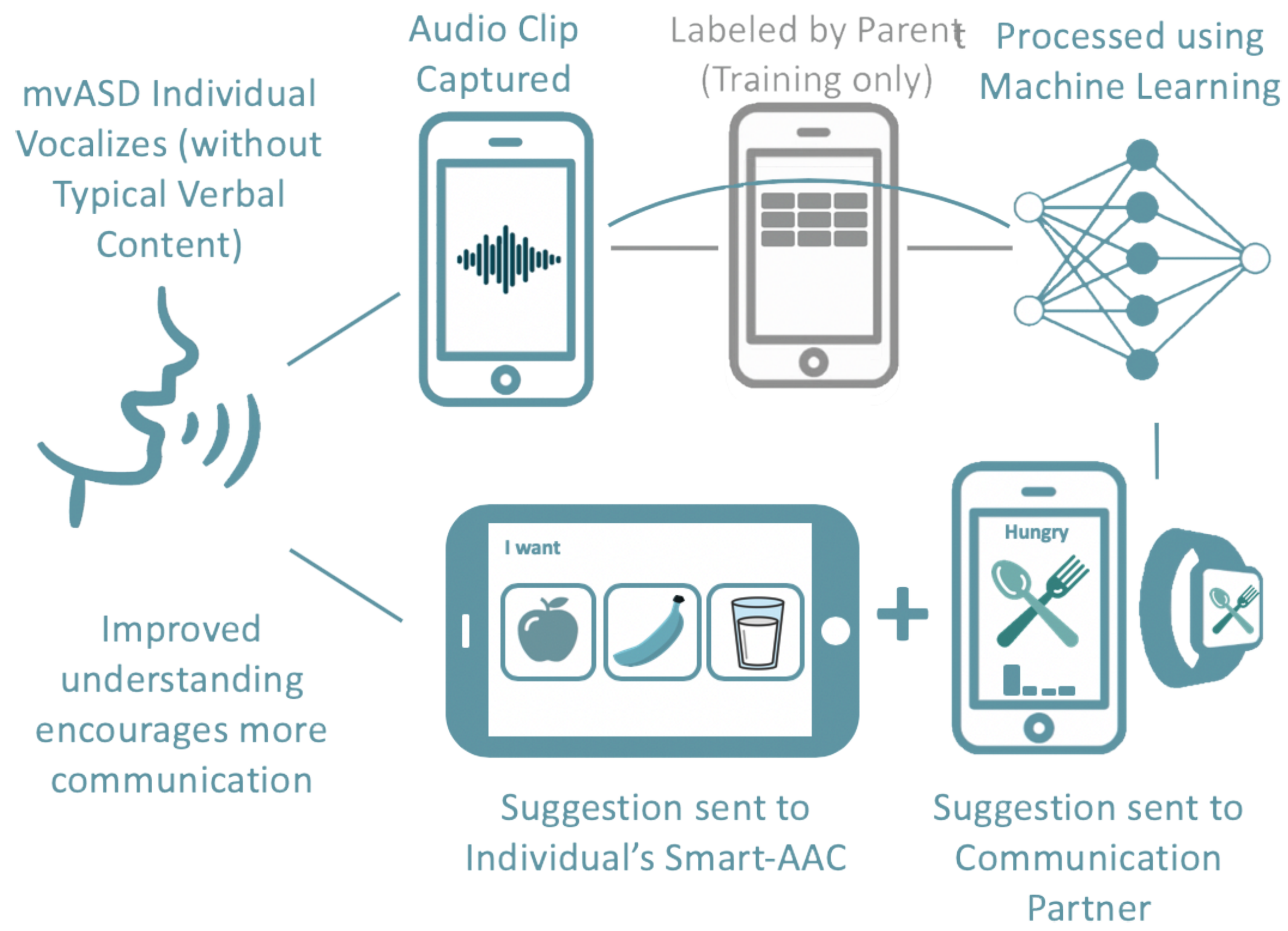
Zero-Shot Transfer Learning to Enhance Communication for Minimally Verbal Individuals with Autism using Naturalistic Data

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

Background & Motivation

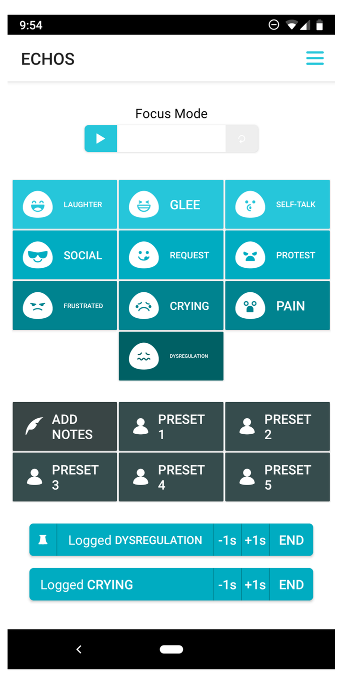
- **Current augmentative communication systems have limited success** in conveying affect and communicative intent of individuals with nonverbal and minimally verbal ASD (mvASD)
- **Vocalizations (without typical verbal content) are affect and content rich** and accessible in any environment
- Our **system uses primary caregivers' unique knowledge** of an individual's vocal sounds to label and train machine learning models in order to build holistic communication technology
- Concept was developed with 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*, to our knowledge the first labeled dataset of vocalizations without typical verbal content from an individual with nonverbal autism
    - Recorded 13 hours of single-channel audio at 16 bits per second with unprompted vocalizations using wireless, wearable microphone
    - Vocalizations are self-motivated communicative and affective exchanges between a nonverbal 8-year-old and his parents
  - Created **custom app for primary caregivers to label sounds** in real time and collected more than 300 caregiver-labeled events
  - Focused on **unobtrusive, affordable data collection** methods so that our protocol can easily be scaled up with a specialized, geographically distributed population
  - Privacy considerations for naturalistic audio remain an open and active field
    - Participant's family could view and delete recordings before sharing them with the research team
    - Planned shared dataset will include only de-identified features (no raw audio) from participants who have opted in

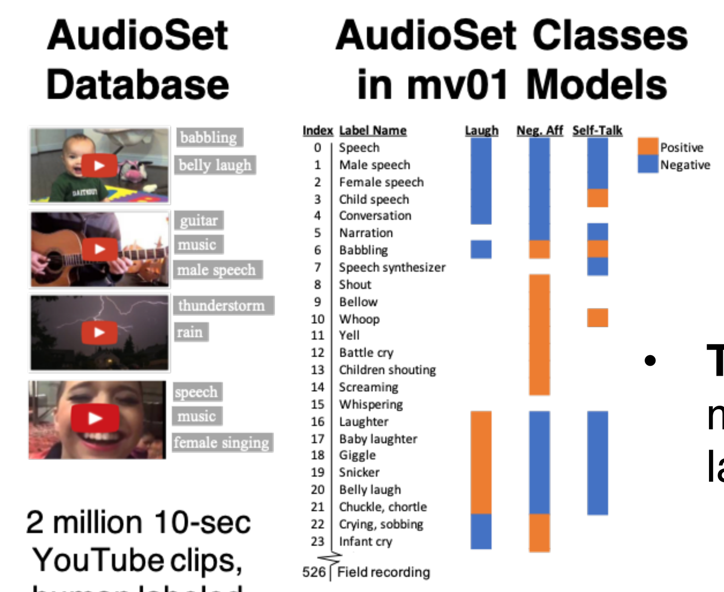


Custom app for "live labeling" by parents

Small, wearable recorder in chest pocket

Parent using labeling app at home

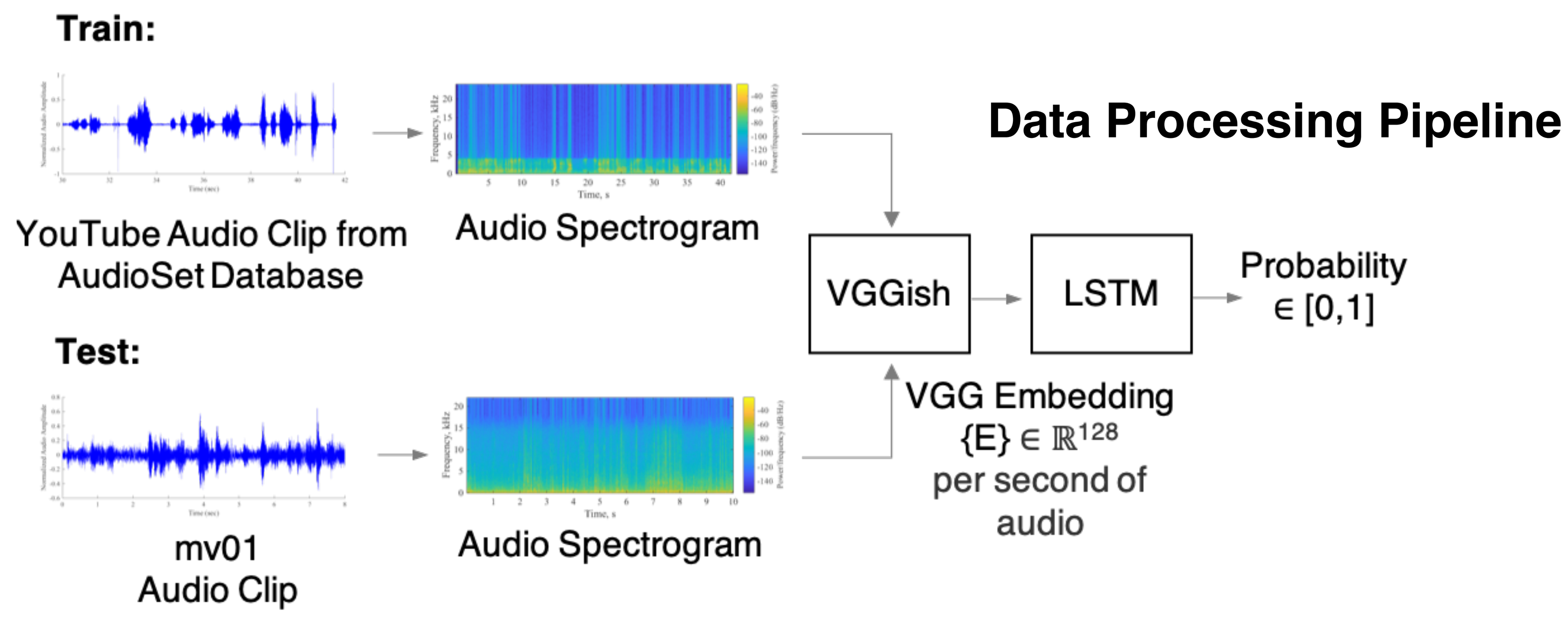
Methods



2 million 10-sec YouTube clips, human labeled using 527 classes

Selected pos/neg classes for balanced training sets

- **Zero-shot transfer learning approach, to:**
  - Examine how models trained with a large, generic audio dataset perform with non-typical vocalization data
  - Inform how to augment limited training data for improved model performance with mvASD individuals
- **Three-layer LSTM** were developed for the meta-classes of self-talk, negative affect, and laughter
  - Trained using AudioSet, a large generic database of YouTube clips human-labeled with 527 classes
  - Validated and tested using *mv01* dataset
- **Training:** Sub-classes of AudioSet were selected as positive and negative training examples
  - Positive classes were picked that sounded similar to the child in *mv01_validation*
  - Negative classes were selected that might confuse the model (e.g. close to *mv01* vocalizations or common in dataset)
- **Testing:** Model was used to predict the probability of the presence of each category in overlapping 9.6s segments of *mv01_test* (data collected after model training and validation), with a window step size of 0.96s
  - Probability thresholds were selected based on the model's performance on *mv01_validation*: 0.80 for Self-Talk, 0.70 for Negative Affect, and 0.95 for Laughter
  - A 5-minute test segment was selected for each meta-class from *mv01_test* using the caregiver's live labels to identify a segment with many instances of that meta-class



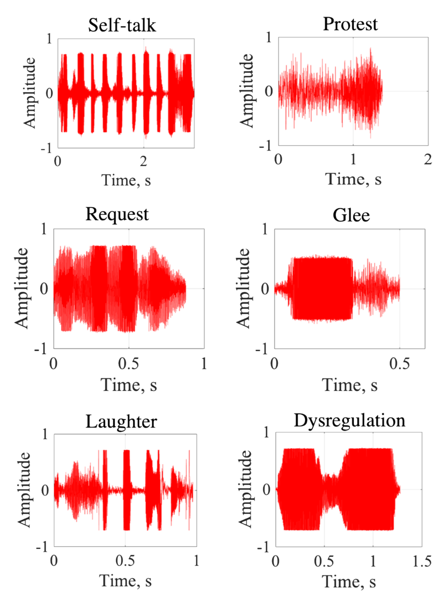
Results

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

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

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

- Low true positive rate may be a consequence of the **highly varied and noisy environment of real-world data**
- Limited accuracy of models may be related to the **availability – or lack thereof – of accurately representative event classes** in the AudioSet dataset, particularly for self-talk



Example, high-quality waveform for audio in six caregiver-labeled classes.

Conclusions

- **Distinct vocalization characteristics are observable** in visualizations
  - Distinctions relate to individuals' emotional or physical state and communication intent; high within-label variance due to natural environment
- **Dataset is one of 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