Towards Automatic Understanding of Parent-Child Interaction Patterns from Family Audio to Monitor Child Mental Health

Presenter: Jialu Li

ECE 590 Seminar, Spring 2023

4/5/2023

Beckman Institute
UNIVERSITY OF ILLINOIS URBANA-CHAMPAIGN

Problem Statement

- Mental, behavioral, and developmental disorders begin in early childhood, and such disorders often go **undiagnosed**

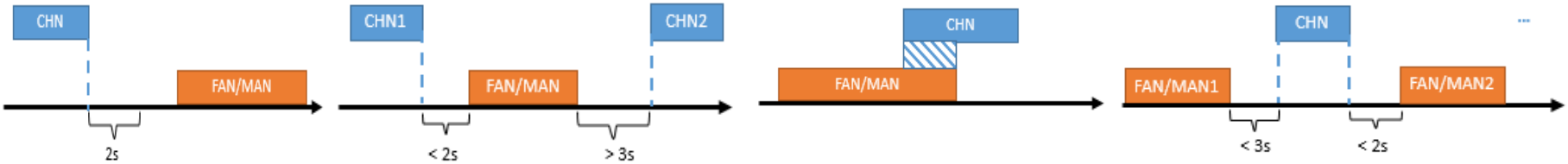
- **Daily interactions** with family members that are **repeated** and reinforced over time are critical to children's emotional well-being

1 in 6 children aged 2-8 years has a mental, behavioral, or developmental disorder.

https://www.cdc.gov/childrensmentalhealth/data.html

Patterns of Parent-Infant Interactions

- **Attachment Theory**
  - The primary caregivers who are available and **respond quickly and consistently** to an infant's needs allow the child to develop a sense of security
  - Children may develop **insecure attachment** styles if parents are often **unavailable, intrusive, or respond inconsistently** to the child's cues, particularly signs of distress



No response | Contingent vocalization | Co-vocalization | Turn-taking

Benefits of Turn-taking Interactions

- Parent/infant engaged in more positive interactions, including turn-taking in vocalizations, lead to better infant stress recovery at 6 and 9 months

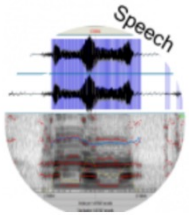
https://onlinelibrary.wiley.com/doi/10.1002/dev.22200. Hu, McElwain, Berry, 2021.

Our Goals

- Develop robust **machine learning** models to automatically predict and analyze **parent-infant** interactions in infants' **home environments**
  - **Speaker diarization** (who speaks when)
  - **Parent/Infant vocalization classifications** (type of vocalization for a given speaker)

LittleBeats™

- LittleBeats™ (LB) is a multimodal infant-wearable device that simultaneously records **audio**, **movement of the child (IMU), and heart-rate variability of the child (ECG)**



https://littlebeats.hdfs.illinois.edu/

LittleBeats™ specially-designed shirt

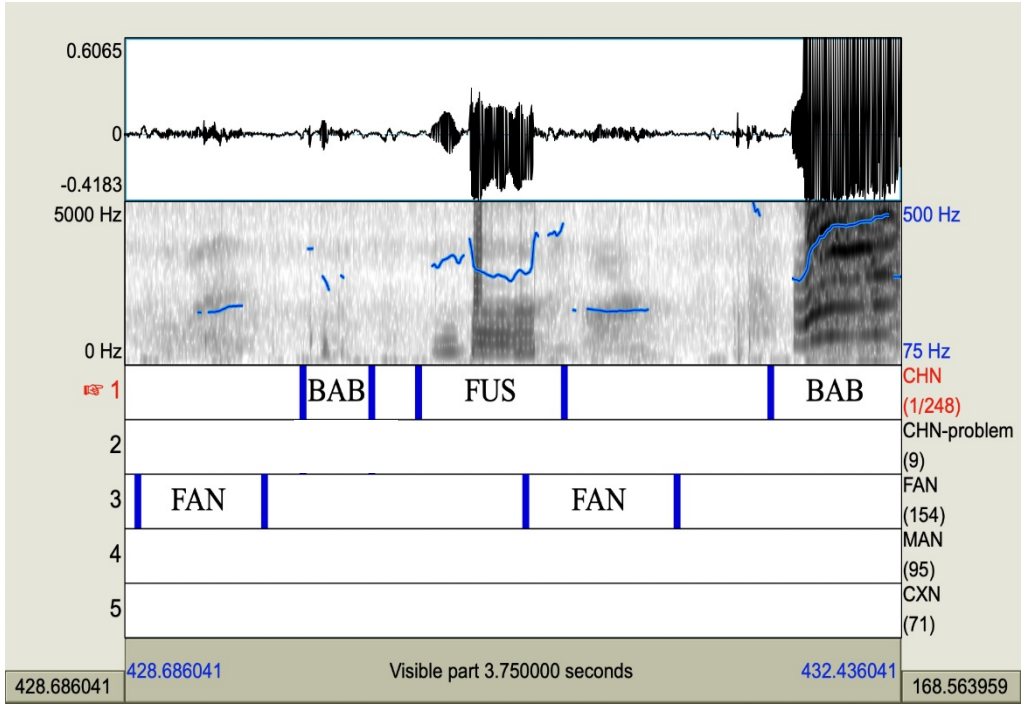
Data Collection

- **Unlabeled data**
  - Child wears the LB or LENA during the day for 2-3 days
- **Labeled data**
  - A few of **10-mins** most vocalized segments are selected for each family to perform manual annotations based on a voice activity detector
  - 10% of collected audio data is double-coded to verify inter-rater reliability scores (all annotated categories have kappa scores > 0.8)

LENA (the Language Environment Analysis device): Another infant-wearable audio-only recording device
https://www.lena.org/

Annotations

- **Speakers**
  - **CHN**: target infant
  - **FAN**: adult female (mother)
  - **MAN**: adult male (father)
  - **CXN**: another child (sibling), if any
- **Infant vocalizations**
  - **CRY**: crying
  - **FUS**: fuss
  - **BAB**: babble
- **Parent (FAN/MAN) vocalizations**
  - **CDS:** child-directed speech, motherese
  - **ADS:** adult-directed speech
  - **LAU:** laugh
  - **SNG:** rhythmic, singing
- ***Annotating 10 mins of family audio requires 2-3 hours of labor!***



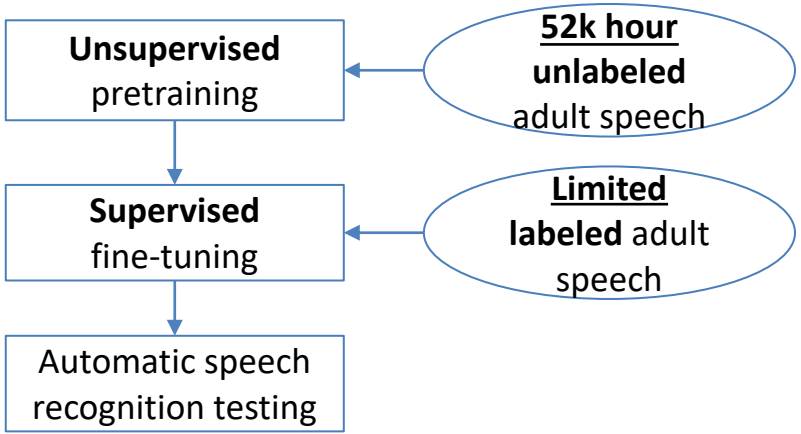
A Praat window of sample labeled segments

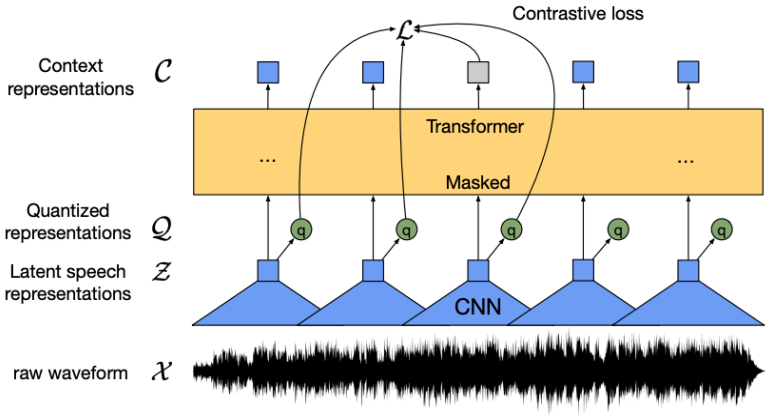
Data Distribution

Device	Type	Domain	Context	Age	# of families	Total dur (h)	Vocal dur (h)
LB	Labeled	In	Home	<14m	22	10.61	4.78h
LENA	Labeled	Out	Home	<24m	30	14.59	9.05h
LB	Labeled	Out	Virtual Visit	<10m	11	1.35	0.8h
Camera	Labeled	Out	Lab Visit	<14m	105	9.94	3.4h
LB	Unlabeled	In	Home	<5y	110	1100h	-
LENA	Unlabeled	Out	Home	<5y	275	3200h	-

Wav2vec 2.0

- The latest cutting-edge self-supervised speech processing model developed by Facebook/Meta AI research

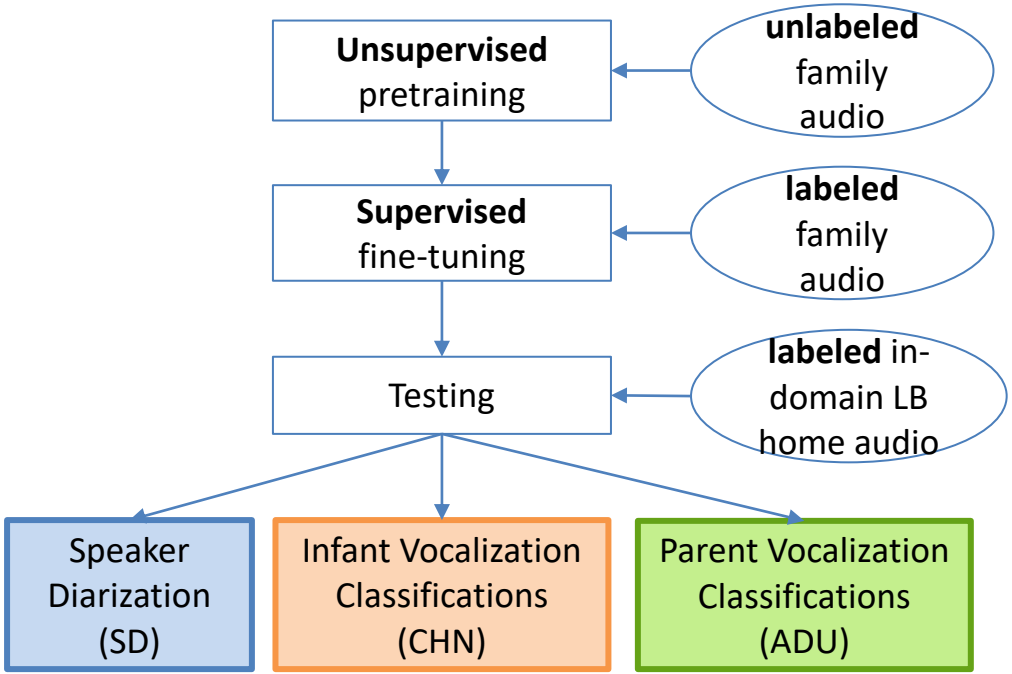




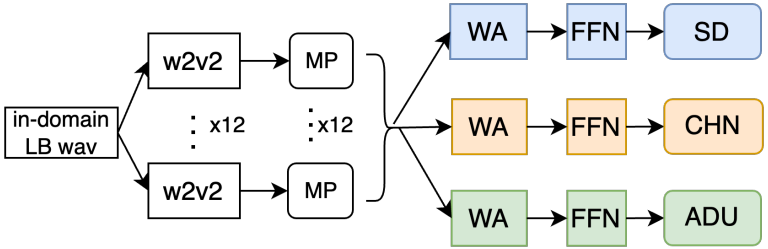
https://ai.facebook.com/blog/wav2vec-20-learning-the-structure-of-speech-from-raw-audio/

Wav2vec 2.0 on LittleBeats

- **Experimental Setup**
  - **W2V2-based** model is used
    - **12** transformer layers
    - **768** hidden feature dimension
  - Data sample is **2s** intervals each for every 0.2s; Label is determined from majority of vocalization labels in centered **1s**
  - **In-domain LB data partition**
    - **Training**: *8.67h* from 15 families
    - **Development**: *1h* from 3 families
    - **Testing**: *2h* from 4 families



Unsupervised Pretraining Comparisons



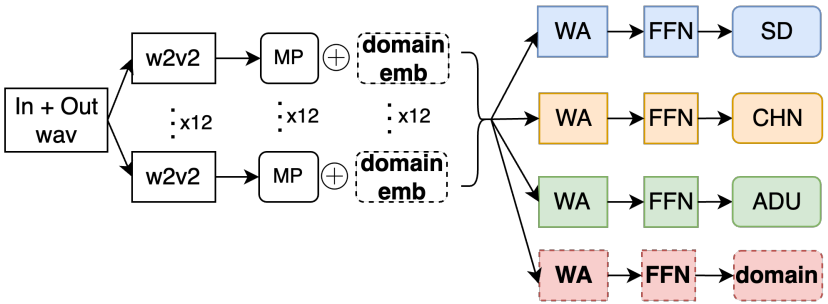
Pretrained unlabeled data	Duration (h)	SD	CHN	ADU	Avg
Oracle	52k	.628	.549	.592	.590
LibriSpeech	960	.625	.465	.535	.542
LB	1100	.682	.663	.630	.658
LB+LENA	4300	**.715**	**.689**	**.669**	**.691**

Baseline architecture

- ***MP****: mean pooling*
- ***WA****: weighted average*
- ***FFN****: feed-forward networks*

F1-scores (%) *for SD, CHN, ADU, and the average of three tiers fine-tuned on* ***in-domain labeled LB*** *data only*

Effects of Adding Out-of-domain Data



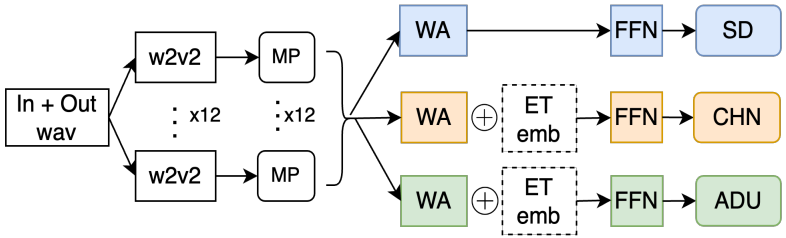
Features used	Domain learning	SD	CHN	ADU	Avg
Layer 12	-	68.7	68.1	67.6	68.1
All layers	-	68.9	69.6	70.6	69.7
Layer 12	One-hot	**70.6**	70.4	**71.9**	**71.0**
All layers	One-hot	69.6	**70.6**	68.7	69.6
Layer 12	Multi-task	66.1	68.3	69.7	68.0
All layers	Multi-task	68.4	68.6	67.3	68.1

F1-scores (%) *trained on both in- and out-of-domain data*

Binary-domain learning

- ***Domain emb****: One-hot learnable embeddings*
- ***Domain multi-task learning***

Effects of Adding Speaker Embeddings



Data	model	ET emb	SD	CHN	ADU	Avg
In	LB1100h	-	70.8	66.6	60.0	65.8
In	LB1100h	In	**73.2**	68.4	62.9	68.2
In+Out	LL4300h	-	68.9	**69.6**	70.6	69.7
In+Out	LL4300h	In+Out	**70.2**	**69.6**	**70.8**	**70.2**

F1-scores (%) *with and w/o ECAPA-TDNN embeddings*

ECAPA-TDNN Speaker Embedding

- ***ET emb****: ECAPA-TDNN speaker embeddings*
- *ECAPA-TDNN F1-scores (%) on LB testing data using training data from*
  - *In-domain: 66.6*
  - *In- + Out-of-domain: 67.7*

Desplanque et al. "*ECAPA-TDNN: Emphasized Channel Attention, Propagation and Aggregation in TDNN Based Speaker Verification*." Interspeech 2020.

Effects of Data Augmentation

Data Augmentation

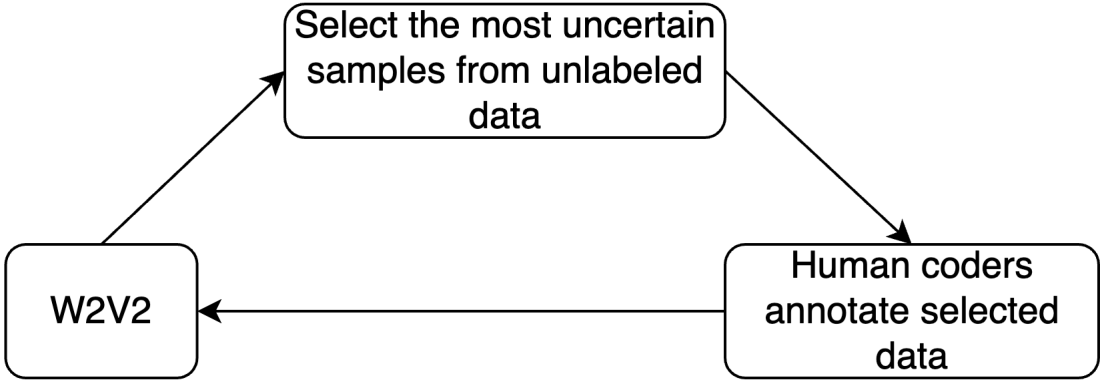
- ***SpecAug***
  - *Random audio chunks dropping*
  - *speed perturbation*
- ***Environmental Speech Corruption***
  - ***Reverberation***
    - *Room impulse responses*
  - ***Additive Noise***
    - *MUSAN (a variety of noises)*
    - *CHiME-Home (domestic noises)*
  - ***Reverberation + Noise***

noise	domain	ET emb	SD	CHN	ADU	Avg
-	-	-	68.9	69.6	70.6	69.7
CH(SD)	-	-	76.5	69.2	67.7	71.1
M(SD)	-	-	77.2	70.8	**71.5**	**73.2**
M(CHN, ADU)	-	-	75.2	71.8	67.8	71.6
M(All)	-	-	76.2	**72.0**	68.7	72.3
M(SD)	One-hot	-	77.4	68.5	64.3	70.1
M(SD)	One-hot	In+Out	**78.2**	70.1	68.8	72.4

Ko, Tom, et al. "A study on data augmentation of reverberant speech for robust speech recognition." *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2017.
Snyder et al. "Musan: A music, speech, and noise corpus." arXiv preprint arXiv:1510.08484 (2015).
Foster, Peter, et al. "Chime-home: A dataset for sound source recognition in a domestic environment." *2015 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*. IEEE, 2015.

Future Work

- Investigate **active learning** to improve W2V2



- Automatically analyze important parent-child interaction patterns
  - **Coordinated**: positive affect, turn-taking
  - **Uncoordinated**: parent unresponsive, parent or child interrupt

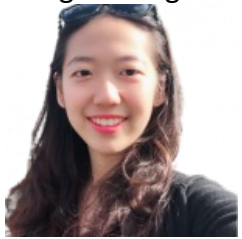
Acknowledgement

- This work is supported by Beckman Graduate Fellowship.
- **Thank YOU** to my supervisors, collaborators, project manager, graduate RAs, undergraduate RAs, and anyone who have helped me conducting research on this amazing project!

Professor Mark Hasegawa-Johnson,
Electrical and Computer Engineering

Professor Nancy McElwain,
Human Development and Family Studies(HDFS)

Professor Bashima Islam,
Worcester Polytechnic Institute



Yannan Hu
Graduate student at HDFS

Jordan Bodway, Project Manager

Lily Gaskin, LittleBeats Coordinator

https://littlebeats.hdfs.illinois.edu/team/

Thank you for listening. Questions?