Interspeech 2018
2-6 September 2018, Hyderabad

On the Difficulties of Automatic Speech Recognition for Kindergarten-Aged Children

Gary Yeung[1], *Abeer Alwan*[1]

[1]Dept. of Electrical and Computer Engineering, University of California, Los Angeles, USA

garyyeung@g.ucla.edu, alwan@ee.ucla.edu

Abstract

Automatic speech recognition (ASR) systems for children have lagged behind in performance when compared to adult ASR. The exact problems and evaluation methods for child ASR have not yet been fully investigated. Recent work from the robotics community suggests that ASR for kindergarten speech is especially difficult, even though this age group may benefit most from voice-based educational and diagnostic tools. Our study focused on ASR performance for specific grade levels (K-10) using a word identification task. Grade-specific ASR systems were evaluated, with particular attention placed on the evaluation of kindergarten-aged children (5-6 years old). Experiments included investigation of grade-specific interactions with triphone models using feature space maximum likelihood linear regression (fMLLR), vocal tract length normalization (VTLN), and subglottal resonance (SGR) normalization. Our results indicate that kindergarten ASR performs dramatically worse than even 1st grade ASR, likely due to large speech variability at that age. As such, ASR systems may require targeted evaluations on kindergarten speech rather than being evaluated under the guise of "child ASR." Additionally, results show that systems trained in matched conditions on kindergarten speech may be less suitable than mismatched-grade training with 1st grade speech. Finally, we analyzed the phonetic errors made by the kindergarten ASR.

Index Terms: kindergarten speech, child speech recognition

1. Introduction

Automatic speech recognition (ASR) for adults has continued to demonstrate substantial improvements over the years. Meanwhile, due to the lack of databases and targeted techniques for child speech, ASR for children is still riddled with errors. The biggest hurdles for improving child ASR are large inter-speaker variability due to differing rates of growth and development and intra-speaker variability due to undeveloped pronunciation skills, especially at very young ages [1, 2, 3, 4, 5, 6]. Additionally, the number of studies investigating the properties of ASR for children is small. Further studies investigating child ASR at a more nuanced and targeted level may be necessary to improve child ASR.

A number of techniques has been proposed to improve the current performance of child ASR. In [7], near and far field ASR systems for children 5-9 and 10-12 years old were investigated using vocal tract length normalization (VTLN), maximum likelihood linear regression (MLLR), and language model adaptation. Similarly, [8] used a variation of VTLN with subglottal resonances (SGRs) as warping factors. In [9], a deep learning approach using a convolutional long short-term memory network (CLDNN) was trained with a large amount of both adult and child speech data. In [10] and [11], deep neural network (DNN) systems were trained on adults and adapted to children, the former using direct fine-tuning of parameters and the latter using VTLN to warp the features or as an additional input to the network. Additionally, [12] proposed the use of adult out-of-domain data with stochastic feature mapping. While the studies above evaluated ASR techniques on children in general, they did not target either specific ages or educational grade levels. Rather, those studies used groupings such as 5-9 years old.

A study by Shivakumar et al. examined child ASR at an age-specific level using acoustic adaptation and pronunciation modeling [13]. That study found that a single year age difference could dramatically affect ASR performance in young children. However, their ASR systems were trained on children 6-14 years old. The effectiveness of training on a specific age was not investigated.

Recently, Kennedy et al. released a study examining the performance of an ASR system of the Aldebaran NAO, a robot interface commonly used in robot interaction research [14]. In that study, speech from 5 year old children was evaluated. They found that the speech recognition performance was unacceptable, even in the most basic tasks, and caused great difficulty in the development of child-to-robot interactions. Furthermore, they evaluated the ability of four additional ASR APIs (Google, Bing, Sphinx, Nuance) to perform an ASR task such that the system could capture the meaning of a sentence. Even the best performing of these ASR systems could only successfully capture the correct meaning 38% of the time.

Another recent study by Safavi et al. discussed the consequences of age-groups in speaker identification and verification [15]. As expected, that study showed that speaker identification performance was dramatically worse for K-2nd grade children than for older children. That study revealed the consequences of evaluating automatic speech processing applications on a larger "child" age group rather than more specific age groups. Even so, their study used 3-4 year groupings rather than a single age or year (which may be reasonable due to the scarcity of data for individual age groups in speaker verification). Similar to our study, [15] also used the OGI Kids' Speech Corpus [16] that we will use in our experiments.

In this study, we will further explore the various differences in ASR performance when both training and testing data are split into specific grade levels (characterized by one year increments of child development). The kindergarten age group (5-6 years old) is of particular interest because many of these children are pre-literate, and applications such as human-robot interaction (HRI) can benefit greatly from a reliable child ASR system for accessibility [14]. As such, we will put additional focus on this age group. We will investigate the effects of grade-specific training data, number of triphones, and feature normalization and adaptation techniques.

The remainder of this paper is organized as follows. Section 2 describes the database used and experimental setup. Section 3 presents and discusses results of the experiments. Finally,

10.21437/Interspeech.2018-2297

Section 4 concludes the paper with a brief summary and plans for future work.

2. Database and Experimental Setup

2.1. Database

The OGI Kids' Speech Corpus [16] was used in this study. This corpus contains speech from approximately 100 speakers per educational grade level, from kindergarten to 10th grade. Both scripted and spontaneous styles of speech were recorded from the speakers. For scripted speech, scripts included single words, sentences, and digit strings. For spontaneous speech, speakers were asked to respond to a series of prompts such as "Tell me about your favorite movie." Recordings had a sampling rate of 16 kHz. To eliminate the confounding factor of a child language model, only single words from the scripted speech task were used in this study to perform word recognition.

Word utterances covered 208 words. These words ranged from easy words, such as "chair," to more difficult words, such as "organization." Notably, the number of utterances of the words "push" and "spoons" was much higher than the other words across all grades. To remove possible biasing effects, we randomly sampled for a subset of the utterances of "push" and "spoons" to be more consistent with the number of utterances of the remaining words. In general, no other words dominated the number of utterances. To eliminate low-quality or misspoken recordings, only the files marked as "1" from verification files in the OGI Kids' Speech Corpus were used. This "1" indicated that the file was judged to contain the target word and was of good quality. After removing the poor-quality files, the remaining database contained 1,654 word utterances from kindergarten children and at least 3,000 word utterances from each remaining grade. Each grade contained at least 88 speakers.

2.2. Matched-Grade Experiments

For the matched-grade experiments, word utterances were randomly sampled from each grade for a total of 1,654 utterances per grade to ensure a fair comparison. A ten-fold cross-validation was performed for each grade with approximately 1,490 utterances used for training a triphone-based ASR and the remainder used for testing. Training and testing data were separated in such a way that any speaker appearing in the training list would not appear in testing. The language model was a multiple-choice single-word selection with all words equally probable.

For all systems, 13 Mel-frequency cepstral coefficients (MFCCs), extracted with a window size of 25 ms, a frame shift of 10 ms, 23 filters, and a lifter coefficient of 22, were used. Cepstral mean normalization (CMN) was applied to the MFCCs. An additional 7 frame linear discriminant analysis (LDA) was then used for a final 40-dimensional feature input. Derivatives of the MFCCs were not used as they were found not to be as helpful as the LDA features.

Due to the small scale of the word recognition task, ASR systems were trained with 250, 500, and 1000 triphones using both Gaussian mixture model (GMM) and DNN-based hidden Markov model (HMM) systems. DNNs were trained on an additional 9 frame LDA and with 2 hidden layers using 2-norm non-linearities with an input dimension of 500 and output dimension of 100 [17]. All systems were trained using the Kaldi ASR tookit [18]. Feature space maximum likelihood linear regression (fMLLR) speaker adaptive training was also used. The use of VTLN or SGR normalization was found not to be helpful in these matched-grade experiments and will not be reported.

2.3. Kindergarten Mismatched-Grade Experiments

For the mismatched-grade experiments, the same systems that were trained using the DNN-HMM 250 triphone system trained on various grades from the matched-grade experiments were used. The systems were tested with kindergarten speech. In addition to fMLLR used in the previous experiments, both conventional piecewise-linear VTLN and feature warping with SGR normalization [8] were also used to reduce the mismatch between age groups. The SGR estimation algorithm in [8] uses a threshold of 11 years old to separate younger children from older children. As such, we chose the division between 5th and 6th grade as our threshold for the SGR algorithm.

3. Results and Discussion

3.1. Matched-Grade Experiments

The results of the matched-grade experiments for both the GMM-HMM and DNN-HMM systems are shown in Table 1. Overall, there were three major jumps in performance between adjacent grades. First, between kindergarten and 1st grade, word error rate (WER) had an absolute decrease of more than 10% and relative decrease of more than 38%. Then, between 1st and 2nd grade, WER had an absolute decrease of an additional 3-7% and a relative decrease of more than 23%. Finally, between 3rd and 4th grade, WER had an absolute decrease of approximately 5% and a relative decrease of approximately 50%. By 4th grade, WER levels stabilized to a value that was likely to represent adult levels. Overall, the WERs indicated four major grade level groups in terms of recognition performance: kindergarten; 1st grade; 2nd and 3rd grade; and 4th grade and above.

Observing the GMM-HMM models, varying the number of triphones used in acoustic modeling had different effects on error rates depending on grade level. For the kindergarten ASR, the WER increased significantly when the number of triphones increased from 500 to 1000 ($p < 0.001$). While at a lower level of significance, the 1st grade ASR also showed significant degradation in performance when the number of triphones increased from 500 to 1000 ($p < 0.05$). For 2nd grade and older, increasing the number of triphones did not seem to have any obvious effect on WER except for the 8th grade system.

The DNN-HMM models seemed to perform comparably to or better than the GMM-HMM models in all cases. Similarly, the DNN-HMM model had the same grade level grouping of kindergarten; 1st grade; 2nd and 3rd grade; and 4th grade and above based on WER. Additionally, the degradation in performance due to increasing the number of triphones vanished for 1st grade and 8th grade when using the DNN-HMM models. However, the kindergarten ASR still degraded significantly when increasing the number of triphones from 500 to 1000 ($p < 0.01$).

The degradation in performance for the youngest grade levels when increasing the number of triphones likely comes from multiple sources. Notably, past studies on child speech suggest that young children do not have the ability to consistently coarticulate in speech production [19, 20, 21]. As such, the inclusion of additional triphones may not provide any additional benefit to the ASR of younger children. Additionally, as the number of triphones increases, the amount of data available to train the model of each triphone decreases rapidly. As younger children tend to be more variable and inconsistent in their pronunciation [1, 2, 3, 4], the models may be encountering poor training con-

Table 1: *Word error rates (WERs) (%) of ASR systems for the **matched-grade** experiments. Each ASR was trained and tested on the same grade level. Systems were trained with fMLLR speaker adaptive training. Both GMM and DNN-based acoustic models are shown with the number of triphones used in parentheses. The kindergarten ASR performed dramatically worse than older grades and was more affected by the number of triphones.*

System	Grade										
	K	1	2	3	4	5	6	7	8	9	10
GMM(250)	28.32	15.39	11.76	11.12	5.98	6.30	4.88	6.24	3.75	4.25	3.85
GMM(500)	30.08	17.28	12.99	11.18	5.61	7.34	5.42	5.89	6.16	4.57	4.23
GMM(1000)	35.93	19.55	14.57	12.86	6.34	7.10	5.85	6.54	6.90	5.28	4.88
DNN(250)	26.91	14.64	10.50	10.39	4.65	4.64	4.78	5.39	3.34	3.58	3.56
DNN(500)	26.34	16.18	9.51	9.54	4.50	5.46	4.15	5.42	3.57	3.39	3.80
DNN(1000)	30.30	16.06	10.69	10.06	5.22	5.09	5.20	5.14	3.65	3.57	4.05

Table 2: *Word error rates (WERs) (%) of ASR systems for the **mismatched-grade** experiments. Each ASR was trained on a single grade level and tested on kindergarten speech. The systems tested were equivalent to the DNN acoustic model ASR systems with 250 triphones and fMLLR in the matched-grade experiments. Additionally, VTLN and SGR feature normalization were used and found to be effective on systems trained on older children. The best performing system (in boldface) was trained on 1st grade speech with no feature normalization.*

Feature Normalization	Training Grade										
	K	1	2	3	4	5	6	7	8	9	10
None	26.91	**23.11**	24.80	25.38	26.45	24.83	28.64	31.64	36.58	39.00	43.62
VTLN	28.49	24.04	26.75	25.40	26.00	24.17	26.11	29.63	31.65	32.25	34.85
SGR	28.07	26.53	26.96	25.79	26.85	25.82	28.02	29.54	32.33	33.36	35.02

ditions when the number of triphones is large. The combination of these factors is likely to be the main cause of performance degradation when increasing the number of triphones. We also note that most of the older grades seemed invariant to the number of triphones in our experiments while training on the same number of words.

Overall, we have observed that modifications to the number of triphones in the acoustic model may affect kindergarten speech more dramatically than some of the older grades. Additional caution should be used when considering ASR design specifications such as the number of triphones when training a kindergarten ASR. We also note that evaluating performance on a larger age range of children (e.g. 6-10 years old) may not be indicative of performance on individual age groups.

We note that the results of this experiment seem comparable to those reported in [14] such that the kindergarten ASR systems have great difficulty on a relatively easy task for older children or adults. That study reported 19% error rate on a clean digit recognition task with a 5 year old average speaker age. The ASR system they tested, provided by Nuance, was the default ASR of the Aldebaran NAO, a popular humanoid robot interface in the social robotics research community. Both the results of our matched-grade experiment and the results from [14] are unacceptable for effective ASR interfaces for kindergarten-aged children and performance must improve dramatically if such interfaces are to be usable.

3.2. Kindergarten Mismatched-Grade Experiments

The results of the mismatched-grade experiments, including VTLN and SGR normalization techniques, are shown in Table 2. The testing data used in Table 2 were always kindergarten speech, and the training data were speech from a specific grade (K-10th grade). The DNN-HMM system with 250 triphones was used for all experiments as this seemed to produce the best results for kindergarten speech (without any obvious compromise for the remaining grades as well). Additionally, both conventional piecewise-linear VTLN and SGR normalization were evaluated as these warping techniques were found to be helpful when dealing with age-mismatched ASR systems [7, 8, 13, 22].

As expected, the ASR systems trained on the older grades (6th-10th grade) rapidly degraded in performance as training grade level increased. Additionally, both VTLN and SGR feature normalization were able to improve performance when kindergarten speech was tested with ASR systems trained on older age groups (6th-10th grade, 11-16 years old), and both normalization techniques were comparable in performance. However, performance still did not reach the level of the ASR systems trained on the younger age groups (1st-5th grade, 6-11 years old).

ASR systems trained on 1st-5th grade did not show any obvious benefit from feature normalization when testing on kindergarten speech, either with VTLN or SGR normalization. This is justifiable as the physiological differences between the younger age groups is smaller than with the older children. Additionally, SGR estimation algorithms rely on formant estimation [8], which is especially difficult for younger children. This may have caused the degradation seen by the systems trained on younger grades when SGR normalization was used. Overall, it seems that the benefit of feature warping only appears when the training and testing data are substantially mismatched.

Notably, the ASR system trained on 1st grade speech performed significantly better on kindergarten speech than even the kindergarten-trained ASR system ($p < 0.05$). This may suggest that kindergarten speech is not as suitable for training ASR systems as 1st grade speech, even when applied to kindergarten children. This is likely due to the large reductions in speech variability as children age [3, 4]. As such, training on 1st grade speech may result in more stable training conditions. Additionally, as 1st grade children are only one year older than kindergarten children, the mismatch in acoustic conditions is small, as

Table 3: *Word error rates (WERs) (%) of ASR systems trained on either kindergarten or 1st grade speech and tested on kindergarten speech, separated by number of syllables in the target word.*

# Syllables	Training Grade	
	K	1st
1	31.39	31.87
2	25.26	19.98
3	26.63	21.53

evidenced by the results.

These results indicate that the training and testing of ASR for kindergarten speech is not as straight-forward as providing more data to train an ASR system. Varying training data by age (and educational level) slightly may significantly help the performance of a young child ASR system. Additional studies must be done to evaluate the effect of age groups across the many applications of child ASR.

3.3. Phonetic Analysis of Kindergarten ASR

To understand the main difficulties of the ASR systems, we performed an analysis of the ASR errors when tested on kindergarten speech. We analyzed the errors from the systems trained on either kindergarten or 1st grade speech with no VTLN or SGR normalization.

The WERs, separated by number of syllables of the target word, were analyzed. As words with 4 or 5 syllables were underrepresented, we only considered words with 1, 2, and 3 syllables. The results are shown in Table 3.

For the ASR systems in Table 3, WERs for words with 1 syllable were approximately 6-12% higher than for words with 2-3 syllables. This could indicate that the ASR systems relied mainly on vowels, which is unsurprising. To investigate this further, we separated vowels into four classes: high-front, high-back, low-front, and low-back. Of the errors where the target word had 1 syllable, we calculated the percentage of trials where the predicted word had a vowel of the same class as the target word.

For the ASR trained on kindergarten speech, 74.42% of the errors made with a one syllable target word had a shared vowel class between the target and predicted word. Similarly, for the ASR trained on 1st grade speech, 78.63% of the errors made with a one syllable target word had a shared vowel class between the target and predicted word. This suggests that consonants are an unreliable source of information for kindergarten ASR, which may be expected.

We note that training on 1st grade speech greatly improved the performance of the system on 2-3 syllable words while providing no performance increase on 1 syllable words. At this time, the reason for this is unclear. However, this does suggest that words with multiple syllables are more suitable for word identification tasks for kindergarten children such as keyword activation or spotting applications.

In the kindergarten-trained ASR system, the three phonemes that caused the highest error rates were /ð/, /ʊ/, and /j/ with WERs of 41.3%, 41.0%, and 36.8%, respectively. In the 1st grade-trained ASR system, the same three phonemes also caused the highest error rates with WERs of 37.0%, 37.3%, and 36.8%, respectively. Words such as 'mouths,' 'bathe,' 'lure,' 'tourist,' 'humor,' and 'mutual' were correctly classified less than 50% of the time. The large number of errors from /ð/ and /j/ is reasonable due to the weak fricative properties of /ð/ and the rapid acoustic changes of /j/, characteristic of semivowels. However, the vowel /ʊ/ is unexpected to be among the difficult phonemes. We believe that varying pronunciations may have had an effect on this phoneme, such as 'lure' pronounced with an /u/. If we consider the ten most difficult phonemes to recognize, both /θ/ and /b/ were also considered difficult by both the system trained on kindergarten speech and the system trained on 1st grade speech.

4. Conclusion

This study investigated different child ASR systems when both training and testing data were separated by specific grade levels. The OGI Kids' Speech Corpus, containing speech from kindergarten to 10th grade children, was used for training and testing data.

The matched-grade experiments revealed two major results necessary for understanding kindergarten ASR. First, the matched-grade ASR system for kindergarten speech performed substantially worse than even the matched-grade ASR system for 1st grade speech. As such, it may be necessary to evaluate ASR systems on kindergarten speech separately instead of grouping several ages together. Second, the selection of the number of triphones used dramatically affected performance for the kindergarten ASR. In contrast, the ASR systems trained on older grades were mostly invariant to the number of triphones. This indicates that additional caution must be taken when selecting the number of triphones for an ASR system suitable for kindergarten speech.

In the mismatched-grade experiments, when testing on kindergarten speech, the ASR system trained on 1st grade speech performed significantly better than even the system trained on kindergarten speech. This suggests that slightly older children may serve as better training data for kindergarten ASR. We also note that feature warping on kindergarten speech, using either conventional piecewise-linear VTLN or SGR normalization, only showed improvements when the systems were trained on 6th grade children or older. These results can be used as an indicator for when to apply such normalization techniques for kindergarten speech.

Finally, an analysis of the errors made by the kindergarten and 1st grade ASR systems (tested on kindergarten speech) revealed additional information. The systems struggled most with 1 syllable words. Interestingly, training on 1st grade speech improved the performance of the ASR on 2-3 syllable words while providing no benefit on 1 syllable words. Finally, we noted several phonemes that the kindergarten ASR systems did not recognize correctly.

For future work, we will consider the additional consequences of continuous speech on the kindergarten ASR systems by considering how the best-performing language models differ across age groups. We will also further explore the best practices for training kindergarten and young child ASR systems for various application usages.

5. Acknowledgements

This work was supported in part by the NSF.

6. References

[1] L. L. Koenig, J. C. Lucero, and E. Perlman, "Speech Production Variability in Fricatives of Children and Adults: Results of Functional Data Analysis," *The Journal of the Acoustical Society of America*, vol. 124, no. 5, pp. 3158–3170, 2008.

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

[3] S. Lee, A. Potamianos, and S. Narayanan, "Acoustics of Children's Speech: Developmental Changes of Temporal and Spectral Parameters," *The Journal of the Acoustical Society of America*, vol. 105, no. 3, pp. 1455–1468, 1999.

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

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

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

[7] 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," in *Proc. of the Workshop on Child Computer Interaction (WOCCI)*, 2014, pp. 21–26.

[8] J. Guo, R. Paturi, G. Yeung, S. M. Lulich, H. Arsikere, and A. Alwan, "Age-Dependent Height Estimation and Speaker Normalization for Children's Speech Using the First Three Subglottal Resonances," in *Proc. of INTERSPEECH*, 2015, pp. 1665–1669.

[9] H. Liao, G. Pundak, O. Siohan, M. K. Carroll, N. Coccaro, Q.-M. Jiang, T. N. Sainath, A. Senior, F. Beaufays, and M. Bacchiani, "Large Vocabulary Automatic Speech Recognition for Children," in *Proc. of INTERSPEECH*, 2015, pp. 1611–1615.

[10] R. Serizel and D. Giuliani, "Deep Neural Network Adaptation for Children's and Adults' Speech Recognition," in *Proc. of the First Italian Computational Linguistics Conference (CLiC-it)*, 2014.

[11] ——, "Vocal Tract Length Normalisation Approaches to DNN-Based Children's and Adults' Speech Recognition," in *Proc. of Spoken Language Technology Workshop (SLT)*, 2014, pp. 135–140.

[12] J. Fainberg, P. Bell, M. Lincoln, and S. Renals, "Improving Children's Speech Recognition through Out-of-Domain Data Augmentation," in *Proc. of INTERSPEECH*, 2016, pp. 1598–1602.

[13] P. G. Shivakumar, A. Potamianos, S. Lee, and S. Narayanan, "Improving Speech Recognition for Children using Acoustic Adaptation and Pronunciation Modeling," in *Proc. of the Workshop on Child Computer Interaction (WOCCI)*, 2014, pp. 15–19.

[14] J. Kennedy, S. Lemaignan, C. Montassier, P. Lavalade, B. Irfan, F. Papadopoulos, E. Senft, and T. Belpaeme, "Child Speech Recognition in Human-Robot Interaction: Evaluations and Recommendations," in *Proc. of the ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, 2017, pp. 82–90.

[15] S. Safavi, M. Russell, and P. Jančovič, "Automatic Speaker, Age-Group and Gender Identification from Children's Speech," *Computer Speech and Language*, vol. 50, pp. 141–156, 2018.

[16] K. Shobaki, J.-P. Hosom, and R. A. Cole, "The OGI Kids' Speech Corpus and Recognizers," in *Proc. of the International Conference on Spoken Language Processing (ICSLP)*, 2000, pp. 258–261.

[17] X. Zhang, J. Trmal, D. Povey, and S. Khudanpur, "Improving Deep Neural Network Acoustic Models Using Generalized Maxout Networks," in *Proc. of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2014, pp. 215–219.

[18] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlíček, Y. Qian, P. Schwarz, J. Silovský, G. Stemmer, and K. Veselý, "The Kaldi Speech Recognition Toolkit," in *Proc. of the IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*, 2011.

[19] M. Gerosa, S. Lee, D. Giuliani, and S. Narayanan, "Analyzing Children's Speech: An Acoustic Study of Consonants and Consonant-Vowel Transition," in *Proc. of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2006, pp. 393–396.

[20] J. A. Sereno, S. R. Baum, G. C. Marean, and P. Lieberman, "Acoustic Analyses and Perceptual Data on Anticipatory Labial Coarticulation in Adults and Children," *The Journal of the Acoustical Society of America*, vol. 81, no. 2, pp. 512–519, 1987.

[21] N. Zharkova, W. J. Hardcastle, F. E. Gibbon, and R. J. Lickley, "Development of Lingual Motor Control in Children and Adolescents," in *Proc. of the International Congress of Phonetic Sciences (ICPhS)*, 2015.

[22] S. Das, D. Nix, and M. Picheny, "Improvements in Children's Speech Recognition Performance," in *Proc. of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 1998, pp. 433–436.