The ECHOS Platform to Enhance Communication for Nonverbal Children with Autism: A Case Study

Kristina T. Johnson*
Jaya Narain*
Craig Ferguson
Rosalind Picard
Pattie Maes
MIT Media Lab
Cambridge, MA 02139 USA
[ktj, jnarain]@mit.edu
craigsonoffergus@gmail.com
[picard, pattie]@media.mit.edu
*Equal Contribution

Abstract
Current augmentative communication technology has had limited success conveying the needs of some individuals with minimally verbal autism spectrum disorder (mvASD). Primary caregivers report being able to better understand these individuals’ non-traditional utterances than those less familiar with the individual such as teachers and community members. We present an eight-month multi-phase case study for a translational platform, ECHOS, that uses primary caregivers’ unique knowledge to enhance communicative and affective exchanges between mvASD individuals and the broader community. Through iterative development and participatory design, we discovered that physiological sensors were impractical for long-term use, on-body audio was content rich and easily accessible, and a custom in-the-moment labeling app was transformative in obtaining accurate labels from caregivers for machine learning advancements. This paper presents the design methodology, results, and reflections from our case study and provides insights into development with and for the special needs community.

Author Keywords
autism spectrum disorder; minimally verbal; augmentative communication; assistive technology; participatory design; affect; audio signal processing; machine learning
Introduction

One in 59 individuals (2.1 million people) in the United States have autism spectrum disorder (ASD). Of those, approximately 34% (~714,000 people) are minimally verbal and 3.7% (~78,000) are nonverbal, meaning they use fewer than 20 or 0 words/word approximations, respectively [1]. These individuals often use non-traditional vocal communication (hums, consonant utterances, babbbling sounds), gestures, or contextual clues (e.g., leading a caregiver to a cabinet to request a snack) to convey affect and intent. Some utterances may have specific meanings: “kakaka” might indicate “car” or “buhbuh” for “go to the playground.” Other vocalizations may have a less clear 1:1 mapping, such as loud, high-pitched squeals to indicate general frustration or low humming sounds to indicate pleasure.

In this paper, we present our vision for ECHOS (Enhancing Communication using Holistic Observations and Sensing), an assistive technology platform designed to enhance communication with autistic individuals who do not use traditional spoken words (see Fig. 1). Using parent-generated affective and communication labels and natural-environment audio captured via a camcorder worn by the individual, we describe the first step towards an unobtrusive, scalable device that can help individuals with mvASD communicate and be more readily understood. Through an iterative longitudinal approach with one family, we present the design process, multi-phase data collection methodology, selected analysis, and results. We conclude with remarks for future work on the platform for this underserved population.

Related Work

Existing communication-based research for ASD or non-traditional language falls broadly into two categories: 1) technology to enhance traditional communication, including generating verbal speech, and 2) studies of non-speech communication, including vocalizations in infants, ASD diagnoses, and affect and physiological analyses.

Research related to augmentative and assistive communication (AAC) devices and autism has largely focused on enhancing typical social and verbal communication abilities in children with ASD [3]. Technology and research targeted for individuals with speech conditions like ALS, cerebral palsy, and stroke often require volitional training with fine motor control and minimal extraneous movement [5]. These interfaces may not be well suited for individuals with nonverbal and mvASD for whom volitional speech-related muscle contraction may be nascent or nonexistent.

Prior work in non-speech vocalizations has focused on classifying infant cries by need (e.g., hunger, pain) using both humans [16] and machines [7], as well as diagnosing ASD using infant cries [14] or naturalistic child vocalizations [10]. Researchers have also explored affect detection in speech with typical verbal content [13], including in ASD populations with verbal abilities in task-driven [2] and natural settings [12]; however, no known work to date has attempted to classify communicative content in vocalizations from non-infant individuals who are minimally or nonverbal. Moreover, previous work towards translational systems for individuals with ASD using machine learning has focused on detecting a single emotional valence in lab-based settings using physiology [11, 8], and most approaches have relied on labels by researchers or therapists rather than caregivers.

CCS Concepts

- Human-centered computing → Accessibility technologies; Participatory design; Field studies;

Figure 1: ECHOS Platform Vision: Audio is acquired in real-time, processed through machine learning algorithms, and “translated” into shareable communicative content.

1We have found preference for person-first [6] and identity-first [4] language to vary throughout the ASD community, so the terms will be used interchangeably in this work.
Design Process
For technology to have utility and value, it must address a deeply rooted need. Hence, we began by interviewing and surveying individuals who have speech or language challenges (n=4 interviews; n=15 survey respondents) and/or their caregivers (n=9; n=49), including 5 interviewed families with nonverbal or mvASD\textsuperscript{2} children and 18 survey respondents with ASD. This early engagement with the community served as the foundation for ECHOS. For example, many respondents (63%; 40/64) reported difficulties using existing communication augmentation devices. All interviewees (13/13) mentioned that devices did not sufficiently meet their communicative or affective needs, and most explicitly reported that miscommunication was a source of stress. The parents of mvASD children, in particular, noted that the existing technology was minimally accessible to their children, and that they understood their offspring’s communicative intent better than others who interacted with their child like teachers and babysitters. As a result, we chose to focus this work on the population of minimally and nonverbal individuals with ASD.

Design Challenges
MvASD individuals are not only are highly heterogeneous in their communication styles and levels, but they are also more likely to have co-occurring intellectual disabilities, genetic conditions, epilepsy, anxiety, and other diagnoses that exclude them from research studies [15]. Sensory sensitivities, difficulty following protocols, and practical considerations like working around school and therapy schedules can further limit participation. Moreover, working with a small, geographically distributed population can hinder local recruitment efforts and the ability for participants to return to a research center repeatedly. Accordingly, studies should work with the needs of the users and should be mindful of placing any additional resource loads — e.g., time, money, logistics — in order to advance science and technology development for and with this specialized population.

Design Approach
Because of the complexity and diversity of these needs, we conducted an eight-month longitudinal case study with one nonverbal child and his family and caregiver network. His mother is both a researcher and one of the authors of this paper, enabling unique insight and access to this specialized group. While most studies take a cross-sectional approach, engaging with many autistic individuals in a controlled, single-session laboratory environment, we built around diverse natural environments and the small network of people surrounding this one individual (parents, grandparents, babysitters, and siblings). These individuals provided verbal feedback 3-4 times per week, which steadily informed the data acquisition process and goals of the project.

Using the oak-tree design approach described in Fig. 2, we are working towards an ECHOS platform that helps enhance individuals’ communicative exchanges using simple, accessible sensors throughout daily life. We envision a system whereby families can participate remotely, use a do-it-yourself kit to collect, label, and upload data, and then benefit from the results of personalized trained communication models.

Although this design approach is not without limitations — for example, usability, comfort, equipment access, privacy, and data transfer will all need to be re-assessed as we move past n=1 — we hope that the combination of firm roots and an upward focus will align this platform with the broader community’s needs as we progress toward scale.

\textsuperscript{2}Since mvASD subsumes nonverbal ASD and is the more recognized term in literature, we will use mvASD to refer to both nonverbal and minimally verbal individuals with autism throughout this text.
Phase 1: Methods
Throughout the interviews and surveys, individuals had expressed interest in devices that could provide information about physical state (e.g., pain, emotions, hunger) without requiring active input from a user. From prior work, we knew that wearable physiological and motion sensors could provide such information [8]. Hence, Phase 1 of the ECHOS system focused on physiological sensing and affect-based detection using wearable sensors.

Setup
Data were collected with a non-speaking autistic boy of elementary school age. He has zero spoken words and limited use of AAC tools. The participant and his family provided IRB-approved informed consent/assent, where assent for this child was determined through body language, behavior, and other holistic forms of communication.

![Image of sensors and electrodes](Image)

Figure 3: A) Phase 1 data collection involved six gelled electrodes and two light adhesives. B) In Phase 2, there was no skin-adhesive contact. First-person audio and video was collected from a small camera in a t-shirt chest pocket.

During this phase, we collected electrodermal activity (EDA; formerly galvanic skin response), electrocardiography (ECG), photoplethysmography (PPG), and acceleration data using wireless wearable sensors (see Fig. 3a). Several sensors were trialed in order to determine which, if any, would be useful for future studies.

The participant wore an E4 sensor ($1700; Empatica, Italy) on each wrist. These watch-like sensors measure skin conductance via changes in sweat gland activity, which are a function of the body’s sympathetic nervous system activation and are often used as a gross proxy for stress. The E4 sensors also record the user’s skin temperature, pulse rate, and 3-axis accelerometry. To reduce motion artifacts and increase data quality, pre-gelled adhesive electrodes with 0.5% NaCl isotropic gel were attached to the user’s distal wrists (see Fig. 3a, inset). Sweatbands were placed over the sensors to further minimize movement and distraction and to ensure comfort. Zephyr Biopatch sensors ($500; Medtronic, USA) were secured to the participant’s chest using adhesive pre-gelled Ag/AgCl ECG electrodes (Cathay, China) to record ECG, respiration, and acceleration. PPG and accelerometer Byteflies sensors (Market price unavailable; Byteflies, Belgium) were attached to the participant’s left and right ankles using non-gelled light adhesives.

During the first few sessions, the participant seemed distressed during the sensor attachment period. Because the participant’s comfort and consent was paramount, multiple adaptations were made, including using a dedicated area of the house for setup and the use of a photo-based schedule of activities. Ultimately, giving the participant the option of watching a preferred video clip during sensor attachment was the most successful at reducing distress during setup.

After attaching the sensors, the participant was free to pursue his regular activities. The researcher captured third-person audio and video but was otherwise minimally engaged in order to capture the most naturalistic record possible. The family participated in 6 sessions over the course of 3 weeks (see Fig. 6), resulting in over 130 hours of multimodal data streams from this phase.

Labeling
Phase 1 data were labeled post hoc by a researcher who was not related to the pilot family. While post hoc labeling by a researcher or other professional is the discipline standard, our intent was always to involve the participant family in this process to reduce bias and enhance accuracy. However, we discovered this process to be impractical (see Phase 1: Results & Discussion) and only the researcher’s labels were used for this phase.

---

3While any interpretation of the participant’s data by someone other than the participant is inherently biased, those closest to the participant have the most experience and therefore the best chance at accurately interpreting the participant’s affect and communication.
First, the collected video streams were labeled continuously for general content such as location, activity, or interactions like pointing to an item. These labels were then used to provide context to the collected signals, which lacked a contextual ground truth due to the natural-environment study design. While exploring the various signals, we discovered that audio was surprisingly rich and appeared to be qualitatively consistent at conveying certain types of affect and communicative intent.

Then, to evaluate the audio signals for affective and communicative information, the researcher manually segmented vocalizations from a 3.5 hour subset of video collected over 3 data sessions. Using contextual information from the video and acquired knowledge from the parents and other caregivers, the researcher assigned these segmented vocalizations to one of seven categories: laughter, crying, self-talk, request, protest, dysregulation, or no label if it was not possible to distinguish. The researcher was able to confidently label some vocalizations (e.g., laughter), but had low confidence in other categories (e.g., dysregulation).

**Figure 4:** A) Time-domain visualization of researcher-labeled Phase 1 data. Each plot shows superimposed semi-transparent audio waveforms, suggesting distinct variations between sound types. B) A t-SNE dimensionality reduction exhibits some clustering of similar sounds, such as crying and self-talk. Only sounds that the researcher could label with high confidence are colored. The feature space included both time- and frequency-domain audio characteristics.

**Analysis Methods**

The overlapping semi-transparent audio waveforms presented in Figure 4a suggested distinct variations between sound types. To explore the feasibility of machine-based classification of vocalizations without typical verbal content, we tested a support vector machine (SVM) with multi-class handling on the researcher-labeled clips using 215 audio segments spanning six classes. For this preliminary analysis, the clips were randomly split 80/20 (train/test), though all data points had been seen by a researcher during the labeling phase. We also explored unsupervised learning based on a t-SNE dimensionality reduction (Fig. 4b).

**Phase 1: Results & Discussion**

The multi-class SVM produced a weighted F1 score of 0.67, suggesting that these researcher-labeled classes contained sufficient similarity to be partially clustered by a machine. Although exploratory, we expected these results to improve if the segments could be labeled with higher confidence by someone more familiar with the child. However, even the primary caregiver found it difficult to label the audio segments with high confidence without context, and watching the full video was too time intensive. It had taken the researcher over 4 hours to continuously label each hour of video and another 1-2 hours to label an hour’s worth of extracted vocal segments. This process was an unrealistic load to place on the participant’s family, and we realized it would limit the ability for a platform built on this labeling procedure to succeed with other families with more limited resources or expertise. This experience prompted the development of the live-labeling app used in Phase 2 (see Fig. 5), enabling real-time labeling by a caregiver.

Qualitative analysis of the physiological signals indicated some post-hoc correlations between labeled context and physiology – e.g., high phasic EDA during challenging activities, low phasic EDA during sick days, and distinct skin conductance responses during stressful or exciting moments – but sensor comfort and long-term use was a concern. While the sensors themselves did not appear to be irritating, the gel electrodes, particularly from the ECG Biopatch sensor, left a sticky residue on the skin that was difficult to remove. The family reported using soapy water, alcohol wipes, and other methods to remove the residue, which led to mild skin irritation. In response, a cloth-based Bioharness (Medtronic, USA) was trialed, but the data quality and fit was poor. The light, non-gelled adheres on the Byteflies sensors seemed comfortable and did not leave a residue, but they occasionally fell off the participant during
Our initial study design was built on the idea that high-quality data and scientific findings would precipitate complementary technology advancements. While this notion is still sensible, the naturalistic physiological data from this phase lacked sufficient context and data quality for reliable interpretations. The data we collected were feature rich and are worth further exploration, but they did not fit our vision for ECHOS as a scalable communication platform that could integrate easily into day-to-day life. As a result, we chose to pivot our approach to explore how more easily observable features such as vocalizations could be used meaningfully in an HCI platform. Such an approach has deployment advantages compared to a platform requiring costly sensors and attachment methods.

### Phase 2: Methods

Phase 2 focused on first-person audio/video data capture for both affective and communicative expressions. This process was found to provide rich data in a comfortable, easy-to-use format. An in-the-moment labeling app enabled over 300 “focused” labels across 13 hours of audio/video data, forming the basis of a dataset for transfer learning and semi-supervised learning approaches.

#### Setup

A lightweight audio/video recorder ($40; BOBLOV, China) measuring 3.7” x 1.0” x 0.4” was placed in a small custom chest pocket on a t-shirt. Single channel audio was recorded at 32 kHz and 16 bits per second. Video was recorded at 30 frames per second and 1920 x 1080 resolution. As before, after donning the shirt, the participant pursued his regular activities. The parents had control over the data and could delete files before sharing them with the research team.

#### Labeling

Using insights gleaned from Phase 1, we developed an Android app\(^4\) that enables simple, intuitive, in-the-moment labeling (see Fig. 5). The app went through five iterative design stages using feedback from every data session. The current version (Fig. 5) includes icons and color gradients for quick visual scanning, customizable label presets, a contextual note option, easily adjustable timestamps, and the ability to “pin” an event to label two things simultaneously. Each label can be registered as an instantaneous marker (one press) or a range (press “End”). Users can review, edit, and delete past events through the hamburger menu button in the upper right. All labels are timestamped and synced to a server at the user’s discretion. These labels can then be integrated with the user audio/video files for further analysis. To enhance label accuracy and synchronization with audio/video data, caregivers were asked to label data in 5-minute focused chunks. A Focus Mode progress bar at the top of the screen serves as a visual reminder, though the timer can be paused at any time.

#### Analysis Methods

To assess the applicability of existing general datasets for this platform, we implemented a zero-shot transfer learning (ZSL) classification approach. The goal of this analysis was to explore whether generic databases could be used to supplement our smaller specialized dataset and minimize the number of caregiver labels needed. The complete details of our method and results are presented in our previous work.\(^4\)

---

\(^4\)This app is open source and available by emailing the authors.
We first trained three models on sounds from the AudioSet database, a large public database of YouTube clips that have been human-labeled with over 500 audio event labels, including speech, music, nature sounds. Each model was built to classify one of three categories of audio from our participant: laughter, self-talk (similar to babbling), and general negative affect. Three-layer LSTM models were trained using batch normalization and Adam optimization with balanced positive and negative AudioSet training sets. The input to the LSTM model was a VGGish embedding of an audio waveform. A past/future split was then used to validate/test: the first 3 days of Phase 2 data was used for validation and optimization, while the last 2 days of data were held out for testing.

Phase 2: Results & Discussion
The ZSL Laughter and Negative Affect models built using the AudioSet database yielded 70% and 69% accuracy, respectively, suggesting that these generic, easily accessible datasets may be helpful in augmenting training or reducing labeling burden on caregivers for certain sounds. However, the ZSL Self-Talk model performed just above chance, and all of the models had high false positive rates, highlighting the need for more targeted approaches and datasets that include unique data from this specialized population.

The hardware used in Phase 2 – a recorder tucked into a t-shirt chest pocket – appeared comfortable for the participant. Collecting data required minimal setup from the parents, and they were able to label a large volume of data during a two week period (see Fig. 6b) with reported ease. Even so, and even during Focus Mode, the caregiver still had to wait for an utterance or interaction, interpret it, and press the appropriate button, resulting in unavoidable and inconsistent delays between the events and labels. The resultant sparsely labeled dataset is well-posed for semi-supervised learning approaches and audio signal processing techniques, which we are currently exploring.

Conclusions & Future Work
This work presented the development and eight-month multi-phase case study for a communication-enhancement platform called ECHOS. We discovered that using multiple physiological sensors with adhesive electrodes was impractical for long-term use due to wearability, cost, and irregular data quality. In contrast, an inexpensive audio recorder provided rich affective and communicative information while maximizing comfort, ease of use, and potential scalability. In-the-moment labeling using a custom-built open-source app enabled critical engagement of primary caregivers, who provided unique insights into the meaning of nonverbal communication. Results from each phase of the study underscored the potential of such a platform to facilitate and empower more meaningful interactions between mvASD individuals and the broader community while functioning effectively at scale.

In the coming months, we will assess how these findings generalize and can be adapted beyond n=1. We are currently developing data processing pipelines to align the naturalistic data with the caregivers’ in-the-moment labels, as well as semi-supervised and transfer learning approaches to analyze these unique audio streams. We are also building – and iteratively testing with target families – accessible user interface platforms to engage the caregiver in the analysis procedure, data transfer protocols to ensure data privacy, and human-computer interfaces to realize the complete ECHOS platform. We hope that this multi-faceted participatory design approach will engender a system that helps us all better understand one another.
Acknowledgements
The authors thank the study participant and his family, as well as Natasha Jaques, Ishwarya Ananthabhotla, Doug Beeferman, and Peter Woford. They gratefully acknowledge support from the MIT Media Lab and Microsoft AI for Accessibility.

REFERENCES