Center for Robust Speech Systems

KU LIFE SPAN INSTITUTE
Juniper Gardens Children's Project

Speech for Social Good Workshop, 2022

INTERSPEECH 2022
September 18 - 22 • Incheon Korea

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





Satwik Dutta[1], Jacob Reyna[1], Jay F. Buzhardt[2], Dwight W. Irvin[2], & John H.L. Hansen[1]

[1]Center for Robust Speech Systems (CRSS), The University of Texas at Dallas
[2]Juniper Gardens Children's Project (JGCP), The University of Kansas

satwik.dutta@utdallas.edu

UT DALLAS

Overview: Early Childhood Language Intervention

- **Early Childhood Language Intervention**  birth  8 years
  - **Limited exposure** to rich and engaging **language environments** ⇒ ⬇ impact **on children's language growth**
  - **Delays in early language acquisition** ⇒
    - 👉 greater need for special education services
    - 👉 lower probability of graduating from high school
    - 👉 fewer employment opportunities
  - Early language delays can be **mitigated through early intervention**

- **Neurological + Socio-behavioral theories of child development**
  - Language-rich home environments ⇒ ⬆ children's language growth
  - Parents/caregivers (diverse backgrounds) can learn skills ⇒ to increase the frequency of **quality language interactions with their children**

Overview: Challenges with LENA

- **L**anguage **EN**vironment **A**nalysis https://www.lena.org
  ⇒ Boost **positive interactions** between **children and caregivers**; give parents **immediate and frequent data** about **daily interactions**

Mobile + Light-weight + Easy-to-use + Durable

- Originally conceived ⇒ tool to benefit low-income families

- **But LENA's cost is prohibitive for most families to use**
  - Mandatory use of their proprietary digital recorder to capture audio that is processed by their speech processing software

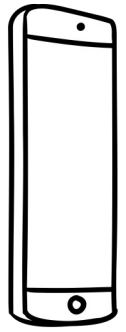
- A **lot has changed** since LENA's development over 12 years ago in **audio/speech processing technology**!

- **For our study**: Audio recording device, not their software!

LENA Foundation has made concerted efforts to make their technology available to diverse communities at little or no cost through various initiatives

Overview: Why Smartphones?

- Advanced audio recording hardware

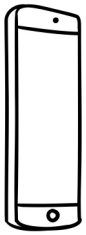


- **85% of Americans** report **owning a smartphone***
  - **76%** of those **earning less than $30,000/year**
  - **85%** of **Latinx** & **83%** of **African-Americans**
- Reduce costs & improve feasibility ⇒ utilizing technology that most families already own
- ⬆ accessibility of this technology for families with **more diverse economic backgrounds**

*https://www.pewresearch.org/internet/fact-sheet/mobile/

Overview

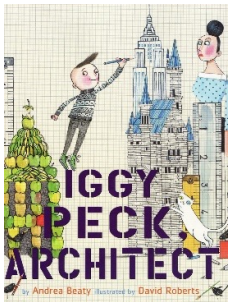
- **TASK:** Purpose of this study was to **examine the technical properties of audio** recorded by parents with the **LENA** digital recorder to the **same audio** recorded by their **personal smartphones**

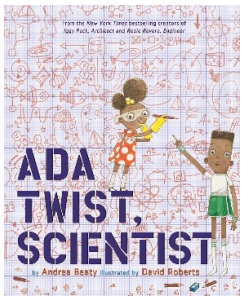



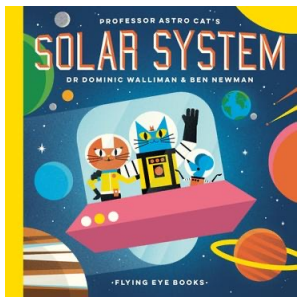
- **MOTIVATION:** Explore low-cost alternatives to **LENA** for **early childhood language intervention**, reducing the need for additional hardware will accelerate the accessibility of this needed technology for families who need it the most.

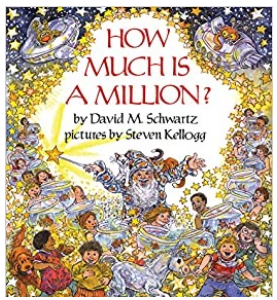
Capturing Parent-Child Booking Reading at Home

- Unsupervised data collection scenario at home of consented participating families
- Audio recording of book reading ⇒ LENA and parent's smartphone
- Recruitment ⇒ email, flyers, and word-of-mouth
- Families were expected to read **10 books** ⇒ age **3 to 8** ⇒ chosen by early childhood researchers



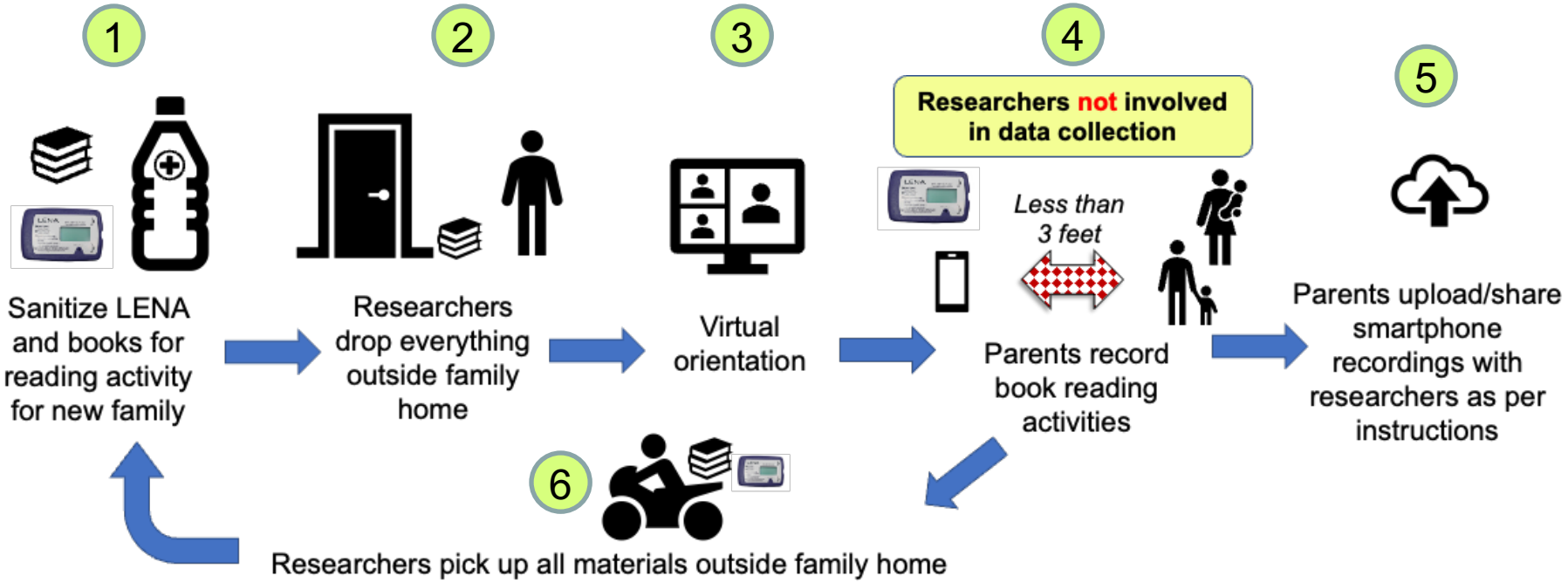






Sample books used for reading activities

Capturing Parent-Child Booking Reading at Home



Unsupervised data collection protocol

Note: ongoing COVID-19 restrictions and restrictions by IRB protocols for human subjects research

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

Details about families for the parent-child book reading activity

Challenges with Un-supervised Data Collection & Pre-processing

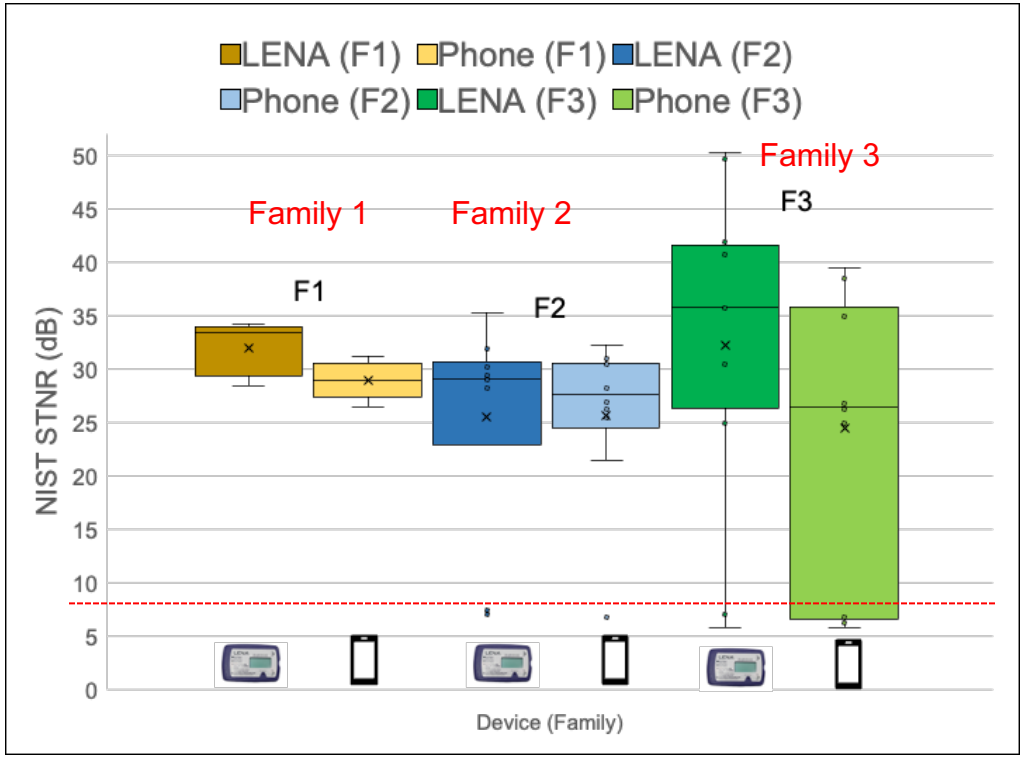
- **Recording** from two devices were not always in sync 
  - Parents responsible ⇒ START & STOP recording manually
  - Parents also unevenly paused and resumed the recordings midway through a recording ⇒ creating more synchronization issues

- **Manual Synchronization** 
  - **Audacity** (open-source) was utilized to help detect these errors
  - Re-synchronize by cropping out parts not found in both recordings
  - 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**
  - All smartphone recordings were down-sampled to 16 kHz 

Experiments: NIST STNR

- Measure speech signal to background **noise** level ⇒ GMM
- Reliable speech processing ⇒ **>8 dB** STNR
- Open-source MATLAB code* + Modifications for MATLAB 2019b



- LENA recordings ≧ Smartphone
- **Factors of impact**: room acoustics, appliances in kitchen, distance from recording devices

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/guardians to amend**

*http://labrosa.ee.columbia.edu/projects/snreval/

Experiments: Adult ASR

- Open-source end-to-end ASR model from **Hugging Face***
  - Librispeech ⇒ Adults reading audiobooks in English
  - RNN-based Language Model
  - Acoustic model ⇒ CNN, bi-LSTM, DNN encoders followed by CTC and attention decoders

- Adult 👨‍🦽👩🏾 **25-40%** WER --vs-- Child 👧🏾👱🏼 **~90%** WER

an ASR system trained on adult speech does not work effectively for children

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 ⇨ Training ❌ Testing ✅

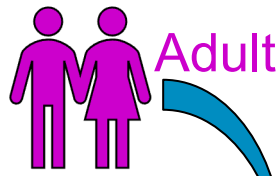
Hugging Face SpeechBrain

*https://huggingface.co/speechbrain/asr-crdnn-rnnlm-librispeech

Challenges with developing ASR for Children

- Adult Speech Recognition systems **will not work** for Children! => *Customized model for children*

- Well developed speech articulation, pronunciation skills
- Knowledge of grammar/language
- Already developed motor skills



- Higher spectral and temporal variability
- Developing knowledge of grammar/language
- Gradually developing motor skills/articulation

Experiments: Child ASR

- Hybrid DNN-HMM model Kaldi ASR toolkit
  - Language Model – RNN-based
  - Acoustic Model – CNN + TDNN-F + Attention

- Datasets:
  - OGI Kids corpus (Shobaki et al., 2000) ⇒ prompted speech of 1100 children between Kindergarten and 10th grade ⇒ head-mounted microphones
  - CMU Kids corpus (Eskenazi et al., 1997) - speech is read aloud by 76 children for an age range of 6 to 11 years using head-mounted microphones
  - [in-house] Spontaneous pre-school children (3-5 yrs) speech captured using LENA in preschool classrooms in a large urban community in a Southern state in US

- Developing ASR systems for spontaneous children speech is very challenging, specially for younger children close to kindergarten age

- Recent research (Shivakumar et.al 2022, Lileikyte et.al 2022) using Hybrid ASR for children have reported WERs (60 to 80%) for prompted and spontaneous kindergarten aged children

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

Therefore, our results reported are not surprising, but we can still see an overall consistency across devices.

Book reading activity data ⇨ Training ❌ Testing ✅

Limitations

- #s of families
- American-English speakers
- Variety of smartphones, including Android devices
- Exact Model of device ⇒ were not captured for the current study
- Exact orientation or position of the devices w.r.t. parent/child ⇒ unknown
- We did not control for or measure the room acoustic properties (e.g., size, density of walls and floors/ceilings, placement of furniture, etc.)
- Recordings in a variety of settings ⇒ how recorders perform under different background and environmental conditions

Conclusions

- **Exploratory study**: Examine properties of unsupervised audio recorded by parents in homes using LENA devices relative to the same audio recorded by parents' personal smartphones
- **Preliminary findings**: Audio recorded by parents' smartphones had similar properties as audio recorded by LENA devices
  - **Experiments**: NIST STNR + Adult ASR + Child ASR
  - **Modern recording hardware used by common smartphones, iPhones in this case, is sufficient for parents to monitor parent and child language in natural settings**
- **Due to its cost**, LENA technology remains out of reach for most low-income families who cannot afford the LENA recorders and computers needed to upload and process audio
- **Future work**:
  - 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 **self-supervised/transfer learning** for training better **end-to-end as well as hybrid ASR models** for adults and children

Acknowledgements

The authors would like to thank all the participating families for supporting the data collection efforts of this study.

NSF CyberLearning Grant: "CSL-MultiAD: Assessing Collaborative STEM Learning through Rich Information Flow based on Multi-Sensor Audio Diarization"

J. Hansen (Univ. Texas-Dallas) [Award Number:1918032]
D. Irvin (Univ. Kansas) [Award Number: 1918012]

Questions?

Satwik Dutta
satwik.dutta@utdallas.edu