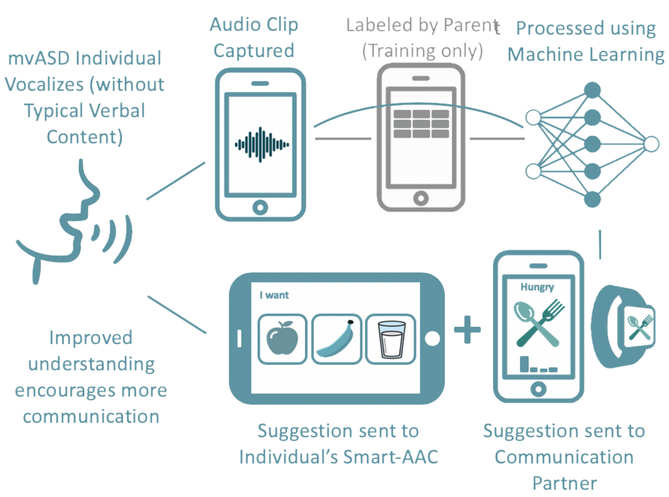
Zero-Shot Transfer Learning to Enhance Communication for Minimally Verbal Individuals with Autism using Naturalistic Data

Jaya Narain* & Kristina Johnson*, Rosalind Picard, Pattie Maes (*Equal contribution)
MIT Media Lab, {jnarain, ktj, picard, pattie}@media.mit.edu

Background & Motivation

- **Current augmentative communication systems have limited success** in conveying affect and communicative intent of individuals with minimally verbal ASD (mvASD)
- **Vocalizations (without typical verbal content) are affect and content rich** and accessible in any environment
- Our **system uses primary caregivers' unique knowledge** of an individual's vocal sounds to label and train machine learning models to build holistic communication technology
- Concept was developed through interviews (n=5) and surveys (n=18) with ASD individuals and their families

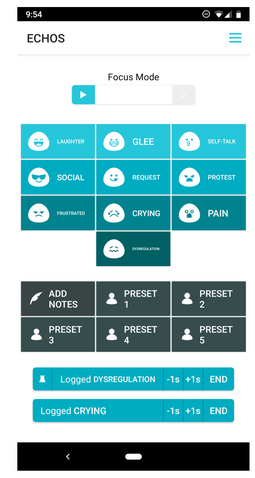


Vision for an augmentative, total communication system for nonverbal individuals using naturalistic vocalizations

Data Collection

- **Spontaneous vocalizations were collected "in the wild" during an eight-month case study** (n=1)
  - Created *mv01*, the **first labeled dataset of vocalizations without typical verbal content** from an individual with nonverbal autism (i.e., has no spoken speech)
    - Recorded **13 hours of single-channel audio** at 16 bits per second using wireless, wearable microphone
    - Vocalizations were **self-motivated communicative and affective exchanges** between the nonverbal 8-year-old and his parents
- Created **custom app** for primary caregivers to label sounds in real time
  - Collected more than **300 caregiver-labeled events**
- Emphasized **unobtrusive, affordable data collection** methods
  - Protocol to be deployed with a specialized, geographically distributed population
- To **protect privacy**, participant's family could view and delete recordings before sharing them with the research team

Small, wearable recorder in chest pocket



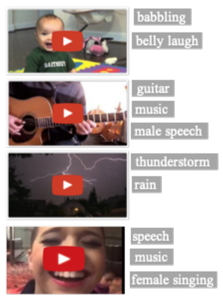
Custom app for "live labeling" by parents

Methods

Why Zero-Shot Transfer Learning?

- Examine how models trained with a large, generic audio dataset perform with non-typical vocalization data
- Inform how to augment limited training data
- Improve model performance for mvASD individuals while minimizing labeling burden by caregivers

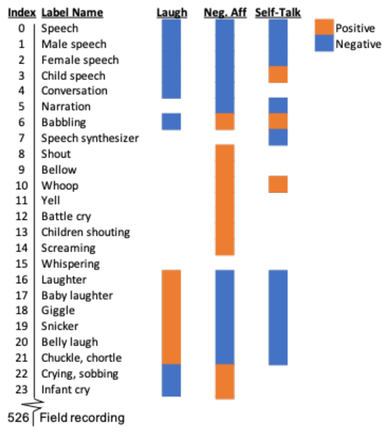
AudioSet Database



- 2 million 10-sec YouTube clips
- Human-labeled using 527 classes

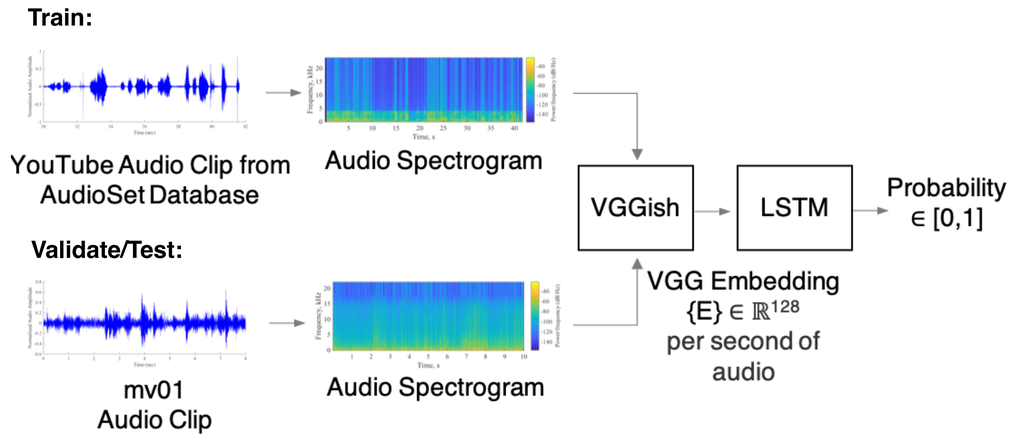
- We examined three meta-classes of vocalizations:
  - Self-talk
  - Negative affect
  - Laughter

Model Training via AudioSet Database



- Sub-classes of AudioSet were selected as pos/neg training examples (balanced sets)
  - Positive classes sounded similar to the child in *mv01*
  - Negative classes might confuse the model (e.g., close to *mv01* vocalizations or common in dataset)

Data Processing Pipeline



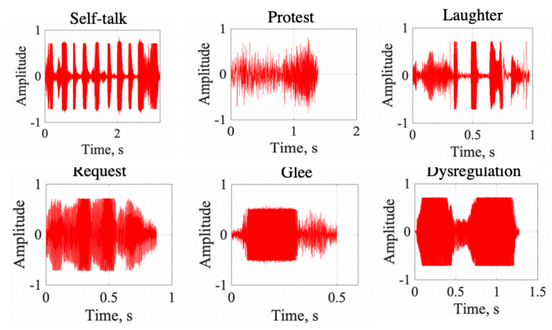
- **Validation data:** first 3 days of data collection
- **Test data:** last 2 days of data collection (held out)
- Used overlapping 9.6s segments of *mv01* with a window step size of 0.96s
- Probability thresholds were selected based on the model's performance on *mv01_validation*
- Ground truth was determined via a 5-minute test segment labeled per second by the caregiver for each category
- Test segments were selected using the caregiver's live labels to identify a segment with many instances of that meta-class

Results

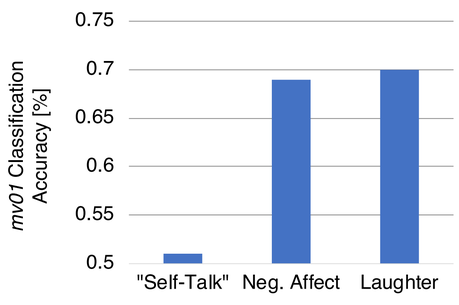
Self-Talk		Actual	
		Yes	No
Pred	Yes	0.33	0.67
	No	0.41	0.59
Accuracy: 0.511			

Neg. Affect		Actual	
		Yes	No
Pred	Yes	0.46	0.54
	No	0.27	0.73
Accuracy: 0.690			

Laughter		Actual	
		Yes	No
Pred	Yes	0.18	0.82
	No	0.06	0.94
Accuracy: 0.703			



Example, high-quality audio waveforms for 6 caregiver-labeled classes.



Near chance self-talk accuracy may reflect lack of appropriate mvASD training data

Conclusions & Future Work

- There is promise in transfer learning approaches for classes like laughter (70% accuracy) and negative affect (69% accuracy)
  - Errors in model accuracy may reflect **low availability of related audio events** in the AudioSet dataset, particularly for self-talk
  - Low true positive rate may be a consequence of the **highly varied and noisy environment of real-world data**
- **Dataset is the first of its kind,** and an important step in developing algorithms that can generalize to sparse, naturalistic data
  - As more data is collected, **direct transfer learning** between the VGGish embedding spaces of AudioSet and *mv01* may improve model performance
  - The live labels were not precisely aligned in the time domain, and future work will include developing methods for **signal-label alignment**
  - The *mv01* dataset is small and sparsely labeled with unique vocalizations, and may be well suited for **semi-supervised algorithms** in the future

Massachusetts Institute of Technology · mit media lab · fluid interfaces · affective computing · AI for Social Good