Modeling Real-World Affective and Communicative Nonverbal Vocalizations From Minimally Speaking Individuals

Jaya Narain, Kristina T. Johnson, Thomas F. Quatieri, *Fellow, IEEE*, Rosalind W. Picard, *Fellow, IEEE*, and Pattie Maes, *Member, IEEE*

Abstract—Nonverbal vocalizations from non- and minimally speaking individuals who speak fewer than 20 words (mv* individuals) convey important communicative and affective information. While nonverbal vocalizations that occur amidst typical speech and infant vocalizations have been studied extensively in the literature, there is limited prior work on vocalizations by mv* individuals. Our work is among the first studies of the communicative and affective information expressed in nonverbal vocalizations by mv* children and adults. We collected labeled vocalizations in real-world settings with eight mv* communicators, with communicative and affective labels provided in-the-moment by a close family member. Using evaluation strategies suitable for messy, real-world data, we show that nonverbal vocalizations can be classified by function (with 4- and 5-way classifications) with F1 scores above chance for all participants. We analyze labeling and data collection practices for each participating family, and discuss the classification results in the context of our novel real-world data collection protocol. The presented work includes results from the largest classification experiments with nonverbal vocalizations from mv* communicators to date.

Index Terms—Affective computing, affect sensing and analysis, nonverbal speech, speech analysis

1 Introduction

In the United States alone, there are over one million people who are non- or minimally speaking with respect to verbal language [1], [2], [3]. Here we focus on a subset of this population, abbreviated as *mv**, who have fewer than 20 words or word approximations and limited expressive language through speech and writing. This includes individuals with Autism, in addition to some individuals with Down Syndrome, Rett Syndrome, Mowat-Wilson, Rubinstein-Taybi syndrome, Pitt-Hopkins syndrome, and other conditions associated with differences in speech and language. Mv* individuals communicate richly through many means including augmentative and alternative communication (AAC) devices, gestures, and vocalizations.

Family members and those close to mv* individuals report that nonverbal vocalizations (i.e., vocalizations that do not have typical verbal content) from mv* individuals often have self-consistent phonetic content and may vary in tone, pitch, and duration depending on the individual's emotional state or intended communication. While these vocalizations contain important affective and communicative information and are understood by close family and friends, they are often poorly understood by those who don't know the communicator well. Improved understanding of nonverbal vocalizations could contribute to the development of technology to augment communication [4], enhance understanding of nonverbal affective expressions broadly, and expand awareness around this form of communication.

Studying nonverbal vocalizations with mv* individuals has unique challenges. Mv* individuals are a small, heterogeneous, and geographically distributed population. The population of mv* communicators includes individuals with diverse and multiple diagnoses; many individuals also have co-occurring intellectual disabilities and other challenges like epilepsy. Moreover, studying vocalizations with this population requires a flexible and thoughtful study design to minimize time burden on families.

Additionally, affective and communicative vocalizations are motivation-driven and cannot be easily elicited in

- *Jaya Narain, Kristina T. Johnson, Rosalind W. Picard, and Pattie Maes are with the Massachusetts Institute of Technology, Cambridge, MA 02139 USA. E-mail: jnarain8@gmail.com, ktj@mit.edu, {picard, pattie}@media.mit.edu.*
- *Thomas F. Quatieri is with MIT Lincoln Laboratory, Lexington, MA 02421 USA. E-mail: quatieri@ll.mit.edu.*

Manuscript received 6 September 2021; revised 22 June 2022; accepted 4 September 2022. Date of publication 21 September 2022; date of current version 15 November 2022.
Approved for public release. Distribution is unlimited. 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.
This work was supported in part by the MIT Media Lab Consortium and the Deshpande Center Technology to Improve Ability Program. The work of Jaya Narain was supported by Apple Scholars in AI/ML and the NSF Graduate Research Fellowship Program. The work of Kristina Johnson was supported by the Hugh Hampton Young Fellowship Program.
This work involved human subjects in its research. Approval of all ethical and experimental procedures and protocols was granted by MIT Committee on the Use of Humans as Experimental Subjects Application No. 1903760614.
(Corresponding author: Jaya Narain.)
Recommended for acceptance by J. Epps.
This article has supplementary downloadable material available at https://doi.org/10.1109/TAFFC.2022.3208233, provided by the authors.
Digital Object Identifier no. 10.1109/TAFFC.2022.3208233

unnatural lab settings. Real-world data collection is critical to capturing vocalizations as they are used organically to express affect and communication. Naturalistic studies often recreate real-life environments and activities in laboratories, but still involve unfamiliar settings and people which might induce anxiety and alter emotional expressions, particularly for mv* individuals who can be sensitive to new sensory environments and experiences [5]. Obtaining ground-truth labels is also an unsolved problem. Many mv* communicators cannot directly provide word-based labels, and external annotators do not have the deep firsthand experience needed to interpret these vocalizations. In our prior work [4], we developed a novel longitudinal data collection process to collect real-world audio with in-the-moment labels provided by a close family member or caregiver. Here we extend our previous work, presenting new analytical approaches applied to a larger number of vocalizations from more individuals.

In this paper, we present the results of the largest real-world nonverbal vocalization classification experiments to date with vocalizations by eight mv* communicators. We show that nonverbal vocalizations can be classified using audio alone for each individual. We present evaluation and sampling strategies to work with messy, real-world data with uneven sample distributions and varying background noise. We implement and evaluate a custom feature set designed for nonverbal vocalizations for mv* individuals. We also analyze the data collection and labeling practices for each participant, and discuss model performance in the context of how data was collected by each participant.

2 Related Work

Prior studies have not focused on mv* individuals specifically, but have studied neurodiverse individuals and/or individuals with developmental differences more generally. Our work is unique in its focus on mv* individuals.

2.1 Affect With Neurodiverse Populations

Prior work on affect recognition with neurodiverse populations has focused on the emotional content of verbal speech [6], [7] and facial expressions [8], [9]. These studies have focused on individuals who communicate using verbal speech, and not on mv* communicators. In this work, we use the term *verbal speech* to specify speech with typical verbal content, which is different from *nonverbal speech* which is also richly expressive and communicative (as from mv* communicators) but may not contain verbal content like words or phrases. Prior work has also examined the relation between affect and physiological signals like electrodermal activity (EDA) and electrocardiography (ECG) [10], [11], [12] in neurodiverse populations. Picard explored using EDA to augment emotional communication with individuals with autism, and suggested approaches for integrating sensors that record autonomic nervous system activation with emotional communication [11]. Kushki et al. explored using an electrocardiogram (ECG) to detect anxiety-related arousal in children with autism [13]. While physiological studies can provide valuable insight into affective expression, they often require wearing uncomfortable sensors and may be difficult to interpret in real-world settings where signals can be affected by many factors. Studying vocalizations with neurodiverse populations is important towards expanding inclusive communication and enhancing understanding on how affect is expressed by diverse populations.

2.2 Other Communication Modalities

Prior work has explored the development and usage of forms of communication, including gestural communication and AAC usage [14], [15], [16], [17], [18], [19], [20]. Like nonverbal vocalizations, these forms of communication are highly expressive and communicative, yet they are less commonly used by the general population. These communication modalities may be personalized to the communicator, and – like any highly individualized communication approach – can require time investment from the communicator and/or the communication partner to learn how to express or receive communication with a given modality [21]. Nonverbal vocalizations are one component of communication from mv* individuals, and it is important to note that these vocalizations occur along with other communication modalities, which may include AAC as well as non-vocal communication like body movements and gaze [21], [22], [23].

2.2.1 AAC Usage

Prior work has highlighted the effectiveness and importance of AAC as a communication modality [24], [25]. Couper et al. studied the use of three types of AAC devices (picture exchange, manual signs, and tablet-based speech generation) with eight Autistic children and found that most of the children preferred the tablet-based device [21] and that children were able to learn to use an AAC device to request preferred stimuli more quickly when using a preferred modality. While AAC is a rich and important communication modality – and one that should be treated equivalently to other types of communication like verbal speech – not all mv* individuals use AAC devices and mv* individuals' AAC vocabularies may vary significantly between one another.

2.2.2 Non-Vocal Communication

Non-vocal communication - including body movements, posturing, gestures, and eye gaze - have been shown to convey affect in both typical and neurodiverse populations [22], [23], [26] though non-vocal communicative expression and reception may differ for neurodiverse individuals [20], [22], [26], [27]. Stone et al. studied non-vocal communication like gestures and eye contact in two- and three-year old children, with the goal of identifying differences in communication styles between children with autism and typically developing children [14]. These researchers observed that Autistic children in the study were more likely to communicate by manipulating the communication partner's hand but less likely to communicate using eye gaze and pointing than typically developing children (n=28). In a retrospective video study, Gordan et al. found a correlation between gestural use and language outcomes among toddlers identified as having a higher likelihood of having ASD (n=42), with less gestural use at age 13-15 months being associated with lower expressive and receptive language outcomes at 20-24 months [15].

Recent studies by Wilson et al. explored interactive communication with neurodiverse individuals [28], [29]. These researchers developed and tested ExpressiBall, a ball with lights, sound, and motion sensors, to study self-expression

and co-design with minimally verbal Autistic children. The research team identified six self-expression modalities: Words, Sound, Bodily Movements, Touch and Gestures, Creativity, and Play. The ExpressiBall encouraged expression and communication through multiple modalities, and their results emphasized that researchers and others should listen and respond to all modalities of expressions [29]. In another study, Wilson et al. used a similar device to identify 'moments of interaction' during which minimally verbal children communicated in ways that extended beyond words [28]. While the research presented here focused on nonverbal vocalizations, understanding that these vocalizations occur as part of a complex communication system is important to contextualizing this work and in developing avenues for further exploration.

2.3 Clinically Oriented Studies on Vocalizations

Nonverbal vocalizations – particulary in infants and young typically developing children – have been studied extensively as part of language development [30], [31], [32], [33], [34], [35], [36], [37]. Donnellan et al. experimentally studied the relation between prelinguistic vocalizations in infants and language development trajectories for typically developing children [31]. McDaniel et al. conducted a meta-analysis on the relationship between prelinguistic vocalizations and expressive language development in children with autism [32]. Bacon et al. [36] created a large naturalistic dataset by manually coding toddler speech in clinic-visit videos to study language development of toddlers with and without autism.

Researchers have explored using nonverbal vocalizations to diagnose autism using infant cries [38] and naturalistic child vocalizations [35], [39], and as a marker for other developmental differences like Fragile X syndrome [40], Down Syndrome [41], and specific language impairments (SLI) [42].

Studies with specialized populations primarily take place in laboratory and clinical settings, and often attempt to elicit vocalizations from participants. In a study with twenty-four children with autism, Chiang et al. found that children produced more spontaneous communication in natural environments than elicited communication and that spontaneous communication had different uses (e.g., requesting) [43]. Oller et al. conducted one of the only known studies of nonverbal vocalizations from non-typically developing children in real-world environments, but focused on diagnosis tasks using toddler speech not on vocalization affect or intent [35]. Tools that track and enhance communication with mv* individuals in real-world settings, which may provide affective and communicative vocalization data not captured by laboratory tests [44], are largely unexplored.

2.4 Nonverbal Vocalizations as Communication

Nonverbal vocalizations include both involuntary (e.g., coughing, hiccupping) and voluntary (e.g., grunting, sighing, screaming) sounds. Nonverbal vocalizations often occur amidst typical verbal speech and can be used to convey an emotion, express intention, and emphasize verbal speech [45], [46], [47].

2.4.1 As an Expression of Affect Amidst Typical Verbal Speech

Nonverbal vocalizations that occur alongside typical language have been studied anthropologically [48], [49] and have been classified affectively with both natural and acted vocalizations across numerous studies [45], [50], [51]. Trouvain and Truong categorized types and usages of nonverbal vocalizations and identified five primary types of nonverbal vocalizations: vegetative sounds (e.g., snoring), affective sounds (e.g., laughter), interjections as semi-words (e.g., "shh"), filler sounds as semi-words (e.g., "uhm"), and melodic utterances (e.g., humming) [52].

Holz et al. found that listeners could reliably identify intensity and arousal in nonverbal vocalizations, but that emotions expressed with maximal intensity were more difficult to categorize than more moderately expressed emotions [50]. Schroder et al. also found that listeners could reliably identify acted nonverbal vocalizations expressing 'affect bursts' across ten categories [53]. Sauter et al. found that vocalizations communicating some basic emotions (anger, disgust, fear, joy, sadness, and surprise) were recognized cross-culturally by both individuals from Western countries and from isolated villages in Namibia [48]. Anikin found that listeners could reliably differentiate between acted and authentic affective nonverbal vocalizations [49], and identified correlations between voice quality and valence in nonverbal vocalizations [51].

2.4.2 As Expressive Communication in Infants

Nonverbal vocalizations have been studied as expressions of affect and communication from infants. In 1964, Wasz-Höckert identified specific meanings – pain, pleasure, and hunger – in infant vocalizations in a study with trained nurses in a hospital [54]. Since then, there has been extensive work on classifying infant cries by need (e.g., hunger, pain) using both humans and machines [55], [56], [57], [58]. Weisman et al. studied the dynamics of vocalizations during infant-father interactions and found that vocalizations played a significant role in regulating social interactions [59]. Recently, Liu et al. used linear predictive coding (LPC), linear predictive cepstral coefficients (LPCC), Bark frequency cepstral coefficients (BFCC), and Mel frequency cepstral coefficients (MFCC) to classify infant cries as being related to hunger, sleepiness, needing a diaper change, a need for attention, or general discomfort [55].

2.4.3 Lack of Studies of Nonverbal Vocalizations as Communication Independent of Typical Verbal Speech

While researchers like Beukelman and Mirenda have noted that mv* individuals use vocalizations to express emotions and communicate, systematic study and tools that can work with these expressions remain undeveloped. For people who are mv*, these nonverbal vocalizations serve a unique linguistic and communicative purpose as they occur independently of verbal speech [60]. These vocalizations include traditional nonverbal cues (e.g., laughter or yells), as well as unique utterances of varying pitch, phonetic content, and

TABLE 1
Demographic and Background Information for Participating mv* Communicators

Participant ID	Gender	Age (years)	Conditions 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, Genetic disorder	56	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

Participants who enrolled in the study but did not collect enough data were not included in the analysis.

tone that do not fall into the typical categories of nonverbal vocalizations.

To our knowledge, no other studies have acquired nonverbal vocalizations from mv* individuals using personalized labels in real-world settings. In our prior work, we describe in detail the participatory design process used to design our approach [61], our novel data collection system [4], [62], and provide preliminary classification results with three mv* communicators [4]. Here we extend this work to include new analytical approaches for classification with eight mv* communicators, and discuss our experimental results in the context of real-world data collection.

3 Methods

3.1 Data Collection and Pre-Processing

Participants were recruited through conversations with community members and word-of-mouth. The study was approved by the MIT institutional review board (IRB). Data was collected as described in [4] and [63] and the dataset used in experiments, ReCANVo (Real-World Communicative and Affective Nonverbal Vocalizations), is presented fully in [63]. The study was conducted entirely remotely, in order to reach a geographically distributed population and minimize time burden. The remote nature of the study enabled data collection even during COVID-19. Data were collected in communicators' natural environments, primarily in and around the home. Participants were encouraged to go about their typical day-to-day activities while recording.

Background and demographic information (provided by families via a web survey) for participating mv* communicators are in Table 1. Additional information on nonverbal communication practices by each participant is summarized in [64] and [65], including details on communication modalities and vocal communication for each participant. In our study, five of the eight participants used AAC with two participants reporting an AAC vocabulary size of twenty words/phrases or greater. In the survey, respondents also reported that the mv* individuals communicated via gestures, picture exchange, and hand leading or pulling.

Audio was recorded using a small recorder (Sony IDC-TX800, recording in 16-bit 44.1 kHz stereo) attached to the communicator's clothing (P01, P02, P03, P06, P11, P16), worn as a necklace (P08) or placed nearby (P05). Recorder placement was flexible to accommodate tactile sensitivities. Vocalizations were labeled in-the-moment by a close family member (i.e., a "labeler") using a custom app (Fig. 1). The labeling process was designed to be highly flexible so that it could integrate into day-to-day life. Labelers provided labels when circumstances allowed – often in short clusters. Labelers were instructed to only label when they could do so confidently. In [65] and [61], we provide additional information on how the labeling system was designed and validated. The labeler used the app to designate both a start and end time for each vocalization, as accurately as possible. The app had six labels that were shared by all families, that were described to the family as below:

- Self-talk: Vocalizations that appear to be associated with being content, happy, or relaxed, and are generally made without apparent communicative intent. (The individual seems to be making them to him/herself.)
- Dysregulated: Vocalizations that appear to be associated with being irritated, upset, agitated, bored, uncomfortable, overstimulated, or distressed. These vocalizations are often (but not always) made without an apparent specific communicative intent.
- Delighted: Vocalizations that appear to be associated with being excited, very happy, or states of glee.

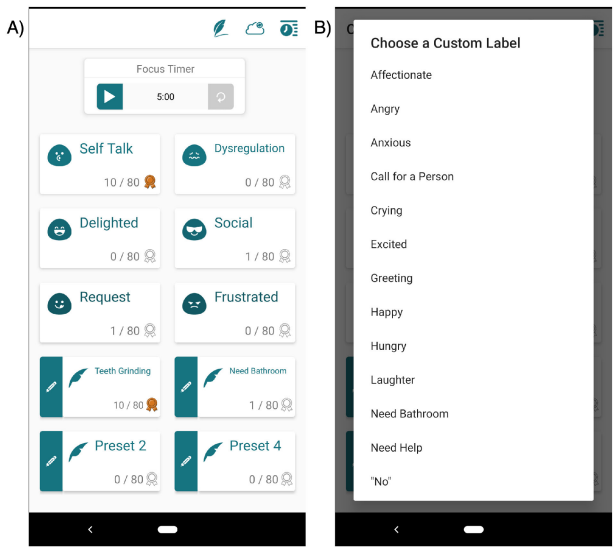


Fig. 1. Custom app for in-the-moment labeling of nonverbal vocalizations [4]. A) Main labeling screen B) Partial list of preset options.

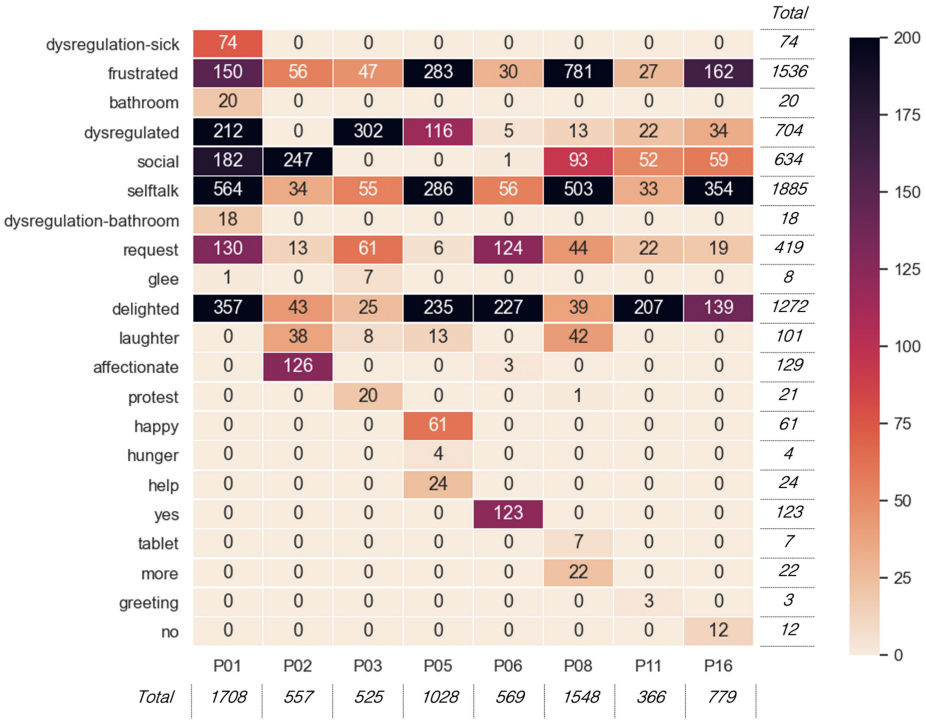


Fig. 2. Number of collected nonverbal vocalization samples per label for each mv* communicator.

- Frustrated: Vocalizations that appear to be associated with being frustrated, angry, or protesting.
- Request: Vocalizations that appear to be associated with making a request.
- Social: Vocalizations that appear to be social in nature (e.g., as part of a back-and-forth vocal exchange while playing a game).

These six labels and descriptions were selected based on conversations with families of mv* communicators and a speech and language pathologist (SLP). The labels span both affective and communicative functions. The labels include positive valence classes (”delighted”, ”selftalk”), negative valence classes (”frustrated”, ”dysregulated”), higher arousal classes (”delighted”, ”frustrated”) and lower arousal classes (”selftalk”, ”dysregulated”). Labels emphasizing communicative functions (e.g., ”request” and ”social”) were included because they capture a critical dimension of how nonverbal vocalizations are used functionally by this population. Importantly, social vocalizations were only labeled as such if they 1) overtly social in nature, such as a back-and-forth vocal exchange while playing a game, and 2) could not be better described by a different label. For example, requesting something is an inherently social exchange, but these vocalizations would be labeled as ”Request”. As in verbal speech, nonverbal vocalizations from mv* individuals can express complex meanings spanning both affective and communicative functions (e.g., a ”social” vocalization might also express happiness and excitement about an interaction, and a ”frustrated” vocalization might be used to communicate that a request has not been adequately met). In the study, families were asked to select the label that they felt would help others best understand how to respond to a vocalization.

Families could also customize four additional labels, by selecting from a drop-down list of twenty-five more specific preset options (e.g., ”hungry”,”greeting”). The app and data collection process were created using a longitudinal participatory design process [61]. The study was designed to give participants flexibility to set the pace, settings, and schedule for data collection. This flexibility resulted in variability in data collection and labeling practices between participants, but was critical in enabling a first-of-its-kind real-world data collection with mv* communicators.

The collected audio recordings were pre-processed as described in [4], [63] to extract labeled vocalizations. Volume-based segmentation was used to find audio segments of interest. Because the recorder was placed close to the communicator, vocalizations were generally louder than other content in the recording. The recorded audio was temporally aligned with label timestamps. Label matching accounted for human delay in labeling and small clock shifts (see [63] and [65] for details on the alignment and segmentation process). Segments that were not temporally near labels were discarded. A researcher listened to each segment to ensure it contained a vocalization from the communicator. As needed, the researcher trimmed noise surrounding the vocalization, leaving approximately 0.15s of buffer at the beginning and end of a vocalization. Fig. 2 shows the number of samples of each vocalization type collected from each participant.

3.2 Analysis of Labeling and Data Collection Practices

Exit interviews were conducted with seven of the eight labelers (the labeler for P02 was not available for an interview) to understand the fidelity of the labeling scheme. Labeling and data collection practices were tabulated for each participant including labeling delay, average recording session length, number of uploaded sessions, average label

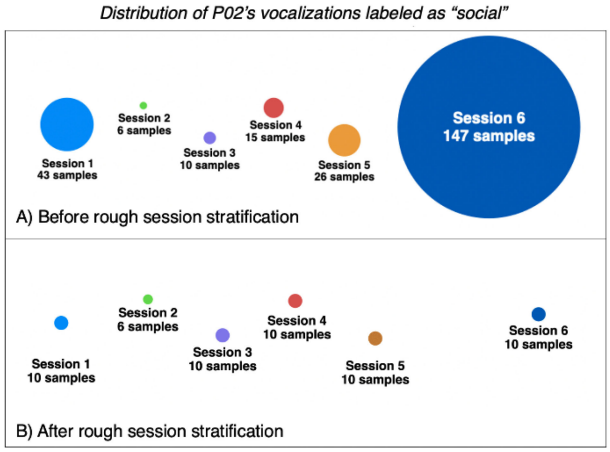


Fig. 3. Distribution of P02's "social" vocalizations by session before and after rough session stratification.

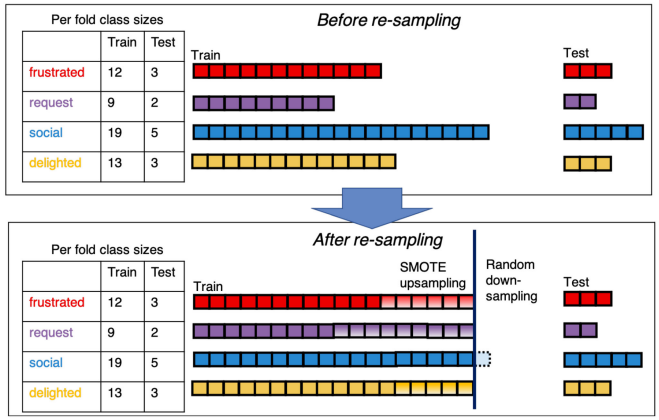


Fig. 4. The class sizes were balanced in each fold using random downsampling and the synthetic minority oversampling technique (SMOTE). In each fold, the number of training samples per class was balanced to the minimum of twice the smallest class size and the largest class size. The figure illustrates the balancing strategy within a fold, using pseudo class sizes. The test data was not balanced, so metrics are reported using macro-averages to weigh each class equally.

duration (using the label start and end time), average number of vocalizations per label, average number of labels per session, and median number of unique labels per session (results are summarized in Table 5). The labeling delay was defined the time passed between the start of a vocalization and the label button push on the app to associated with that vocalization. To minimize the effect of any small clock drifts between the recorder and labeling app, only the first two weeks of labeling data were used in calculating the labeling delay. For each participant, the range of delays were similar at the start and end of the two-week period.

3.3 Machine Learning Evaluation and Sampling Strategies

Throughout the evaluation and results sections the term "sample" is used to refer to a labeled vocalization used for model training or evaluation. Because of the heterogeneity of the participants (Table 1), and known differences in nonverbal vocal expressions between mv* individuals, personalized models were trained for each participant. Rough session stratification was used to reduce fitting to background noise. For a given label and participant, the distribution of labels across sessions were often skewed. To avoid a model learning to associate a label with the soundscape from a dominant session, the maximum number of vocalizations having the same label per session per participant was limited to 10 via random undersampling. To illustrate this, Fig. 3 shows the distribution of P02's vocalizations labeled as "social" by session before and after rough session stratification. Models were trained and evaluated with and without rough session stratification.

Models were evaluated using two strategies: 5-fold cross validation and a leave one session out (LOSO) evaluation. A session was a single uploaded recording from a participant and contained many vocalizations. Sessions were time-separated and correlated to a day or a specific activity. Background soundscapes between sessions were generally distinct from each other. The LOSO approach was implemented to further prevent model fitting to the background soundscapes of the vocalization recordings. The 5-fold cross validation evaluations were run with 3 distinct random seeds. The average metric and 95% confidence interval for 5-fold cross validation are reported using each fold and random seed.

Within each fold/session, the outer loop (or held out session) was used for evaluation and the inner loop was used for regularization parameter selection. The training data in each loop was balanced to the minimum of twice the smallest class size and the largest class size using random downsampling and the synthetic minority oversampling technique ("SMOTE") [66] (Fig. 4). SMOTE creates synthetic data points by selecting randomly along lines in the feature space between nearest neighbor points in the minority classes [66]. In each loop, the oversampler was fit only on the training data for that loop. This sampling scheme ensured that no class had more synthetic than real data and that, in each loop, at least one class has no synthetic data. This bound was selected empirically after systematic experimentation with different allowed synthetic class proportions for its consistency in performance and interpretability across loops with varying class distributions.

Tables 2 and 3 show the number of training samples per class per fold for each evaluation strategy. The classes were balanced within each fold. Because of the difference in vocalization distribution between sessions, the balanced per class training size varies between folds for the LOSO evaluation strategy. Because the test dataset is not balanced, every sample is used as a test sample exactly once. Reported metrics are macro-averaged across classes to weigh each class equally.

3.4 Multi-Class Classification

Classes were selected for analysis for each participant based on the number and distribution of samples. The selected classes for each participant are shown in gray in Tables 2 and 3. For a participant, a label was included in the analysis if there were at least 30 vocalization samples spread across at least 3 sessions. These thresholds were chosen because they allowed for inclusion of at least 4 labels per participant, while maintaining sample volume (by requiring at least 30 samples) and diversity (by requiring the samples to come from at least 3 recording sessions). Additionally, in this

TABLE 2
Number of Samples for Each Class for Each Participant With Rough Session Stratification

	P01	P02	P03	P05	P06	P08	P11	P16
Multi-class with session stratification								
delighted	109	41		120	91	38	143	47
dysregulated	31		52	65				
frustrated	61	35	34	62	22	58	27	102
request	31		52		69	33		
selftalk	79	34	25	142	47	158	32	148
social		56				71	43	47
yes					84			
5-fold cv training samples per class (balanced per fold)	50	44	40	98	36	54	44	74
LOSO training samples per class (balanced per fold)	42-62	46-56	30-94	104-124	24-44	46-56	42-54	74-94
Binary with session stratification								
postive valence	167	69	25	186	99	166	148	200
negative valence	92	35	71	97	22	58	27	103
LOSO training samples per class (balanced per fold)	157-167	50-67	30-50	174-186	24-44	96-116	42-54	184-195

The classes selected for evaluation for each participant are shown in gray. The number of per class training samples per fold (balanced with SMOTE upsampling and random downsampling) are shown.

analysis, if a participant had two closely related labels (e.g., "delighted" and "happy"), only one was included in the model because of the increased difficulty of differentiating between closely overlapping categories. As more samples are collected, models could be trained for more fine-grained classification. Multi-class models were trained using four or five classes for each participant.

3.5 Binary Valence Classification

Binary valence classification experiments were conducted to evaluate the model performance on larger meta-classes. "Delighted" and "selftalk" vocalizations were merged into a positive valence class and "dysregulated" and frustrated" samples were merged into a negative valence class. The selected classes, number of training samples per class, and the number of training samples per class per fold with rough session stratification are given in Table 2. Binary valence models were only evaluated with LOSO and rough session stratification, the most conservative evaluation approach.

Only binary valence experiments were conducted because mapping to binary arousal states was not well-defined for the selected labels. The labels included lower and higher arousal states within a particular valence: "dysregulated" is generally lower arousal than "frustrated" and "selftalk" is generally lower arousal than "delighted". Within a valence, the arousals have a general relative relationship but are not necessarily independently "low" or "high". For instance, for some participants, "dysregulated" may be associated with high arousal and "frustrated" may be associated with very high arousal. There is not a clear distinction in relative arousals between "dysregulated" and "delighted" and between "selftalk" and "frustrated". Additionally, labelers were not asked to consider broad arousal mappings among labels.

3.6 Feature Extraction and Modeling

Experiments were conducted with Random Forest models, a support vector machine (SVM) model with a radial basis function (RBF) kernel, and a linear SVM with stochastic gradient descent (SGD) training. These models were selected after experimenting with a broader selection of models and based on their use in the literature for similar classification tasks. Models were evaluated with statistical features aggregated for each vocalization, bag-of-phones features, and data-learned features extracted using auDeep [67].

Aggregate features were extracted for each vocalization with the following feature sets:

- the extended Geneva minimalistic acoustic parameter set extracted using openSMILE [68], size 88 [69]
- a custom feature set, size 63
- mean mel frequency cepstral coefficients (MFCC), size 13
- mean gammatone cepstral coefficients (GTCC), size 13
- means of a mel-base filter bank applied to a short-time Fourier transform (STFT), size 40
- means of an ERB-based filter bank applied to a STFT, size 40

Table 4 lists the features included in the custom feature set. The custom feature set was designed using the research team's observations, feedback from practicing speech and language pathologists, and prior work with related classification tasks [70], [71]. Consulted speech and language pathologists suggested that the prosody of the vocalizations might be particularly informative in nonverbal vocalizations. As such, the custom feature set includes features that capture characteristics of the pitch contour like number of peaks and polynomial fit coefficients. Parents of mv* communicators suggested that the phonetic content of vocalizations might vary with usage and so the custom feature set applies a functional that extracts the longest constant value of the first and second formants, which are related to the vowel content of a vocalization. The custom set also includes mean GTCC and BFCC values (which have been used in the related task of cry classification [55]), the audio amplitude duration and mean auto-correlation, and

TABLE 3
Number of Samples for Each Class for Each Participant Without Rough Session Stratification

Multi-class without session stratification

	P01	P02	P03	P05	P06	P08	P11	P16
delighted	357	43		235	227	39	206	132
dysregulated	212		302	116				
frustrated	150	56	47	283	30	781	27	161
request	130		61		124	44		
selftalk	564	34	55	286	56	502	33	339
social		247				93	52	59
yes					123			
5-fold cv training samples per class (balanced per fold)	216	56	76	198	52	64	44	94
LOSO training samples per class (balanced per fold)	152-260	48-68	48-94	144-232	30-60	46-78	42-54	84-118

The classes selected for evaluation for each participant are shown in gray. The number of per class training samples per fold (balanced with SMOTE upsampling and random downsampling) are shown.

functionals applied to power, cepstral peak prominence (related to voice quality), and harmonics.

The custom feature set includes features used in automatic speech recognition that are related to lexical content (e.g., cepstral coefficients), features with a clear interpretable mapping to nonverbal phonetic content (e.g., formant values), and features that capture aspects of speech often associated with affect (e.g., prosody and voice quality measures). Implementation details for the custom feature set are provided in Appendix A.1, which can be found on the Computer Society Digital Library at http://doi.ieeecomputersociety.org/10.1109/TAFFC.2022.3208233. Details on the extraction of the cepstral coefficients and filter-bank coefficients are provided in Appendices A.2 and A.3, respectively, available in the online supplemental material.

The eGeMAPs feature sets and cepstral coefficient features have been used extensively in prior work in speech emotion recognition and other speech modeling tasks. Filterbank-based features of size 40 were evaluated because they have been used effectively in noise-robust speech emotion recognition in prior work [72]. The bag-of-phones feature set had size 50 and consisted of language independent phones extracted from each vocalization using *allosaurus*, a Python library for phone extraction [73]. The bag-of-phones included the 50 phones that appeared most frequently in the dataset. Similarly to bag-of-words approaches, the bag-of-phones encoded phone multiplicity but not order. Data-learned features were extracted using the auDeep autoencoder to learn a feature representation via unsupervised learning with a deep neural net (DNN). The autoencoding was generated for each fold split for 5-fold nested cross validation, before session stratification or class balancing. Because of the fold-based learning structure, results with the auDeep feature set are only reported with 5-fold nested CV evaluation and not LOSO evaluation. A parameter specifying the length of vocalization used in training the autoencoder and the parameters defining the DNN architecture were optimized for each participant. The selected parameters are provided in Appendix B, available in the online supplemental material. Because of the computationally-intensive nature of selecting hyperparameters for a given fold split, the evaluation metrics with auDeep features are provided using one set of folds.

4 Results and Discussion

4.1 Analysis of Labeling and Data Collection Practices

In exit interviews, labelers generally reported high confidence in the fidelity of their labels and cited that contextual information (e.g., gestures and setting) helped them label confidently [65]. Fig. 5 shows the distribution of labels by session for each participant. Fig. 6 shows box-and-whisker plots of the labeling delay for each participating labeler. For each participant, Table 5 shows the average session length, number of sessions, average label duration, average number of vocalizations per label, the average number of labels per session, and the median number of unique labels per session. Additionally, Appendix C, available in the online supplemental material, shows the average and standard deviations for label duration per class per participant. The

TABLE 4
Features and Applied Functionals Used in the Custom Feature Set (Additional Implementation Details are Provided in Appendix A)

Feature	Applied functionals
Audio amplitude	Duration; Mean auto-correlation
Formants 1, 2, and 3	Mean; variation; frequency and duration of longest constant value (formants 1 & 2)
Power	Freq associated with max. power; Variation; Interquartile range
Pitch (fundamental freq.)	Mean; range; max; min; quartiles 1-3; number of peaks; overall rise/fall; quartile 1; quartile 2; quartile 3; Fit coefficients for polynomials of order 1, 2, 3
GTCC,13 coeff.	Mean
BFCC, 13 coeff.	Mean
Cepstral peak prominence	Mean
H1-H2; H2-H4	Mean

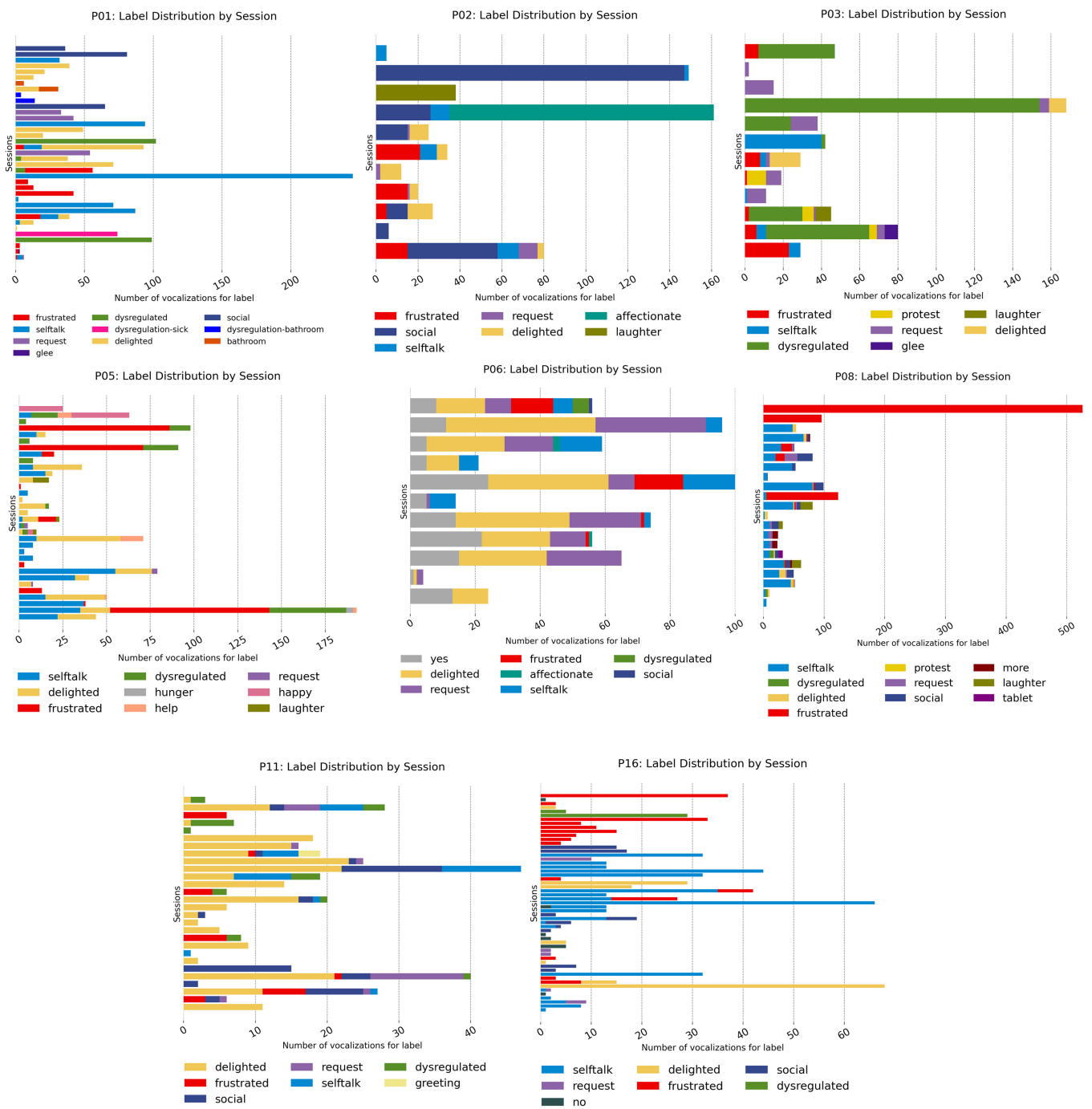


Fig. 5. Each horizontal bar shows the distribution of labels in a particular session for a participant.

presented statistics can inform the design of other in-the-moment tagging systems.

Labelers for P01 and P16 often deviated from the preferred labeling protocol by assigning labels to entire recordings, based on the general mood or communicative intent, instead of individual vocalizations. As a result, they often have only a single label in a given session (Fig. 5) and were not included in the delay analyses and some of the tabulated statistics could not be calculated for P01 and P16. Still, these handwritten labels could be aligned with the recordings and used in classification models. The other participants more closely followed the preferred collection protocol, designating labels for vocalizations that occurred within a recording.

The average labeling delay ranged from 3.5-7 seconds and were significantly different between labelers. The Kruskal-Wallis nonparametric test found that differences in label delay between labelers were significant ($p = 4.4^{*} 10^{-9}$). Data collection practices varied between participants. In future studies, data collection could utilize an integrated recording and labeling app (as shown in [65]). In such a system, the clock for the recording system would be coupled directly to the clock for the labeling system thereby eliminating any clock drift and enabling in-depth analysis on the effect of labeling delay on model performance.

P08 had significantly longer average label durations than other participants. Very long labels risk encompassing

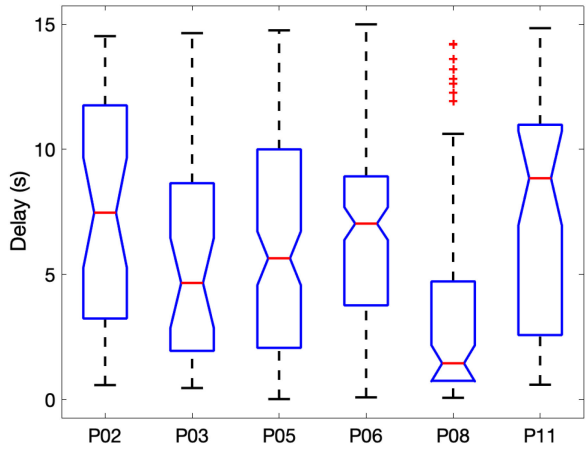


Fig. 6. Distribution of labeling delays for each participant. The labeling delay was defined as the time passed between a button push on the app to start a label and the start of the first vocalization associated with that label. P01 and P16 were not included in the delay analysis because they deviated from the preferred labeling protocol and did not assign labels using the app. Still, the labelers for P01 and P16 did provide handwritten labels for each file that could be aligned with the data and used in classification models.

vocalizations of multiple categories, and could be associated with mislabeled data. Generally, for a given participant, the average label duration was within one standard deviation across class types (Appendix C), available in the online supplemental material. Frustrated had the highest average labeling duration for four participants. P02, P03, P05, P06, P08, and P11 all tended to have multiple unique labels in a given session. Having multiple unique labels per session may reduce the model's likelihood of learning to associate a label with a particular background soundscape.

4.2 Classification

Evaluation metrics are reported using macro-averages across classes. If a device or human were to listen and respond to nonverbal vocalizations, having both high recall and precision would be important for enabling consistent appropriate responses. For this reason, the F1 score (the harmonic mean of recall and precision which is $\frac{2*precision*recall}{precision*recall}$) was used to evaluate and compare models. The unweighted average recall (UAR) is also reported in the results tables.

4.2.1 Multi-Class Classification

Multi-class classification results for each model and evaluation strategy with each aggregated feature set with and without session stratification are provided in Appendix D in Tables D1, D2, D3, and D4, available in the online supplemental material. Results with the feature-learned auDeep features and the bag-of-phones feature set are provided in Appendix D in Tables D5 and D6, respectively, available in the online supplemental material.

Fig. 7 shows the highest F1 score across aggregate feature sets and model types for each evaluation strategy. Fig. 8 shows model performance for only the LOSO evaluation with session stratification, the strategy least susceptible to fitting to background sounds. The best performing models for each evaluation strategy had multi-class F1 scores higher than chance for all participants (Fig. 7).

4.2.2 Evaluation Strategies

Generally, model performance is higher for models evaluated using 5-fold cross validation. Model performance with the 5-fold evaluation scheme, particularly without session stratification, is likely artificially inflated due to fitting to background noise. While the LOSO evaluation scheme is less likely to classify samples correctly by fitting to the background soundscape, it cannot correctly classify vocalizations that were expressed uniquely in one session. For each participant, each vocalization label encompassed many different sounds. For example, for P05 "selftalk" included laughter, sighs, and complex phonetic expressions with multiple constant-vowel components and transitions. The LOSO evaluation scheme cannot classify unique sub-types of a vocalization category that appeared only in one session.

For some participants (P01, P02, P03, P11), models with session stratification had better performance than models without session stratification even though session stratification reduced the number of available training samples. Without session stratification a model is more likely to fit to the background soundscape. For LOSO evaluations, the test data for each split has distinct background soundscapes from the training data and so fitting to background noise can reduce model performance. In these cases, session stratification can improve model performance - i.e., for P01 models with LOSO evaluation. The P01 data tended to have a single label for an entire session which may have made models for P01 particularly susceptible to fitting to background noise (Fig. 5).

4.2.3 Differences in Model Performance Between Participants

All F1 scores are above chance, but there are large variations in performance between participants (Figs. 7 and 8). These

TABLE 5
Statistics Describing Labeling and Data Collection Practices

	P01	P02	P03	P05	P06	P08	P11	P16
Avg. session length (s)	943.6	2151.4	6704.7	5515.6	1641.4	1165.6	804.8	142.0
Number of sessions	38	11	12	33	11	21	28	57
Avg. label duration (s)	N/A	9.7	0.9*	13.5	10.7	50.3	3.3	N/A
Avg. vocalizations per label	47.4	8.0	2.5	4.7	2.4	7.2	2.1	13.7
Avg. number of labels per session	2.6	11.3	26.1	11.7	30.0	16.7	11.7	1.1
Unique labels per session (med.)	1	3	3.5	2	5	4	3	1

P03 collected some data using a previous version of the app which did not require specifying an 'end' time and assumed a 2 second label duration.

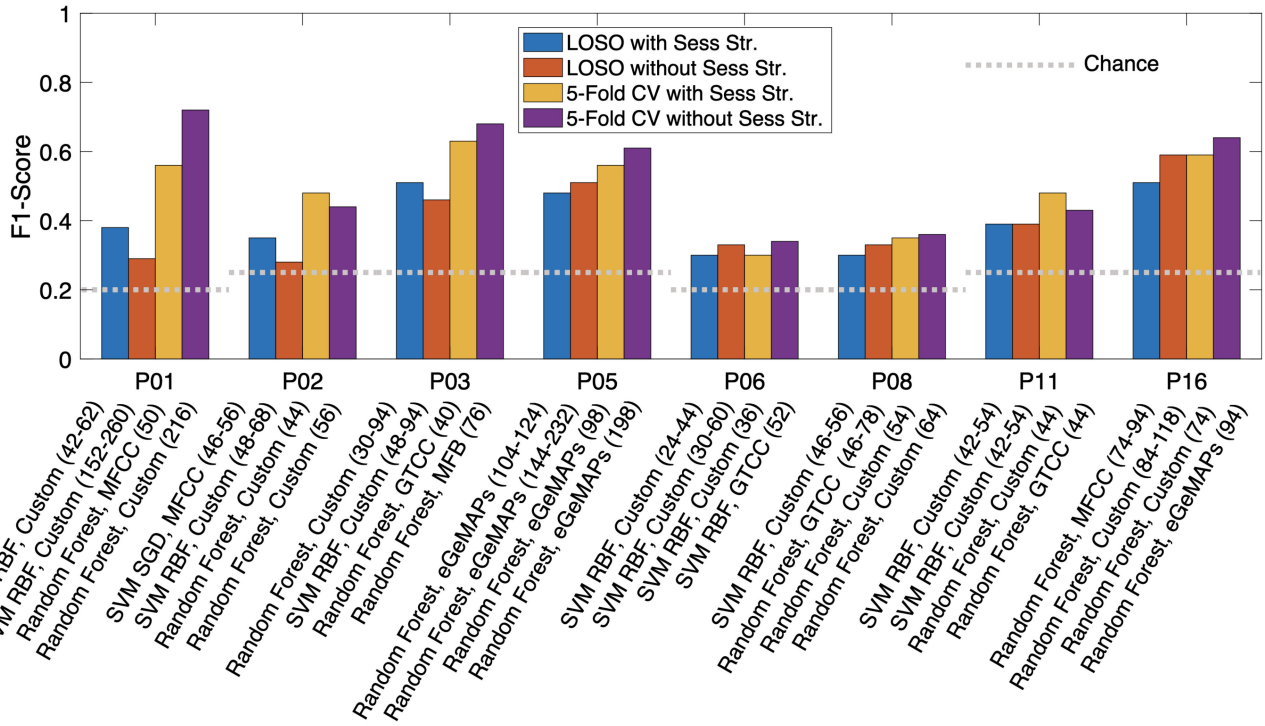
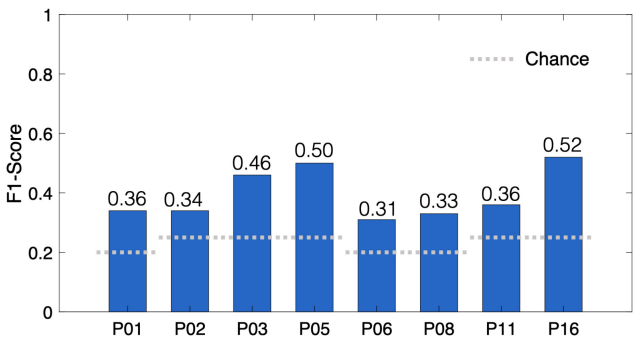


Fig. 7. F1 score for best performing model and feature set for each participant, evaluated using leave-one-session-out ("LOSO") and 5-fold cross-validation ("CV") with and without session stratification ("Sess Str."). The labels under each bar indicate the model and feature set. The number of training samples per class is shown in parentheses. A range is provided for LOSO evaluations, where the number of training samples varied between folds. The confidence intervals for the 5-fold CV evaluation are provided in Tables D1 and D2 in the Appendix, available in the online supplemental material.

variations could be due to differences in data collection and labeling as well as inherent differences in vocal communication.

The number of samples used in training varied between participants. P05 and P16 had the largest number of training samples per class and relatively high model performances (Fig. 8). Still, the variations are not due to training sizes alone – P03 also had a relatively high model performance but had one of the fewest number of training samples per class. Differences in labeling quality and style may have also affected model performance. During follow-up interviews, the labelers for P06 and P08 both mentioned forgetting to end a label when the vocalization ended on occasion. This could have led to mislabeled vocalizations in the dataset that affected model performance. P03 had a high model performance despite a low number of training samples per class and had the lowest average label duration (Table 5). A low labeling duration indicates a close mapping between vocalizations and labels and a lower likelihood of mislabels. P08 had a relatively low modeling performance and the highest average label duration (Table 5).



Feature Set	Custom	Custom	Custom	eGeMAPs	Custom	Custom	Custom	eGeMAPs
Model	SVM RBF	SVM SGD	Random Forest	Random Forest	SVM RBF	SVM RBF	SVM RBF	Random Forest
Samples per class	42-62	46-56	30-50	104-124	24-44	46-56	42-54	74-94

Fig. 8. F1 score for best performing model and feature set for each participant, evaluated with LOSO and session stratification.

Inherent differences in vocal communication between participants may also contribute to variations in model performance. The age, diagnoses affecting speech and language, and number of spoken words or word approximations varied between participants (Table 1). Future work with additional participants and revised data collection procedures could allow for the decoupling of labeling practices, model performance, and demographics, providing valuable insights for clinical and modeling practices.

Speaker-independent models were trained for each participant (using only data from other participants') to explore the performance of non-personalized models. F1-scores for speaker independent models were below chance, with the exception of models evaluated with P05 data. Speaker-independent models evaluated with P05 data had F1-scores around 0.30, above chance but much lower than the F1-scores for P05's personalized models (Fig. 7).

4.2.4 Labels

Fig. 9 shows multi-class confusion matrices for the leave one session out with session stratification evaluation.

Fig. 9. Multi-class confusion matrices for best performing model and feature set for each participant with LOSO with session stratification evaluation. The diagonal entries of the matrix are the recall for each class.

Differences in how well a particular label could be classified could be characteristic of the vocalization itself the number and quality of available training samples for that label. "Frustrated" had high recall even when it had relatively few samples, like for P03 (Table 2). For P06, "frustrated" had a high recall even though it had fewer than half of the number of samples of the other classes. For P11, "delighted" vocalizations had the highest recall and the largest number of training samples (Table 2).

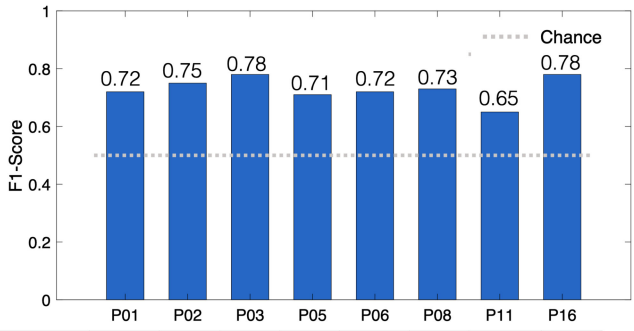
Some classes might have poor classification performance because vocalizations in that class tended to have multiple meanings (i.e., a "frustrated request") for an mv* communicator, in which case the class predicted by the model might be accurate even if it didn't match the single given label. We experimented with multiple labels while piloting the study but found that asking labelers to designate multiple classes imposed too high of a cognitive load and reduced the overall fidelity of the marked labels. Labelers were asked to choose the most representative label for a vocalization, and to only label a vocalization if they were confident in its label. Still, understanding that a vocalization could fall into multiple categories is important when interpreting the results. Future studies could revisit developing labeling methods to allow for assigning multiple labels to a vocalization, which would enable further analyses.

For many participants, there was a label that had more ambiguity than the others. For instance, removing the "selftalk" class for P01 and P06 improved the F1 score of the best performing model (with LOSO and session stratification) to 0.50 (+0.12) and 0.37 (+0.07), respectively. Removing the "delighted" label from P05 and P08 improved the F1 scores to 0.62 (+0.13) and 0.39 (+0.09), respectively.

4.2.5 *Model and Feature Set Performance*

The nonlinear models (Random Forest with RBF kernel) generally had better performance for the multi-class classification task. The custom feature set had the best model performance for five of the eight participant with LOSO and session stratification (Fig. 8), suggesting that the distinct features and applied functionals chosen for the custom feature set capture the unique differences between nonverbal vocalizations of different types. The cross-validation approach utilized here was appropriate for the small, highly varying dataset. However, future work with sufficient data for unique training/validation/evaluation splits would enable investigations of feature performance. These results could then inform the development of an improved custom feature set and broader applications to nonverbal vocalization classification models.

The bag-of-phones feature set generally had poor performance compared to the other feature sets. For some participants - particularly, P03, P05, and P16 - the bag-of-phones feature set had classification performance greater than chance. This may indicate a clearer variation in phonetics between vocalizations of different types for these participants. The utilized phone extraction model was not trained for nonverbal vocalizations and had performance limitations even on typical verbal speech [73], which likely contributed to the poor performance of the bag-of-phones feature set. The data-learned features extracted using auDeep generally performed similarly to the aggregate feature sets. For 5-fold cross-validation with session stratification, the features extracted using auDeep had the highest F1 score for P01 and P02 by 3% and 4% respectively, compared to the best performing aggregate feature set. Generating data-learned features (as in auDeep) is computationally intensive compared to the other approaches explored in this paper but was evaluated because such features have been shown in the literature to contribute to significant performance improvements for some audio classification tasks [74]. The presented results suggest that, if additional data were collected, further explorations of data-learned features, including other autoencoder architectures and self-supervised learning, may be beneficial.



	P01	P02	P03	P05	P06	P08	P11	P16
Feature Set	MFB	MFB	Custom	MFCC	GTCC	Custom	Custom	GTCC
Model	SVM RBF	Random Forest	SVM SGD	SVM RBF	Random Forest	SVM RBF	Random Forest	Random Forest
Samples per class	157-167	50-67	30-50	174-186	24-44	96-116	42-54	184-195

Fig. 10. F1 score for best performing binary valence model and feature set for each participant, evaluated with LOSO and session stratification.

4.2.6 *Binary Valence Classification*

Binary valence classification results are provided for the leave-one-session-out with session stratification evaluation method, the method least susceptible to fitting to background soundscapes. Results for each model with each aggregated feature set are provided in Table D7 in Appendix D, available in the online supplemental material. Fig. 10 shows the best performing model and feature set for binary valence classification, evaluated with LOSO and session stratification.

Variations in model performance between participants for the binary classification task were less pronounced (Fig. 10). This may be because there were a larger number of training samples available per class for each participant for the binary task. The linear SVM with stochastic gradient descent (SGD) training had the highest performance for four of the eight participants for binary classification. The linear model may have had comparatively better performance for the binary classification task than the multi-class task because of the simpler nature of the task and the larger amount of training data.

4.3 Limitations

The amount of data available for the presented analysis was limited due to the time-intensive nature of the data collection process. As such, the presented results may have been affected by overfitting due to experimenting with different architectures and feature sets, particularly for participants P01, P02, and P03. However, the other participants were not used in architecture and feature development, so their results reflect a more conservative representation of model fitting with this data.

Custom descriptive labels were an option in the labeling app, but could be difficult to create in-the-moment. Additionally, while labelers often incorporated cues from the communicator's gestures, body language, and other communication when selecting a label, the labeling system did not directly capture feedback from the communicator. In the future, the labeling system could be improved by allowing for higher complexity and diversity of labels, and by integrating feedback directly from the communicator.

5 Conclusion

In this paper, we presented results from the largest study of nonverbal vocalizations with mv* communicators to date along with modeling approaches appropriate for dealing with real-world, messy data. Nonverbal vocalizations from mv* communicators contain important communicative and affective information but are understudied and not well understood by those who don't know a communicator well. Developing methods to classify nonverbal vocalizations by affect and intent is an important step towards improved understanding of this unique type of communication. The F1 score for each participant for multi-class classification was above chance, even with conservative evaluation schemes. This result suggests that there are inherent acoustic differences between vocalizations of different types for the eight mv* communicators in this study. This result is important because it shows, for the first time, that it is possible for models to classify nonverbal vocalizations by affect and intent with mv* individuals using audio alone.

There were large variations in model performance between mv* communicators, which may have been due to presented differences in labeling and data collection practices between participating families and due to inherent differences in communication practices between participants. Additional training data from each participant would likely improve multi-class classification results - models had relatively high performance with data from participants with many training samples, even for participants with potentially lower labeling fidelity (i.e., P01). High quality labels (i.e., P03) may allow for accurate classification even with a smaller number of training samples. In future work, vocalizations from more mv* communicators could also be used to explore whether there are subgroups of communicators with similar vocalization practices.

We contribute to an improved understanding of nonverbal vocalizations from mv* communicators. There has been little prior work with this unique and understudied population. Better understanding of nonverbal vocalizations with mv* communicators could lead to improved communication technology: for instance, this understanding could be used to develop real-time vocalization classification systems and educational tools for individuals who don't know a mv* communicator well, and vocalization controlled AAC devices for mv* communicators. Real-world studies of vocal communication can also contribute to answering scientific questions around language development, such as how and when phonemes are used across development, especially within early affective and communicative expression.

While many families and clinical practitioners understand that nonverbal vocalizations are communication, individuals who are new to communicating with mv* communicators often do not know to listen for this type of communication. We hope that our study of nonverbal communication with mv* communicators will also lead to improved awareness among the community-at-large that

nonverbal vocalizations that occur without typical verbal speech are communication that should be acknowledged and responded to appropriately.

Acknowledgments

The authors would like to thank the participants in this study. Amanda O'Brien and Ayelet Kershenbaum provided feedback on the study design and data analysis. Michelle Luo and Yuji Chan contributed to instructional materials used with the study. Kristina Johnson and Thomas Quatieri were participants in the study.

References

[1] H. Tager-Flusberg and C. Kasari, "Minimally verbal school-aged children with autism spectrum disorder: The neglected end of the spectrum," *Autism Res.*, vol. 6, no. 6, pp. 468–478, 2013.

[2] P. M. Dietz, C. E. Rose, D. McArthur, and M. Maenner, "National and state estimates of adults with autism spectrum disorder," *J. Autism Develop. Disord.*, vol. 50, no. 12, pp. 4258–4266, Dec. 2020.

[3] M. D. Kogan et al., "The prevalence of parent-reported autism spectrum disorder among US children," *Pediatrics*, vol. 142, no. 6, 2018, Art. no. e20174161.

[4] J. Narain et al., "Personalized modeling of real-world vocalizations from nonverbal individuals," in *Proc. Int. Conf. Multimodal Interaction*, 2020, pp. 665–669.

[5] A. Batten, "Inclusion and the autism spectrum," *Improving Sch.*, vol. 8, no. 1, pp. 93–96, 2005.

[6] E. Marchi, B. Schuller, A. Batliner, S. Fridenzon, S. Tal, and O. Golan, "Emotion in the speech of children with autism spectrum conditions: Prosody and everything else," in *Proc. 3rd Workshop Child Comput. Interaction*, 2012.

[7] E. Marchi et al., "Typicality and emotion in the voice of children with autism spectrum condition: Evidence across three languages," in *Proc. 16th Annu. Conf. Int. Speech Commun. Assoc.*, 2015.

[8] O. Rudovic et al., "CultureNet: A deep learning approach for engagement intensity estimation from face images of children with autism," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst.*, 2018, pp. 339–346.

[9] D. J. Faso, N. J. Sasson, and A. E. Pinkham, "Evaluating posed and evoked facial expressions of emotion from adults with autism spectrum disorder," *J. Autism Develop. Disord.*, vol. 45, no. 1, pp. 75–89, Jan. 2015.

[10] K. K. Hyde et al., "Applications of supervised machine learning in autism spectrum disorder research: A review," *Rev. J. Autism Develop. Disord.*, vol. 6, no. 2, pp. 128–146, 2019.

[11] R. W. Picard, "Future affective technology for autism and emotion communication," *Philos. Trans. Roy. Soc. B: Biol. Sci.*, vol. 364, no. 1535, pp. 3575–3584, 2009.

[12] S. Sarabadani, L. C. Schudlo, A. A. Samadani, and A. Kushski, "Physiological detection of affective states in children with autism spectrum disorder," *IEEE Trans. Affective Comput.*, vol. 11, no. 4, pp. 588–600, Oct.–Dec. 2020.

[13] A. Kushki, A. Khan, J. Brian, and E. Anagnostou, "A kalman filtering framework for physiological detection of anxiety-related arousal in children with autism spectrum disorder," *IEEE Trans. Biomed. Eng.*, vol. 62, no. 3, pp. 990–1000, Mar. 2015.

[14] W. L. Stone, O. Y. Ousley, P. J. Yoder, K. L. Hogan, and S. L. Hepburn, "Nonverbal communication in two-and three-year-old children with autism," *J. Autism Develop. Disord.*, vol. 27, no. 6, pp. 677–696, 1997.

[15] R. G. Gordon and L. R. Watson, "Brief report: Gestures in children at risk for autism spectrum disorders," *J. Autism Develop. Disord.*, vol. 45, no. 7, pp. 2267–2273, 2015.

[16] S. E. Colgan, E. Lanter, C. McComish, L. R. Watson, E. R. Crais, and G. T. Baranek, "Analysis of social interaction gestures in infants with autism," *Child Neuropsychol.*, vol. 12, no. 4/5, pp. 307–319, 2006.

[17] A. de Marchena and I.-M. Eigsti, "Conversational gestures in autism spectrum disorders: Asynchrony but not decreased frequency," *Autism Res.*, vol. 3, no. 6, pp. 311–322, 2010.

[18] C.-H. Chiang, W.-T. Soong, T.-L. Lin, and S. J. Rogers, "Nonverbal communication skills in young children with autism," *J. Autism Develop. Disord.*, vol. 38, no. 10, pp. 1898–1906, 2008.

[19] H. Sowden, J. Clegg, and M. Perkins, "The development of co-speech gesture in the communication of children with autism spectrum disorders," *Clin. Linguistics Phonetics*, vol. 27, no. 12, pp. 922–939, 2013.

[20] J. Hashemi et al., "Computer vision analysis for quantification of autism risk behaviors," *IEEE Trans. Affective Comput.*, vol. 12, no. 1, pp. 215–226, 1st Quart. 2021.

[21] L. Couper et al., "Comparing acquisition of and preference for manual signs, picture exchange, and speech-generating devices in nine children with autism spectrum disorder," *Develop. Neurorehabilitation*, vol. 17, no. 2, pp. 99–109, 2014.

[22] C. C. Peterson, V. Slaughter, and C. Brownell, "Children with autism spectrum disorder are skilled at reading emotion body language," *J. Exp. Child Psychol.*, vol. 139, pp. 35–50, 2015.

[23] K. Schindler, L. Van Gool, and B. De Gelder, "Recognizing emotions expressed by body pose: A biologically inspired neural model," *Neural Netw.*, vol. 21, no. 9, pp. 1238–1246, 2008.

[24] C. Holyfield, K. D. Drager, J. M. Kremkow, and J. Light, "Systematic review of AAC intervention research for adolescents and adults with autism spectrum disorder," *Augmentative Altern. Commun.*, vol. 33, no. 4, pp. 201–212, 2017.

[25] M. Romski, R. A. Sevcik, A. Barton-Hulsey, and A. S. Whitmore, "Early intervention and AAC: What a difference 30 years makes," *Augmentative Altern. Commun.*, vol. 31, no. 3, pp. 181–202, 2015.

[26] E. Zamagni, C. Dolcini, E. Gessaroli, E. Santelli, and F. Frassinetti, "Scared by you: Modulation of bodily-self by emotional body-postures in autism," *Neuropsychology*, vol. 25, no. 2, 2011, Art. no. 270.

[27] R. Cañigueral and A. F. D. C. Hamilton, "The role of eye gaze during natural social interactions in typical and autistic people," *Front. Psychol.*, vol. 10, 2019, Art. no. 560.

[28] C. Wilson, M. Brereton, B. Ploderer, and L. Sitbon, "Co-design beyond words: 'moments of interaction' with minimally-verbal children on the autism spectrum," in *Proc. CHI Conf. Hum. Factors Comput. Syst.*, 2019, pp. 1–15.

[29] C. Wilson, L. Sitbon, B. Ploderer, J. Opie, and M. Brereton, "Self-expression by design: Co-designing the expressiball with minimally-verbal children on the autism spectrum," in *Proc. CHI Conf. Hum. Factors Comput. Syst.*, 2020, pp. 1–13.

[30] M. Gratier and E. Devouche, "Imitation and repetition of prosodic contour in vocal interaction at 3 months," *Develop. Psychol.*, vol. 47, no. 1, 2011, Art. no. 67.

[31] E. Donnellan, C. Bannard, M. L. McGillion, K. E. Slocombe, and D. Matthews, "Infants' intentionally communicative vocalizations elicit responses from caregivers and are the best predictors of the transition to language: A longitudinal investigation of infants' vocalizations, gestures and word production," *Develop. Sci.*, vol. 23, no. 1, 2020, Art. no. e12843.

[32] J. McDaniel, K. D. Slaboch, and P. Yoder, "A meta-analysis of the association between vocalizations and expressive language in children with autism spectrum disorder," *Res. Develop. Disabilities*, vol. 72, pp. 202–213, 2018.

[33] L. Morgan and Y. E. Wren, "A systematic review of the literature on early vocalizations and babbling patterns in young children," *Commun. Disord. Quart.*, vol. 40, no. 1, pp. 3–14, 2018.

[34] S. R. Morris, "Clinical application of the mean babbling level and syllable structure level," *Lang. Speech Hear. Serv. Sch.*, vol. 41, pp. 223–230, 2010.

[35] D. K. Oller et al., "Automated vocal analysis of naturalistic recordings from children with autism, language delay, and typical development," *Proc. Nat. Acad. Sci. USA*, vol. 107, no. 30, pp. 13 354–13 359, 2010.

[36] E. C. Bacon, S. Osuna, E. Courchesne, and K. Pierce, "Naturalistic language sampling to characterize the language abilities of 3-year-olds with autism spectrum disorder," *Autism*, vol. 23, no. 3, pp. 699–712, 2019.

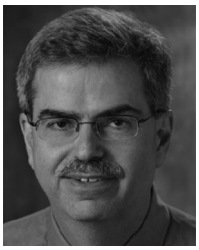
[37] A. Gregory, M. Tabain, and M. Robb, "Duration and voice quality of early infant vocalizations," *J. Speech Lang. Hear. Res.*, vol. 61, no. 7, pp. 1591–1602, 2018.

[38] S. J. Sheinkopf, J. M. Iverson, M. L. Rinaldi, and B. M. Lester, "Atypical cry acoustics in 6-month-old infants at risk for autism spectrum disorder," *Autism Res.*, vol. 5, no. 5, pp. 331–339, 2012.

[39] E. J. Tenenbaum et al., "A six-minute measure of vocalizations in toddlers with autism spectrum disorder," *Autism Res.*, vol. 13, no. 8, pp. 1373–1382, 2020.
[40] L. R. Hamrick, A. Seidl, and B. L. Tonnsen, "Acoustic properties of early vocalizations in infants with fragile x syndrome," *Autism Res.*, vol. 12, no. 11, pp. 1663–1679, 2019.
[41] G. E. Martin, J. Klusek, B. Estigarribia, and J. E. Roberts, "Language characteristics of individuals with Down syndrome," *Topics Lang. Disord.*, vol. 29, no. 2, 2009, Art. no. 112.
[42] L. Rescorla and N. B. Ratner, "Phonetic profiles of toddlers with specific expressive language impairment (SLI-E)," *J. Speech Lang. Hear. Res.*, vol. 39, no. 1, pp. 153–165, 1996.
[43] H.-M. Chiang, "Differences between spontaneous and elicited expressive communication in children with autism," *Res. Autism Spectr. Disord.*, vol. 3, no. 1, pp. 214–222, 2009.
[44] F. H. Wilhelm and P. Grossman, "Emotions beyond the laboratory: Theoretical fundaments, study design, and analytic strategies for advanced ambulatory assessment," *Biol. Psychol.*, vol. 84, no. 3, pp. 552–569, 2010.
[45] D. A. Sauter, F. Eisner, A. J. Calder, and S. K. Scott, "Perceptual cues in nonverbal vocal expressions of emotion," *Quart. J. Exp. Psychol.*, vol. 63, no. 11, pp. 2251–2272, 2010.
[46] I. Poggi, A. Ansani, and C. Cecconi, "Sighs in everyday and political communication," in *Proc. Laughter Workshop*, 2018, pp. 50–53.
[47] M. L. Knapp, *Essentials of Nonverbal Communication*. San Diego, CA, USA: Harcourt School, 1980.
[48] D. A. Sauter, F. Eisner, P. Ekman, and S. K. Scott, "Cross-cultural recognition of basic emotions through nonverbal emotional vocalizations," *Proc. Nat. Acad. Sci. USA*, vol. 107, no. 6, pp. 2408–2412, 2010.
[49] A. Anikin and C. F. Lima, "Perceptual and acoustic differences between authentic and acted nonverbal emotional vocalizations," *Quart. J. Exp. Psychol.*, vol. 71, no. 3, pp. 622–641, 2018.
[50] N. Holz, P. Larrouy-Maestri, and D. Poeppel, "The paradoxical role of emotional intensity in the perception of vocal affect," *Sci. Rep.*, vol. 11, no. 1, pp. 1–10, 2021.
[51] A. Anikin, "A moan of pleasure should be breathy: The effect of voice quality on the meaning of human nonverbal vocalizations," *Phonetica*, vol. 77, no. 5, pp. 327–349, 2020.
[52] J. Trouvain and K. P. Truong, "Comparing non-verbal vocalisations in conversational speech corpora," in *Proc. LREC Workshop Corpora Res. Emotion Sentiment Soc. Signals*, 2012, pp. 36–39.
[53] M. Schröder, "Experimental study of affect bursts," *Speech Commun.*, vol. 40, no. 1/2, pp. 99–116, 2003.
[54] O. Wasz-Höckert, T. Partanen, V. Vuorenkoski, K. Michelsson, and E. Valanne, "The identification of some specific meanings in infant vocalization," *Experientia*, vol. 20, no. 3, pp. 154–154, 1964.
[55] L. Liu, W. Li, X. Wu, and B. X. Zhou, "Infant cry language analysis and recognition: An experimental approach," *IEEE/CAA J. Automatica Sinica*, vol. 6, no. 3, pp. 778–788, May 2019.
[56] I.-A. Bănică, H. Cucu, A. Buzo, D. Burileanu, and C. Burileanu, "Automatic methods for infant cry classification," in *Proc. Int. Conf. Commun.*, 2016, pp. 51–54.
[57] S. Sharma and V. K. Mittal, "Infant cry analysis of cry signal segments towards identifying the cry-cause factors," in *Proc. IEEE Region 10 Conf.*, 2017, pp. 3105–3110.
[58] T. Fuhr, H. Reetz, and C. Wegener, "Comparison of supervised-learning models for infant cry classification/vergleich von klassifikationsmodellen zur säuglingsschreianalyse," *Int. J. Health Professions*, vol. 2, no. 1, pp. 4–15, 2015.
[59] O. Weisman et al., "Dynamics of non-verbal vocalizations and hormones during father-infant interaction," *IEEE Trans. Affective Comput.*, vol. 7, no. 4, pp. 337–345, Oct.–Dec. 2016.
[60] D. R. Beukelman et al., *Augmentative and Alternative Communication*. Baltimore, MD, USA: Paul H. Brookes, 1998.
[61] K. T. Johnson, J. Narain, C. Ferguson, R. Picard, and P. Maes, "The ECHOS platform to enhance communication for nonverbal children with Autism: A case study," in *Proc. Extended Abstr. CHI Conf. Hum. Factors Comput. Syst.*, 2020, pp. 1–8.
[62] J. Narain, K. T. Johnson, R. Picard, and P. Maes, "Zero-shot transfer learning to enhance communication for minimally verbal individuals with autism using naturalistic data," in *Proc. AI Soc. Good Workshop NeurIPS*, 2019.
[63] K. T. Johnson, J. Narain, T. Quatieri, R. Picard, and P. Maes, "ReCANVo: A database of real-world communicative and affective nonverbal vocalizations," *Submitted for publication*, 2022.
[64] K. T. Johnson et al., "Phonemic content of nonverbal vocalizations from individuals with 0-10 spoken words," *INSAR*, 2022.
[65] J. Narain, "Interfaces and models for improved understanding of real-world communicative and affective nonverbal vocalizations by minimally speaking individuals," Ph.D. dissertation, Dept. Mech. Eng., Massachusetts Inst. Technol., Cambridge, MA, 2021.
[66] G. Lemaître, F. Nogueira, and C. K. Aridas, "Imbalanced-learn: A python toolbox to tackle the curse of imbalanced datasets in machine learning," *J. Mach. Learn. Res.*, vol. 18, no. 1, pp. 559–563, 2017.
[67] M. Freitag, S. Amiriparian, S. Pugachevskiy, N. Cummins, and B. Schuller, "auDeep: Unsupervised learning of representations from audio with deep recurrent neural networks," *J. Mach. Learn. Res.*, vol. 18, no. 1, pp. 6340–6344, 2017.
[68] F. Eyben, F. Weninger, F. Gross, and B. Schuller, "Recent developments in openSMILE, the Munich open-source multimedia feature extractor," in *Proc. 21st ACM Int. Conf. Multimedia*, 2013, pp. 835–838.
[69] F. Eyben et al., "The Geneva minimalistic acoustic parameter set (GeMAPS) for voice research and affective computing," *IEEE Trans. xAffective Comput.*, vol. 7, no. 2, pp. 190–202, Apr.–Jun. 2016.
[70] D. D. Mehta, D. Rudoy, and P. J. Wolfe, "Kalman-based autoregressive moving average modeling and inference for formant and antiformant tracking," *J. Acoustical Soc. Amer.*, vol. 132, no. 3, pp. 1732–1746, 2012.
[71] Y.-L. Shue, "The voice source in speech production: Data, analysis and models," Univ. California, Los Angeles, Los Angeles, CA, 2010.
[72] M. Jaiswal and E. M. Provost, "Best practices for noise-based augmentation to improve the performance of emotion recognition "in the wild"," 2021, *arXiv:2104.08806*.
[73] X. Li et al., "Universal phone recognition with a multilingual allophone system," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, 2020, pp. 8249–8253.
[74] B. W. Schuller, "Speech emotion recognition: Two decades in a nutshell, benchmarks, and ongoing trends," *Commun. ACM*, vol. 61, no. 5, pp. 90–99, 2018.

Jaya Narain received the BS, MS, and PhD degree from the Massachusetts Institute of Technology (MIT), in 2015, 2017, and 2021, respectively. She conducted her doctoral research in the Fluid Interfaces group in the MIT Media Lab. She was named an Apple Scholar in AI/ML, in 2020 and an NSF Graduate research fellow, in 2016. Her research interests are machine learning and human-centered design for accessibility and health applications, and particularly real-world data collection and personalized systems.

Kristina T. Johnson received the bachelor's and master's degrees with highest honors in physics, and the PhD degree from the Massachusetts Institute of Technology (MIT), in 2021, where she was a member of the Affective Computing Group, MIT Media Lab. She is a multi-disciplinary researcher whose work lies with the intersection of neuroscience, engineering, computer science, affective sciences, technology development, and clinical applications, especially for individuals with complex neurodevelopmental differences, genetic disorders, or autism. She was twice named an MIT Hugh Hampton Young Fellow and three times named an MIT Media Lab Learning Innovation Fellow during her doctoral studies. As a physicist, she was awarded the national Barry M. Goldwater Scholarship.



Thomas F. Quatieri (Fellow, IEEE) received the BS degree (summa cum laude) from Tufts University, and the SM, EE, and ScD degrees from the Massachusetts Institute of Technology (MIT). He is a senior member of the technical staff with MIT Lincoln Laboratory, Lexington, focused on speech and auditory signal processing and neuro-biophysical modeling with application to detection and monitoring of neurological, neurotraumatic, and stress conditions. He holds a faculty appointment in the Harvard-MIT Speech and Hearing Bioscience and Technology Program. He is an author on more than 200 publications, holds 12 patents, and authored the textbook Discrete-Time Speech Signal Processing: Principles and Practice. He is a recipient of four IEEE Transactions best paper awards and the 2010 MIT Lincoln Laboratory Best Paper Award. He led the Lincoln Laboratory team that won the 2013 and 2014 AVEC Depression Challenges and the 2015 MIT Lincoln Laboratory Team Award for their work on vocal and facial biomarkers. He has served on many IEEE signal processing and speech technical committees and currently is an associate editor of *Computer, Speech, and Language*. He is a member of Tau Beta Pi, Eta Kappa Nu, Sigma Xi, ICSA, ARO, ASA, and SFN.

Rosalind W. Picard (Fellow, IEEE) received the BS degree in electrical engineering from the Georgia Institute of Techology, and an SM and ScD degrees in electrical engineering and computer science from MIT. She is professor of media arts and sciences with MIT, founder and director of the Affective Computing Research Group, MIT Media Lab, and co-founder of the startups Affectiva and Empatica. In 2019, she received one of the highest professional honors accorded an engineer, election to the National Academy of Engineering for her contributions on affective computing and wearable computing. She is credited with starting the branch of computer science known as affective computing with her 1997 book of the same name. She is author of the book Affective Computing, and author or co-author of over 350 scientific articles (more than 66k citations).

Pattie Maes (Member, IEEE) received the PhD degree in computer science from the University of Brussels, Belgium. She is a professor in MIT's Program in media arts and sciences and until recently served as its academic head. She runs the Media Lab's Fluid Interfaces research group, which aims to radically reinvent the human-machine interaction. Coming from a background in artificial intelligence and human-computer interaction, she focuses on immersive and wearable systems that can actively assist people with memory, attention, learning, decision making, communication, and well-being. She is the editor of three books, and is an editorial board member and reviewer for numerous professional journals and conferences. Fast Company named her one of 50 most influential designers (2011); Newsweek picked her as one of the "100 Americans to watch for" in the year 2000; TIME Digital selected her as a member of the "Cyber Elite," the top 50 technological pioneers of the high-tech world; the World Economic Forum honored her with the title "Global Leader for Tomorrow"; Ars Electronica Awarded her the 1995 World Wide Web category prize; and in 2000 she was recognized with the "Lifetime Achievement Award" by the Massachusetts Interactive Media Council.