Nonverbal Vocalizations as Speech: Characterizing Natural-Environment Audio from Nonverbal Individuals with Autism

Jaya Narain*,[a], Kristina T. Johnson*,[a], Amanda O'Brien[a], Peter Wofford[a], Pattie Maes[a], Rosalind Picard[a] (*Equal Contribution)

[a]Massachusetts Institute of Technology, Cambridge, Massachusetts, United States
{jnarain, ktj, amobrien, peterwof}@mit.edu; {pattie, picard}@media.mit.edu

Abstract

The study of nonverbal vocalizations, such as sighs, grunts, and monosyllabic sounds, has largely revolved around the social and affective implications of these sounds within typical speech. However, for individuals who do not use any traditional speech, including those with non- or minimally verbal (nv/mv) autism, these vocalizations contain important, individual-specific affective and communicative information. This paper outlines the methodology, analysis, and technology to investigate the production, perception, and meaning of nonverbal vocalizations from nv/mv individuals in natural environments. We are developing novel signal processing and machine learning methods that will help enable augmentative communication technology, and we are producing a nonverbal vocalization dataset for public release. We hope this work will expand the scientific understanding of these exceptional individuals' language development and the field of communication more generally.

1 Introduction

In the United States alone, over one million people are non- or minimally verbal (nv/mv), meaning they use zero or fewer than 20 words/word approximations, respectively (Tager-Flusberg and Kasari 2013). This category includes, but is not limited to, people with autism spectrum disorder (ASD) and/or certain genetic disorders. These individuals often experience stress, frustration, and isolation when communicating in a society largely constructed around typical verbal speech. Through non-speech vocalizations, nv/mv individuals organically express rich affective and communicative information for a range of functions (e.g., protesting, requesting, commenting). These vocalizations are highly idiosyncratic to the speaker, which makes interpretation of these utterances challenging for unfamiliar listeners. Despite the prevalence of these vocalizations among people who are nv/mv, and the importance of interpreting these vocalizations to reduce frustration and support communication for these individuals, the production, meaning, and usage of these sounds has, to our knowledge, not been systematically investigated.

Here we present a study of nonverbal vocalizations that occur independent of traditional speech. We summarize the meanings of vocalizations reported by families of nv/mv individuals, detail our methodology for collecting spontaneous natural vocalizations labeled in real-time by close family members, and present audio samples. We are currently developing machine learning methods to classify subsets of vocalizations, both for the general advancement of nonverbal vocalization characterization from nonverbal individuals and for use in an augmentative communication system that enhances interaction and dialogue between individuals without traditional speech and the wider community (see Figure 1). The labeled vocalization segments will be released in the first-ever database of vocalizations from nv/mv individuals.

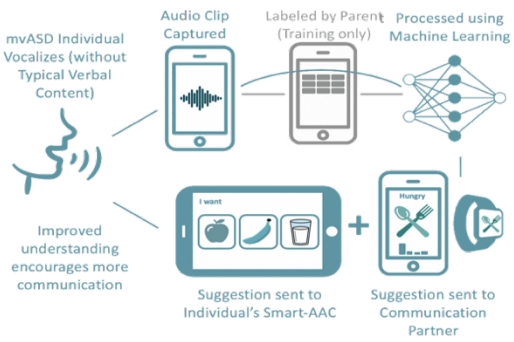

Figure 1: Vision for a nonverbal vocalization-based augmentative communication system.

2 Related Works

Non-speech vocalizations include both involuntary (e.g., coughing, hiccupping) and voluntary (e.g., laughing, sighing, screaming) sounds. Within the context of traditional speech, many people use nonverbal vocalizations to augment speech, including to convey an emotion, express an intention, or emphasize verbal speech (Knapp 1980; Poggi, Ansani, and Cecconi 2018; Sauter et al. 2010). For people who are nv/mv, however, these non-speech vocalizations serve a larger linguistic and communicative purpose by serving as a substitute for speech (Beukelman and Mirenda 2013). Here we study nonverbal vocalizations in populations that use minimal or no typical speech. These vocalizations include traditional nonverbal cues (e.g., laughter or yells), as well as unique utterances of varying pitch, phonetic content, and tone that do not fall into the usual categories of nonverbal vocalizations.

Prior work in classifying vocalizations that are not accompanied by typical speech has focused on classifying typically developing infant cries by need (e.g., hunger, pain) using both humans and machines (Liu et al. 2019; Wasz-Höckert et al. 1964). There has been extensive prior work on affect detection in speech with typical verbal content (Schuller 2018) including in ASD populations with verbal abilities in task-driven (e.g., Baird et al. 2017) and natural settings (e.g., Ringeval et al. 2016). However, no known work to date has attempted to classify communicative content in vocalizations from non-infant children or adults who are nv/mv.

In addition, nv/mv children are primarily tested and characterized in laboratory and clinical settings. Tools that track and enhance communication with nv/mv individuals in naturalistic settings, which may provide communicative not captured by laboratory tests (Wilhelm and Grossman 2010), are unexplored. Our approach is novel in that it characterizes vocalizations that are not accompanied by typical speech in natural settings using personalized labels provided in situ from individuals who know the speaker well.

3 Methods to Characterize Nonverbal Vocalizations from nv/mv Individuals

The methods for this work comprise two major phases: (1) Categorization of nonverbal vocalizations through interviews, and (2) acquisition and analysis of natural-environment audio recordings with real-time labels.

First, we conducted conversational interviews with three pilot families who have a nv/mv family member. The nv/mv individuals were all male and were 8 (P0), 16 (P1), and 23 (P3) years old. They all had a diagnosis of ASD, and P0 also had a rare genetic disorder. P0 and P1 were nonverbal, while P2 had four word approximations (“boi,” “eri(c),” “hi,” and “five”)

Families were asked to describe how the nv/mv individual generally communicates, including sounds, gestures, physical communication (e.g., hand-leading), assistive augmentative communication (AAC) modalities, and any other commonly used techniques. They were then specifically asked to describe the ways in which the nv/mv individual used vocalizations.

Second, in order to capture and systematically investigate the linguistic patterns of nonverbal vocalizations produced organically throughout daily life, we recorded audio in the individual’s natural environment (e.g., home, playground) using a small, wireless recorder (Sony IDC-TX800) over the course of several weeks. Caregivers attached the recorder to the individual’s shirt near the shoulder using strong magnets. The 16-bit, 44.1 kHz stereo recordings were then transferred at the participants’ discretion to IRB-approved researchers using a cloud-based service.

In order to identify in situ, real-time affect and communicative cues from the participants, we built an open-source, easy-to-use Android app to collect labels from a primary caregiver (see Figure 2). Caregivers were instructed to select labels corresponding to four affective states (self-talk, dysregulation, frustration, delight) and two

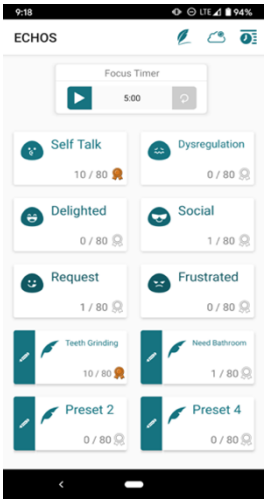

Figure 2: Open-source Android app enables in-the-moment labeling of affective and communicative sounds. These timestamped labels are synced to a server and combined with the audio recordings for further analysis.

Vocalization Category	Vocalization Interpretation	P0	P1	P2
Self-talk	Self-talk	X	X	X
Dysregulation	Dysregulation	X		X
Frustration	Frustration	X	X	X
	Impatient		X	
	Protest	X		
	Upset		X	
Delight	Glee	X		X
	Happy		X	
	Affection		X	
	Excited		X	
Request	General request	X		X
	Request for a person		X	X
	Request for a high five			X
	Request for the bathroom		X	
Social exchange	Social (general)	X		X
	"Fun participation"		X	
	Greeting			X
Other	Teeth grinding	X	X	
	Singing		X	
	Laughing	X	X	

Table 1: Family-reported uses of nonverbal vocalizations by three nv/mv individuals. The reported presence of a given vocalization type in an individual's non-speech repertoire is indicated by an "X."

communicative/interaction states (request, social exchange) whenever their child made a vocalization corresponding to one of these states.

Affective state labels were associated with general states-of-being. For example, "dysregulation" was used to describe a state of passive over- or under stimulation that corresponded with generalized negative affect. In contrast, interaction states were related to informational exchanges that were often in anticipation of a tangible response from the communication partner (e.g., requesting a high five). These labels were selected based on the previously described caregiver interviews and were designed to capture an array of frequent emotional and communicative vocalizations occur often but may be ambiguous to those who do not know the individual well. Caregivers could also set four customizable label options on the app to further personalize the system. We collected more than 300 labeled vocalizations from the three participants using this protocol.

4 Results and Discussion

Table 1 summarizes the reported intentions of nonverbal vocalizations used by the nv/mv individual from the three interviewed families. The vocalizations tended to vary in pitch, duration, and

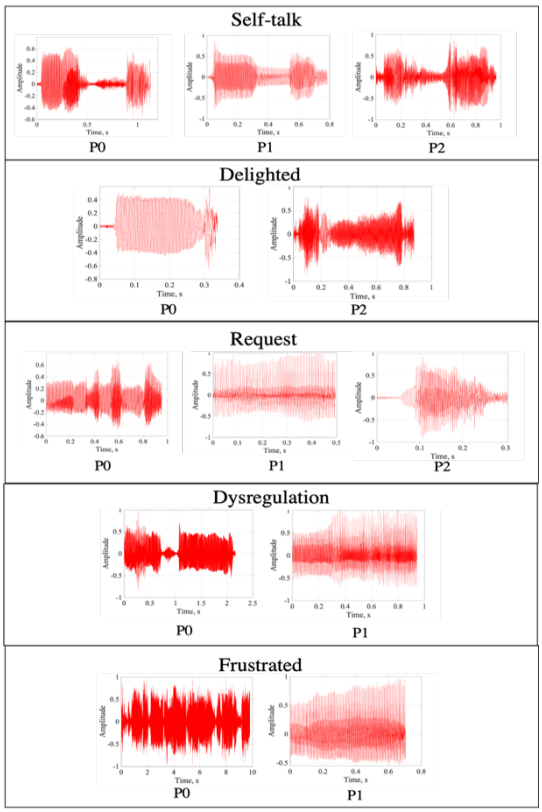

Figure 3: Representative time-domain waveforms of vocalizations in five labeled categories from three participants (P0, P1, P2).

other audible characteristics depending on the individual's emotion, physical state, and/or intent. Families reported a range of nonverbal vocalizations including word approximations used for more general purposes (e.g., "bɔɪ" for a social interaction), consonant-vowel syllables (e.g., "ma" used for multiple functions), vocalizations without phonetic cues (e.g., laughter), and sounds that were often associated with an affective state (e.g., teeth grinding related to dysregulation).

Figure 3 shows canonical waveforms for different nonverbal expressions from the participants. Because the in-situ labeling and recording procedure results in a sparsely labeled dataset, we used volume-based segmentation to find audio events corresponding to given labels. The process accounts for the human delay in labeling. In previous work, we explored whether generic databases could be used to supplement our smaller specialized dataset and minimize the required number of caregiver labels. We used a zero-shot transfer learning approach with LSTM models to classify three types of nonverbal vocalizations from one participant. The input to the LSTM model was a VGGish embedding of an audio waveform. The complete details of this approach are available in our previous work

(Narain/Johnson et al. 2019). While this approach exhibited some success (~70% accuracy) for more common nonverbal sounds like laughter and frustration, it was less successful (51% accuracy) for idiosyncratic vocalizations like self-talk.

5 Future Work

We are currently expanding our data collection to include more participants, with a cumulative target of over 24,000 labeled vocalizations covering six affective or communicative states from approximately 40 participants. In addition to releasing this dataset, we will examine the vocal characteristics of nonverbal communicative and emotional expressions within and across nv/mv individuals. Given the heterogeneity of this population, we hope to assess the universal and/or idiosyncratic nature of nonverbal vocalizations independent of typical speech. The dataset will also be used to probe scientific questions related to language and development across ages, genders, and diagnoses, such as the development of specific phonemes from nonverbal vocalizations.

Collecting the audio in natural environments and labeling in real time produces data that most closely represent real life and benefits both the theoretical understanding of these sounds and the development of ecologically valid machine learning models. Yet, this approach also results in sparse, noisy data that are imprecisely aligned with labels. In order to improve the vocalization detection and labeling, future work will leverage unlabeled segments via semi-supervised learning, simulate vocalizations to use as additional training data, and train models across individuals with similar characteristics. We will also cluster various vocalization types within a function to expand understanding of these sounds and improve algorithmic classification methods. We hope that the novel methodology and dedicated demographic focus presented here will expand the understanding of nonverbal vocalizations for individuals without typical speech.

Acknowledgments

The authors would like to thank the study participants, the MIT Media Lab Consortium, and the Microsoft AI for Accessibility program.

References

Baird, Alice, Shahin Amiriparian, Nicholas Cummins, Alyssa M. Alcorn, Anton Batliner, Sergey Pugachevskiy, Michael Freitag, Maurice Gerczuk, and Björn Schuller. 2017. “Automatic Classification of Autistic Child Vocalisations: A Novel Database and Results.”

Beukelman, David R. and Pat Mirenda. 2013. *Augmentative and Alternative Communication: Supporting Children and Adults with Complex Communication Needs*. 4th ed. Baltimore.

Knapp, Mark L. 1980. *Essentials of Nonverbal Communication*. Holt, Rinehart and Winston.

Liu, Lichuan, Wei Li, Xianwen Wu, and Benjamin X. Zhou. 2019. “Infant Cry Language Analysis and Recognition: An Experimental Approach.” *IEEE/CAA Journal of Automatica Sinica*.

Narain*, Jaya, Kristina T. Johnson*, Rosalind Picard, and Pattie Maes. 2019. “Zero-Shot Transfer Learning to Enhance Communication for Minimally Verbal Individuals with Autism Using Naturalistic Data.” (*Equal Contribution) in *NeurIPS 2019 Joint Workshop on AI for Social Good*. Vancouver.

Poggi, Isabella, Alessandro Ansani, and Christian Cecconi. 2018. “Sighs in Everyday and Political Communication.” in *Laughter Workshop*. Paris.

Ringeval, Fabien, Erik Marchi, Charline Grossard, Jean Xavier, Mohamed Chetouani, David Cohen, and Björn Schuller. 2016. “Automatic Analysis of Typical and Atypical Encoding of Spontaneous Emotion in the Voice of Children.” in *INTERSPEECH*. San Francisco.

Sauter, Disa A., Frank Eisner, Andrew J. Calder, and Sophie K. Scott. 2010. “Perceptual Cues in Nonverbal Vocal Expressions of Emotion.” *Quarterly Journal of Experimental Psychology* 63(11):2251–72.

Schuller, Björn W. 2018. “Speech Emotion Recognition: Two Decades in a Nutshell, Benchmarks, and Ongoing Trends.” *Communications of the ACM* 61(5):90–99.

Tager-Flusberg, Helen and Connie Kasari. 2013. “Minimally Verbal School-Aged Children with Autism Spectrum Disorder: The Neglected End of the Spectrum.” *Autism Research* 6(6):468–78.

Wasz-Höckert, O., T. J. Partanen, V. Vuorenkoski, K. Michelsson, and E. Valanne. 1964. “The Identification of Some Specific Meanings in Infant Vocalization.” *Experientia* 20(3):154.

Wilhelm, Frank H. and Paul Grossman. 2010. “Emotions beyond the Laboratory: Theoretical Fundaments, Study Design, and Analytic Strategies for Advanced Ambulatory Assessment.” *Biological Psychology* 84(3):552–69.