DEVELOPING NEURAL REPRESENTATIONS FOR ROBUST CHILD-ADULT DIARIZATION

Suchitra Krishnamachari[1], *Manoj Kumar*[1], *So Hyun Kim*[2], *Catherine Lord*[3], *Shrikanth Narayanan*[1]

[1]Signal Analysis and Interpretation Laboratory, University of Southern California, Los Angeles
[2] Center for Autism and Developing Brain, Weill Cornell Medicine
[3]Semel Institute of Neuroscience and Human Behavior, University of California, Los Angeles

2021 IEEE Spoken Language Technology Workshop (SLT) | 978-1-7281-7066-4/20/$31.00 ©2021 IEEE | DOI: 10.1109/SLT48900.2021.9383488

ABSTRACT

Automated processing and analysis of child speech has been long acknowledged as a harder problem compared to understanding speech by adults. Specifically, conversations between a child and adult involve spontaneous speech which often compounds idiosyncrasies associated with child speech. In this work, we improve upon the task of speaker diarization (determining who spoke when) from audio of child-adult conversations in naturalistic settings. We select conversations from the autism diagnosis and intervention domains, wherein speaker diarization forms an important step towards computational behavioral analysis in support of clinical research and decision making. We train deep speaker embeddings using publicly available child speech and adult speech corpora, unlike predominant state-of-art models which typically utilize only adult speech for speaker embedding training. We demonstrate significant reductions in relative diarization error rate (DER) on DIHARD II (dev) sessions containing child speech (22.88%) and two internal corpora representing interactions involving children with Autism: excerpts from ADOS Mod3 sessions (33.7%) and combination of full-length ADOS and BOSCC sessions (44.99%). Further, we validate our improvements in identifying the child speaker (typically with short speaking time) using the recall measure. Finally, we analyze the effect of fundamental frequency augmentation and the effect of child age, gender on speaker diarization performance.

Index Terms— speaker diarization, child speech processing, autism spectrum disorder

1. INTRODUCTION

Automated child speech understanding is an inherently harder problem when compared to adult speech understanding. Various contributing factors for this has been identified and studied over the years, such as the wide range and variability in the acoustic speech characteristics across age-groups attributed to a developing vocal tract [1, 2, 3], and syntactic and pronunciation differences and errors in spoken language production [4, 5] of children. Furthermore, child speech collected in naturalistic settings especially spontaneous speech tends to encompass increased variability in speaking style and background conditions when compared to prompted speech [6]. An important use-case in analyzing spontaneous child speech relates to child-adult conversations from the autism spectrum disorder (ASD) domain in the context of screening, diagnostics and understanding treatment progress.

Autism Spectrum Disorder (ASD) is a neuro-developmental disorder affecting social and communication abilities. The predominant symptoms of ASD manifest as difficulties in language and non-verbal comprehension and expression, and anomalies in expressive vocal prosody patterns [7]. The symptoms become apparent in early years of an individual, hence early diagnosis in children is considered an important step towards effective treatment and intervention. One of the most common observation tools in support of diagnosis consists of clinically-administered interactions between the child and a trained clinician [8, 9]. These dyadic interactions often consist of multiple activities which are designed to observe various socio-communicative behaviors [10, 11]. An important technology component of supporting automated analysis of such interactions is child/adult speaker diarization, namely identifying regions of child speech and adult speech in the interaction. Sessions containing child speech were identified as one of the most challenging domains for speaker diarization in recent evaluations [12]. Moreover, at younger ages or in cases of ASD with verbal ability differences, a child's vocabulary mainly contains vocalizations and other sounds which makes it difficult for the models to learn and distinguish between regions of speech activity and regions of noise.

In this work, we develop deep neural network embeddings customized to the task of child/adult speaker diarization. We use a combination of child speech and adult speech to train embeddings similar to the state-of-the-art x-vector [13] representations. We combine two publicly available corpora, My Science Tutor (MyST) corpus [1] and Voxceleb [2]. Our approach is inspired from the ASR domain, where a combination of child speech and adult speech has been consistently shown to return the best performance [14, 15]. We experiment with various training hyper-parameters, artificial data augmentation

[1]https://boulderlearning.com/request-the-myst-corpus/
[2]http://www.robots.ox.ac.uk/ vgg/data/voxceleb/

techniques and speaker clustering methods to understand the effect on diarization error rate (DER). In addition to DER, we compute the recall measure for child speech to validate that sufficient amounts of child speaking duration are retrieved using the speaker diarization system.

The rest of the paper is organized as follows: Section 2 outlines previous approaches for extracting speaker representations from child speech audio. Section 3 provides a description of the different steps involved in training and evaluating the models. Section 4 gives details of the various corpora utilized in this work. Section 5 describes the various experiments conducted with the trained models. Section 6 reports the key results obtained. Section 7 presents the conclusions and possible future directions and improvements.

2. BACKGROUND

Early works on feature representations for child speech have used a variety of measures such as Mel-Frequency Cepstral Coefficeints (MFCC), Perceptual Linear predictive coding(PLP) and i-vectors [16, 17, 18]. These studies illustrate commonly faced issues with processing "in-the-wild" child speech obtained in natural interaction settings: short speaker turns, varied noise sources and a large fraction of overlapping speakers. Recently, x-vectors trained on adult speech were used for child/adult diarization [12], where the child speech was augmented to adult speech during PLDA (Probabilistic Linear Discriminant Analysis) training. A combination of gender-specific and speaker-specific PLDA scoring trained with child speech were found to improve diarization performance. Alternatively, in [19], pre-trained x-vectors were fine-tuned for child/adult speaker diarization using a meta-learning paradigm, namely prototypical networks. Child speech from multiple adaptation sessions were used to train meta-learned embeddings, which learnt a robust speaker representation invariant to (child) developmental and channel factors. Meta-learned models returned better performance on speaker clustering and speaker classification tasks.

3. SYSTEM DESCRIPTION

This section describes the x-vector diarization system that has been implemented by closely following the VOXCELEB V2 recipe available with the Kaldi speech processing toolkit [3].

3.1. Pre-Processing

The training data was sampled at 16 KHz. 30-dimensional MFCC features were extracted at the frame-level, using a frame length of 25 ms with an overlap of 10 ms. Cepstral mean normalization was applied to the training data using a sliding window of 3 seconds. The training utterances were filtered by removing utterances that fell below the minimum utterance length threshold (< 300 seconds) and speakers that had very few utterances (< 8 utterances).

3.2. Training model architecture

In our work, time delay neural network (TDNN) [20] models, 7 layers deep with an input layer and an output softmax layer, are trained to separate and classify the different speakers in a given audio recording. The initial layers of the TDNN operate at the frame level while the final layers operate at the segment level. The context width of these models increases as we go deeper into the network. The TDNN models are typically used to model long-term temporal dependencies from short term speech signals. In keeping with the high speech correlation assumption, an important aspect of the TDNN is sub-sampling the neighboring input context activation. ReLu non-linearity is used for learning the distinct data characteristics. In order to get a single representation of the entire utterance, a statistics pooling layer followed by two feed-forward layers are used. X-vector embeddings are extracted after the statistics pooling layer. The final softmax layer has nodes equal to the number of speakers present in the training data. Model hyper-parameters are varied during the training process to improve DER.

3.3. Evaluation

For the purpose of speaker diarization, x-vectors are extracted from the test session using a uniform segmentation of 1.5 second duration in width and overlap of 0.75 seconds. Two methods were used for speaker clustering. The first one is the commonly used agglomerative hierarchical clustering (AHC) method scored with a probabilistic linear discriminant analysis (PLDA) model. The second method is a recently proposed variant of spectral clustering [21] which uses negative cosine distances as an affinity measure.

3.3.1. Probabilistic Linear Discriminant Analysis (PLDA)

PLDA is a generative model that learns the projection space, maximizing the separation between different speakers while minimizing the variability of the same speaker [12]. One of the major advantages of PLDA is its ability to learn probability distributions for unseen classes based on just a single example. PLDA transforms have been extensively used in speaker diarization [22] and speaker verification tasks [23].

3.3.2. Agglomerative Hierarchical Clustering (AHC)

This is a bottom up clustering algorithm that is initialized with number of clusters same as the number of sub-segments, while clustering them together based on a certain clustering threshold or a stopping criterion. The expected number of final clusters are equal to the total different speakers that are

[3] https://github.com/kaldi-asr/kaldi/tree/master/egs/voxceleb/v2

present in the corpus. AHC is used with average linking [13] in which the distance between two clusters is calculated as the average of the distances between each point in one cluster and each point in the other cluster.

3.3.3. Auto Tuning Spectral Clustering

Auto Tuning Spectral Clustering is a variation of Spectral Clustering, a graph based clustering method using an affinity matrix. Auto tuning spectral clustering is adaptive of the dataset under consideration and thus adjusts the hyperparameters automatically, reducing sensitivity to the affinity matrix [21]. The threshold for the stopping criterion and the number of clusters are both determined by this algorithm itself.

3.3.4. Evaluation Metric

Diarization error rate is the most common metric used for evaluating the performance of a model. It takes into consideration the number of times speech segments have been assigned to the wrong speaker, *speaker ID error*, the number of times speech activity has been treated as non-speech activity, *missed speech*, and the number of times non-speech segments are treated as speech segments, *false alarm*. We target and report reductions in DER for our proposed models in this work [24]. Additionally, we compute the recall measure for child speech. A higher recall measure implies that we identify a significant fraction of the child speech correctly.

4. DATASET

Several different corpora were used for training and evaluating the models. This section gives a description of each corpus used.

4.1. Training Corpora

The models are trained using different combinations of publicly available child and adult audio data from the MyST - My Science Tutor corpus and the Voxceleb corpus, respectively.

The MyST corpus [25] is an outcome of the MyST project, a 13 year long project conducted by Boulder Learning. This corpus includes speech of students from the third, fourth and fifth grades. Each student converses with a virtual science tutor in a strict turn-taking fashion ensuring no overlap between the two parties. The child audio used for modeling are the recordings of the answers given by the child to a question posed by the tutor. This corpus consists of a total of 499 hours of child speech. For the purposes of our experiments, each audio file has been treated as one single utterance.

For adult speech data, audio from Voxceleb1 and Voxceleb2 datasets are used together. These publicly available datasets have been used in a wide range of tasks including speech recognition, speaker verification, face and emotion generation. This combined corpus, referred to as Voxceleb in this work, consists of over 2000 hours of adult speech collected from celebrity interviews uploaded on YouTube.

Similar to the Voxceleb recipe, we augment both corpora with noise, music and babble speech using the MUSAN corpus [26], and reverberation using the RIR_NOISES [4] corpus.

Table 1. Original training data specifications

Parameters	Training Corpora	
	MyST	Voxceleb
Number of speakers	1372	7323
Number of utterances	244069	1104264

4.2. Testing Corpora

We use a combination of freely available and specially-obtained corpora for evaluation: *DIHARD-kids* consists of 47 sessions from the *Child* and *Clinical* domains of DIHARD-II dev set, and two internal datasets representing specific Autism related clinical interactions using ADOS and BOSCC protocols, further explained below.

Autism Diagnostic Observation Schedule (ADOS) interactions [8] are considered the gold standard in support of assessing and diagnosing ASD. An ADOS session consists of multiple activities (ranging from 10 to 15 depending on the child language level) targeted to analyze specific characteristics of ASD. The typical duration of an ADOS session is between 40-60 minutes. An ADOS session is used primarily by a clinician for diagnostic purposes. Brief Observation of Social and Communication Changes (BOSCC) interactions [9] are used to track behavioral changes in terms of social and communication skills over the course of treatment of an individual with ASD. Each BOSCC session is about 12 minutes long with typically a 2 minute conversational session and two 4 minute play sessions where in the child plays with a toy during the course of the interaction.

The test corpus considered in this work consists of the Emotions and Social Difficulties and Annoyance activities from Module 3 of ADOS. These activities are most likely to contain spontaneous responses elicited from children, including when they are under cognitive stress. The sessions were collected at the University of Michigan Autism and Communication Disorder Center (UMACC) and Cincinnati Children's Medical Center (CCHMC). A total of 346 sessions from 165 different child speakers (86 affected by ASD) are present in this corpus. The other corpus considered consists of a combination of ADOS (n=3) and BOSCC (n=24) sessions collected from four different centers. In contrast with

[4] http://www.openslr.org/28

the aforementioned *ADOS-3* data subset, we use the audio from all activities in the sessions.

Table 2. Test corpora specifications

Corpus	Duration *(min)* (mean $\pm$ std)	Child Speaking fraction (mean $\pm$ std)
CARE	17.77 ± 11.99	0.399 ± 0.083
ADOS-3	3.23 ± 1.61	0.464 ± 0.18
DIHARD-kids	5.23 ± 0.31	n.a

5. EXPERIMENTAL SETUP

This section describes the ways the training corpora have been utilized to experiment with the DER results. Information about the number utterances and number of speakers in the original child and adult data can be found in Table 1.

5.0.1. Approach 1: Reducing adult speech data

The different histogram plots of the MyST corpus revealed that there were a large number of utterances falling below the threshold for minimum utterance length but a very small number of speakers falling below the threshold for minimum number of utterances, both of which have been described in section 3. We anticipated a loss of nearly half the utterances during the process of filtration while the number of speakers were expected to remain fairly close to that of the original data. In order to ensure sufficient amount of child utterances in the training data, the amount of augmented data consisting of noise, music, background speech and reverberation was chosen to be three times that of the original data. This was added to the original clean data. Thus, the size of the child speech corpus was increased to four times the number of utterances as of the original, with the expectation of having final number of utterances twice that of the clean utterances, after filtration. The adult data was randomly chosen to mimic the anticipated child data in terms of speaker number and number of utterances. The combined child-adult training data after filtration comprised a total of 2731 speakers and over a million utterances. The specifications of the final training data can be found in Table 3.

Table 3. Training Data Specifications for Approach 1

Parameters	Data before filtration		Data after filtration	
	MyST	Voxceleb	MyST	Voxceleb
Number of speakers	1372	1372	1359	1372
Number of utterances	960204	479323	545570	479076

5.0.2. Approach 2: Increasing child data

Adding pitch and speed perturbation is a data augmentation technique [27, 28] used to increase the amount of training data at hand. In our work, we select pitch variation to augment the MyST corpus. Our reasoning is as follows: the vocal apparatus growth, including the laryngeal aspects, is an integral aspect within child linguistic development. Furthermore, there are inherent pitch variations in natural speech across contexts. Hence, simulating pitch variations while keeping the phonetic component fixed can potentially mimic a larger sample size of children and some of the expected speech variability.

Pitch variations in the range of 0.9 to 1.1 times the original fundamental frequency were randomly introduced in the child audio. Pitch values in cents were chosen based on the formula $f2/f1 = 2^{cents/1200}$, where f2 and f1 are the frequencies associated with the pitch varied audio file and the original audio file, respectively. This process was repeated five times. After adding the newly generated data to the clean data, the new child data had speakers and utterances six times that of the original. Along with added reverberation, noise, speech and music, the combined child data had utterances 12 times that of the original. A subset of speakers from 8232 speakers was taken to closely match the speakers in the adult data. This new child-adult training data after filtration consists of 14534 speakers and over 3.5 million utterances. The specifications for this approach can be found in Table 4.

Table 4. Training Data Specifications for Approach 2

Parameters	Data before filtration		Data after filtration	
	MyST	Voxceleb	MyST	Voxceleb
Number of speakers	7324	7323	7211	7323
Number of utterances	2549773	2270149	1396688	2269038

6. RESULTS

The mean Diarization Error Rate results (DER) have been calculated for the different trained models. In the work presented in this paper, the model called Pretrained is the baseline model against which all the newly trained models have been compared.

Table 5. Model Hyperparameters:

Model	Learning rate	Frames	Repeats	Epochs
Model1	0.002	100-200	15	5
Model2	0.002	100-200	15	8
Model3	0.005	100-200	15	5
Model4	0.002	100-250	15	5
Model5	0.002	100-300	15	5
Pitch Variation	0.002	100-200	5	5

The newly trained TDNN models namely, Model1,

Model2, Model3, Model4 and Model5, closely resemble each other and the baseline model with slight variations in their hyperparameters, aiming to improve the diarization performance. The final model called Pitch Variation is the model where the pitch variations were introduced to the child data. The details of the model hyperparameters can be found in Table 5. The mean DER results have been tabulated in Table 6.

Table 6. Mean DER Results of different corpora

Model	CARE		ADOS-3		DIHARD-kids	
	PLDA	SC	PLDA	SC	PLDA	SC
Pretrained	27.89	16.60	19.61	13.52	26.57	26.21
Model 1	16.48	11.70	14.06	11.42	20.89	23.67
Model 2	19.37	11.66	14.72	11.83	21.85	23.79
Model 3	**15.34**	11.40	15.23	**11.14**	21.71	22.13
Model 4	18.85	12.11	14.41	11.33	21.28	22.25
Model 5	19.59	11.61	14.39	11.76	**20.49**	**21.27**
Pitch Variation	16.43	**10.71**	**13.00**	11.88	24.97	21.76

From the obtained results, we observe, not surprisingly, that incorporating child audio as a part of our training data significantly improves the DER results as compared to the baseline model. The proposed models are trained by treating each child and adult from the combined corpora as an individual speaker. Hence, we hypothesize that the embeddings learn both developmental variabilities within the child group, and the differences in the speech between children and adults, both of which are useful in improving speaker diarization performance. Among the proposed models, even the ones with the highest DER results outperform the baseline model with a relative improvement of nearly 30%, 22.36% and a little over 6% for the CARE, ADOS-3 and the DIHARD-kids corpora, respectively.

Child speech diarization is especially important in the ASD domain, since it is the primary target during automated interaction analysis of clinical sessions. Even though the performance of our models is superior to that of the baseline model, the improved mean DER values fall short in providing concrete evidence of improvement specifically in child speech diarization. In order to confirm the effectiveness of our models from this point of view, recall values have been generated. This gives an idea of what fraction of the original child speech has been correctly classified by our models as the class of interest, in our case the child class. The recall values have thus been computed for child speakers of the two internal test datasets. Ground truth Rich Transcribed Times Marked (RTTM) files were used along with the RTTM file generated as a part of evaluation, to compute the recall values. As can be seen in Table 7 there is a significant increase in the recall values for both the datasets, which confirms the performance improvement of our models. The improvements on the PLDA score are higher than that of the SC scores. Furthermore, by comparing the best values in Table 6 we infer

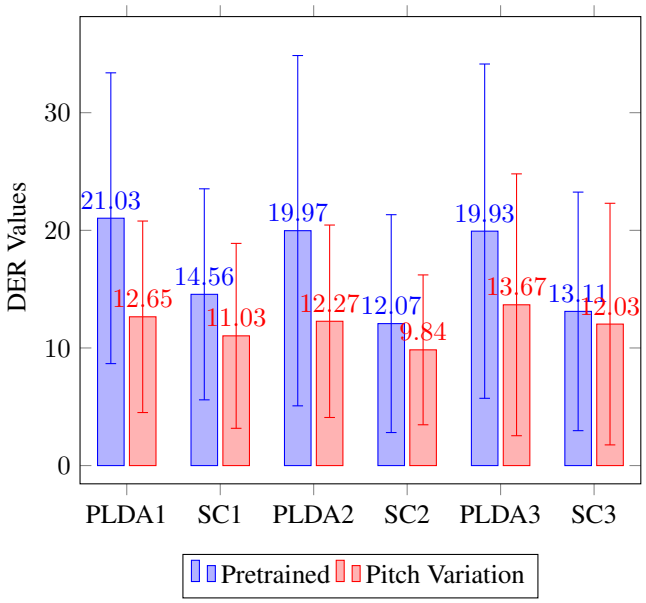

Fig. 1. Mean and Std DER for the age experiments conducted on the ADOS-3 corpus

that the models with the best recall values aren't necessarily the best performing model in terms of mean DER values, justifying our reasons to calculate recall.

Table 7. Recall values generated for model performance validation

Model	CARE		ADOS-3	
	PLDA	SC	PLDA	SC
Pretrained	0.497	0.702	0.820	**0.869**
Model1	0.694	**0.835**	0.852	0.860
Model2	0.726	0.829	0.850	0.835
Model3	0.713	0.828	0.863	0.855
Model4	0.693	0.803	0.868	0.855
Model5	0.644	0.830	0.859	0.853
Pitch Variation	**0.732**	0.831	**0.919**	0.842

6.1. Age oriented experiments on the ADOS-3 corpus

The improvement of our model performance reinforces the study in [1, 2, 29] about the effect of age on speech characteristics. To further investigate the performance of our models, we conducted the following experiment.

The ADOS-3 corpus consists of children in the age groups between 3.5 years to 13 years of age. The corpus was divided into three categories such that each category contained equal number of sessions in them. Since the DER improvement was the highest for the model with added pitch variations, the age-focused experiments were performed using this

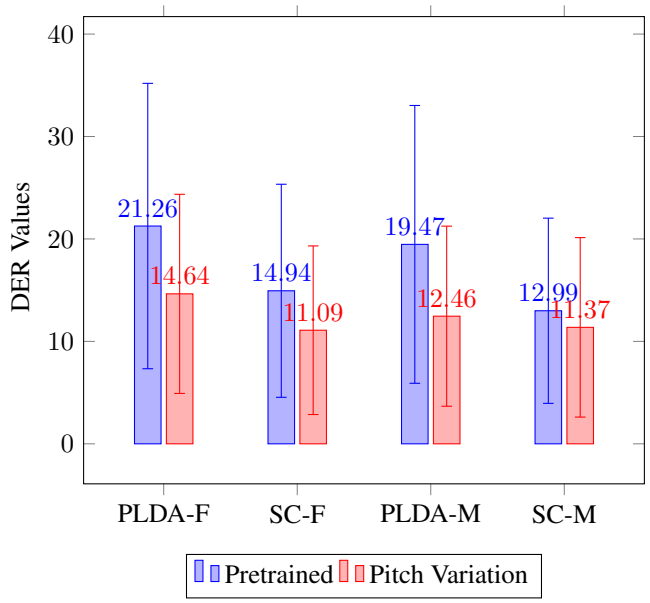

Fig. 2. Mean and DER for the gender experiments of the ADOS-3 corpus

model, compared against the Pretrained model. The results are shown in Figure 1. In this figure, PLDA1/SC1 correspond to PLDA and SC DER scores for children of ages between 43 months to 91 months, PLDA2/SC2 correspond PLDA and SC DER scores for children of ages between 92 months to 118 months, PLDA3/SC3 correspond PLDA and SC DER scores for children of ages between 119 months to 158 months.

The relative improvement in the PLDA DER scores are 39.85%, 38.56% and 31.41%, respectively, for the three age groups (youngest to oldest). This improvement is higher for the two lower age groups, jointly consisting of children in the ages between 3 years and around 10 years of age, as compared to children older than 10 years of age. The larger improvement in the lowest age groups are particularly interesting, since: 1) Speech produced by younger children is especially challenging for automated speech analysis, hence it is important to correctly distinguish from adult speech 2) Better diarization can assist downstream analysis of speech patterns for clinically meaningful information.

6.2. Gender-focused performance on the ADOS-3 corpus

Analyses of developmental changes in speech have revealed gender differences in speech characteristics, especially post puberty [1]. Moreover, reported prevalence statistics for ASD show inherent distributional differences across gender [30]. In the following, we consider gender dependent analysis of diarization performance. The ADOS-3 corpus has been divided into two subsets, male (M) and female (F), based on the available gender information. The data has 244 male and 84 female individuals, consistent with prevalence trends in ASD. For the same reasons stated in the age experiments, comparisons are drawn between the Pretrained model and the model with added pitch variations as depicted in Figure 2. The relative improvement in the DER is 36% and 31.14% for the male and female subsets, respectively. The differences in performance improvements may be due to both inherent speech pattern differences, and also inherent data distribution biases.

7. CONCLUSION

In this work we experimentally investigated the effects of using child and adult data to improve diarization in recordings involving interactions with children, including in clinical settings involving children with Autism. The results obtained confirm our hypothesis of benefiting from incorporating child speech together with adult speech, across both age and gender dimensions of variability. The effects of added pitch variation and different data augmentation techniques are also analyzed.

Ground truth reference files containing clean annotations for the different speakers, especially for the clinical interactive session data sets are not always easily accessible. To perform diarization in such cases where the regions of speech and silence are unknown, speech activity detection systems generating voice activity detection (VAD) files are used resulting in a more naturalistic use setting. The DER values thus generated in those cases are higher than those obtained in presence of ground truth speech activity. Our future work can extend to training models to improve diarization performance in the absence of reference RTTM files to improve DER further. In addition to this, more focus can be directed towards improving diarization involving infants and toddler speech. Toddlers typically do not enunciate and their vocabulary contains many pre-verbal sounds and other nonverbal vocalizations, in contrast to fully well-formed words, distinguishing between regions of speech and non-speech, and speakers, becomes even more challenging. These are topics for future study.

8. ACKNOWLEDGMENT

This research was supported by the Simons Foundation and National Institute of Mental Health (NIMH Grant No. 1R01 MH114925-01).

9. REFERENCES

[1] Sungbok Lee, Alexandros Potamianos, and Shrikanth 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.

[2] Sungbok Lee, Alexandros Potamianos, and Shrikanth S. Narayanan, "Developmental acoustic study of american

english diphthongs," *J. Acoust. Soc. Am.*, vol. 136, no. 4, pp. 1880–1894, oct 2014.

[3] Matteo Gerosa, Diego Giuliani, and Shrikanth Narayanan, "Acoustic analysis and automatic recognition of spontaneous children's speech," in *Proceedings of the 9th International Conference on Spoken Language Processing*, 2006, pp. 1886–1889.

[4] Valerie Hazan and Sarah Barrett, "The development of phonemic categorization in children aged 6-12," *Journal of Phonetics*, vol. 28, no. 4, pp. 377 – 396, 2000.

[5] Alexandros Potamianos and Shrikanth Narayanan, "Spoken dialog systems for children," in *Proceedings of ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, 1998, vol. 1, pp. 197–200.

[6] Matteo Gerosa, Diego Giuliani, Shrikanth Narayanan, and Alexandros Potamianos, "A review of ASR technologies for children's speech," in *Proceedings of the 2nd Workshop on Child, Computer and Interaction*, 2009, pp. 1–8.

[7] Leo Kanner, "Autistic disturbances of affective contact," *Nervous Child*, vol. 2, pp. 217–250, 1943.

[8] Catherine Lord et al., "The autism diagnostic observation schedule—Generic: A standard measure of social and communication deficits associated with the spectrum of autism," *Journal of autism and developmental disorders*, vol. 30, no. 3, pp. 205–223, 2000.

[9] Rebecca Grzadzinski et al., "Measuring changes in social communication behaviors: preliminary development of the brief observation of social communication change (BOSCC)," *Journal of autism and developmental disorders*, vol. 46, no. 7, pp. 2464–2479, 2016.

[10] Natacha Akshoomoff, Christina Corsello, and Heather Schmidt, "The role of the autism diagnostic observation schedule in the assessment of autism spectrum disorders in school and community settings," *The California School Psychologist*, vol. 11, no. 1, pp. 7–19, 2006.

[11] Carla A Mazefsky and Donald P Oswald, "The discriminative ability and diagnostic utility of the ADOS-G, ADI-R, and GARS for children in a clinical setting," *Autism*, vol. 10, no. 6, pp. 533–549, 2006.

[12] Jiamin Xie1, Leibny Paola Garcia-Perera, Daniel Povey, and Sanjeev Khudanpur, "Multi-PLDA diarization on children's speech," in *Interspeech*, 2019, pp. 376–380.

[13] Gregory Sell, David Snyder, Alan McCree, Daniel Garcia-Romero, Jesús Villalba, Matthew Maciejewski, Vimal Manohar, Najim Dehak, Daniel Povey, Shinji Watanabe, and Sanjeev Khudanpur, "Diarization is hard: Some experiences and lessons learned for the jhu team in the inaugural dihard challenge," in *Proc. Interspeech 2018*, 2018, pp. 2808–2812.

[14] Alexandros Potamianos and Shrikanth S. Narayanan, "Robust recognition of children's speech," *IEEE Transactions on Speech and Audio Processing*, vol. 11, no. 6, pp. 603–616, nov 2003.

[15] Prashanth Gurunath Shivakumar and Panayiotis Georgiou, "Transfer learning from adult to children for speech recognition: Evaluation, analysis and recommendations," 2018.

[16] Xin Wang, Jun Du, Lei Sun, Qing Wang, and Chin-Hui Lee, "A progressive deep learning approach to child speech separation," in *ISCSLP*. IEEE, 2018, pp. 76–80.

[17] Alejandrina Cristia, Shobhana Ganesh, Marisa Casillas, and Sriram Ganapathy, "Talker diarization in the wild: The case of child-centered daylong audio-recordings," in *Interspeech*, 2018, pp. 2583–2587.

[18] Tianyan Zhou, Weicheng Cai, Xiaoyan Chen, Xiaobing Zou, Shilei Zhang, and Ming Li, "Speaker diarization system for autism children's real-life audio data," in *ISCSLP*, Oct 2016, pp. 1–5.

[19] N. R. Koluguri, M. Kumar, S. H. Kim, C. Lord, and S. Narayanan, "Meta-learning for robust child-adult classification from speech," in *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020, pp. 8094–8098.

[20] Vijayaditya Peddinti, Daniel Povey, and Sanjeev Khudanpur, "A time delay neural network architecture for efficient modeling of long temporal contexts," in *Sixteenth Annual Conference of the International Speech Communication Association*, 2015.

[21] T. J. Park, K. J. Han, M. Kumar, and S. Narayanan, "Auto-tuning spectral clustering for speaker diarization using normalized maximum eigengap," *IEEE Signal Processing Letters*, vol. 27, pp. 381–385, 2020.

[22] Daniel Garcia-Romero, David Snyder, Gregory Sell, Daniel Povey, and Alan McCree, "Speaker diarization using deep neural network embeddings," in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2017, pp. 4930–4934.

[23] David Snyder, Daniel Garcia-Romero, Daniel Povey, and Sanjeev Khudanpur, "Deep neural network embeddings for text-independent speaker verification.," in *Interspeech*, 2017, pp. 999–1003.

[24] Neville Ryant, Kenneth Church, Christopher Cieri, Alejandrina Cristia, Jun Du, Sriram Ganapathy, and Mark Liberman, "The second dihard diarization challenge: Dataset, task, and baselines," *arXiv preprint arXiv:1906.07839*, 2019.

[25] Wayne Ward, Ron Cole, and Sameer Pradhan, "My science tutor and the myst corpus," 2019.

[26] David Snyder, Guoguo Chen, and Daniel Povey, "Musan: A music, speech, and noise corpus," 2015.

[27] Justin Salamon and Juan Pablo Bello, "Deep convolutional neural networks and data augmentation for environmental sound classification," *IEEE Signal Processing Letters*, vol. 24, no. 3, pp. 279–283, 2017.

[28] Tom Ko, Vijayaditya Peddinti, Daniel Povey, and Sanjeev Khudanpur, "Audio augmentation for speech recognition," in *Sixteenth Annual Conference of the International Speech Communication Association*, 2015.

[29] Matteo Gerosa, Sungbok Lee, Diego Giuliani, and S Narayanan, "Analyzing children's speech: An acoustic study of consonants and consonant-vowel transition," in *2006 IEEE International Conference on Acoustics Speech and Signal Processing Proceedings*. IEEE, 2006, vol. 1, pp. I–I.

[30] Jake Gockley, A Jeremy Willsey, Shan Dong, Joseph D Dougherty, John N Constantino, and Stephan J Sanders, "The female protective effect in autism spectrum disorder is not mediated by a single genetic locus," *Molecular autism*, vol. 6, no. 1, pp. 1–10, 2015.