scientific data

Check for updates

OPEN

DATA DESCRIPTOR

ReCANVo: A database of real-world communicative and affective nonverbal vocalizations

Kristina T. Johnson[1,3] ✉, Jaya Narain[1,3] ✉, Thomas Quatieri[2], Pattie Maes[1] & Rosalind W. Picard[1]

Nonverbal vocalizations, such as sighs, grunts, and yells, are informative expressions within typical verbal speech. Likewise, individuals who produce 0–10 spoken words or word approximations ("minimally speaking" individuals) convey rich affective and communicative information through nonverbal vocalizations even without verbal speech. Yet, despite their rich content, little to no data exists on the vocal expressions of this population. Here, we present ReCANVo: Real-World Communicative and Affective Nonverbal Vocalizations - a novel dataset of non-speech vocalizations labeled by function from minimally speaking individuals. The ReCANVo database contains over 7000 vocalizations spanning communicative and affective functions from eight minimally speaking individuals, along with communication profiles for each participant. Vocalizations were recorded in real-world settings and labeled in real-time by a close family member who knew the communicator well and had access to contextual information while labeling. ReCANVo is a novel database of nonverbal vocalizations from minimally speaking individuals, the largest available dataset of nonverbal vocalizations, and one of the only affective speech datasets collected amidst daily life across contexts.

Background & Summary

Nonverbal vocalizations, such as grunts, yells, and squeals, are an important part of communication[1]. Traditionally, human-based studies of affect and communication using nonverbal vocalizations have focused on pre-verbal vocalizations in infants[2,3] or on nonverbal vocalizations that occur amidst typical word-based speech like moans and sighs[4,5]. Yet, for non- and minimally speaking individuals who produce zero or only a handful of spoken words (denoted here as mv* individuals), nonverbal vocalizations convey important communicative and affective information. Note that we use the term mv* ("M-V-star") to refer to a sub-population of non- and minimally speaking individuals. These individuals have limited expressive language through verbal speech, alternative and augmentative communication (AAC) devices, and signed languages, though they use vocalizations and other nonverbal expressions such as gestures, facial expressions, and vocalizations as effective modes of communication. To our knowledge, nonverbal vocalizations as communication from mv* individuals have not been systematically studied, due in part to a lack of access to data from this community. Here, we present the first dataset of nonverbal vocalizations from mv* individuals labeled for affect and communicative function. The goal of this dataset is to spur further investigation into the acquisition, analysis, and reciprocation of non-speech vocalizations from minimally speaking individuals.

The study of nonverbal vocalizations with mv* individuals presents unique challenges. The population is relatively small, comprising approximately 1–2 million in the United States[6–8], and they are geographically distributed[9,10]. The resource burden on this population is high[11,12], so studies must be designed thoughtfully to minimize the time, effort, and inconvenience of participation. Additionally, this population is highly heterogeneous, including diagnoses of autism spectrum disorder (ASD), genetic neurodevelopmental disorders, cerebral palsy (CP), and other global developmental delays, and the specific etiologies of certain behaviors and symptoms are often not known[13–15]. For example, a person may not speak due to motor planning difficulties, cognitive delays, differences in social motivation, some combination thereof, or alternative causes. In addition, the abilities and/

[1]Massachusetts Institute of Technology, MIT Media Lab, Cambridge, MA, USA. [2]Massachusetts Institute of Technology, Lincoln Laboratory, Lexington, MA, USA. [3]These authors contributed equally: Kristina T. Johnson, Jaya Narain. ✉e-mail: ktj@mit.edu; jnarain@alum.mit.edu

Participant ID	Gender	Age (year range)	Diagnoses affecting speech and/or language	Time span of included data (weeks)	Number of spoken words or word approximations (parent report)
P01	M	18–25	Autism, Down syndrome (DS)	64	0
P02	M	18–25	Autism	7	4
P03	M	6–9	Autism, Rare genetic disorder	16	0
P05	F	9–12	Autism	11	0
P06	M	9–12	Autism, Cerebral palsy (CP)	4	3
P08	F	6–9	Autism	20	0
P11	M	9–12	CP	19	1
P16	M	6–9	Autism	10	5–8

Table 1. Participant demographics.

or behaviors affecting communication can evolve over time, further augmenting the heterogeneity of this group and motivating the need for quantitative longitudinal data from each individual.

Furthermore, an individual may also communicate in one way in their home or family but completely differently or not at all in a laboratory setting or with examiners[16,17], underscoring the need for *in-situ* environmental contexts, familiar people, and real-world data collection. The sparsity and diversity of vocalizations requires a longitudinal approach to collect a representative sample of vocalizations from each person and necessitates a data processing methodology that accounts for the spontaneity of the vocalizations and the noisy, variable audio environment of real life. Finally, understanding these vocalizations requires familiarity and camaraderie with the communicator. Since most mv* communicators cannot directly provide word-based labels, labels from a person with a long-term relationship with the communicator are the closest obtainable ground truth. Moreover, labels denoted in-the-moment have access to the full multimodal context of the communication exchange, such as body language, gestures, and environment, increasing the fidelity of the labels.

Previously collected available datasets of nonverbal vocalizations have focused on vocalizations that occur amidst typical verbal speech using actors[18,19] or recordings scraped from the web[5,20]. There is also a body of work analyzing infant vocalizations[21–25], though few available datasets exist[5]. Likewise, affective speech datasets have predominantly been collected in lab environments with actors[26,27]. Naturalistic speech datasets have only been collected with typical verbal speech and are often only collected during specified activities[28–30], limiting their ability to capture the breadth of affective expressions that occur across the varied experiences of daily life. To our knowledge, the ReCANVo dataset is the first dataset of affective speech vocalizations collected fully "in the wild," across settings and activities.

This dataset presents over 7000 samples of labeled real-world vocalizations from eight mv* communicators. It is, to our knowledge, the only dataset of nonverbal vocalizations from non-speaking individuals, the largest available dataset of nonverbal vocalizations, and one of the only datasets collected in real-world settings with personalized labels with any population. In addition, basic demographic information and communication profiles are provided for each individual to offer additional insight into how nonverbal vocalizations are used by mv* communicators. Improved understanding of nonverbal vocalizations could contribute to the development of technology to augment communicative interactions[31] and help answer critical questions around the emergence of language and communication across all stages of human development and expression. We hope that the published dataset will engage other researchers in this critical field of study.

Methods

Participants. Participants were recruited through conversations with community members and word of mouth for a larger study examining how mv* individuals use nonverbal vocalizations to communicate and how communication exchanges might be augmented by technology[32]. In the ReCANVo dataset presented here, we included participants who had collected data for at least ten recording sessions to ensure a sufficient number of captured vocalizations across a diversity of settings. These participants ranged in age from 6–23 years old and included diagnoses of autism spectrum disorder (ASD), cerebral palsy (CP), and genetic disorders. They all had fewer than 10 spoken words or word approximations, per parent report (see Table 1). The gender distribution of this sample (6 males, 2 females) reflects the gender distribution among the larger diagnostic categories (e.g., approximately 3.8 males are diagnosed with ASD for every 1 female[33]). No participants were excluded on the basis of age, diagnosis, or other measures in order to capture a broad cross section of this unique and understudied population of communicators.

Importantly, the focus of this initial work and dataset release was on capturing deep, longitudinal, ecologically valid data from a range of participants. This process involved creating new real-world data acquisition methodologies and post-processing signal analysis techniques. Following best practices for novel research with specialized populations[34,35], we utilized a highly iterative and participatory co-design process with a small number of participants[36]. Our dataset includes a variety of different recording settings over time spans of months, along with personalized labels for each participant. Given the limited prior work on real-world vocalizations from minimally speaking individuals, this depth-focused approach was a critical first step towards understanding the heterogeneity of this population, and we look forward to future work expanding our understanding of vocalizations from mv* communicators.

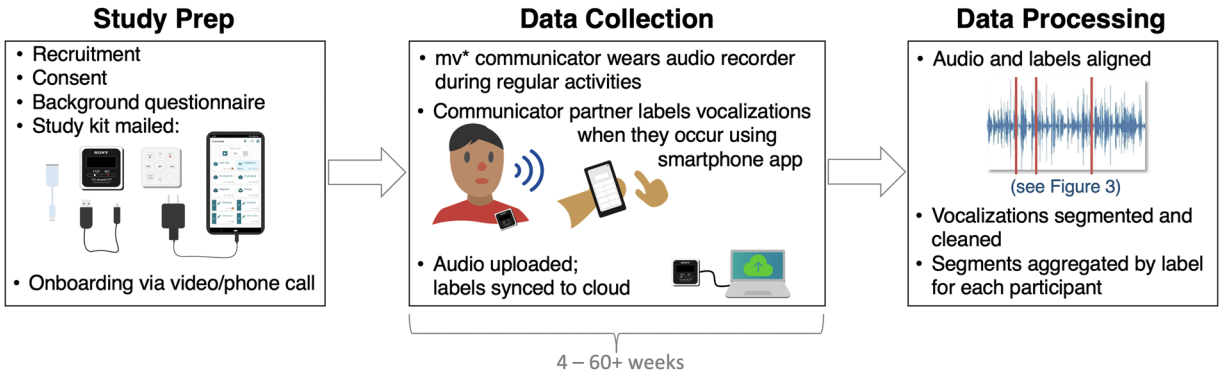

Fig. 1 Schematic for the ReCANVo dataset creation.

Table 1 provides basic demographic information on the eight mv* communicators included in this dataset. Age ranges are provided to bolster anonymity. Figure 1 provides a schematic of the overall study setup and data processing. The study and data collection protocol was approved by the Committee on the Use of Humans as Experimental Subjects (COUHES), the institutional review board (IRB) at the Massachusetts Institute of Technology (MIT). Informed consent or assent was obtained from all participants. A parent or legal guardian provided consent for mv* participants, who were all considered minors for the purposes of consent, and special attention was given to the assent of the non-speaking communicators throughout the study. Families were given flexibility to record audio when it was most convenient for them and could terminate the recording session at any time. They were also given control over when and whether to share the audio data with the researchers and were specifically asked if they wanted to opt in to sharing the de-identified vocalization clips publicly, instead of an opt-out policy. All data clips were manually checked to ensure no identifying information (such as spoken names) remained.

Terminology. In this paper, we use the term *communicator* to refer to the mv* individual who was producing the vocalizations of interest. We use the term *communication partner* or *labeler* to refer to the individual — often a parent or family member — who knew the communicator well and was providing the real-time labels of the vocalizations. The term *participant* refers to both communicators and communication partners as they worked jointly during data collection. Finally, we use the word *communication* to refer to the exchange of information between people. This definition includes, but is not limited to, non-speech vocalizations that convey information to a listener or communication partner. These sounds do not need to be intentional to be informative (similar to the way a shout or cry may convey affect or other information to listeners – whether the sound was produced intentionally or not, and whether it was directed to another person or not).

Real-world audio data. Audio data were collected using a Sony IDC-TX800 wearable audio recorder in 16-bit, 44.1 kHz stereo (see Fig. 1). Magnets were attached to the back of the recorder so that it could be comfortably attached to the communicator's clothing (see previous work for a detailed discussion of this methodology[36]). Some participants had tactile sensitivities that prevented even the lightweight recorder from being attached to the clothing and were instructed to place or hold the recorder near the communicator.

All data were collected and labeled remotely, in the participants' homes and natural environments. This remote administration allowed us to reach a geographically distributed population and produced highly naturalistic data. Data collection kits were mailed to participating families. These kits included the Sony audio recorder, a mobile phone with a custom app for in-the-moment labeling, peripheral cables, and instructions (see Fig. 1). Participants were encouraged to go about their typical daily activities while recording and to label at their convenience to lower the burden of integrating data collection into daily life. This naturalistic data acquisition method resulted in intentionally sparsely labeled recordings.

Real-time labeling app. Vocalizations were labeled in real-time using a custom-built smartphone application (see Fig. 2). The app included 6 labels that were identical for all participants, as well as 4 labels that could be customized by each family from a list of 25 preset options (see Fig. 2b). These labels were selected for this study based on interviews with families of minimally speaking communicators and conversations with speech-language pathologists. They were designed to span a range of common affective states that might be associated with a sound (e.g., Frustration, Delight), as well communicative expressions that many mv* individuals conveyed via vocalizations (e.g., Request). While broad, the category of "social" vocalizations was included because it is important for unfamiliar communication partners to recognize and understand social calls from the communicator even if they were not able to precisely identify a more specific meaning. Descriptions of the 6 pre-determined labels were provided to labelers and are outlined here:

- **Frustration:** Vocalizations that are associated with being frustrated or angry. These vocalizations are typically made in response to a specific situation (e.g., not getting what is wanted).
- **Delight:** Vocalizations that are associated with being excited, very happy, or gleeful.
- **Dysregulation:** Vocalizations that are associated with being irritated, upset, agitated, bored, uncomfortable, understimulated, overstimulated, or generally distressed. These vocalizations may be made involuntarily or

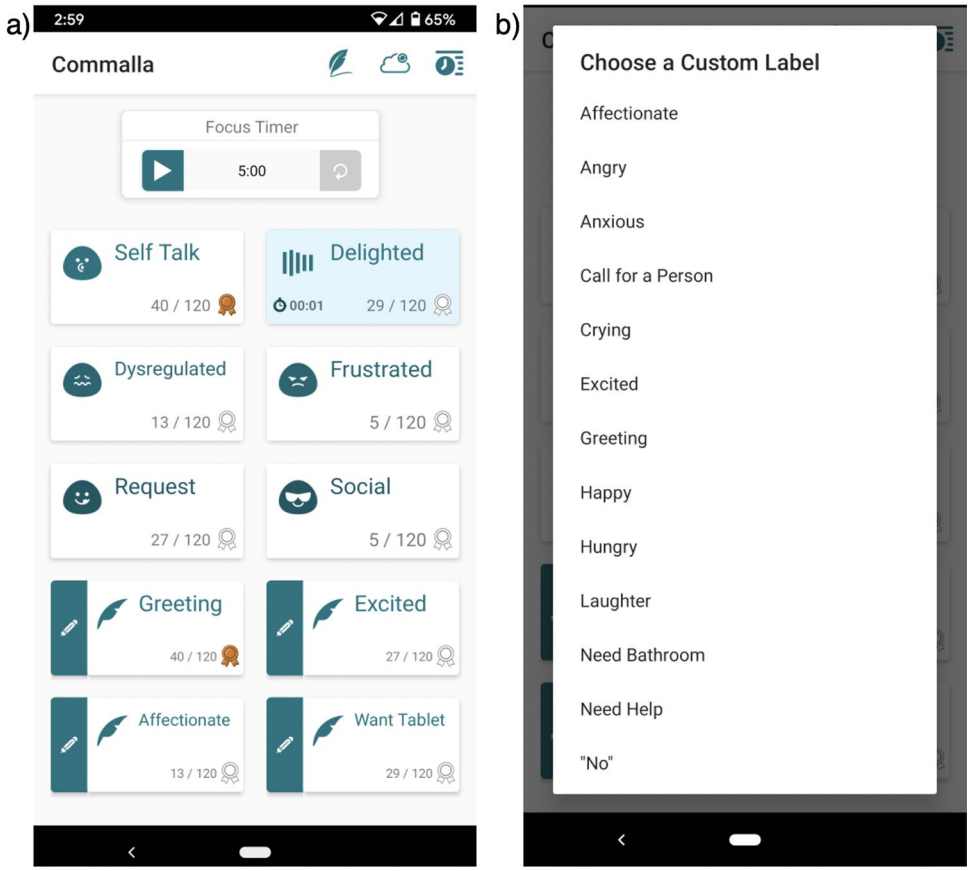

Fig. 2 Custom labeling smartphone application provided to participants. (**a**) Main interface for in-the-moment labeling. Labels were tapped to indicate the start of a vocalization and tapped again to indicate the end of a vocalization. After a label was pressed, an animation appeared on the label (shown on the "Delighted" button) to remind the user which label was active. Labels and audio from a time-synchronized wearable recorder were aligned during post-processing. The six labels at the top of the screen were the same for all participants, while the four labels at the bottom of the screen could be customized for each participant. The "Focus Timer" was provided so participants could keep track of how long they had been labeling. (**b**) Partial list of preset options for the four customizable labels. See Table 1 for the complete list of labels used by participants in this dataset.

without a known communicative function; however, they convey a dysregulated affective state and are well understood by listeners who know the communicator well, making them deeply informative and important to capture.

- **Self-talk:** Vocalizations that are associated with being content, happy, or relaxed and often seem playful or exploratory in nature. These vocalizations generally appear to be made without an overt communicative function (i.e., the individual seems to be making the vocalizations to him/herself). For some individuals (of any age), these vocalizations may sound similar to canonical babbling, singing, or other vocal play heard in young typically-developing children.
- **Request:** Vocalizations that are associated with making a request.
- **Social:** Vocalizations that are social in nature and are not more accurately described by a different term or more specific social term (e.g., "greeting," "call for a specific person").

Label descriptions were shared with families. Particularly, the distinctions between "dysregulation" and "frustration" and between "self-talk" and "delight" were discussed in depth with families. For example, dysregulated vocalizations tend to be more general and less specific to a situation than other negative affective expressions like frustration (e.g., being frustrated that you cannot have a snack versus being dysregulated due to malaise, under/overstimulation, or some broader cause). Likewise, self-talk vocalizations differ from delight vocalizations in both function (e.g., delight vocalizations are more likely to be made in response to pleasurable circumstances to convey delight whereas self-talk vocalizations are generally made to one's self and may have no specific prompt or obvious meaning) and arousal ("very happy or gleeful" versus "context, happy, or relaxed"). These states also often differ in contextual information and other nonverbal cues, such as facial expressions and body language.

During the intake screening, parents indicated the ways in which their mv* communicator consistently used vocalizations. Participants were instructed to use only those labels for which the communicator had a vocalization. Not every communicator produced vocalizations associated with every category of sound. For example,

Vocalization Label	P01	P02	P03	P05	P06	P08	P11	P16
delighted	357	43	25	235	227	39	207	139
dysregulated	212	0	302	116	5	13	22	34
frustrated	150	56	47	283	30	781	27	162
request	130	13	61	6	124	44	22	19
self-talk	564	34	55	286	56	503	33	354
social	182	247	0	0	1	93	52	59
affectionate	0	126	0	0	3	0	0	0
bathroom	20	0	0	0	0	0	0	0
dysregulation-bathroom	18	0	0	0	0	0	0	0
dysregulation-sick	74	0	0	0	0	0	0	0
glee	1	0	7	0	0	0	0	0
greeting	0	0	0	0	0	0	3	0
happy	0	0	0	61	0	0	0	0
help	0	0	0	24	0	0	0	0
hunger	0	0	0	4	0	0	0	0
laughter	0	38	8	13	0	42	0	0
more	0	0	0	0	0	22	0	0
no	0	0	0	0	0	0	0	12
protest	0	0	20	0	0	1	0	0
tablet	0	0	0	0	0	7	0	0
yes	0	0	0	0	123	0	0	0

Table 2. The number of vocalizations included in the dataset, organized alphanumerically by participant and communicative function. The first six vocalization labels were the same across all participants while the rest of the labels were optional semi-customizable labels chosen by each participant.

some communicators did not have "social" vocalizations and others did not produce "dysregulated" vocalizations. Thus, these labels were not used for those individuals.

The app also included the option for 4 semi-personalizable labels. By using a combination of both pre-determined labels that were consistent across all participants and labels that could be customized for each individual, we were able to capture a personalized representative sample of the types of vocalizations that these individuals produced. We were also able to capture vocalization functions that might be uncommon across all participants but very meaningful to that participant, such as "help", "yes", "request tablet", or "hungry". Not all participants used the preset options; they were only used if the family indicated that the communicator had additional specific vocalizations that they wanted to capture. Hence, the these additional labels were specific to each participant (see Table 2 for the labels used per participant). Participants were guided through selecting these semi-personalizable labels from a preset list of words during app setup, after which these labels were not changed. Thus, each participant had a fixed set of 6–10 labels to use throughout their study.

Note that vocalizations produced by mv* individuals might have multiple simultaneous meanings (i.e., a "frustrated request") or ambiguous meanings. Labelers were asked to only label a vocalization if they had high confidence in their interpretation of the function of the sound and to assign the most appropriate label. As a result, there is one intended label per recorded vocalization.

Vocalization labeling procedure. While the communicator was wearing the recorder, the communication partner labeled vocalizations as they were produced. For example, a communicator might request a drink by vocalizing and gesturing toward a cup. The communication partner would then tap the "Request" label on the smartphone labeling app.

Labelers were asked to achieve as close to a 1:1 mapping between a vocalization and a label as possible. However, not all participants followed this instruction closely; some participants designated long periods of time containing multiple vocalizations as a single label. These labeling techniques are further discussed in the preprocessing methods (e.g., Alignment of Audio and Labels) below. The app required labelers to indicate a 'start' and 'end' time for a vocalization by tapping the corresponding label. A color change and animation appeared on a label that had been 'started' to visually indicate which label was active (see Fig. 2a).

Additional study details. At the beginning of the study, each participant had a personal meeting or call with the research team to review the study protocol, ask any questions, and set up the labeling app. Consent and/or assent was acquired from each participant. In addition, participants were provided with multimodal instructions to aid understanding and reliable data acquisition, including a series of video instructions (https://bit.ly/commalla-youtube) as well as a website with written and illustrated instructions. Step-by-step instructions were also included in each mailed data collection kit and provided as a PDF for each participant. Finally, participants were given the researcher's contact information and encouraged to reach out with any questions as the study progressed.

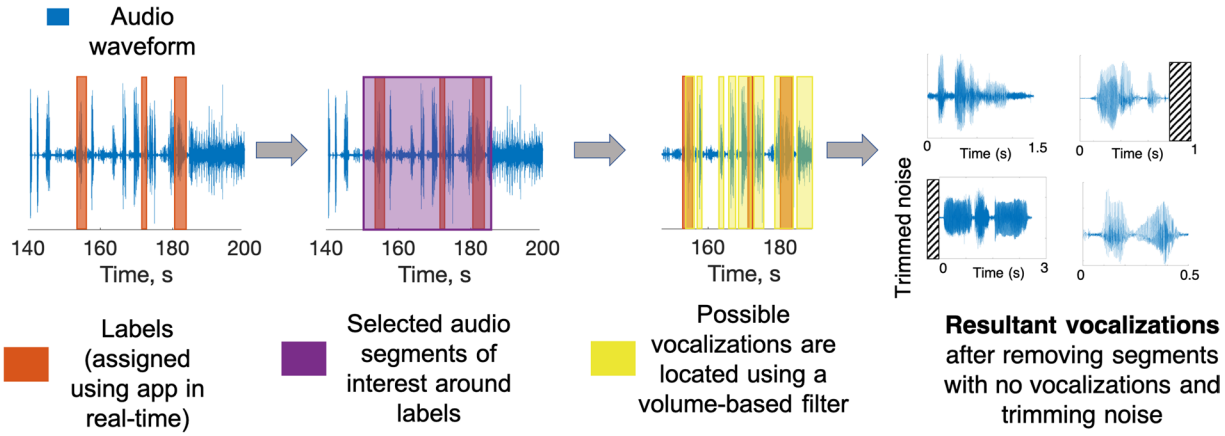

Fig. 3 The audio data and real-time labels from the app were processed post-hoc to align labels with vocalizations. A volume-based filter was used to isolate audio segments of interest. Segments temporally near a label were assigned to that label. A researcher listened to each segment to ensure it contained a vocalization and, if necessary, trimmed excess noise around the vocalization.

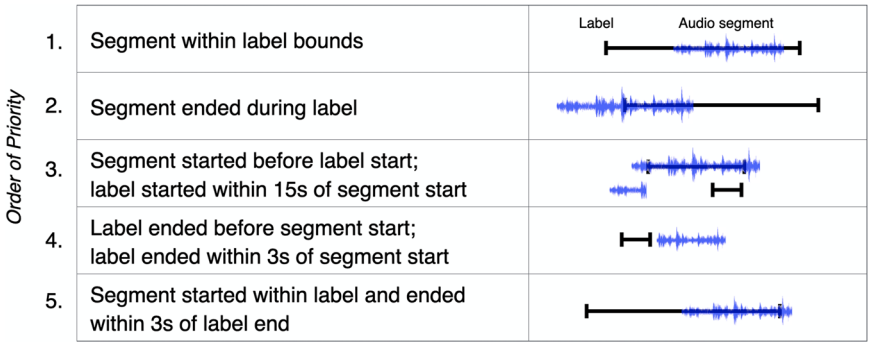

Fig. 4 Illustration of rules for assigning labels to segments. The rule numbers in the figure correspond to the descriptions in the body of the paper.

The study was designed to be flexible and minimize the time and effort burden on participating families. Participants could choose the pace, location, and activities for data collection. While this flexibility resulted in some variability in the collected data between participants, it was critical in enabling this real-world first-of-its-kind data collection with a specialized population.

Alignment of audio and labels. Participants uploaded recorded audio files via a cloud-based file sharing platform. Labels from the app were synced directly to a web server managed by the research team. The clock on the recorder and the app were synced to the same internet-accessible atomic clock (https://time.is) prior to shipping the equipment.

The audio recordings and label information were then processed to isolate vocalizations of interest and align them with the assigned vocalization labels. Because participants were instructed to record and label at their discretion, we first isolated regions of audio that were temporally near labels (see Fig. 3, purple regions). Then, because the recorder was attached to the communicator's clothing or placed nearby, a volume-based filter was used to isolate smaller audio segments within these regions that were likely to be vocalizations (see Fig. 3, yellow regions). The volume filter thresholds were selected for each session based on the recording levels during that session; they ranged between −20 and −45 dB. Vocalizations were considered distinct (separate vocalizations) if they were separated by approximately 250–450 ms of silence, determined heuristically based on the volume levels and background noise of that session's recording. Additional information on alignment and segmentation is detailed in other work[37].

Isolated segments were then assigned a label based on the following rules (see Fig. 4):

1. The audio segment was within the label bounds.
2. The audio segment ended during a label. This timing occurred naturally when a labeler pressed the label after hearing and recognizing a vocalization.
3. The audio segment started before the label started, and the label began within 15 s of the segment start. This threshold was determined after listening to hundreds of raw audio files. It accounts for the human labeling

delay associated with in-the-moment labeling in real-world setting. In many cases, labels could be assigned to vocalizations even with this long delay due to the sparse nature of vocalizations from the mv* population (i.e., no other vocalizations were made during that time).
4. The label ended 3 seconds or less before the segment started. This alignment strategy was primarily necessary for series of vocalizations that occurred amidst multiple identical labels. For example, a communicator might produce four or five frustrated vocalizations sequentially, but owing to human delay and the realities of attending to the communicator's needs during real-world data collection, only a few vocalizations might be labeled. However, the temporal proximity of the labels and the vocalizations still allowed for label assignment.
5. The segment started within a label and ended within 3 s of the label end. Because some labels encompassed multiple vocalizations, some segments began after the label had been pressed. This timing threshold is shorter to account for the possibility that a labeler may have ended a label because the vocalization type had changed and the current label was no longer accurate.

These rules were determined heuristically by comparing the timings of labels to the full-length audio files. The background audio and conversational exchange in the audio files provided context to determine if a label matched a given vocalization. If multiple distinct labels satisfied the rules above, a single label was selected, prioritizing the label with the rule with the lowest number in the list above. Note, by design, that not every vocalization in a recording was assigned a label since participants were instructed to label during free moments while recording. Unlabeled vocalizations were not included in this dataset. Provided labels were comprehensively included in the dataset. On average, each label corresponded to 2–3 final vocalization segments[37].

After labels had been assigned to audio segments, a researcher listened to each labeled audio segment. Segments that did not contain vocalizations were discarded. Segments that contained additional noise or voices before or after a vocalization were manually trimmed. Vocalizations were defined as any clear sound from the communicator that could be associated with a label. This definition encompassed both voiced and non-voiced vocalizations, including word-like approximations with clear vowel-consonant sounds, as well as sounds like grunts, moans, yells, laughter, and breathy vocalizations. Every audio file included in the dataset has been manually confirmed to contain a vocalization. The ReCANVo dataset was intended to be representative of real-world data, so some trimmed vocalizations contain background sounds.

Data Records

All dataset files described below, including raw data files, can be found on Zenodo[38]: https://doi.org/10.5281/zenodo.5786859.

The dataset contains audio recordings of segmented vocalizations, labeled by vocalization meaning or function. The vocalizations are 16-bit, 44.1 kHz .wav files that are organized by assigned label. A .csv file is provided that has the name of each vocalization file and the corresponding participant ID and vocalization label. In addition, communication profiles are provided for each participant in a separate .csv file. This background information was shared by each mv* communicator's parent as part of a study intake questionnaire. The communication profile includes the communication modalities used by the participant (e.g., AAC use, gestures, vocalizations), the number of spoken word and word approximations produced by the communicator, and feedback on if and how the communicator uses vocal sounds across various communicative and affective categories.

The filenames of each audio recording have the following format:

YYMMDD_HHMM_SH_SM_SS.ss – EH_EM_ES.ss

where YYMMDD_HHMM indicates the year (YY), month (MM), day (DD), hour (HH), and minute (MM) of an audio file, respectively. The start and end times of a vocalization *relative to the file start time* are given by SH, SM, SS, ss and EH, EM, ES, ss, indicating the vocalization start or end hour, minute, second, and sub-second, respectively. These times are included in the filename to provide additional information regarding the longitudinal nature of the dataset. Users of the dataset should note that these times are approximate and were determined using the segmentation and cleaning process described above. For P01 specifically, the start and end times of the vocalizations were estimated post-hoc using an autocorrelation and have known errors.

The ReCANVo dataset includes 7,077 vocalizations collected longitudinally with 8 mv* communicators. Table 2 shows the number of vocalizations in the dataset for each participant and vocalization type. To our knowledge, the ReCANVo dataset is the first dataset of nonverbal communication that occurs independent of typical verbal speech, the largest existing dataset of nonverbal vocalizations, and the first public dataset of affective speech collected longitudinally during day-to-day life across settings.

Technical Validation

We identified three possible sources of labeling error:

1. Accidental labels (e.g., a labeler accidentally tapping the wrong label on the app)
2. Inaccurate vocalization-label alignment (e.g., labels being incorrectly matched with a vocalization audio during post-processing)
3. Inaccurate interpretation of a vocalization by a communication partner.

To mitigate the first two sources of error, a researcher listened to the audio surrounding each labeled vocalization. A researcher also listened to full-length audio recordings for each participant at least every two weeks of collected data. The surrounding context from the audio recording, such as spoken dialogue that confirmed an

emotional label (E.g., "I know you want to go outside and we can't. That's frustrating.") or answered a communicator's request (E.g., "Is this the snack you wanted?"), was used to confirm that the assigned labels matched the audio context near the label. Because of the longitudinal nature of the study, some clock drift (~10 seconds or less) was observed for some participants. This drift was manually determined and accounted for when aligning the labels with vocalizations.

To mitigate the third source of error (i.e., incorrect interpretation by the communication partner), only communication partners who were deeply familiar with the mv* communicator and their communication style provided labels. In addition, partners were instructed to only label vocalizations that they felt like they could confidently interpret. However, any interpretation of a vocalization remains, at best, an interpretation. We hope that as additional knowledge and communication technology for mv* communicators becomes available, it will be possible to obtain ground truth meaning of these vocalizations directly from communicators.

In addition, there were expected sources of noise associated with real-world data, including environmental noise (e.g., wind, movement, background toys and electronics), overlapping voices, and intensity changes due to variable location of the recorder. Many extraneous sources of noise were removed during the segmentation process or through manual trimming; however, vocalization segments of all qualities were included here to ensure naturalistic, real-world data transfer.

Code availability

We used the Python programming language for the data processing described above. Volume segmentation was implemented using the *pydub* libary. The label assignment algorithm is summarized in Fig. 3. The code is available as part of our dataset in Zenodo: https://doi.org/10.5281/zenodo.5786859.

Received: 14 February 2022; Accepted: 24 July 2023;
Published online: 05 August 2023

References

1. Sauter, D. A., Eisner, F., Calder, A. J. & Scott, S. K. Perceptual cues in nonverbal vocal expressions of emotion. *The Quarterly Journal of Experimental Psychology* **63**, 2251–2272, https://doi.org/10.1080/17470211003721642 (2010).
2. Liu, L., Li, W., Wu, X. & Zhou, B. X. Infant cry language analysis and recognition: an experimental approach. *IEEE/CAA Journal of Automatica Sinica* **6**, 778–788, https://doi.org/10.1109/JAS.2019.1911435 (2019).
3. Oller, D. K. *et al*. Automated vocal analysis of naturalistic recordings from children with autism, language delay, and typical development. *Proceedings of the National Academy of Sciences* **107**, 13354–13359, https://doi.org/10.1073/pnas.1003882107 (2010).
4. Anikin, A. A moan of pleasure should be breathy: the effect of voice quality on the meaning of human nonverbal vocalizations. *Phonetica* **77**, 327–349, https://doi.org/10.1159/000504855 (2020).
5. Parsons, C. E., Young, K. S., Craske, M. G., Stein, A. L. & Kringelbach, M. L. Introducing the oxford vocal (OxVoc) sounds database: A validated set of non-acted affective sounds from human infants, adults, and domestic animals. *Frontiers in Psychology* **5**, 562, https://doi.org/10.3389/fpsyg.2014.00562 (2014).
6. Rose, V., Trembath, D., Keen, D. & Paynter, J. The proportion of minimally verbal children with autism spectrum disorder in a community-based early intervention programme. *Journal of Intellectual Disability Research* **60**, 464–477, https://doi.org/10.1111/jir.12284 (2016).
7. Anderson, D. K. *et al*. Patterns of growth in verbal abilities among children with autism spectrum disorder. *Journal of consulting and clinical psychology* **75**, 594, https://doi.org/10.1037/0022-006X.75.4.594 (2007).
8. Tager-Flusberg, H. & Kasari, C. Minimally verbal school-aged children with autism spectrum disorder: The neglected end of the spectrum. *Autism research* **6**, 468–478, https://doi.org/10.1002/aur.1329 (2013).
9. Chiarotti, F. & Venerosi, A. Epidemiology of autism spectrum disorders: a review of worldwide prevalence estimates since 2014. *Brain sciences* **10**, 274, https://doi.org/10.3390/brainsci10050274 (2020).
10. Hoffman, K. *et al*. Geographic patterns of autism spectrum disorder among children of participants in nurses' health study ii. *American journal of epidemiology* **186**, 834–842, https://doi.org/10.1093/aje/kwx158 (2017).
11. Hayes, S. A. & Watson, S. L. The impact of parenting stress: A meta-analysis of studies comparing the experience of parenting stress in parents of children with and without autism spectrum disorder. *Journal of autism and developmental disorders* **43**, 629–642, https://doi.org/10.1007/s10803-012-1604-y (2013).
12. Kogan, M. D. *et al*. A national profile of the health care experiences and family impact of autism spectrum disorder among children in the united states, 2005–2006. *Pediatrics* **122**, e1149–e1158, https://doi.org/10.1542/peds.2008-1057 (2008).
13. Geschwind, D. H. & Levitt, P. Autism spectrum disorders: developmental disconnection syndromes. *Current opinion in neurobiology* **17**, 103–111, https://doi.org/10.1016/j.conb.2007.01.009 (2007).
14. Geschwind, D. H. & State, M. W. Gene hunting in autism spectrum disorder: on the path to precision medicine. *The Lancet Neurology* **14**, 1109–1120, https://doi.org/10.1016/S1474-4422(15)00044-7 (2015).
15. Masi, A., DeMayo, M. M., Glozier, N. & Guastella, A. J. An overview of autism spectrum disorder, heterogeneity and treatment options. *Neuroscience bulletin* **33**, 183–193, https://doi.org/10.1007/s12264-017-0100-y (2017).
16. Baroková, M. D., Hassan, S., Lee, C., Xu, M. & Tager-Flusberg, H. A comparison of natural language samples collected from minimally and low-verbal children and adolescents with autism by parents and examiners. *Journal of Speech, Language, and Hearing Research* **63**, 4018–4028, https://doi.org/10.1044/2020_JSLHR-20-00343 (2020).
17. Bernard-Opitz, V. Pragmatic analysis of the communicative behavior of an autistic child. *Journal of Speech and Hearing Disorders* **47**, 99–109, https://doi.org/10.1044/jshd.4701.99 (1982).
18. Lima, C. F., Castro, S. L. & Scott, S. K. When voices get emotional: A corpus of nonverbal vocalizations for research on emotion processing. *Behavior Research Methods* **45**, 1234–1245, https://doi.org/10.3758/s13428-013-0324-3 (2013).
19. Holz, N., Larrouy-Maestri, P. & Poeppel, D. The paradoxical role of emotional intensity in the perception of vocal affect. *Scientific Reports* **11**, 1–10, https://doi.org/10.1038/s41598-021-88431-0 (2021).
20. Anikin, A. & Persson, T. Nonlinguistic vocalizations from online amateur videos for emotion research: A validated corpus. *Behavior research methods* **49**, 758–771, https://doi.org/10.3758/s13428-016-0736-y (2017).
21. Harding, C. G. & Golinkoff, R. M. The origins of intentional vocalizations in prelinguistic infants. *Child development* 33–40, https://www.jstor.org/stable/1129038 (1979).
22. Oller, D. K. *The emergence of the speech capacity* (Psychology Press, 2000).
23. Oller, D. K. *et al*. Preterm and full term infant vocalization and the origin of language. *Scientific Reports* **9**, 1–10, https://doi.org/10.1038/s41598-019-51352-0 (2019).
24. Nathani, S., Ertmer, D. J. & Stark, R. E. Assessing vocal development in infants and toddlers. *Clinical linguistics & phonetics* **20**, 351–369, https://doi.org/10.1080/02699200500211451 (2006).

25. Jhang, Y. & Oller, D. K. Emergence of functional flexibility in infant vocalizations of the first 3 months. *Frontiers in Psychology* **8**, 300, https://doi.org/10.3389/fpsyg.2017.00300 (2017).
26. Busso, C. *et al*. IEMOCAP: interactive emotional dyadic motion capture database. *Language Resources and Evaluation* **42**, 335–359, https://doi.org/10.1007/s10579-008-9076-6 (2008).
27. Livingstone, S. R. & Russo, F. A. The ryerson audio-visual database of emotional speech and song (RAVDESS): A dynamic, multimodal set of facial and vocal expressions in north american english. *PloS One* **13**, e0196391, https://doi.org/10.1371/journal.pone.0196391 (2018).
28. Ringeval, F., Sonderegger, A., Sauer, J. & Lalanne, D. Introducing the RECOLA multimodal corpus of remote collaborative and affective interactions. In *IEEE International Conference on Automatic Face and Gesture Recognition*, 1–8, https://doi.org/10.1109/FG.2013.6553805 (IEEE, 2013).
29. Lotfian, R. & Busso, C. Building naturalistic emotionally balanced speech corpus by retrieving emotional speech from existing podcast recordings. *IEEE Transactions on Affective Computing* **10**, 471–483, https://doi.org/10.1109/TAFFC.2017.2736999 (2019).
30. Canavan, A., Graff, D. & Zipperlen, G. *CALLHOME American English Speech LDC97S42* https://doi.org/10.35111/exq3-x930 (1997).
31. Narain, J. *et al*. Personalized modeling of real-world vocalizations from nonverbal individuals. In *Proceedings of the 2020 International Conference on Multimodal Interaction*, 665–669, https://doi.org/10.1145/3382507.3418854 (2020).
32. Narain, J. *et al*. Nonverbal vocalizations as speech: Characterizing natural-environment audio from nonverbal individuals with autism. In *Proceedings of Laughter and Other Non-Verbal Vocalisations Workshop*, https://doi.org/10.4119/lw2020-923 (2020).
33. Maenner, M. J. *et al*. Prevalence and characteristics of autism spectrum disorder among children aged 8 years–autism and developmental disabilities monitoring network, 11 sites, united states, 2020. *MMWR. Surveillance Summaries* **72** (2023).
34. Biller, M. F. & Johnson, C. J. Examining useful spoken language in a minimally verbal child with autism spectrum disorder: a descriptive clinical single-case study. *American journal of speech-language pathology* **29**, 1361–1375 (2020).
35. Wilson, C., Brereton, M., Ploderer, B. & Sitbon, L. Co-design beyond words: 'moments of interaction'with minimally-verbal children on the autism spectrum. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, 1–15 (2019).
36. Johnson, K. T., Narain, J., Ferguson, C., Picard, R. & Maes, P. The ECHOS platform to enhance communication for nonverbal children with autism: A case study. In *Extended Abstracts of the 2020 CHI Conference on Human Factors in Computing Systems*, 1–8, https://doi.org/10.1145/3334480.3375206 (2020).
37. Narain, J. Interfaces and models for improved understanding of real-world communicative and affective nonverbal vocalizations by minimally speaking individuals. *Massachusetts Institute of Technology* https://dspace.mit.edu/bitstream/handle/1721.1/140101/narain-jnarain-phd-meche-2021-thesis.pdf (2021).
38. Narain, J., Johnson, K., Quatieri, T., Picard, R. & Maes, P. ReCANVo: A dataset of real-world communicative and affective nonverbal vocalizations, *Zenodo*, https://doi.org/10.5281/zenodo.5786859 (2021).

Acknowledgements

Funding for this research was provided by the MIT Media Lab Consortium and the MIT Deshpande Center for Technological Innovation. In addition, K.T.J. was funded by the MIT Hugh Hampton Young Memorial Fellowship, and J.N. was funded by Apple Scholars in AI/ML and the NSF Graduate Research Fellowship program. For T.Q., this material is based upon work supported by the Under Secretary of Defense for Research and Engineering under Air Force Contract No. FA8702-15-D-0001. Any opinions, findings, conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the Under Secretary of Defense for Research and Engineering. The authors thank Craig Ferguson for helping develop the custom labeling app, Ayelet Kershenbaum and Amanda O'Brien for their feedback and insights, and Michelle Luo and Yuji Chan for contributing to the design of instructional materials. They also thank the participants of this study for their time, effort, feedback, and data contributions. Approved for public release. Distribution is unlimited.

Author contributions

K.T.J. and J.N. conceived the study, conducted the experiments and data collection, analyzed the results, and wrote the manuscript. J.N. designed the processing algorithms, and aligned and segmented the vocalizations. K.T.J. and T.Q. were participants in the study. T.Q., R.P. and P.M. provided technical mentorship. All authors reviewed the manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Correspondence and requests for materials should be addressed to K.T.J. or J.N.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.