.3%	89.1%
TDNN-F + Aug	**16.0**%	89.0%

As seen in Table 3, GMM outperforms both neural models in the mismatched scenario as expected because it has less parameters. Though showing no significant advantage over TDNN in the mismatched setting, augmentation does compensate for the shortage in training data, and improve the WER on the matched test set.

4.3. Error Analysis

Table 2 shows that TDNN-F is generally a better model than the baseline, but it gives limited information about the kind of mistakes that the models made. Table 4 presents the details of the three types of mistakes measured in WER: insertion, deletion and substitution. For each model, the first row shows the number of mistaken words that fall into each category, followed by the percentage of each category in the total WER in the second row. For TDNN-F, we also show their relative improvement compared to the baseline.

Table 4: *ASR Error Breakdown of Children's Speech*

	-	Total	Ins	Del	Sub
TDNNN (baseline)	# Errors	11560	2168	3210	6182
	Fraction	—	18.8%	27.8%	53.5%
TDNN-F	# Errors	8552	1681	2794	4077
	Fraction	—	19.7%	32.7%	47.7%
	Improvement	26%	22%	13%	34%

Compared to ASR systems for adults that have a similar WER as our TDNN-F system, a larger portion of the total word errors in childrens' speech comes from insertions and deletions. For example, in one of the example recipes for LibriSpeech from Kaldi, the total WER is around 12%, and substitutions make up 80% of the errors. This is not surprising, considering that both the test sets we use contain children trying to read, and most of them are still learning how to read. Here is an example transcript given by the TDNN-F system and its reference transcript provided by the CMU_Kids corpus:

> **TDNN-F** : If a lightning storm comes there are four things you can do to *say stay sick help stay help* stay safe.
> **Reference**: If a lightning storm comes there are four things you can do to stay safe.

When measuring WER, this sentence is ruled to have 6 insertion errors. But when we listened to the recording, we found that the speaker was clearly struggling with the phrase "stay safe", as she attempted it multiple times, and hesitated between saying "stay health (healthy)" and "stay safe". This points to potential transcription issues and their impact, albeit likely to be small, on the measured WER in the two corpora we are using.

5. Conclusions and Future Work

By comparing factored time delay neural networks (TDNN-F) with different traditional and state-of-the art systems, we demonstrate the efficacy of TDNN-F for the task of automatically recognizing child speech. We build a TDNN-F system that outperforms its alternatives in datasets with various sizes. We explore the impacts of vocal track length normalization (VTLN) and data augmentation on the performance of TDNN-F system. Though effective with traditional models like GMM-HMM, VTLN has no significantly impacts on TDNN-F. When trained with an extremely small dataset, data augmentation helps improve the performance of TDNN-F on test data in the same channel condition as the training set.

For future research, we would like to explore the potential improvements brought by the acoustic characteristics of child speech. Formants, for example, can be particularly useful in recognizing the vowels in the speech. Another problem that needs to be solved is to tackle the inaccuracy of evaluating the models with WER when there is a mismatch between the actual recordings and the provided scripts. A biased language model, or a sub-word-level adjustment might be some good directions to go for identifying and cleaning up the reference transcripts of child speech. We are also interested in strengthen the training data sets by adding in adult speech.

6. References

[1] B.-H. Juang and L. R. Rabiner, "Automatic speech recognition - a brief history of the technology development," no. 1, 2004.

[2] G. Hinton, L. Deng, D. Yu, G. E. Dahl, A. Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, T. N. Sainath, and B. Kingsbury, "Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups," *IEEE Signal Processing Magazine*, vol. 29, no. 2, pp. 82–97, Nov 2012.

[3] J. Vajpai and A. Bora, "Industrial applications of automatic speech recognition," *International Journal of Engineering Research and Applications*, vol. 6, no. 3, pp. 88–95, 2016.

[4] J. Mostow, "Why and how our automated reading tutor listens," no. 4, 2012.

[5] K. Zechner, K. Evanini, and C. Laitusis, "Using automatic speech recognition to assess the reading proficiency of a diverse sample of middle school students," in *Third Workshop on Child, Computer and Interaction*, 2012.

[6] K. Evanini and X. Wang, "Automated speech scoring for non-native middle school students with multiple task types." in *INTERSPEECH*, 2013, pp. 2435–2439.

[7] A. Potamianos and S. Narayanan, "Robust recognition of children's speech," *IEEE Transactions on Speech and Audio Processing*, vol. 11, no. 5, pp. 603–616, Nov 2003.

[8] H. Hermansky, "Perceptual linear predictive (plp) analysis of speech," *the Journal of the Acoustical Society of America*, vol. 87, no. 4, pp. 1738–1752, 1990.

[9] P. Gurunath Shivakumar and P. Georgiou, "Transfer Learning from Adult to Children for Speech Recognition: Evaluation, Analysis and Recommendations," *arXiv e-prints*, no. 6, p. arXiv:1805.03322, May 2018.

[10] 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.

[11] F. Claus, H. G. Rosales, R. Petrick, H.-U. Hain, and R. Hoffmann, "A survey about databases of children's speech." in *INTERSPEECH*, 2013, pp. 2410–2414.

[12] H. Liao, G. Pundak, O. Siohan, M. K. Carroll, N. Coccaro, Q.-M. Jiang, T. N. Sainath, A. W. Senior, F. Beaufays, and M. Bacchiani, "Large vocabulary automatic speech recognition for children," in *INTERSPEECH*, no. 9, 2015.

[13] Q. Li and M. J. Russell, "An analysis of the causes of increased error rates in children's speech recognition," in *INTERSPEECH*, no. 7, 2002.

[14] 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*, ser. WOCCI '09, no. 8. New York, NY, USA: ACM, 2009, pp. 7:1–7:8.

[15] M. Qian, I. McLoughlin, W. Quo, and L. Dai, "Mismatched training data enhancement for automatic recognition of children's speech using dnn-hmm," in *2016 10th International Symposium on Chinese Spoken Language Processing (ISCSLP)*, no. 10, Oct 2016, pp. 1–5.

[16] D. Povey, G. Cheng, Y. Wang, K. Li, H. Xu, M. Yarmohamadi, and S. Khudanpur, "Semi-orthogonal low-rank matrix factorization for deep neural networks," in *Proceedings of the 19th Annual Conference of the International Speech Communication Association (INTERSPEECH 2018), Hyderabad, India*, no. 13, 2018.

[17] A. Waibel, T. Hanazawa, G. Hinton, K. Shikano, and K. J. Lang, "Phoneme recognition using time-delay neural networks," *Backpropagation: Theory, Architectures and Applications*, no. 13, pp. 35–61, 1995.

[18] 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*, no. 15, 2015.

[19] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[20] D. Povey, V. Peddinti, D. Galvez, P. Ghahremani, V. Manohar, X. Na, Y. Wang, and S. Khudanpur, "Purely sequence-trained neural networks for asr based on lattice-free mmi." in *Interspeech*, 2016, pp. 2751–2755.

[21] 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," IEEE Signal Processing Society, Tech. Rep., 2011.

[22] M. Eskenazi, J. Mostow, and D. Graff, "The cmu kids corpus," *Linguistic Data Consortium*, no. 11, 1997.

[23] 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.

[24] M. Eskenazi and J. Mostow, "The cmu kids speech corpus (ldc97s63)," 2006.

[25] K. Shobaki, J.-P. Hosom, and R. Cole, "Cslu: Kids speech version 1.1," *Linguistic Data Consortium*, 2007.

[26] C. Cieri, W. Andrews, J. P. Campbell, G. Doddington, J. Godfrey, S. Huang, M. Liberman, A. Martin, H. Nakasone, M. Przybocki *et al.*, "The mixer and transcript reading corpora: Resources for multilingual, crosschannel speaker recognition research," MASSACHUSETTS INST OF TECH LEXINGTON LINCOLN LAB, Tech. Rep., 2006.

[27] T. Ko, V. Peddinti, D. Povey, M. L. Seltzer, and S. Khudanpur, "A study on data augmentation of reverberant speech for robust speech recognition," in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2017, pp. 5220–5224.