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

Improving ASR Systems for Children with Autism and Language Impairment Using Domain-Focused DNN Transfer Techniques

Robert Gale, Liu Chen*, Jill Dolata, Jan van Santen, Meysam Asgari**

Center for Spoken Language Understanding (CSLU)
Oregon Health & Science University (OHSU), Portland, OR
{galer, chliu, dolataj, vansantj, asgari}@ohsu.edu

Abstract

This study explores building and improving an automatic speech recognition (ASR) system for children aged 6-9 years and diagnosed with autism spectrum disorder (ASD), language impairment (LI), or both. Working with only 1.5 hours of target data in which children perform the Clinical Evaluation of Language Fundamentals Recalling Sentences task, we apply deep neural network (DNN) weight transfer techniques to adapt a large DNN model trained on the LibriSpeech corpus of adult speech. To begin, we aim to find the best proportional training rates of the DNN layers. Our best configuration yields a 29.38% word error rate (WER). Using this configuration, we explore the effects of quantity and similarity of data augmentation in transfer learning. We augment our training with portions of the OGI Kids' Corpus, adding 4.6 hours of typically developing speakers aged kindergarten through 3rd grade. We find that 2nd grade data alone — approximately the mean age of the target data — outperforms other grades and all the sets combined. Doubling the data for 1st, 2nd, and 3rd grade, we again compare each grade as well as pairs of grades. We find the combination of 1st and 2nd grade performs best at a 26.21% WER.

Index Terms: speech recognition, children speech recognition, autism spectrum disorder, language impairment, deep neural network, transfer learning.

1. Introduction

Autism spectrum disorder (ASD) is a behaviorally defined neurodevelopmental disorder that is characterized by impairments in social communication and patterns of restricted and repetitive behavior. Children with ASD may or may not have a comorbid language impairment. To determine the presence of a language impairment, it is necessary to perform a comprehensive clinical language evaluation. The Clinical Evaluation of Language Fundamentals edition 4 (CELF-4) [1] is one example of a standardized, norm-referenced language assessment used in a comprehensive evaluation. Administration of this test, however, is quite time consuming; the CELF-4 includes up to 18 subtests, all of which are currently scored on paper by speech language pathologists [2]. Automation of such a test would allow professionals more time to complete other aspects of the comprehensive assessment.

Recent advancements in automatic speech recognition (ASR) have caused the field of human-machine interactions to evolve to a new era in which innovative voiced-based technologies such as voice-activated digital assistants (e.g., Microsoft's Cortana and Amazon's Alexa) have become part of daily life. Recent improvements of ASR systems are mostly due to the advent of deep neural networks (DNN) that have revolutionized many areas of machine learning including speech recognition and natural language processing (NLP). In addition, recent parallel processing advancements, made available through graphics processing units, have significantly boosted the computational power required in intensive training tasks, for example, in ASR systems [3, 4].

A wealth of data is the key to successfully training a DNN-based ASR system. As suggested in prior research, there exists a strong correlation between the performance of ASR systems and both the quantity and quality of training samples. This is particularly important in scenarios with limited training resources, for example, in medical applications. Similar challenges remain in development of ASR systems for low-resource languages such as Zulu [5]. To cope with data limitations, efforts have been made in recent years for creating speech corpora that are suitable for training and evaluating ASR systems. As an example, LibriSpeech [6], a corpus of 1000 hours of English read speech derived from audiobooks, is now freely available to the public. While large, open source corpora for adults have become widely available, there are few such corpora for children's speech. The differences between adult and child speech preclude adult corpora from being directly used in training ASR models used in applications that interact with children. Moreover, datasets for children with language impairment and/or neurodevelopmental disorders are rare indeed.

In our study, developing a specialized ASR system is a core component as part of a larger system to automate scoring of children with language impairment (SLI), autism (ALN), and both autism and language impairment (ALI). Our target dataset includes less than 1.5 hours of speech from children performing a sentence repetition diagnostic task. In addition to challenges we encounter due to data scarcity, ASD and language impairments compound the difficulty of ASR training. Among the unique ASR challenges of this dataset are short vocal tracts, a variety of non-verbal vocalizations, and the wide phonological variance of both typical and atypical speech development. With the hope of overcoming some of these challenges, our aim is to employ a "transfer learning" approach to improve the performance of accuracy of our ASR systems built on no more than a couple hours of data. In particular, we explore a number of weight transfer configurations to empirically determine the optimal DNN architecture.

The structure of our experiments in this study are as follows: first we explore strategies in transfer learning to optimize the hyperparameters for the CSLU Autism Speech Corpus, setting up an effective baseline approach for later experiments. Second, we incorporate the OGI Kids' Corpus [7], with scenarios designed to search the augmentation data broadly, learning how data from each age and grade contributes to ASR performance. Finally, we narrow our focus based on the

*Equal contribution

http://dx.doi.org/10.21437/Interspeech.2019-3161

findings from the second experiment, exploring the effects of adding larger quantities of the best performing segments.

1.1. Transfer Learning

Early attempts to improve the performance of small dataset ASR systems focused on adaptation techniques, in which acoustic parameters of the source model (the larger dataset) are transformed to mimic the acoustic properties of the target speakers. Mustafa et al. [8] successfully employed a maximum a posteriori (MAP) adaptation [9] technique to overcome the data scarcity in development of an ASR system for speakers with speech articulation disorders. Building the source model on two larger unimpaired TIMIT [10] and impaired TORGO [11] speech corpora, they achieved a significant improvement in performance of adapted ASR model on recognizing impaired speech. More recently, adaptation techniques based on "transfer learning" have become popular for DNN-based ASR systems. Transfer learning aims to reuse the learned knowledge of the source model in a related target domain [12]. The method is especially useful in data-scarce scenarios where sufficient training data is unavailable for learning the source model. One can categorize several kinds of proposed transfer learning methods into three major approaches: *weight transfer*, *multi-task training*, and *domain assimilation*. Assuming the low-level features of the source and the target data are similar (e.g., lines or dots in various object recognition tasks), the *weight transfer* approach initializes the parameters (weights) of the target DNN model using the learned parameters of the source model. It has been empirically shown that given a small amount of data in the target domain, *weight transfer* can greatly improve the performance in addition to speeding up the training process due to the small number of parameters required to be fine-tuned [13]. In *multi-task training*, the shared hidden layers of the DNN model are simultaneously learned by multiple tasks with the hope of benefiting from correlated information shared across tasks [14, 15]. Unlike the two aforementioned approaches, *domain assimilation* selectively utilizes a subset of target data that is more similar to the target domain. Although it is one of the popular topics in image processing [16], *domain assimilation* has scarcely been explored in the ASR field.

2. Data

2.1. CSLU Autism Speech Corpus: Recalling Sentences

The primary target of our ASR efforts, the CSLU Autism Speech Corpus, is a dataset gathered between 2005 and 2012 at Oregon Health & Science University (OHSU) during the clinical administration of ASD evaluations. Each member of the study group was evaluated using the CELF-4. [1] The CELF-4 is an individually administered language assessment, published by The Psychological Corporation and used by speech language pathologists to determine if a child has a language disorder or delay. The specific task that was recorded and provided the speech data for this experiment is the *Recalling Sentences* task. In this task, the examiner reads sentences of increasing length and syntactic complexity (32 sentences in English) (the *prompt*), and the child is asked to repeat the sentence verbatim. These audio recordings amount to 1.5 hours of recorded speech.

The study group includes 45 children, ages 6 to 9 years (mean age of 7 years 4 months), 17 of whom were diagnosed as having an ASD as well as a language impairment (ALI), 15 with ASD but no language impairment (ALN), and 13 with specific language impairment (SLI) but no ASD. The study group included 36 males and 9 females.

Table 1: *Breakdown of training data selected from the CSLU Autism and OGI Kids' corpora*

Source	Speakers	Utterances	Hours
CSLU Autism	**45**	**1022**	**1.5**
ALN (ASD without LI)	15	438	0.7
ALI (ASD with LI)	17	335	0.4
SLI (LI without ASD)	13	249	0.4
OGI Kids	**340**	**2796**	**4.6**
Kindergarten	85	699	1.1
1st grade	85	699	1.2
2nd grade	85	699	1.2
3rd grade	85	699	1.2

2.2. Data Augmentation: OGI Kids' Corpus

Our data augmentation set, the OGI Kids's Speech Corpus [7], was created as an effort by the Oregon Graduate Institute of Science and Technology's (OGI) Center for Spoken Language Understanding (CSLU), an institution which has since been become a part of OHSU. This corpus was recorded at the Northwest Regional School District in the Portland, Oregon area and consists of two distinct examples of spoken language: one spontaneous and one read. One portion of the OGI corpus — the sentence reading task — was of particular interest for its similarity to our target data from Section 2.1, Recalling Sentences. Each recording has been evaluated for quality by the corpus authors, and we limited the data to only those rated at the best audio quality rating of "good." The corpus includes voice recordings of 1,100 participants in kindergarten through 10^{th} grade, but only data from grades kindergarten through 3^{rd} grade were considered to limit the age ranges to those in the CSLU Autism Speech corpus. In order to control for speaker variance, we limited the number of speakers and utterances from each grade to the counts from the smallest grade (the kindergartners) at 85 speakers and 699 utterances. Details of the data selected from the OGI Kids' Corpus, as well as the data selected from the CSLU Autism Corpus are shown in Table 1.

2.3. Comparing age statistics of the two corpora

While the CSLU Autism corpus has precise age information at the time of testing, the OGI kids' corpus only has each participant's grade in school. To get a sense of how the age demographics compare, we used the following algorithm to convert participants' ages into grades:

$$g = \text{round}(a - 5)$$

where a is the participant's age at the time of the exam, and g is the whole number representing the grade, with kindergarten as 0, 1^{st} grade as 1, and so on. As a result of this calculation on the CSLU Autism corpus, we found a grade range of 1 through 3, with a mean of grade of 1.89, or, just under 2^{nd} grade. Table 2 presents the demographics of participants.

3. Experiments

For each experiment, the data from the CSLU Autism Speech Corpus was split into five folds to balance the sets by the

Table 2: *Mean age and mean calculated grade demographics of the participants from the CSLU Autism Speech Corpus. Kindergarten would be grade zero.*

Diagnosis	Age (std)	Grade (std)
any	7.31 (0.81)	1.89 (0.80)
ALN	7.26 (0.85)	1.80 (0.86)
ALI	7.23 (0.88)	1.88 (0.86)
SLI	7.49 (0.69)	2.00 (0.71)

speaker's ALN/ALI/SLI diagnosis. Rotating through the five folds, a model was trained on four folds and tested on the remaining fifth fold. For experiments augmented with the data from the OGI Kids' Corpus, the supplemental material was simply combined with the train set. The reported word error rate (WER) is calculated as the sum of errors across all test sets divided by the total number of correct words.

3.1. Exp. 1: Finding an optimal weight transfer strategy

3.1.1. Methods

In order to establish the weight transfer approach for later experiments, we set up an exploratory experiment to determine the optimal hyperparameter configurations. This experiment was conducted on the CSLU Autism Speech corpus only.

We pre-trained a time-delay neural network (TDNN) model on the freely available LibriSpeech corpus [6], which consists of 1000 hours of volunteers reading books aloud. This model was trained using the lattice-free maximum mutual information (MMI) ("chain") method as described in [17], and was built using the pre-configured scripts included in the Kaldi toolkit [18]. The DNN architecture consists of a pre-trained and fixed linear discriminant analysis (LDA) input layer, six TDNN hidden layers, and a fully connected output layer. Batch normalization is employed throughout the input and hidden layers in addition to L2 regularization and cross-entropy normalization.

We used the transfer learning approach described in [13] to adapt the LibriSpeech model to our target data. First, high-resolution mel-frequency cepstral coefficient (MFCC) features were extracted (with 40 mel bins and 40 cepstra) in addition to 100-length i-vectors calculated based on the same model trained for the LibriSpeech model. The LibriSpeech model is used to generate frame-level acoustic alignments for training. For the lattice-free MMI objective, we weighted the training values at a 10 to 1 ratio to the LibriSpeech data.

The weights were copied from the LibriSpeech model to initialize the target model, and various learning configurations for the layers were strategically explored. The general strategy was to first try all combinations of entirely trainable at the overall learning rate (1.0) and entirely frozen (0.0). We trained with these configurations over 5 epochs with an overall learning rate starting at 0.001 for epoch 0, decreasing linearly at each iteration to end at 0.0001. With those results, we tried a number of configurations between 0.0 and 1.0. Upon decoding, we applied a weighted finite state transducer (WFST) language model built from the training data.

3.1.2. Results

On our dataset, it appears that the output layer contained the most important weights to retrain. The best results were consistently the experiments with the output layer set to learn at the full rate of 1.0. Only training the output layer resulted in a 40.00% WER. Training the hidden layers as well as the output layer at a rate of 1.0 gave a WER of 30.85%. Restricting the hidden layers to a partial rate of 0.3 yielded the best performance of 29.38%. Retraining the input (LDA) layer had a disadvantageous effect, performing at 81.04%, which was worse than the pure LibriSpeech weights' performance of 80.75%. Highlights of the experimental results are presented in Table 3.

Table 3: *Optimizing learning rate factors (LRFs) to get the lowest word error rate, trained and tested on only the CSLU Autism data. A layer's learn rate is equal to the LRF times the overall learn rate.*

Learning Rate Factors			**Test WER**
Input	Hidden	Output	
0.0	0.3	1.0	**29.38%**
0.0	0.1	1.0	30.24%
0.0	1.0	1.0	30.85%
0.0	0.0	1.0	37.87%
0.0	1.0	0.0	58.11%
1.0	0.0	0.0	78.09%
0.0	0.0	0.0	83.19%

3.2. Exp. 2: Data augmentation and age domain selection

3.2.1. Methods

In the next phase of our study, we looked at the effect of augmenting the CSLU Autism dataset with the OGI Kids' dataset. We explored the effects of training on the full K-3 set versus augmenting with each individual grade separately. We also randomly selected 25% of each grade for a smaller K-3 set that matches the individual grades in quantity. Using the approach we settled on in Section 3.1.1, we transferred the weights from the LibriSpeech model, and retrained with a combination of the CSLU Autism data and each OGI Kids' segment.

Testing was performed on the whole of the CSLU Autism data using five-fold cross validation, as described in Section 3. We also tested the performance of each model on the three diagnoses separately for a finer-grained look at domain similarities between the CSLU Autism data and the OGI Kids' data.

3.2.2. Results

Training with the full augmentation set yielded a 28.60% WER, a 0.78% absolute improvement on the unaugmented performance from Section 3.1.2. Training with each smaller, age-specific portion of the augmentation set consistently performed better than the larger K-3 set. The smaller K-3 set was second best at 27.61%. The best result was found augmenting only with the 2nd graders, with a WER of 27.07%. This was an absolute improvement of 1.53% versus the larger augmentation set, and a 2.31% absolute improvement versus the unaugmented performance. Note how we determined in 2.3 that our target data had an approximate mean age of 2nd grade.

Testing the diagnoses separately, we found the 2nd grade subset performed best across all test splits except the subset of children with both diagnoses (ALI), where 700 utterances of the combined grades had a WER of 29.95% versus 2nd grade's 30.05%. Considering children with only an autism diagnosis (ALN), we found a WER of 20.51%, a 2.64% absolute improvement on the unaugmented model. The subset of children without an autism diagnosis but with LI (SLI) had a WER of 36.60%, improving on the unaugmented model by 3.55% abso-

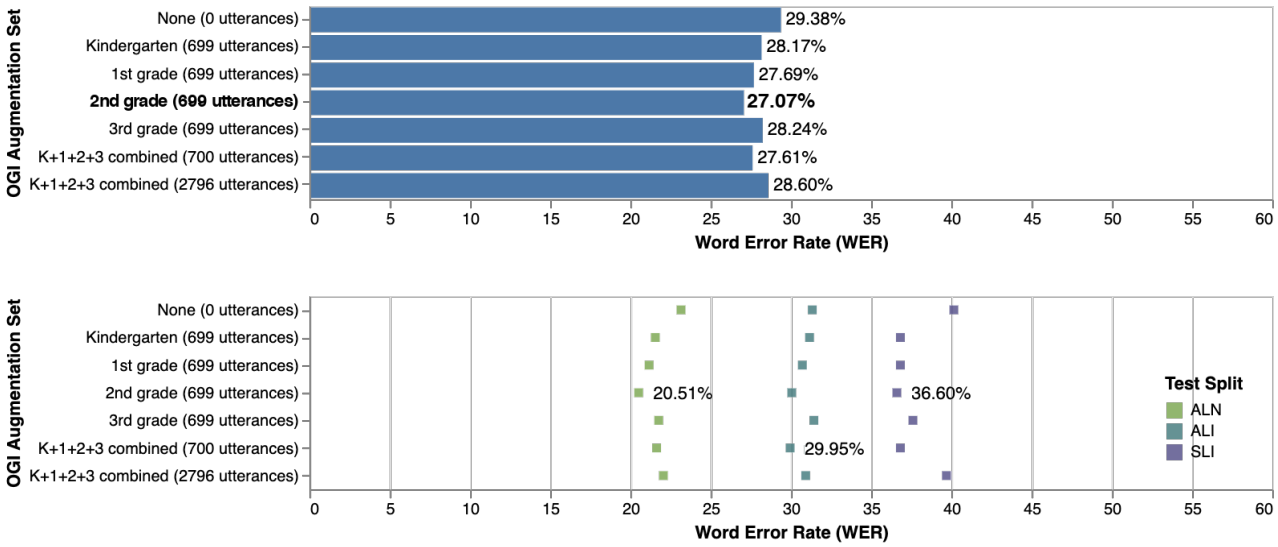

Figure 1: *Results for Experiment 2; WERs for transfer learning on the CSLU Autism Speech Corpus when augmenting with various age groups of OGI Kids' Corpus, then WERs calculated separately for each diagnosis.*

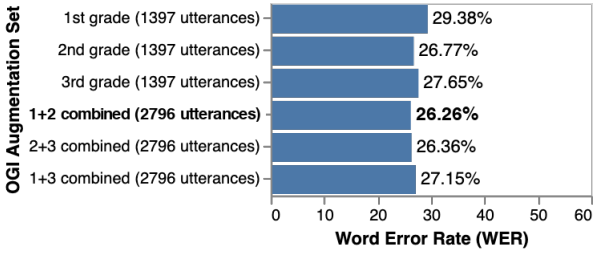

Figure 2: *Results for Experiment 3; WERs for transfer learning on the CSLU Autism Speech Corpus when augmenting with larger sets of the age groups of OGI Kids' Corpus and their pairwise combinations.*

lute. Results are presented in Figure 1. The subset of children with both diagnoses (ALI) showed a word error rate of 30.05%, a 1.28% absolute improvement on the unaugmented model.

3.3. Exp. 3: Strategically adding more data

3.3.1. Methods

Using the knowledge gained from our previous experiment, we wanted to see the effect of larger quantities of the best-performing data. We doubled the number of OGI Kids' Corpus utterances to 1,398 per grade. With insufficient data to double the kindergarten set, we eliminated the youngest speakers from the augmentation dataset altogether, leaving the best-performing 2nd grade, as well as the adjacent 1st and 3rd grades. Following the methods from previous experiments, we transferred the weights from the pre-trained LibriSpeech model and retrained to the CSLU Autism Corpus and each of these new augmentation sets. We also re-trained the model on the CSLU Autism Corpus combined with each pairwise combination of the new 1st, 2nd, and 3rd grade augmentation sets.

3.3.2. Results

The overall best WER of 26.21% was the combination of 1st and 2nd grades — an improvement of 0.71% from the previous experiment's best — followed closely by the combination of 2nd and 3rd grades at 26.36%. Of the individual grade sets, 2nd grade still performed better than the others at 26.77%, a 0.30% improvement on 2nd grade at half the quantity from the previous experiment.

4. Conclusions

In our scenario of data scarcity and domain specificity, even seemingly small changes in quantity and similarity of supplementary data can have an impact on model performance. Experiment 2 demonstrated the effect of data quality over quantity, where a narrow age range similar to the target data outperformed other combinations with both greater variance and greater quantity. Experiment 3 showed how larger quantities of data can still improve model performance if domain similarity can be successfully identified and expanded upon.

Our experiments focused on the age of speakers in training data for children's ASR, which can have a profound impact on acoustic properties, but other forms of domain similarity can be much harder to define. While we show how important an age match in data augmentation with young children can be, age is only one facet of an ASR domain. Note how the different diagnoses yield significantly different WERs in Figure 1; this hints at some possibilities for further refinement of a model based on that domain aspect. Previous work on the CSLU Autism Speech Corpus has shown that the ALN and SLI groups can be distinguished from one another using only pitch information [19]. Future work will look at unsupervised methods of "fine tuning" weight transfer, using techniques like those in [16], in which low-level feature similarity can automatically drive data selection for optimized augmentation of scarce data.

5. Acknowledgements

This research was supported by NIH awards 5R01DC013996 and 5R21AG055749. Any opinions, findings, conclusions or recommendations expressed in this publication are those of the authors and do not reflect the views of the funding agencies.

6. References

[1] E. Semel, E. Wiig, and W. Secord, "Clinical evaluation of language fundamentals (CELF-4). the psychological corporation," *San Antonio, TX*, 2003.

[2] T. Paslawski, "The clinical evaluation of language fundamentals, (CELF-4) a review," *Canadian Journal of School Psychology*, vol. 20, no. 1-2, pp. 129–134, 2005.

[3] A. Neustein, *Advances in speech recognition: mobile environments, call centers and clinics*. Springer Science & Business Media, 2010.

[4] A. Senior, V. Vanhoucke, P. Nguyen, T. Sainath *et al.*, "Deep neural networks for acoustic modeling in speech recognition," *IEEE Signal Processing Magazine*, 2012.

[5] M. J. Gales, K. M. Knill, A. Ragni, and S. P. Rath, "Speech recognition and keyword spotting for low-resource languages: Babel project research at CUED," in *Spoken Language Technologies for Under-Resourced Languages*, 2014.

[6] 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)*. IEEE, 2015, pp. 5206–5210.

[7] 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.

[8] M. B. Mustafa, S. S. Salim, N. Mohamed, B. Al-Qatab, and C. E. Siong, "Severity-based adaptation with limited data for ASR to aid dysarthric speakers," *PloS one*, vol. 9, no. 1, p. e86285, 2014.

[9] J.-L. Gauvain and C.-H. Lee, "Maximum a posteriori estimation for multivariate gaussian mixture observations of markov chains," *IEEE transactions on speech and audio processing*, vol. 2, no. 2, pp. 291–298, 1994.

[10] V. Zue, S. Seneff, and J. Glass, "Speech database development at MIT: TIMIT and beyond," *Speech communication*, vol. 9, no. 4, pp. 351–356, 1990.

[11] F. Rudzicz, A. K. Namasivayam, and T. Wolff, "The TORGO database of acoustic and articulatory speech from speakers with dysarthria," *Language Resources and Evaluation*, vol. 46, no. 4, pp. 523–541, 2012.

[12] D. Wang and T. F. Zheng, "Transfer learning for speech and language processing," in *2015 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA)*. IEEE, 2015, pp. 1225–1237.

[13] P. Ghahremani, V. Manohar, H. Hadian, D. Povey, and S. Khudanpur, "Investigation of transfer learning for ASR using LF-MMI trained neural networks," in *2017 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, Dec 2017, pp. 279–286.

[14] R. Sahraeian and D. Van Compernolle, "Using weighted model averaging in distributed multilingual DNNs to improve low resource ASR," *Procedia Computer Science*, vol. 81, pp. 152–158, 2016.

[15] G. Heigold, V. Vanhoucke, A. Senior, P. Nguyen, M. Ranzato, M. Devin, and J. Dean, "Multilingual acoustic models using distributed deep neural networks," in *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, 2013, pp. 8619–8623.

[16] W. Ge and Y. Yu, "Borrowing treasures from the wealthy: Deep transfer learning through selective joint fine-tuning," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 1086–1095.

[17] 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.

[18] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz, J. Silovsky, G. Stemmer, and K. Vesely, "The Kaldi speech recognition toolkit," in *IEEE 2011 Workshop on Automatic Speech Recognition and Understanding*. IEEE Signal Processing Society, Dec. 2011, iEEE Catalog No.: CFP11SRW-USB.

[19] G. Kiss, J. P. v. Santen, E. Prud'Hommeaux, and L. M. Black, "Quantitative analysis of pitch in speech of children with neurodevelopmental disorders," in *Thirteenth Annual Conference of the International Speech Communication Association*, 2012.

[20] L. D. Shriberg, R. Paul, J. L. McSweeny, A. Klin, D. J. Cohen, and F. R. Volkmar, "Speech and prosody characteristics of adolescents and adults with high-functioning autism and Asperger syndrome," *Journal of Speech, Language, and Hearing Research*, vol. 44, no. 5, pp. 1097–1115, 2001.