Can Smartphones be a cost-effective alternative to LENA for Early Childhood Language Intervention?

Satwik Dutta[1], *Jacob Reyna*[1], *Jay Buzhardt*[2], *Dwight Irvin*[2], *and John H.L. Hansen*[1]

[1]Center for Robust Speech Systems, The University of Texas at Dallas, Richardson, Texas, USA
[2]Juniper Gardens Children's Project, The University of Kansas, Kansas City, Kansas, USA

satwik.dutta@utdallas.edu, jaybuz@ku.edu, john.hansen@utdallas.edu

Abstract

Although non-profit commercial products such as LENA can provide valuable feedback to parents and early childhood educators about their children's or student's daily communication interactions, their cost and technology requirements put them out of reach of many families who could benefit. Over the last two decades, smartphones have become commonly used in most households irrespective of their socio-economic background. In this study, conducted during the COVID-19 pandemic, we aim to compare audio collected on LENA recorders versus smartphones available to families in an unsupervised data collection protocol. Approximately 10 hours of audio evaluated in this study was collected by three families in their homes during parent-child science book reading activities with their children. We report comparisons and found similar performance between the two audio capture devices based on their speech signal-to-noise ratio (NIST STNR) and word-error-rates calculated using automatic speech recognition (ASR) engines. Finally, we discuss implications of this study for expanding this technology to more diverse populations, limitations and future directions.

Index Terms: parent-child book reading, smartphone, speech recognition, early childhood

1. Introduction

Limited exposure to rich and engaging language environments in the first few years of life has an immediate [1, 2, 3] and lasting effect on children's language growth [4, 5]. Delays in early language acquisition have been linked to future outcomes, such as a greater need for special education services, lower probability of graduating from high school, and fewer employment opportunities [6, 7, 8]. However, early language delays can be mitigated through early intervention [9]. Derived from neurological and socio-behavioral theories of child development [10, 11], an increasing body of research has demonstrated that language-rich home environments can improve children's language growth and that parents/caregivers from diverse cultural and socio-economic backgrounds can learn the skills needed to increase the frequency of quality language interactions with their children [12, 13, 14].

Providing feedback to parents about talk (adult words, child vocalizations, and adult-child turns) has shown to be an effective approach to improve parent talk and, in turn, children's early language growth [15, 16, 17]. However, measuring talk outside of lab settings and providing meaningful feedback to parents is not feasible without technology solutions. In the spirit of the Fitbit and other wearable devices that give users ongoing feedback about their daily physical activity, the Language ENvironmental Analysis System[1] (LENA) in Fig. 1(a) was designed to give parents immediate and frequent data about the amount of adult language and adult-child interactions their child experiences on a day-to-day basis [18]. Several empirical studies in diverse settings with diverse populations have shown that families that use LENA show significantly greater growth in parent-child conversational turns and children's expressive communication compared to comparison families who did not use LENA [19, 20].

Unfortunately, although originally conceived as a tool to benefit families from low-income backgrounds, LENA's cost is prohibitive for most families to use. A key component of the LENA system is mandatory use of their proprietary digital recorder to capture audio that is processed by their speech processing software [21]. However, since LENA's development over 12 years ago, smartphone technology and internal recording hardware has advanced, and their use has become nearly ubiquitous[2] with 85% of Americans reporting owning a smartphone, including 76% of those earning less than $30,000 year, 85% of Latinx individuals, and 83% of African-Americans. Using smartphones to record audio rather than proprietary devices, such as LENA's digital recorder, would reduce the costs and improve the feasibility of similar systems by utilizing technology that most families already own. This would increase the accessibility of this technology for families with more diverse economic backgrounds who would likely benefit the most from it. Therefore, the purpose of this study was to examine the technical properties of audio recorded by parents with the LENA digital recorder to the same audio recorded by their personal smartphones. To our knowledge, this is the first study of this kind.

The primary goal of this exploratory study was not to conduct an exhaustive study of the feasibility of LENA versus smartphones, but to report basic technical properties of audio recorded by smartphones by a small number of families in their homes relative to the same audio recorded by LENA devices. Parents read and engaged with their child during reading sessions at home while recording with both a LENA and their own smartphone. This paper is structured as follows: in Sec. 2 we describe the dataset and the data collection protocol, in Sec. 3 we elaborate and discuss the results of speech processing: speech signal-to-noise ratio and automatic speech recognition, in Sec. 4 we describe the limitations and finally in Sec. 5 we conclude the study and provide directions for extending this study.

[1] https://www.lena.org/

[2] https://www.pewresearch.org/internet/fact-sheet/mobile/

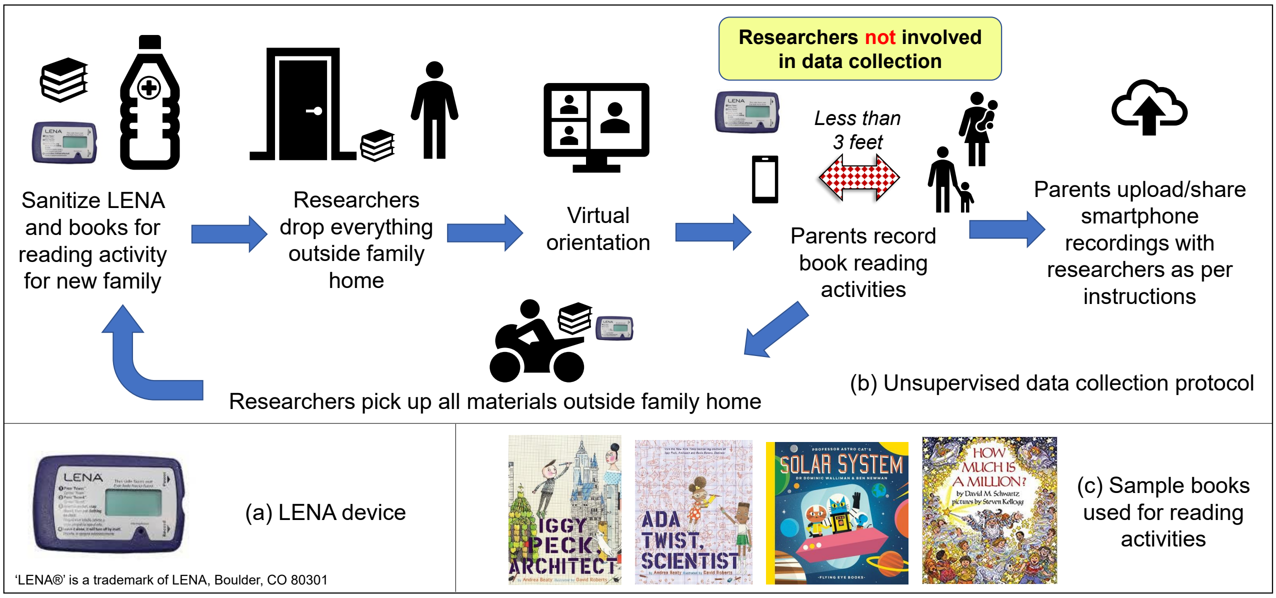
Researchers not involved in data collection
Sanitize LENA and books for reading activity for new family
Researchers drop everything outside family home
Virtual orientation
Less than 3 feet
Parents record book reading activities
Parents upload/share smartphone recordings with researchers as per instructions
Researchers pick up all materials outside family home
(b) Unsupervised data collection protocol
LENA
(a) LENA device
'LENA®' is a trademark of LENA, Boulder, CO 80301
IGGY PECK ARCHITECT
ADA TWIST, SCIENTIST
SOLAR SYSTEM
HOW MUCH IS A MILLION?
(c) Sample books used for reading activities

Figure 1: *(a) Language Environmental Analysis (LENA) device, (b) Unsupervised contactless data collection protocol, (c) Sample books used for reading activities by the families.*

2. Dataset: Parent-Child Book Reading Activity

Due to ongoing COVID-19 restrictions and following IRB protocols for human subjects research, we formulated an unsupervised data collection scenario at home of consented participating families for audio recording of book reading using both LENA recorders and parent's smartphone, as shown in Fig. 1(b). Families were recruited through email, flyers, and word-of-mouth. Parents or caregivers contacted the researchers to learn more about the study and consent if they agreed to the procedures. All materials including books and LENA units were properly sanitized and dropped-off at the parent's home by contactless delivery. A feedback procedure was established to allow participating parents and children to request help from research assistants with the recordings through email or phone. An instruction manual was provided to the primary caregiver for each family with details for recording using LENA and Android or iPhone smartphones, and the procedure to upload the smartphone audio to a secure file-sharing platform. LENA recorders and books were picked up after they completed the recording sessions, and sanitized again for reuse for another family. For this study, we included multiple recordings of 3 families. The 3 participating families used iPhones and were expected to read a total of 10 books each, which were provided to families. These were science-based books (Fig. 1(c)) and included: 'Rocket Science for Babies' (Chrise Ferrie), 'One Day On Our Blue Planet: In the Rainforest' (Ella Bailey), 'What do you do with a problem' (Kobi Yamada), 'National Geographic Little Kids First Big Book of Why' (Amy Shields), etc. Most books were designed for ages 3 to 8, and were chosen by early childhood researchers. Parents were asked to conduct readings in a quiet location with minimal distractions and to encourage interactions and conversations during reading sessions. They were asked to place the LENA and smartphone within 3 feet of them and the child. For each recording, parents reported contextual information, such as location/setting, book title, voices from people captured in the recording, background noise/disturbance, etc.

Table 1: *Details about families for the parent-child book reading activity.*

Family #	Child Age	Location or Settings	Audio (hrs)	# Words Adult:Child
1	7	Bedroom, living room	4	88:12
2	5	Kitchen	4	79:21
3	5	Bedroom	2.4	95:5

There were also several challenges with this un-supervised audio recording procedure. Parents were responsible for starting and stopping both the LENA and smartphone recorders manually, so the two recordings were not in perfect sync. Furthermore, parents also unevenly paused and resumed the recordings midway through a recording, creating more synchronization issues. Audacity[3], an open-source audio editing program, was utilized to help detect these errors and re-synchronize the files by cropping out parts not found in both recordings. The transcription files were similarly trimmed in order to match the newly synchronized LENA and smartphone recordings. LENA recorded at a sample rate of 16 kHz, while smartphones (here iPhones) recorded at 44.1 - 44.8 kHz. For further processing, all smartphone recordings were down-sampled to 16 kHz.

3. Experiment Setup, Results & Discussion

3.1. NIST Speech Signal-to-Noise Ratio

NIST speech signal-to-noise ratio[4] (STNR) is a signal-to-noise measurement method for precise measurement of speech sig-

[3] https://www.audacityteam.org/

[4] https://www.nist.gov/itl/iad/mig/nist-speech-signal-noise-ratio-measurements

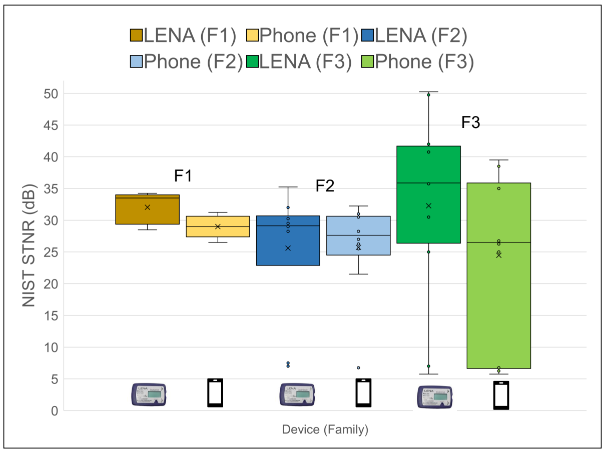

Figure 2: *Box plots reporting NIST STNR across all families for two recording devices: LENA and Smartphone. "F*" stands for family and # as provided in Table 1.*

nal strength with relatively high background noise levels, and is based on Gaussian Mixture Models. Audio for reliable speech processing should have STNR values higher than 8 dB. The available open-source MATLAB code[5] for calculating NIST STNR was modified for execution using latest version of MATLAB 2019b due to various installation issues and version updates.

In Fig. 2 we report NIST STNR calculated across multiple recordings for the three families. Irrespective of the family, LENA recordings tend to have a slightly higher NIST STNR values than Smartphone. For Family #s 1 & 2, most of the recordings are higher than 22 dB NIST STNR. But for the third family, some recordings were in the range of 7 dB to 20 dB for the smartphone. The third family reported - "Our dogs starting playing/growling/running around, so there is some background noise", which might be one of the reasons for lower NIST STNR values. Overall factors like room acoustics, appliances in kitchen, and distance from recording devices can also impact the recorded audio quality. Since the data was recorded without observation of researchers, it is difficult to precisely identify factors affecting NIST STNR for individual recordings. However, if smartphones are used in future by parents to record, we do expect that such factors will prevail, and might not be at the hands of either the researchers or parents to amend.

3.2. Automatic Speech Recognition

For the ASR experiments, we considered two models: one open-source end-to-end ASR model trained on adult speech and another fine-tuned child speech Hybrid ASR model.

3.2.1. Adult ASR

For the adult ASR experiments, we used an open-source end-to-end ASR model from Hugging Face[6] trained using the SpeechBrain[22] toolkit. This model consisted on three components: an unigram tokenizer trained using Librispeech transcriptions, RNN-based language model, and acoustic model also trained on Librispeech composed on CNN, bi-LSTM, DNN encoders followed by CTC and attention decoders.

Table 2: *WER reported for LENA versus Smartphone for the adult, child speech segments and both across three families of the parent-child book reading activity for the pre-trained end-to-end Adult ASR.*

Family	**Device**	**WER (%)**		
#		Adult	Child	Both
1	LENA	38.84	90.88	45.27
	Smartphone	35.86	87.43	42.22
2	LENA	27.09	91.13	30.00
	Smartphone	26.84	90.47	29.72
3	LENA	35.1	87.6	46.1
	Smartphone	35.3	88.6	46.5

Book reading activity data was not used in training any model, and only for test

Results in Table 2 indicate that for family #1 the WER is slightly lower for smartphone than that for LENA, while for family #s 2 & 3, the WERs are very close for both the devices (note: lower WER is better). While WER for adult speech range in between 25-40%, for children's speech the reported WERs across all families are higher and close to 90%. Although ASR systems are accurate enough for most purposes, an ASR system trained on adult speech does not work effectively for children due to large spectral and temporal variability of the characteristics of children's speech [23]. Also developing ASR systems for spontaneous child speech is far more challenging than that for adults. In the next Section, we compare these recordings using a Hybrid ASR model fine-tuned for children speech.

3.2.2. Child ASR

Language Environmental Analysis (LENA) device might often be out of reach to families due to its cost, even though it was aimed to assist low-income families for child language development. Most American household have access to a smartphone which can record quality audio. In this study, we show that smartphones can be used by families/early childhood researchers for data collection when LENA is not available. NIST STNR, a speech signal-to-noise ratio metric, shows that most of the recorded audio are higher than the expected threshold for further speech processing applications. Using a pre-trained Adult ASR model we notice that WERs across all families for smartphones are slightly lower than that for LENA. However for children speech segments the WERs are high and closer irrespective of the device. For the child ASR experiments, we trained a Hybrid DNN-HMM model using Kaldi [24] toolkit. Three corpora were used: (1) OGI Kids corpus[25] ($\approx$ 60 hours) contains both prompted and spontaneous speech of 1100 children between Kindergarten and 10th grade, collected using head-mounted microphones while interacting with a computer using prompts, (2) CMU Kids corpus[26] ($\approx$ 9 hours) in which speech is read aloud by 76 children for an age range of 6 to 11 years using head-mounted microphones, and (3) Spontaneous pre-school children speech captured using LENA in preschool classrooms in a large urban community in a Southern state using LENA recorders attached to subjects. MUSAN dataset [27] was used to augment noise to the OGI and CMU corpora.

Results in Table 3 indicate that irrespective of the device, for an Adult ASR (as shown in Sec. 3.2.1) WERs are as high as 89-90% for the child speech segments in the book reading activity. Using our Hybrid DNN-HMM Child ASR model, LENA shows a WER of 80% while Smartphone reports 82.43%. A

[5] http://labrosa.ee.columbia.edu/projects/snreval/

[6] https://huggingface.co/speechbrain/asr-crdnn-rnnlm-librispeech

Table 3: *WER reported for LENA versus Smartphone only for the child speech segments for all families of the parent-child book reading activity for the trained Hybrid DNN-HMM Child ASR.*

Device	**WER (%)**	
	Adult ASR	Child ASR
LENA	89.87	80.05
Smartphone	88.83	82.43

Book reading activity data was not used in training any model, and only for test

reason for LENA to have a slightly lower WER is that the one of the datasets (#3) used to train the hybrid ASR model was collected using LENA. Developing ASR systems for spontaneous children speech is very challenging, specially for younger children close to kindergarten age. Younger children are still developing their speech sound skills (articulion/pronunciation) until the age of 8 [28]. Their grammar or language skills are also under-developed, therefore they are prone to make mistakes as compared to spoken English (by adults). Recent research using Hybrid DNN-HMM ASR for children[29, 30], have reported WERs as high as 60 to 80% for prompted and spontaneous kindergarten aged children. Therefore, our results reported are not surprising, but we can still see an overall consistency across devices.

4. Limitations

There are several limitations of this study, which warrant further research. Importantly, we need recordings from a larger sample of families who use a variety of smartphones, including Android devices. The exact model of the devices (iPhones) used for collecting the data were not captured for the current study, which is important as audio recording hardware evolves with every new version of smartphone. Due to the unsupervised data collection protocol and without support from research staff during data collection, the exact orientation or position of the devices from the parent and the child is unknown. Also, we did not control for or measure the room acoustic properties (e.g., size, density of walls and floors/ceilings, placement of furniture, etc.) of the settings in which recordings were made. Finally, because these recordings were limited to parent-child reading activities, we need recordings collected in a variety of settings to better understand how recorders perform under different background and environmental conditions.

5. Conclusions & Future Work

In this exploratory study, we sought to examine the properties of unsupervised audio recorded by parents in their homes using LENA devices relative to the same audio recorded by parents' personal smartphones. These preliminary findings suggest that audio recorded by parents' smartphones had similar properties as audio recorded by LENA devices. Audio from two of the three families had nearly equivalent NIST STNR outcomes. Audio from the third family, which reported that the audio was recorded in a noisy environment, had better NIST STNR outcomes from the LENA suggesting that LENA reduced noise more effectively than the smartphone. However, this improved NIST STNR did not translate to substantially lower WER for LENA relative to smartphone recordings. Indeed, across all three families and both ASR models, WER's were nearly identical between the two devices. These findings provide evidence that the modern recording hardware used by common smartphones, iPhones in this case, is sufficient for parents to monitor parent and child language in natural settings.

Although this study has several limitations (see Sec. 4), these findings support further investigation of the feasibility of parents using their own devices to measure their child's language environment. As prior research has demonstrated, parents who receive data on the amount of adult language and parent-child interactions that their child experiences increase their interactions with their children, resulting in improved language growth [19, 20, 3]. However, due to its cost, this technology remains out of reach for most low-income families who cannot afford the LENA recorders and desktop computers needed to upload and process audio. Although the LENA Foundation has made concerted efforts to make their technology available to diverse communities at little or no cost through various initiatives, alternatives that reduce the need for additional hardware will accelerate the accessibility of this needed technology for families who need it the most.

In future work we aim to collect more data across a diverse population (e.g., bilingual, non-native English speakers and culturally diverse population) and diverse range of smartphones available to families. Having collected enough data, we aim to leverage transfer learning for training better end-to-end as well as hybrid ASR models for adults and children.

6. Acknowledgements

The authors express our sincere gratitude to the participating families for taking part in this study, especially during the pandemic. We would also like to thank Kyle Consolver for setting up the data collection protocol. This study is supported by NSF Cyberlearning Grant #1918032 award to Hansen and #1918012 awarded to Irvin.

7. References

[1] A. L. Ford, M. Elmquist, A. M. Merbler, A. Kriese, K. K. Will, and S. R. McConnell, "Toward an ecobehavioral model of early language development," *Early Childhood Research Quarterly*, vol. 50, pp. 246–258, 2020.

[2] B. Hart and T. Risley, "Meaningful differences in the everyday life of america's children baltimore," *MD: Paul Brookes.[Google Scholar]*, 1995.

[3] D. L. Suskind, K. R. Leffel, E. Graf, M. W. Hernandez, E. A. Gunderson, S. G. Sapolich, E. Suskind, L. Leininger, S. Goldin-Meadow, and S. C. Levine, "A parent-directed language intervention for children of low socioeconomic status: A randomized controlled pilot study," *Journal of child language*, vol. 43, no. 2, pp. 366–406, 2016.

[4] G. J. Whitehurst and C. J. Lonigan, "From prereaders to readers," *Handbook of early literacy research*, vol. 1, p. 11, 2003.

[5] D. Walker, S. J. Sepulveda, E. Hoff, M. L. Rowe, I. S. Schwartz, P. S. Dale, C. A. Peterson, K. Diamond, S. Goldin-Meadow, S. C. Levine *et al.*, "Language intervention research in early childhood care and education: A systematic survey of the literature," *Early Childhood Research Quarterly*, vol. 50, pp. 68–85, 2020.

[6] D. K. Dickinson, R. M. Golinkoff, and K. Hirsh-Pasek, "Speaking out for language: Why language is central to reading development," *Educational Researcher*, vol. 39, no. 4, pp. 305–310, 2010.

[7] J. Gilkerson, J. A. Richards, S. F. Warren, D. K. Oller, R. Russo, and B. Vohr, "Language experience in the second year of life and language outcomes in late childhood," *Pediatrics*, vol. 142, no. 4, 2018.

[8] A. C. Payne, G. J. Whitehurst, and A. L. Angell, "The role of home literacy environment in the development of language ability in preschool children from low-income families," *Early Childhood Research Quarterly*, vol. 9, no. 3-4, pp. 427–440, 1994.

[9] M. R. Burchinal, J. E. Roberts, R. Riggins, Jr, S. A. Zeisel, E. Neebe, and D. Bryant, "Relating quality of center-based child care to early cognitive and language development longitudinally," *Child development*, vol. 71, no. 2, pp. 339–357, 2000.

[10] E. Hoff, "How social contexts support and shape language development," *Developmental review*, vol. 26, no. 1, pp. 55–88, 2006.

[11] N. R. Council *et al.*, "From neurons to neighborhoods: The science of early childhood development," 2000.

[12] C. H. Biel, J. Buzhardt, J. A. Brown, M. K. Romano, C. M. Lorio, K. S. Windsor, L. A. Kaczmarek, R. Gwin, S. S. Sandall, and H. Goldstein, "Language interventions taught to caregivers in homes and classrooms: A review of intervention and implementation fidelity," *Early Childhood Research Quarterly*, vol. 50, pp. 140–156, 2020.

[13] J. Buzhardt, C. R. Greenwood, F. Jia, D. Walker, N. Schneider, A. L. Larson, M. Valdovinos, and S. R. McConnell, "Technology to guide data-driven intervention decisions: Effects on language growth of young children at risk for language delay," *Exceptional children*, vol. 87, no. 1, pp. 74–91, 2020.

[14] M. Y. Roberts and A. P. Kaiser, "The effectiveness of parent-implemented language interventions: A meta-analysis," 2011.

[15] K. Leffel and D. Suskind, "Parent-directed approaches to enrich the early language environments of children living in poverty," in *Seminars in speech and language*, vol. 34, no. 04. Thieme Medical Publishers, 2013, pp. 267–278.

[16] C. Sacks, S. Shay, L. Repplinger, K. R. Leffel, S. G. Sapolich, E. Suskind, S. Tannenbaum, and D. Suskind, "Pilot testing of a parent-directed intervention (project aspire) for underserved children who are deaf or hard of hearing," *Child Language Teaching and Therapy*, vol. 30, no. 1, pp. 91–102, 2014.

[17] D. Suskind, K. R. Leffel, M. W. Hernandez, S. G. Sapolich, E. Suskind, E. Kirkham, and P. Meehan, "An exploratory study of "quantitative linguistic feedback" effect of lena feedback on adult language production," *Communication Disorders Quarterly*, vol. 34, no. 4, pp. 199–209, 2013.

[18] J. A. Richards, D. Xu, J. Gilkerson, U. Yapanel, S. Gray, and T. Paul, "Automated assessment of child vocalization development using lena," *Journal of Speech, Language, and Hearing Research*, vol. 60, no. 7, pp. 2047–2063, 2017.

[19] C. C. Beecher and C. K. Van Pay, "Investigation of the effectiveness of a community-based parent education program to engage families in increasing language interactions with their children," *Early Childhood Research Quarterly*, vol. 53, pp. 453–463, 2020.

[20] M. Elmquist, L. H. Finestack, A. Kriese, E. M. Lease, and S. R. McConnell, "Parent education to improve early language development: A preliminary evaluation of lena starttm," *Journal of child language*, vol. 48, no. 4, pp. 670–698, 2021.

[21] D. Xu, U. Yapanel, and S. Gray, "Reliability of the lena language environment analysis system in young children's natural home environment," *Boulder, CO: Lena Foundation*, pp. 1–16, 2009.

[22] M. Ravanelli, T. Parcollet, P. Plantinga, A. Rouhe, S. Cornell, L. Lugosch, C. Subakan, N. Dawalatabad, A. Heba, J. Zhong *et al.*, "Speechbrain: A general-purpose speech toolkit," *arXiv preprint arXiv:2106.04624*, 2021.

[23] M. Gerosa, D. Giuliani, and F. Brugnara, "Acoustic variability and automatic recognition of children's speech," *Speech Communication*, vol. 49, no. 10-11, pp. 847–860, 2007.

[24] 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," in *IEEE 2011 workshop on automatic speech recognition and understanding*, no. CONF. IEEE Signal Processing Society, 2011.

[25] K. Shobaki, J.-P. Hosom, and R. Cole, "The ogi kids' speech corpus and recognizers," in *Proc. of ICSLP*, 2000, pp. 564–567.

[26] M. Eskenazi, J. Mostow, and D. Graff, "The cmu kids corpus ldc97s63," *Linguistic Data Consortium database*, 1997.

[27] D. Snyder, G. Chen, and D. Povey, "Musan: A music, speech, and noise corpus," *arXiv preprint arXiv:1510.08484*, 2015.

[28] L. D. Shriberg, "Four new speech and prosody-voice measures for genetics research and other studies in developmental phonological disorders," *Journal of Speech, Language, and Hearing Research*, vol. 36, no. 1, pp. 105–140, 1993.

[29] P. G. Shivakumar and S. Narayanan, "End-to-end neural systems for automatic children speech recognition: An empirical study," *Computer Speech & Language*, vol. 72, p. 101289, 2022.

[30] R. Lileikyte, D. Irvin, and J. H. Hansen, "Assessing child communication engagement and statistical speech patterns for american english via speech recognition in naturalistic active learning spaces," *Speech Communication*, 2022.