Interfaces and Models for Improved Understanding of Real-World Communicative and Affective Nonverbal Vocalizations by Minimally Speaking Individuals

by

Jaya Narain

S.B., Massachusetts Institute of Technology (2015)

S.M., Massachusetts Institute of Technology (2017)

Submitted to the Department of Mechanical Engineering
In Partial Fulfillment of the requirements for the Degree of

Doctor of Philosophy

at the

Massachusetts Institute of Technology

September 2021

Author……..
Department of Mechanical Engineering
August 4, 2021

Certified by…….
Pattie Maes
Professor of Media Arts and Sciences
Thesis Supervisor

Accepted by…….
Nicolas Hadjiconstantinou
Professor of Mechanical Engineering
Graduate Officer

Interfaces and Models for Improved Understanding of Real-World Communicative and Affective Nonverbal Vocalizations by Minimally Speaking Individuals

by

Jaya Narain

Submitted to the Department of Mechanical Engineering
on August 4, 2021, in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy in Mechanical Engineering

Abstract

This work focuses on a sub-group (denoted by mv*) of non- and minimally speaking individuals who have fewer than 10 words or word approximations and limited expressive language through speech and writing. In the United States alone, this group comprises over one million individuals. Their nonverbal vocalizations (i.e., vocalizations that do not have typical verbal content) often have self-consistent phonetic content and 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 these nonverbal vocalizations could contribute to the development of technology to augment communication. This thesis aims to help the community at-large better understand and communicate with mv* individuals by utilizing families' unique understanding of nonverbal vocalizations.

For this work, families provided personalized labels for vocalizations, which were then used to compile a novel dataset and train machine learning models. The thesis contributes (1) the design and evaluation of a novel data collection protocol for real-world audio with personalized in-the-moment labels, (2) a new dataset, ReCANVo, of over 7,000 nonverbal vocalizations from eight mv* communicators, collected longitudinally in real-world settings, (3) machine learning evaluation strategies and algorithms suitable for messy, real-world data that can classify vocalizations from mv* individuals with F1-scores above chance, and (4) the design of a novel communication interface, based on these interviews, surveys, and data analyses. The presented dataset ReCANVo is the only dataset of nonverbal vocalizations from mv* individuals, the largest dataset of nonverbal vocalizations, and one of the first datasets capturing real-world emotions across settings. The presented data analyses show, for the first time, that it is possible for models to classify nonverbal vocalizations by mv* individuals by function using audio alone. While this work was motivated by impact for a small, specialized population, the results can inform the design of real-world data collection and modeling approaches more broadly.

Thesis Supervisor: Pattie Maes
Title: Professor of Media Arts and Sciences

Acknowledgements

- Thank you to the individuals and families who participated in research studies, interviews, and survey for sharing your time and your perspective. I am very grateful for your engagement and involvement. This work would not have been possible without you.

- Thank you to my advisor Pattie Maes, doctoral committee chair Maria Yang, and committee members Rosalind Picard and John Leonard for their support and guidance throughout this work. I have learned a lot from working with all of you - thank you for your mentorship!
    - Pattie, thank you so much for giving me the freedom to explore research questions that I am passionate about while providing advice and guidance along the way. I deeply appreciate all your support and encouragement. I have learned a lot from working with you both technically and beyond. I feel very lucky to have had you as a PhD advisor, and to have been a part of Fluid.
    - Maria, thank you for your support as my committee chair and for providing a home base for me in MechE. I really appreciate all your advice and feedback on this work, and it has been awesome being part of the Ideation lab community.
    - Roz, thank you for your feedback and support. Thank you for all your careful feedback on papers and this thesis. I really appreciate all your encouragement in pursuing this field of work and your support of this research, it has meant a lot to me!
    - John, thank you for your advice and for your support of my work here and with ATHack. I'm really glad to have had you on my committee and appreciate your feedback and engagement!

- I have been lucky to work with a lot of wonderful collaborators and undergraduate research students along the way. Thank you for being part of this work.
    - This thesis includes a lot of work conducted jointly with Kristy Johnson. Kirsty, thank you for being a wonderful collaborator and friend! I have really enjoyed working with you and building this research project together. I will miss our long online chats and our pre-pandemic impromptu office sync-ups. I am so glad I got to work with you!
    - Tom Quatieri, Amanda O'Brien, Ayelet Kershenbaum, and Janet Baker - thank you for being part of this work and for sharing your ideas and feedback. It has been a pleasure working with you.
    - Hyeyoung Shin, Peter Wofford, Yuji Chan, and Michelle Luo - thank you for M. Eng-ing and UROPing on this project.

- Thank you to all the members of the Fluid Interfaces research group - Nataliya, Angela, Neo, Camilo, Utkarsh, Guillermo, Arnav, Mina, Adam, Abhi, Pat, Joanne, Caitlin, Guarav, Aubrey, Judith, Holly, Janet, Eric, Tomas, Oscar, Jess, Kai, Vlad, Nora, and Nirmita. I have had so much fun working and sharing ideas with all of you.

- Thank you to the mechanical engineering and Media Lab communities - and especially the Ideation Group, the Affective Computing Group, and GEAR Lab - for your support along the way.

- Thank you to the organizations who have sponsored this work, particularly:
    - The Apple Scholars in AI/ML program and my Apple research mentor, Shirley
    - The NSF Graduate Research Fellowship program
    - The Deshpande Center Technology to Improve Ability Program
    - The Microsoft AI for Accessibility program
    - The MIT Media Lab

- Thank you to my friends and family for all your support and being part of this adventure - Alexandriya, Ishwarya, Felicia, Vivian, Anisha, Ayushi, Easha, Linda, Derek, David, Michael, Kaitlyn, Hosea, Erin, and many others!

- Thank you to Memo, my wonderful husband, for your love and support in this journey. Thank you for helping me through difficult moments and celebrating with me in happy ones. I love you!

Table of Contents

List of Figures

List of Tables

1 Introduction

1.1 Motivation

In the United States alone, over one million people are non- or minimally speaking with respect to verbal language meaning they use zero or fewer than 20 words/word approximations, respectively [1], [2]. This thesis focuses specifically on those individuals who have 10 or fewer spoken word/word approximations. This population includes approximately 30% of individuals with autism [1], in addition to some individuals with Down Syndrome [3], Rett Syndrome [4], Mowat-Wilson [5], Rubinstein-Taybi syndrome [6], Pitt-Hopkins syndrome [7], and other conditions associated with differences in speech and language. Individuals who are non-speaking communicate richly through many means, including augmentative and alternative communication (AAC) devices, gestures and vocalizations.

Nonverbal vocalizations from mv* individuals often have self-consistent phonetic content and others vary in tone, pitch, and duration depending on the individual's emotional or physical state or intended communication. Mv* individuals are understudied in both clinical and design literature; many have co-occurring intellectual disabilities and genetic conditions and are excluded from research studies. This thesis includes, to my knowledge, the first studies of intent and affect in real-world vocalizations that do not have typical verbal content from mv* individuals.

Augmentative and Alternative Communication (AAC) devices and tools are technologies that help facilitate communication, ranging from low-tech systems like a pen and paper to high-tech systems that use tablets and computers to generate speech [8]. At the start of this research, I conducted interviews and surveys with people with differences in speech and language abilities and their families to understand the variety in user needs for AAC devices and identify trends. Common problems reported with available commercial devices included not being able to make input selections due to motor difficulties, exhaustion due to the input mechanism, limited use environments for devices (e.g., can't be used outdoors), and difficulty discriminating between options. Survey respondents reported that that their communication was often misunderstood, and said that communicating in unfamiliar settings was particularly challenging [9]. Many respondents – particularly those for whom existing AAC devices have limited functionality – expressed interest in a device to help communicate emotions, stress, and pain.

Furthermore, interviewed families of mv* individuals said that they have a unique ability to understand the mv* communicator's nonverbal communication style – including vocalizations, gestures, and body language. However, others (e.g., teachers, new caregivers) often do not understand nonverbal communication well and may not respond appropriately to the communicator's emotions and needs. Parents of mv* children cited this miscommunication with people who do not know their child well as a source of stress.

Nonverbal vocalizations can be a form of language within a family culture. Those who do not know a communicator well are often unaware that they should listen to and respond to nonverbal vocalizations. By collecting real-world data, and making the data and study results publicly available, I hope to contribute to expanding awareness about a critical but often unfamiliar

communication modality. This work was primarily motivated by creating technology to help the community at-large better understand and communicate with mv* individuals by utilizing families' unique understanding of nonverbal vocalizations from mv* communicators. This thesis also contributes a new dataset of nonverbal vocalizations from mv* individuals with personalized affect and intent labels, the design and evaluation of a novel data collection protocol and communication interfaces, and machine learning models and evaluation strategies to classify real-world vocalizations from mv* individuals.

While research towards addressing this gap via technology development for mv* individuals is the focus of this thesis, inclusive cultures and social structures are also critical components of inclusive communication. These directions are not explored in depth in this thesis, but some citations are provided for interested readers: [10]–[16].

1.2 Mv* Communicators

The presented work was motivated by developing technology specifically for individuals who are non- or minimally speaking with respect to verbal language, who use fewer than 10 spoken words or word approximations, and who have limited expressive language through speech and writing. This population is abbreviated using "mv*" throughout this work. The term "mv*" is meant to be more specific than "non- and minimally speaking" which can include individuals who do not speak but have no differences in expressive language abilities via writing. The population of mv* individuals may include individuals with no differences in receptive language abilities, and individuals with expressive language that is expressed through AAC devices.

Mv* communicators included in this work had fewer than ten spoken words or word approximations. Many of the communicators had fewer than ten word or word approximations

communicated via AAC. Some mv* communicators had 100+ words through communication apps like Proloquo2Go and TouchChat. Participating families self-identified whether they felt the research and technologies explored in this thesis were applicable for the mv* individual's communication style.

The designation "with respect to verbal language" is to acknowledge that mv* individuals do communicate via speech. The term "nonverbal vocalizations" is used to reference speech from mv* individuals that does not have typical verbal content. All forms of communication – whether there is a clear mapping to language or not – are rich and meaningful.

The population of mv* individuals includes some individuals with autism. Some Autistic individuals prefer identify-first language that emphasizes and respects Autism as an integral part of identity, and not a disability [17]; others prefer person-first language that emphasizes the person first (e.g., "person with autism"), with other differences as secondary qualifiers [18]. Because identity is deeply personal, to the best of my knowledge, I mirror the language chosen by the individuals and families referenced in a particular section.

1.3 Design Approach

Human-centered design and participatory co-design design approaches were core to the design and development work presented herein. The work began with a human-centered design approach through interviews and a web survey with people with differences in communication abilities, and their loved ones and communication partners. Interviews with AAC communicators and their families were conducted throughout the design and development process to get feedback on the utility and appropriateness of the research topics and proposed designs. When possible, first-

person feedback was obtained directly from the AAC communicators. For instance, interviews early in the design process included adults with muscular dystrophy and cerebral palsy who communicated via typing, writing, eye gaze, and other modalities and were conducted directly with the AAC device users. While first-person feedback from AAC communicators is preferable, it was not always possible. As the focus on mv* communicators emerged, it was often necessary to rely on parents to convey communication practices and preferences because of the limited utility of existing communication tools and strategies and because most participants were young children. This is a limitation of this work but was necessary to enable working with the underserved population of mv* communicators.

A large portion of the development of this work was conducted with a participatory co-design approach with one mv* communicator and their family, who tested processes and technology in their day-to-day lives. The mv* communicator provided feedback on study designs through facial expressions, body language, and gestural communication. The communicator's family also provided extensive feedback on the feasibility of study activities with respect to time commitment, on the usability of tools and apps used in the study, and on the utility and appropriateness of both study and interface designs. Interviews (represented by small branches extending from the trunk in Figure 1) with other families were conducted throughout this development process, as "check-ins" to understand whether and how the feedback and insights provided from the family engaged in the longitudinal case study might extend to more mv* communicators.

Data collection was then scaled up slowly, and eventually included eight families. All participating families also provided feedback on the study design and overall research and design directions throughout their engagement in the longitudinal study. Through this engagement, the designs and

interfaces presented here have been updated iteratively to address the needs and suggestions expressed by participating families.

The designs explored in this thesis include communication interfaces that use nonverbal vocalizations as well as a system to collect and label data in real-world settings. Families participating in the longitudinal study recorded continuous audio in real-world settings and provided personalized affect and intent labels in-the-moment using a custom cell phone app. Machine learning models were trained using parent labels to classify the affect and intent of recorded nonverbal vocalizations. The resulting dataset, ReCANVo, provide insights into understanding this type of communication in real-world settings. The collected data could be used to track the evolution of a communicator's speech over time. The presented methods and analysis could enable data collection at a larger scale which could provide insights into language development

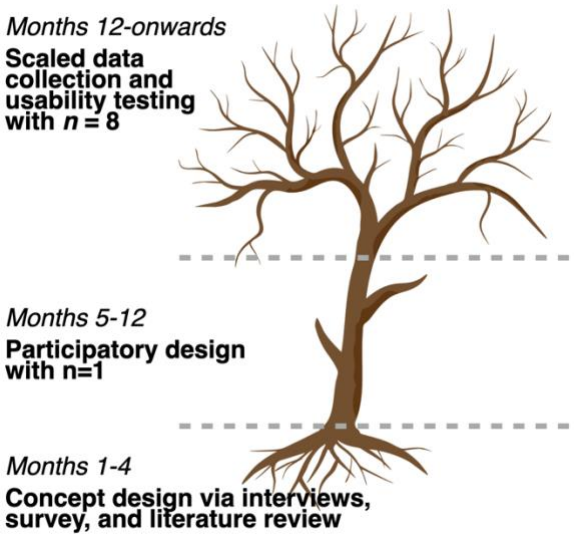


Figure 1: A tree illustrating the participatory design process.

The roots at the bottom of the tree represent interviews and survey outreach at the beginning of the research process. Afterwards, we engaged in a long-term participatory design process with one family, in parallel with continued interviews with other families (represented by the branches on the tree trunk). Data collection and usability was slowly scaled up, with a focus on minimizing time burden and collecting representative feedback and data from mv communicators and families.*

trajectories and could be used to explore similarities in communication practices between mv* individuals.

Mv* communicators have diverse needs and communication practices. This work focused on a subset of mv* communicators, and the designs and approaches described in this thesis will not be relevant for all mv* communicators. Some mv* communicators do not communicate using vocalizations and prefer to use other communication modalities including gestures and AAC devices.

1.4 Thesis Summary

This thesis begins by exploring AAC usage and opportunities for improved AAC technologies through exploratory interviews and a web survey with AAC device users and their loved ones and communication partners. The interviews and survey examined what aspects of currently available technologies work well, and what aspects need improvement. The interviews and surveys included anyone who uses or might consider using AAC devices and were not focused solely on mv* communicators. From the interview and survey results, it was apparent that there are many ways in which communication technology and practices can be improved for AAC device users. Based on what I learned from AAC device users and their families and from a literature view on available technologies and previous studies, I chose to focus on the population of mv* communicators.

To further develop ideas and expand my experience with mv* communication, I co-conducted a seven-month case study with Kristy Johnson with an 8-year-old mv* communicator and his family. The case study began with exploring the relationship between physiology and communication in the real-world, an extension of numerous laboratory studies with Autistic people and physiological data

[19]–[24]. We observed some interesting intersections between communication and physiology, but the sensors were too uncomfortable and expensive for day-to-day life. The mv* communicator did not like wearing the sensors, and the communicator's families were concerned the equipment might be easily damaged when the child was playing. During the study, we observed that the mv* individual frequently vocalized when playing and interacting with others. As an external observer, I often did not understand what a vocalization was expressing, but the mv* communicator's family could consistently differentiate vocalizations based on their affect and intent. Interviews with other families confirmed these observations – wearable sensors are very uncomfortable for many mv* individuals, and nonverbal vocalizations are communicative but are hard to understand by people who don't know the communicator well – were relevant for other mv* communicators as well.

We then explored recording both audio and video to capture vocalizations and detailed contextual information. Early in the design process, we placed a strong emphasis on low-cost recording technologies for scalability. As we continued the case study, we learned that it was too intrusive to ask a family to record video in their day-to-day life, and that it was more important to optimize for high data quality than low equipment cost for a first study in this topic. We then began refining a protocol to collect real-world vocalization data by mv* communicators with personalized in-the-moment labels, and slowly scaled up the study to include more families. Kristy and I continued working closely together. Much of the work presented in this thesis on designing and developing interfaces and technologies to study and understand vocalizations is collaborative work conducted with Kristy.

After identifying that helping the community at-large understand nonverbal vocalizations from mv* communicators was an impactful direction for research and technology development, we looked for

related work from researchers and clinicians. While some practicing speech and language pathologists (SLPs) reference encouraging and responding to nonverbal vocalizations as part of a total communication approach [25], there was no prior work in the literature that collected and analyzed vocalization data by mv* communicators labeled for affect and intent. We began the long-term project of collecting the ReCANVo dataset, the first dataset of nonverbal vocalizations from mv* communicators, labeled by close family members and caregivers in-the-moment for affect and intent. In designing the data collection process, we prioritized ease of data collection and flexibility, to place minimal time burden on participating families, who are part of a specialized population that is often time-burdened and underserved by available resources. The data collection process was also designed to be done remotely, to allow us to reach a geographically distributed population. While these design choices had some trade-offs with respect to consistency in labeling and data collection practices across participants, they were critical to collecting data with a small, specialized population, particularly amidst a global pandemic. The ReCANVo dataset includes nonverbal vocalizations from eight mv* communicators, collected amidst day-to-day life. To our knowledge, it is: the only dataset collected with mv* communicators, the largest available dataset of nonverbal vocalizations, and one of the only affective audio datasets collected amidst day-to-day life, capturing organic motivation-driven communicative and emotional expressions.

I used the collected dataset to analyze labeling and data collection practices amongst participating families. Studying how families labeled and collected data can help inform other studies on recording and labeling data in real world settings, with mv* communicators and beyond. There were many differences in how families chose to label and record data with respect to environments and activities, number of sessions, labeling delays, and labeling frequencies. In-the-moment labels can be affected by multi-tasking, mishits, and human delays in reacting to the occurrences of day-

to-day life. Audio recorded in the real-world has varying background sounds and recording soundscapes, which can also affect data analysis. The labels used in this study were designed for affect and intent and not to have a 1:1 mapping to particular sound – for instance, there can be many vocalizations that express frustration but are distinct from each other. The collected dataset is relatively small compared to the size of data often used in machine learning tasks. The dataset also consists of a heterogeneous population of mv* communicators, spanning different ages, genders, and diagnoses.

For these reasons, the analysis presented in this thesis focused on personalized supervised learning methods suitable for dealing with a small, heterogenous, and messy dataset. Evaluation strategies were developed and implemented to address noise and variability in the data. For each participant, models are evaluated using leave one session out and cross-validation approaches. Because participants upload sessions delineated by time and/or activity, leave one session out evaluation reduces the model's ability to perform well by fitting to background noise in a recording. Rough session stratification was also used to reduce the model's propensity of fitting to background noise. Random forest and support vector machine (SVM) models were trained using commonly used acoustic feature sets and a custom feature set designed for this task. The trained supervised models consistently had f1 scores above chance for multi-class classification, indicating that there are acoustic features in the recorded vocalizations that can separate vocalizations by class. There were large variations in model performance between participants, which could be due to differences in the number of samples available for training, data quality, labeling fidelity and practices, and inherent differences in vocal communication practices. Experiments with TrAdaboost and with transfer learning with deep neural networks (DNN) suggested that, for the participants in this study, fully personalized models had generally better performance compared to

transfer learning approaches. Interviews with participating families informed differences in labeling and data collection processes between families.

An integrated app with classification and listening features was designed and tested iteratively. The app was designed collaboratively with Kristy Johnson and Craig Ferguson. The app went through numerous revisions during pilot testing with a family with an mv* communicator. A first-generation version of the app was able to classify vocalizations recorded using the app with 90% accuracy as positive or negative valence. These results are preliminary and further evaluation is needed to evaluate the app's performance longitudinally and in diverse environmental settings. An integrated version of the app was then created to integrate data labeling and recording. The integrated version of the app was iteratively tested and updated via pilot testing with a family with an mv* communicator who tried the app in day-to-day life. During interviews, seven families of mv* individuals provided feedback on the design concepts.

In the future, additional work that builds from this thesis could include further evaluation of the utility of the app and the accuracy of the classification feature. Additional implementations beyond a standalone app could also be explored, like incorporating personalized nonverbal vocalization classification and controls as an add-on to more general voice control system (e.g., Siri, Alexa) and/or integrated with communication apps (e.g., Proloquo2Go, TouchChat). Allowing for further customization with respect to the available labels could help make the technology more useful for a larger population beyond mv* communicators. For the classification feature, implementing active learning strategies to automate the training and updating of models longitudinally could be explored. The ReCANVo dataset could be used to explore prosodic and acoustic patterns towards an improved understanding of language development and vocal communication by mv* individuals.

The foundations of this work – looking at the syllabic structure and prosodic contours of vocalizations by function – is underway, along with collaborators. Continuing data collection in real-world settings with mv* communicators would allow for an expanded exploration of scientific questions, including exploring whether there are groupings of individuals with overlapping nonverbal vocalization communication practices.

1.5 Research Questions and Contributions

This thesis explores research questions related to the design of augmentative communicative interfaces and nonverbal communication, including:

1. What are limitations of available AAC technology? Who do those limitations affect and how?
2. How do families record and label data in real-time using a novel data collection system with a wearable recorder and a labeling app?
    a. Are data collection practices between users different?
    b. How are labels distributed among sessions?
    c. How long do assigned labels last, and how many vocalizations does each label include?
    d. How long does it take a parent or close family member to assign a label to a vocalization in real-time using a custom labeling app (i.e., what is the delay between the mv* communicator vocalizing and the labeler assigning a label)?
3. Can vocalizations from mv* individuals be classified using audio alone (using parent labels)?
    a. How do differences in data collection practices inform modeling results?
    b. How can models be evaluated with small datasets and noisy, real-world data? How does the evaluation method affect model performance?

    c. How does the model performance vary with extracted features? Particularly, how does a proposed custom feature set tailored to vocalizations from mv* individuals perform compared to traditional acoustic feature sets?
4. Do transfer learning approaches improve model performance? What do the results of modeling approaches with transfer learning suggest regarding overlap in expression among mv* individuals and between mv* individuals and others?
5. Are mv* individuals and their families interested in communication interfaces that use nonverbal vocalizations? What device features might be useful to families, based on interview feedback?

RQ1 is primarily explored in Chapter 2, RQ2 in Chapter 7, RQ3 in Chapter 8, RQ4 in Chapter 9, and RQ5 in Chapter 10. Additionally, this thesis presents a novel dataset of nonverbal vocalizations, machine learning evaluation strategies for messy real-world data, and the design and evaluation of data collection and labeling systems and communication interfaces. The contributions of this thesis include:

- The identification of user needs and research opportunities for AAC via interviews, survey, and a case study along with the compilation and publication of survey results for use by other researchers
- The development and evaluation of a novel data collection protocol for real-world audio with personalized in-the-moment labels
- The creation of the first database of nonverbal vocalizations with affective and communicative labels from mv* individuals

- The development of machine learning evaluation strategies and algorithms suitable for real-world data that can classify nonverbal vocalizations from mv* individuals with F1 scores above chance
- The design and evaluation of novel communication interfaces towards enhanced understanding of nonverbal vocalizations

1.6 Thesis Roadmap

Chapter 2 provides background on AAC technology and presents interviews and a web survey towards understanding AAC usage and identifying areas for further exploration. Chapter 3 provides an overview of related work on technology development for mv* communicators, nonverbal vocalizations, and speech and emotions. References to related work are also included in each chapter as appropriate. Chapter 4 presents and discusses a seven-month longitudinal case study with one mv* communicator and their family. Chapter 5 presents designs for novel communication interfaces and a system for collecting real-world audio data with in-the-moment labels. Chapter 6 presents the ReCANVo dataset that was collected as part of this thesis. The ReCANVo dataset is the first dataset of nonverbal vocalizations from mv* communicators with affect and intent labels. Chapter 7 presents an analysis of the real-world data collection system including statistics describing labeling practices and interviews with participating families. Chapter 8 and Chapter 9 present machine learning models to classify nonverbal vocalizations, evaluation strategies for messy, real-world data, and a discussion of the results in the context of real-world data collection. Chapter 10 presents a prototype of a novel communication interface, and preliminary evaluation of the interface via interviews and real-world testing. Chapter 11 describes future directions of research, presents final conclusions, and summarizes the contributions of this thesis.

2 AAC Usage and Needs

The American Speech-Language-Hearing Association defines Augmentative and Alternative Communication (AAC) as “all the ways we share our ideas and feelings without talking”. AAC includes nonverbal communication like gestures, expressions, and body language. AAC devices include technologies that help facilitate communication, ranging from low-tech systems like a pen and paper to high-tech systems that use tablets and computers to generate speech [8].

Dedicated AAC devices are most often used by people who have differences in speech and language abilities. Speech-related differences may be related to articulation, raspy speech, and stuttering, but not problems rooted in language comprehension and production. Conditions affecting speech include laryngectomy, muscular dystrophies, amyotrophic lateral sclerosis (ALS), multiple sclerosis (MS), cerebral palsy (CP), stroke, and traumatic brain injuries. In contrast, differences in language abilities include difficulties with the form, content, and use of language. This includes differences in understanding and forming sentences, different timelines when developing language, and differences in interacting with and understanding social communication norms [26]. Language-related differences can be associated with stroke, aphasia, autism, and other developmental and neurological differences. Differences in language abilities may affect expressive language (expressing ideas with language, receptive language (comprehending language), or both [27].

This chapter includes an overview of commercially available AAC devices (Section 2.1), a summary of prior research on AAC usage habits and user needs (Section 2.2), results from exploratory interviews with AAC users (Section 2.3), results from a survey on AAC user needs conducted with AAC users and/or their loved ones (Section 2.4), and a discussion of possible directions for further work in AAC development based on the presented interviews, survey results, and prior work (Section 2.5 and 2.6) [9].

2.1 AAC Devices: Commercially Available Technology

Many commercially available AAC devices ask a user to select from labeled icons and symbols to generate speech, form words by selecting individual characters, or select from buttons and switches to generate specified outputs (Figure 2). AAC devices can also include keyboards and tablets for communication via written text, and alternative inputs for electronic devices (e.g., computers, tablets, mobile phones) like eye-tracking. A summary of some commonly used commercial devices is provided below. The categories and devices described below are not exhaustive or mutually exclusive – there are many available devices and tools, and often products can be used in multiple ways.

- Icon and symbol-based AAC: Popular apps like Proloquo2Go [28] and TouchChat [29] allow users to output speech from a general-use tablet. Other symbol-based AACs like NOVA chat [30] use a dedicated device. App and electronic-based AACs generally have expansive vocabularies and often allow users to create custom vocabulary groupings. Icon and symbol-based communication can also include non-electronic tools, like the Picture Exchange Communication System (PECS) [31].

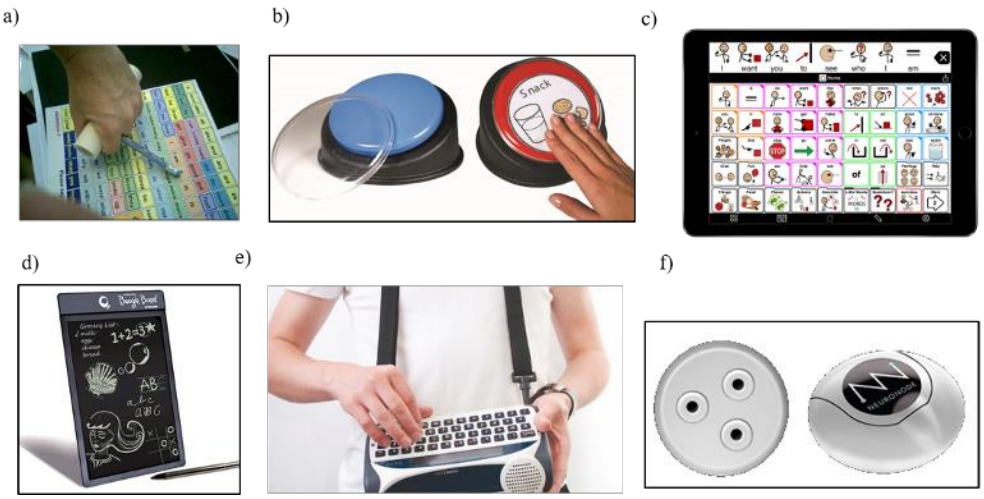


Figure 2: Images of various AAC devices

a. a physical word board, b. BigTalk audio buttons, c. Proloqou2Go app, d. erasable tablet and stylus, e. Lightwriter SL40, f. NeuroNode; Images a, d from amyandpals.com, b from enablingdevices.com, c from prmac.com, e from tobbidynavox.com, and f from controlbionics.com

- Typing and writing devices: AAC also includes devices that enable communication via text and writing, like general purpose keyboards, smartphones, and stylus writing tools. Some devices include features like text to speech. There are a variety of dedicated communication devices and add-on sensors that allow for varied motor inputs when typing, like Tobii eye tracking technology [32].
- Buttons and tactile devices: Switches and buttons are used as an alternative control for electronic devices and to output specific messages and actions [33], [34].
- Gestural input: Gestural input devices can be used to interact with electronics and trigger specified outputs. For example, NeuroNode [35] and Pison [36] are wearable devices that recognize gestural inputs and interface with e-mail, text messaging, and social media. For instance, a user can train a device to send a specific message if it recognizes a particular arm gesture.

2.2 Background and Prior Work

To better understand AAC usage and opportunities for research, I conducted interviews and a web survey with individuals who use AAC devices and their families [9]. When possible, interviews were conducted with the individual who communicated via AAC. Family members were included as interviewees and survey responders because they have experience communicating with an AAC user and can provide valuable feedback and insights particularly around communication practices for children and for individuals with differences in expressive language abilities. Responses from family members of a device user were and should be interpreted with care, as they are inherently biased and may not represent the needs and opinions of device users themselves. Working directly with users from early in the design stage is irreplaceable. The interviews and survey were primarily meant to help the researcher learn more about AAC usage and diverse communication practices and motivate possible areas for further research and evaluation. The study was approved by MIT's Committee on the Use of Humans as Experimental Subjects.

Prior studies have explored the usage and design of AAC devices. Elsahar et. al. [37] provided a comparison of the strengths and limitations of signal sensing modalities used in high-tech AAC. Elsahar concluded that affordability, inflexibility of technologies to be used for different needs, and high support requirements negatively impact the adoption of high-tech AAC. Moorcroft et. al. [38] conducted a qualitative analysis of 43 articles and found large effects on AAC usage with environmental factors (e.g. attitudes of friends, family and society), personal factors (e.g. socio-economic status) and body factors (e.g. cognitive ability). Baxter et. al. [39] analyzed 27 studies on AAC usage and found barriers to usage included attitudes of family members, complex device input, and device voice. They highlighted the need to include AAC device users and their families early in the process.

A number of prior studies have examined the usability of AAC devices from the perspective of device users [40], their families [41], [42], and speech language pathologists [43]. Here, I show results from interviews and surveys with both device users and their families – with a focus on understanding technologically what designs might be most usable and acceptable by the respondents – to help inform research efforts.

2.3 Exploratory Interviews

Fifteen conversational interviews of approximately 45-minutes each were conducted. The interviews included nine with families of people who regularly use AAC devices and six with the AAC device user. Self-identified conditions included autism, cerebral palsy (CP), amyotrophic lateral sclerosis (ALS), and rare genetic conditions. An outline of the questions discussed in interviews is provided below.

General day-to-day

1. What is a normal day like for you?
2. What devices/technology do you most enjoy using (not specifically assistive technology)?
3. What are the things that frustrate you most throughout the day?
4. How do you enjoy spending your time?

Communication process and tools

1. How do you communicate now? What works well about your current process? Are there any limitations?

2. Do you use any devices or technology for communication? What are their strengths and limitations?
3. What are any physical limitations you have with respect to speech?
4. What are you most interested in having a device help you communicate? (e.g., natural language, interacting with devices/media, controlling physical objects like TV, general words/phrases, physical and emotional states)
5. Who do you communicate with most frequently? Does any part of your communication changes based on who you are communicating with?
6. Where do you have most difficulty communicating (e.g., at the dinner table, in a coffee shop)? Are there places it is more important for you to use a communication device?
7. What is important to you in terms of aesthetics?
8. What would your ideal communication device look like?

Motor abilities

1. Do you have any difficulties related to motor control? If so, what types of activities are difficult or tiring?
2. What types of motions are easiest? Most comfortable?
3. Is your motor control changing or developing in the long term? Are there differences over the course of a day?
4. Do you communicate via speech or sounds? When is it easiest? When is it most tiring?

Several themes emerged that were common amongst at least four interviewees:

- More difficulty communicating with strangers: People reported that communication was easier with familiar people, who could understand unique speech patterns and nonverbal cues.

- Difficulty using available AAC devices due to motor limitations: Interviewees reported issues with eye gaze technology because of spastic eye motions, and with tablet and physical buttons which require relatively fine motor control.
- Difficulty communicating needs and wants quickly: Interviewees reported having occasional difficulty quickly communicating pressing need and wants (e.g., hunger, annoyance), particularly with non-family members.
- Communication augmented by gestures: Many interviewees reported using gestures to augment communication but did not have AAC devices with gestural inputs.

2.4 Web-based Survey

The conversational interviews were used to design a survey around AAC usage practices and needs. Survey participants were recruited using a list of e-mails of people who had previously contacted the research lab, by contacting NGOs, and through web forums and social media. An "Other" text box was provided for most multiple choice and check-box questions to help capture variability. The survey was tested with three family members of people with conditions affecting speech or language before it was distributed. The full survey text is provided in Appendix A.

64 people completed the survey, with a median completion time of 26.4 minutes. 23.1% (n=15) of respondents reported that they were responding to the survey on their own behalf, and 76.9% (n=49) of respondents reported that they were taking the survey on behalf of someone else. 30 of

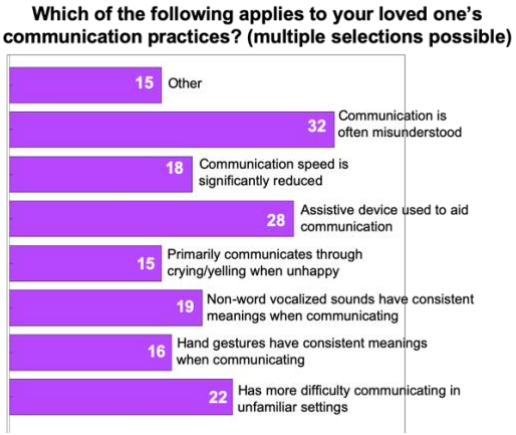


Figure 3: Reported communication practices

the 49 respondents who took the survey on behalf of someone else were parents of a child with a speech or language difference, and 29 (45.3%) were the family member of an individual with an intellectual disability. The survey subjects reported having conditions related to speech and language difficulties including autism (n=18), cerebral palsy (13), stroke (6), ALS (5), and other (21). There were 16 conditions reported in the "Other" category, highlighting the heterogeneity of the population.

The survey included questions about the communicator's speech and language practices (Figure 3). The typed free responses that corresponded with the "other" category and related open response questions provided rich information and personal stories. One respondent reported that they can communicate verbally fluently but only in some circumstances, and that people think they are being rude or lying when they are neurologically incapable of speaking. Half of respondents indicated that their communication was often misunderstood by others.

Common sentiments amongst responses were that communication was often tied to emotions and stress, and that often communication partners did not fully understand what the communicator was expressing. Respondents reported that misunderstood communication was often accompanied by

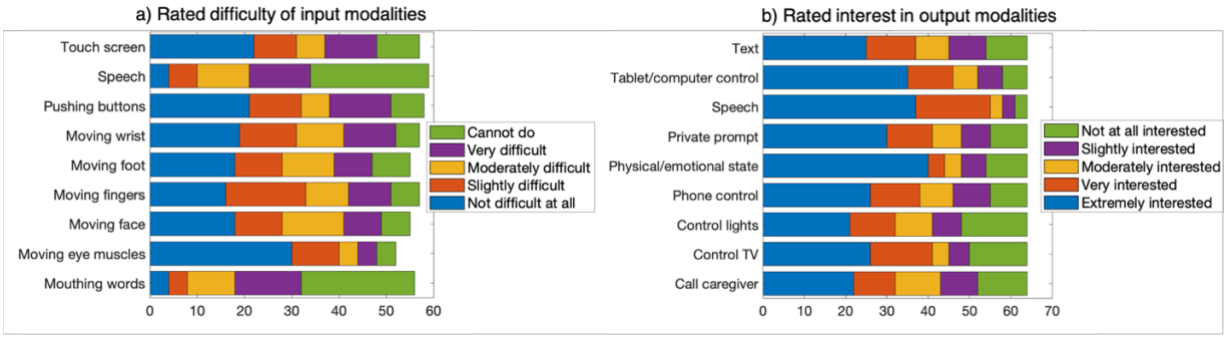


Figure 4: Responses to Likert scale questions on the (a) difficulty of input mechanisms and (b) the utility of output preferences

frustration and anxiety, which made communication more difficult. These problems were exacerbated by a lack of understanding by respondents' communication partners as to their verbal abilities, intellectual abilities, and communication practices. Respondents conveyed frustration that available communication tools and cultural practices around communication often affected their ability to participate in social events and community exchanges.

The survey also asked respondents about their experiences and preferences related to AAC devices, including input and output preferences (Figure 4), aesthetic preferences and important device features (Figure 5), and feedback on the utility of hypothetical devices (Figure 6). The hypothetical devices suggested were based on feedback from AAC device users and their families during conversational interviews. Summaries of the provided descriptions are provided below.

Trained Speech Generation: The user trains a wearable headset-like device to recognize the intent of specific facial muscle movements. Then when they make a facial motion the device recognizes, it outputs the corresponding word as auditory speech, or as text input to a device (e.g., a phone or tablet).

Context-based Speech Generation: The wearable device has a camera and microphone to record

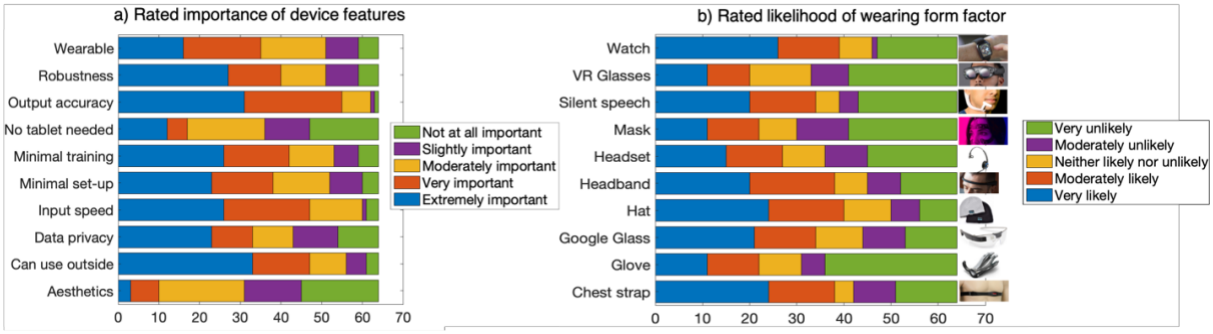


Figure 5: Responses to Likert scale questions on the (a) importance of device features and (b) aesthetic preferences

From top to bottom images from macworld.com, gizmodo.com, news.mit.edu, media. mit.edu, www.headsets.com, amazon.com, sportpro221.com, x.company/glass, teamsoletics.com, audiocardiometrics.com

context. The device uses context from environmental audio and video to suggest words or phrases to the user, who can accept, deny, or edit the suggestions via trained muscle movements and then output the message as speech or as device input.

Speech Prompting: A wearable headset-like device that can prompt the wearer in conversations using context from onboard sensors. For instance, if the wearer says the word "sandwich" and they are in a kitchen, the device will ask them easy yes/no questions to better understand their goals (e.g., Do you want a peanut butter sandwich? Do you want a sandwich now?). The device will process the given information and will prompt them to say an integrated sentence like "Can you please make me a ham sandwich now?")

Physical State Messaging: A glove or armband that records basic physiological data that can be correlated to physical states like pain, tiredness, excitement, or hunger. The device can be programmed to communicate with pre-specified users or/or with individuals in the wearer's vicinity,

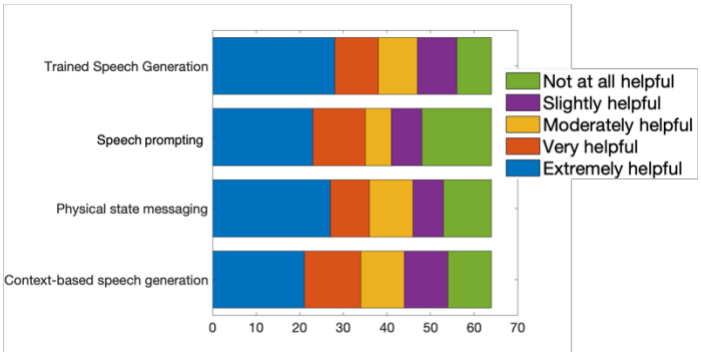


Figure 6: Responses to Likert scale question on the helpfulness of proposed hypothetical devices

and set up to share information only with the wearer's permission. (Note: this is closest to the research direction later selected for this thesis)

In aggregate, the helpfulness ratings for each device appear similar (Figure 6) however most users selected different preferences for different hypothetical devices. 18% of respondents said that all the proposed devices would be "Extremely helpful" with other respondents generally rating one or two devices as "Extremely helpful". One respondent who said that the *Speech prompting* device would be "Extremely helpful" wrote, "I kind of look at the prompting device as a crutch - when he's nervous he sometimes can use that support and believe it would give him confidence". Another respondent said that "Trained Speech Generation" would be very helpful wrote, " I don't have much coordination, either. I can type, but only much more slowly than I used to be able to type. So many solutions wouldn't work well at all for me."

81.25% of respondents (52 people) reported using AAC devices consistently. People who used AAC used an average of 1.4 devices. Respondents who reported no AAC usage included those who did not find available options effective, did not need AAC, and those who said they had not tried AAC but would like to in the future. Reported AAC tools were electronic word/symbol selection (25); gaze or head tracking (14); phone, computer, or tablet (13); physical keyboard (11);

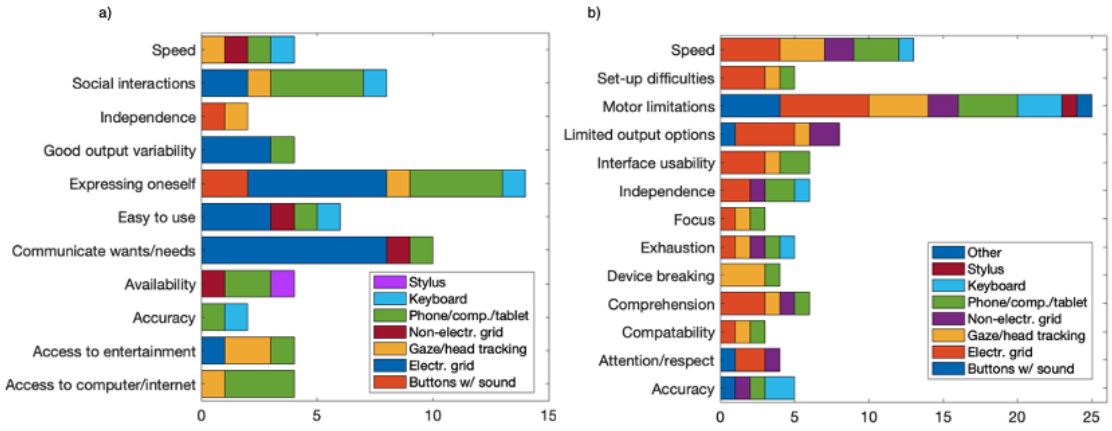


Figure 7: Identified features participants reported as a) aspects they like about existing AAC devices and b) usage difficulties reported with existing devices

stylus/pen + writing surface (6); non-electronic grid (8); buttons that produce pre-specified sounds (4); eye movement (1); tongue movement (1); virtual keyboard (1); and voice prosthetic (1).

40 people responded to open response questions to describe usage difficulties with their AAC devices, and 45 people responded to an open response question to describe features that they like (Figure 7). The most common difficulties reported were motor limitations and speed. The published data includes respondents' ratings of the difficulties of various motor tasks. Respondents reported difficulty selecting small icons, buttons, and using switches due to spasticity and fine motor control limitations. Communication speed limits were reported as particularly problematic with electronic word/symbol selections, gaze tracking, and phones. Devices were described as having slow, cumbersome input processes that made it difficult to maintain a conversation flow both in-person and using web-based and mobile chat systems. Device breakage was particularly a concern for gaze and tracking systems, high-tech AAC devices. Respondents mentioned cost as a prohibiting factor to testing and evaluating available products to select the most effective technology.

Respondents liked that the AAC devices better allowed them to express themselves and communicate wants and needs. AAC devices also enabled access to critical tools like work-related software and entertainment (e.g., music, games, television).

2.5 Discussion

As noted in Section 2.2, the responses of this survey should be interpreted with care, as many respondents were not the device users themselves. Firsthand perspectives from device users are critical in technology design and research. This exploration was intended to explore trends from a larger group of people with some experience and knowledge of AAC and motivate questions for further exploration. Quantitative survey results and an interactive GUI to view survey results have been published online and are similarly intended to help other researchers familiarize themselves with a breadth of possible user needs and formulate questions for further exploration directly with AAC communicators.[1].

There were some notable differences in responses between surveys filled out by a device user themselves and surveys filled out by a loved one. For example, those who filled out the survey for themselves generally reported being able to do more motor-related tasks independently. Respondents from the device user itself reported a higher preference for devices with a short training time, robustness, and a low set-up time. There were also differences in demographic information between the two groups - 61.2% of responses from people who took the survey on behalf of someone else were about a minor child and 45.3% were about a person with an intellectual disability. Of the responses filled out directly by a device user, 6.7% were from a minor

[1] See https://www.media.mit.edu/projects/aac-survey for the complete survey questions, and downloadable code for a MATLAB GUI that visualizes quantitative survey results

child and 20% were filled out by someone with an intellectual disability. In the future, collecting additional data directly with device users themselves could help identify areas where device users' needs are not well understood by those close to them.

A principal component analysis (PCA) was conducted to explore possible groupings of needs using a normalized data matrix with quantitative responses and numbers representing categorical variables. 11 factors were needed to explain 95% of the variance in the dataset. The relatively high number of factors needed to explain variance in the data, and low and distributed coefficients indicated that there are no clear underlying factors that can be used to create groups in the dataset. Hierarchical clustering led to high remaining within-cluster variances [44] with 2-8 hierarchical clusters, and even larger variances when grouping by the respondent's reported condition affecting speech or language. This is consistent with the known heterogeneity of populations with speech and language conditions and emphasizes the need for personalized technologies. We discuss three areas of active research in AAC [39], [45], in the context of meeting user needs identified in our survey and interviews: (1) customizable input options, (2) context-aware output, and (3) personalized devices.

1. Customizable input options: Many respondents reported that their device usage was limited by their motor abilities. Communicating complex thoughts was particularly difficult for people who had motor difficulties that prohibited use of letter-by-letter input methods. Tiredness and fatigue were frequently reported sources of difficulty. New easier input mechanisms for AAC devices [46]–[50] could improve accessibility and comfort for people with motor control challenges. Usability could also be improved with devices that can easily be adapted to use a variety of input mechanisms

(e.g., eye gaze, switches, foot control) and switch fluidly between those mechanisms without interrupting a conversation.

2. Context-aware output: Communication speed was a major difficulty cited by survey respondents. Speed could be improved by more accurate word predictions [51]. Most commercial technologies do not use 'smart' word and phrase suggestion mechanisms, and those that do generally provide suggestions based only on previously outputted speech. Combining sensed contextual information (e.g., activity detection using images and audio from the environment, GPS) with machine learning algorithms could allow devices to make appropriate complex suggestions. Researchers have implemented context-aware word prediction using location and time on AAC systems [52] with Bluetooth [53], context history [54], with speech recognition on conversational partners [55], and using acoustic cues [56]. Incorporating context-aware word prediction into commercial devices could allow for smaller word selection boards that incorporate the most relevant options. A smaller device could be more easily incorporated into a transportable device like heads up displays and smart watches, design factors and aesthetics valued by survey respondents.

3. Personalized devices: Many respondents were frustrated by the lack of appropriate output options. Numerous respondents reported the use of self-consistent vocalizations or gestures for communication. Implementations of some personalization features in research studies have shown improved usability [57]–[59]. Devices trained to recognize an individual's unique audio and gestural cues – for instance, using personalized machine learning – and then curate appropriate communication options could improve communication speed and allow more complex output. Multiple output options from a single device based on situation could enhance use flexibility. Devices could also be trained to adapt to fluctuating motor abilities - for example, adapting to

progressive motor difficulties or stiffness in mornings or evenings. Some level of personalization has traditionally been incorporated into commercial AAC devices, though many respondents said setting up these features in their devices was difficult and time-consuming.

2.6 Conclusions

Several directions for innovations – and related research studies – that meet the needs identified from interviews and surveys are discussed. There is a gap between technologies used in research and what is available in commercial devices. Most respondents felt that the available resources, both social and technological, limited their ability to communicate. Many available commercial technologies require arduous motor input, have limited and/or juvenile output options, cannot be used outside or in new environments, and are not affordable. The respondents were generally interested in a range of output options, and a subset of respondents expressed high enthusiasm for each hypothetical device, highlighting the need for newer, smarter devices. In additional to technological innovation, there is a need for greater societal support for AAC. Many respondents cited a lack of understanding and accommodation from society as a large source of stress.

Additionally, AAC devices are sold in low volumes, which often means devices are sold with high margins. Purchasing devices can require navigating insurance agencies, which may have restricted coverage and can impose additional financial burdens. Recent advancements in sensing and computation improve the feasibility of implementing research projects commercially [60]–[62]. A push for open-source software platforms [63], [64], and the creation of low-cost, open-source hardware platforms [65], [66] could enable the creation of do-it-yourself kits and help bridge the divide between research and technologies available to users.

The feedback from interview and survey respondents highlighted that many individuals who are mv* (see Section 1.2 for a definition of the terminology) are underserved by available technologies and research. Particularly, there are some individuals with neurological and motor-planning differences for whom existing commercial devices have limited effectiveness. The results of the survey and literature review also highlighted that non-touch-based inputs (e.g., tablets, buttons, and switches), like nonverbal speech and physiological signals, are particularly underexplored. These results motivated the case study described in Chapter 4, which led to my eventual focus on studying nonverbal vocalizations by mv* individuals.

3 Related Work

Previous studies on vocalizations by mv* individuals have largely focused on tracking language development and using vocalizations as a diagnostic tool. Speech and language pathologists (SLPs) employing a total communication approach may track and encourage vocalization-based communication [25], yet no prior work has systematically recorded and analyzed vocalizations recorded in real-world settings along with personalized labels on the underlying communication or affect. Nonverbal vocalizations from infants and pre-verbal children, nonverbal vocalizations that occur amidst traditional speech, and affect in typical speech have been studied extensively. This thesis presents a first study of real-world nonverbal vocalizations from mv* individuals as a communication modality, through the lens of human-centered design towards improved communication technology. The chapter includes work created collaboratively with Kristy Johnson.

3.1 Technology Development and Neurodiverse Individuals

3.1.1 Adapting to Typical Social-Communication Styles

Technologies targeted to neurodiverse individuals often focus on facilitating communication via prevalent typical social-communication styles [67], [68]. For example, researchers from Stanford created and evaluated the Superpower Glass, a wearable designed to facilitate social communication for Autistic children by giving the wearer feedback on social cues and emotions of conversation partners. The researchers' reported improvements on the Social Responsiveness Scale, as evaluated by participant's parents, after a study with 14 families over two months [69].

Zhao et. al. created a VR-based interactive game designed to improve social communication abilities for children with autism and found that collaboration between partnered children (n=12) improved as they played the game, which required players to cooperatively control a ball in a maze. Other research in this space has included virtual reality job training simulations [70], AI-based emotional interpretation in video calls [71], and exploring uses of conversational agents for Autistic adolescents [72].

3.1.2 Translating Physiology

Previous work towards translational systems for individuals with autism using machine learning has focused on detecting a single state in lab-based settings using physiology [19]–[24], and most approaches have relied on labels by researchers or therapists rather than the individual themselves or someone who knows them well. 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 [22]. Goodwin et. al. used electrodermal activity (EDA) to model what he called 'challenging behavior' in individuals with autism in a clinic [23]. Kushki et. al. explored using an electrocardiogram (ECG) to detect anxiety-related arousal in children with autism [24]. Generally, these approaches have been limited to a single state or dimension, which has limited utility for communication. While such approaches are promising and may integrate with communication technology in the future, further research is needed towards obtaining consent to share and record physiological activity from mv* communicators, and towards creating sensors comfortable for individuals with different tactile preferences to wear in day-to-day life.

3.1.3 Design for and with Mv* Individuals

Most human-computer interaction (HCI) and design research on communication technology focuses on individuals with typical expressive language abilities. Some recent studies have explored the use of novel interactive technologies with mv* individuals. Wilson et. al. explored self-expression with minimally verbal children with autism using the ExpressiBall, a ball with lights, sound, and motion sensors, to better understand how to include them in co-design. 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 stressed that researchers and others should listen to expressions that occur through all modalities [73]. In another study, Wilson et. al. used a similar artifact to identify 'moments of interaction' during which minimally verbal children communicate in ways that extend beyond words [74].

3.2 Clinically Oriented Studies on Nonverbal Communication with Mv* Individuals

Studies with mv* individuals often focus on exploring clinical research questions in laboratory environments, and particularly on identifying differences in communication practices between mv* individuals and typically developing individuals [75]–[79]. Stone et. al. studied nonverbal 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 [78]. Gordon et. al. and Colgan et. al. separately studied patterns of gesture use in infants with autism [79], [80].

3.2.1 Clinical Studies on Vocalizations with Mv* Individuals

Nonverbal vocalizations – particularly in preverbal children – have been studied extensively as a marker of language development [81]–[85]. Donnellan et. al. experimentally studied the relation between prelinguistic vocalizations in infants and language development trajectories for typically developing children [82] and McDaniel et.al. conducted a meta-analysis on the relationship between prelinguistic vocalizations and expressive language development in children with autism [83]. Bacon et al. [86] created a large naturalistic dataset by manually annotating toddler speech in clinic-visit videos to study language development of toddlers with and without autism.

Researchers have explored using nonverbal vocalization acoustics to diagnose autism using infant cries [87] and naturalistic child vocalizations [88], [89] and as a marker for other developmental differences like Fragile X syndrome [90], Down Syndrome [3], and specific language impairments (SLI) [91].

3.2.2 Limitations of Lab-Based Studies

Studies with mv* individuals primarily take place in laboratory and clinical settings. 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) [92]. Oller et. al. conducted one of the only known studies of nonverbal communicators from non-typically developing toddlers in real-world environments but focused on diagnosis tasks and not on vocalization affect or intent. Tools that track and enhance communication by mv* individuals in real-world settings, which may provide communicative not captured by laboratory tests [93], are largely unexplored.

3.3 Nonverbal Vocalizations

Nonverbal 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 [94]–[96].

3.3.1 As an Expression of Affect Amidst Typical Verbal Speech

Nonverbal vocalizations that occur alongside traditional language have been studied anthropologically [97], [98] and have been classified by affective content with both natural and acted vocalizations across numerous studies [99]–[101]. 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 [101]. 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 [97]. Anikin found that listeners could reliably differentiate between affective acted and authentic nonverbal vocalizations [98], and identified correlations between voice quality and valence in nonverbal vocalizations [102].

3.3.2 As Expressive Communication in Infants

Nonverbal vocalizations have been studied as expressions of affect and communication with infants. In 1964, Wasz-Höckert identified specific meanings – pain, pleasure, and hunger – in infant vocalizations in a study with trained nurses in a hospital [103]. Since then, there has been extensive work on classifying infant cries by need (e.g., hunger, pain) using both humans and machines [104]–[107]. 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 [104].

3.3.3 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 to communicate these expressions remain undeveloped. For people who are mv*, these nonverbal vocalizations serve a larger linguistic and communicative purpose as they occur independently of verbal speech [108]. 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. To my knowledge, no other studies have acquired nonverbal vocalizations from both child and adult mv* individuals using personalized labels in real-world settings.

3.4 Speech and Emotions

There has been extensive prior work on affect detection in speech with typical verbal content. Emotion is often modeled using the "big six" classes defined by Ekman (anger, disgust, fear, happiness, and sadness) [109] or by using scales of arousal, valence, and dominance [110].

3.4.1 Aggregated Statistical Features

Feature extraction often focuses on spectral features like formants and on cepstral features like cepstral coefficients. Often, aggregate statistics across a segment are calculated. These aggregate statistics often include many derived functionals and can include thousands of features [111]. Such an approach is common even when the size of the feature set becomes very large compared to the amount of training data. ComParE, the computational paralinguistic feature set (size: 6373) [112] and the extended Genova minimalistic acoustic parameter set (size: 88) [113]

are often used in speech emotion classification, extracted with the openSMILE toolbox [114]. Feature extraction can also include the linguistic content of vocalizations, using approaches based both on word sentiment and linguistic structure. Support vector machines (SVM) and random forest models are often used for speech emotion classification tasks.

3.4.2 Deep Learning Approaches

Recent trends in speech emotion analyses have included using weakly supervised learning approaches, multi-target data generation, transfer learning, data-learned features, and modeling confidence [111], [115]–[121]. auDeep, DeepSpectrum, and VGGish are frameworks that extract data-learned features from audio data. auDeep uses sequence to sequence autoencoders to learn features from spectrograms of audio data using deep neural nets [116]. DeepSpectrum extracts features using convolutional neural nets (CNN) with spectrograms and chromagrams [121]. The VGGish model was trained on Google's AudioSet (a labeled database of audio from YouTube, not specific to speech and language), and can be used as a feature extractor with audio data [119], [120]. WaveNet, a generative model by DeepMind, has also been used as an autoencoder to create features from audio data [118]. Gerczuk et. al. evaluated transfer learning across many emotion datasets using a CNN architecture based on ResNet, and found improvements in model performance with cross-corpora training for 21 of 26 evaluated datasets [117]. Zhang proposed and evaluated semi-supervised and active learning strategies to mitigate data label scarcity as well as multi=task approaches to leverage task relatedness in the speech emotion recognition domain [122]. Zhang et. al. proposed the PaNDA method (Paralinguistic Non-metric Dimensional Analysis) to identify relatedness between tasks and found that using PaNDA along with multi-talk methods and DNN models was successful across 18 tasks [122]. Deep learning-based modeling and feature extraction approaches often require large amounts of training data. Even amidst new

modeling approaches utilizing neural networks, classical machine learning models using aggregated features often still perform competitively for emotion classification tasks [111], [117].

4 Longitudinal Case Study

This chapter presents an eight-month case study with one mv* communicator and his family. The work presented in this chapter was done collaboratively with Kristy Johnson, and parts of this chapter are published in [123].

This chapter presents a design process, multi-phase data collection methodology, and data analysis through an iterative longitudinal approach that originated with physiological sensing and progressed to focus on nonverbal vocalizations. The chapter concludes with remarks on how the presented case study informed the next steps for research and development.

4.1 Design Process

As presented in Chapter 2, the research process began by interviewing and surveying individuals who have differences in speech or language abilities and/or their families. 5 of the interviewed families had children who were mv* communicators, and 18 of the survey responses were about a communicator with autism. This early engagement served as the foundation for subsequent studies.

The parents of mv* communicators, in particular, noted that the existing technology was minimally accessible to their children, and that they understood their offspring's communicative intent better than others who interacted with their child like teachers and babysitters. As a result, we chose to focus this work on the population of mv* communicators (as defined in 1.2).

4.2 Design Approach

Because of the complexity and diversity of needs of mv* communicators, we conducted an eight-month longitudinal case study with one nonverbal child and his family and caregiver network. His mother is both a researcher and one of the authors of [123], enabling unique insight and access to this specialized group. While most studies take a cross-sectional approach, engaging with many Autistic individuals in a controlled, single-session laboratory environment, we built around diverse natural environments and the small network of people surrounding this one individual (parents, grandparents, babysitters, and siblings). These individuals provided verbal feedback 3-4 times per week, which steadily informed the data acquisition process and goals of the project.

Using the design approach described in Figure 1, we began work with the goal of building a platform to enhance individuals' communicative exchanges using simple, accessible sensors throughout daily life. We envisioned a system whereby families can participate remotely, use a do-it-yourself kit to collect, label, and upload data.

Although this design approach (focusing on a single communicator for an extended period) is not without limitations – for example, usability, comfort, equipment access, privacy, and data transfer all required re-assessment when we moved past n=1 – the combination of firm roots and an upward focus helped ensure alignment of the platform with the broader community's needs. Interviews with families with mv* communicators were conducted during the case study (symbolized in Figure 1 by the branches extending from the trunk), allowing us to understand if and what we learned during the case study was also applicable for other families.

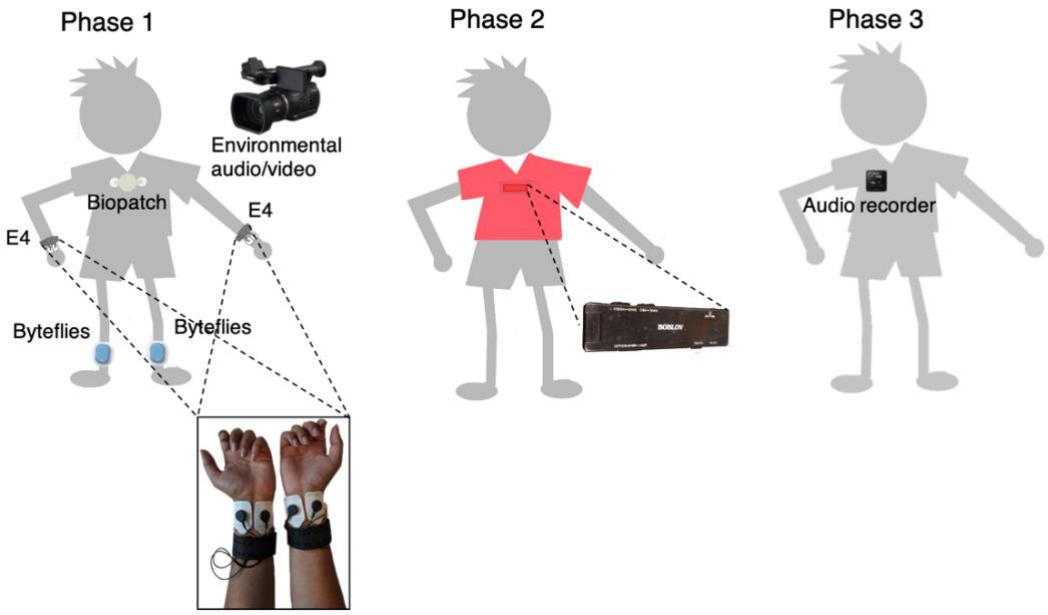


Figure 8: Phases of the longitudinal case study. Phase 1 data collection involved six gelled electrodes and two light adhesives. In Phase 2, there was no skin-adhesive contact. First-person audio and video were collected from a small camera in a t-shirt chest pocket. Phase 3 involved data collection with a high-quality audio recorder magnetically attached to the participant's shirt and was developed and scaled beyond n=1.

4.3 Phase 1: Methods

In the interviews and survey, individuals had expressed interest in devices that could provide information about physical state (e.g., pain, emotions, hunger) without requiring active input from a user. From prior work, we knew that wearable physiological and motion sensors could provide such information [19]. Hence, Phase 1 of the case study focused on physiological sensing and affect-based detection using wearable sensors.

4.3.1 Setup

Data were collected with a non-speaking autistic boy of elementary school age. He has zero spoken words and limited use of AAC tools. The participant and his family provided IRB-approved informed

consent/assent, where assent for this child was determined through body language, behavior, and other holistic forms of communication.

During this phase, we collected electrodermal activity (EDA; formerly galvanic skin response), electrocardiography (ECG), photoplethysmography (PPG), and acceleration data using wireless wearable sensors (Figure 8). Several sensors were trialed to determine which, if any, would be useful for future studies. The cost of the components ranged from approximately $500 (Biopatch) to $1700 (E4s) per sensor.

The participant wore an E4 sensor ($1700; Empatica, Italy) on each wrist. These watch-like sensors measure skin conductance via changes in sweat gland activity, which are a function of the body's sympathetic nervous system activation and are often used as a gross proxy for stress. The E4 sensors also record the user's skin temperature, pulse rate, and 3-axis accelerometry.

The E4 was too big for a small child, and to reduce motion artifacts and increase data quality, pre-gelled adhesive electrodes with 0.5% NaCl isotropic gel were attached to the user's distal wrists (see Figure 8). Sweatbands were placed over the sensors to further minimize movement and distraction and to ensure comfort. Zephyr Biopatch sensors ($500; Medtronic, USA) were secured to the participant's chest using adhesive pre-gelled Ag/AgCl ECG electrodes (Cathay, China) to record ECG, respiration, and acceleration. PPG and accelerometer Byteflies sensors (Market price unavailable; Byteflies, Belgium) were attached to the participant's left and right ankles using non-gelled light adhesives.

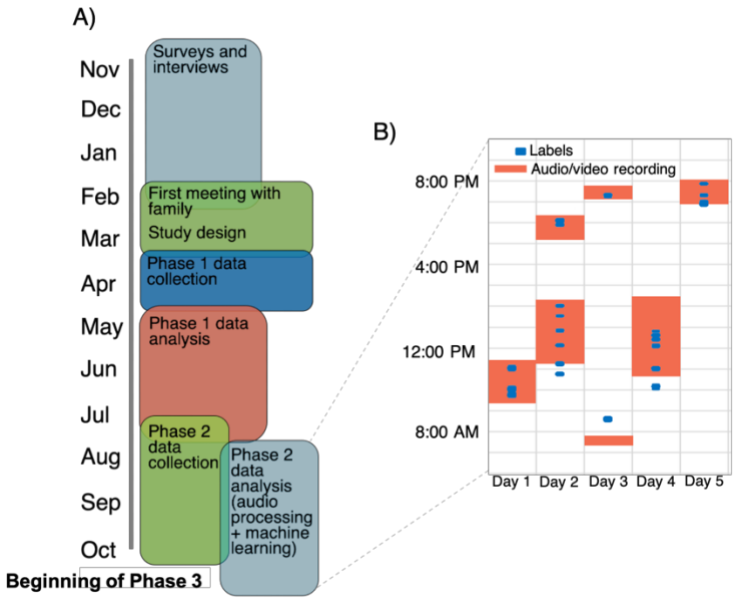


Figure 9: A) Case study timeline for Phase 1 and 2. B) Distribution of data collected in Phase 2 of the case study. Data collection included 13 hours of audio/video data with 300+ in-the-moment labels. Caregivers labeled data in focused chunks, yielding a sparsely labeled

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

After attaching the sensors, the participant was free to pursue his regular activities.

The researcher captured third-person audio and video but was otherwise minimally engaged to capture the most naturalistic record possible. The family participated in 6 sessions over the course of 3 weeks (Figure 9) resulting in over 130 hours of multi-modal data streams from this phase.

4.3.2 Labeling

Phase 1 data were labeled post hoc by a researcher (myself) who was not related to the pilot family. We note that while any interpretation of the participant's data by someone other than the

participant is inherently biased, those closest to the participant have the most experience and therefore the best chance at accurately interpreting the participant's affect and communication. Our intent was always to involve the participant family in this process to reduce bias and enhance accuracy. In this phase, we discovered that asking the communicator's family for post-hoc labels was impractical because of the laboriousness of the task (see 4.4) and so only the researcher's labels were used for this phase. An in-the-moment labeling app was developed for later phases so that data could be labeled quickly and easily by the communicator's family. The analysis with researcher labels was preliminary exploration with the collected Phase I data.

First, the collected video streams were annotated continuously for general content such as location, activity, or interactions like pointing to an item. These annotations were then used to provide context to the collected signals. The context was used to label vocalizations, along with insights on how to interpret vocalizations that were provided by the family during data collection. The researcher only included vocalizations they could label confidently in the analysis, which may have biased the analysis.

While exploring the various signals, we discovered that audio was surprisingly rich and appeared to be qualitatively consistent at conveying certain types of affect and communicative intent. To evaluate the audio signals for affective and communicative information, I manually segmented vocalizations from a 3.5 hour subset of video collected over 3 data sessions.

Using contextual information from the video and acquired knowledge from the parents and other caregivers, I assigned these segmented vocalizations to one of seven categories: laughter, crying, self-talk, request, protest, dysregulation, or no label if it was not possible to distinguish. I was able to confidently label some vocalizations (e.g., laughter), but had low confidence in other categories (e.g., dysregulation).

4.3.3 Analysis Methods

The overlapping semi-transparent audio waveforms and spectrograms presented in Figure 10 suggested distinct variations between sound types.

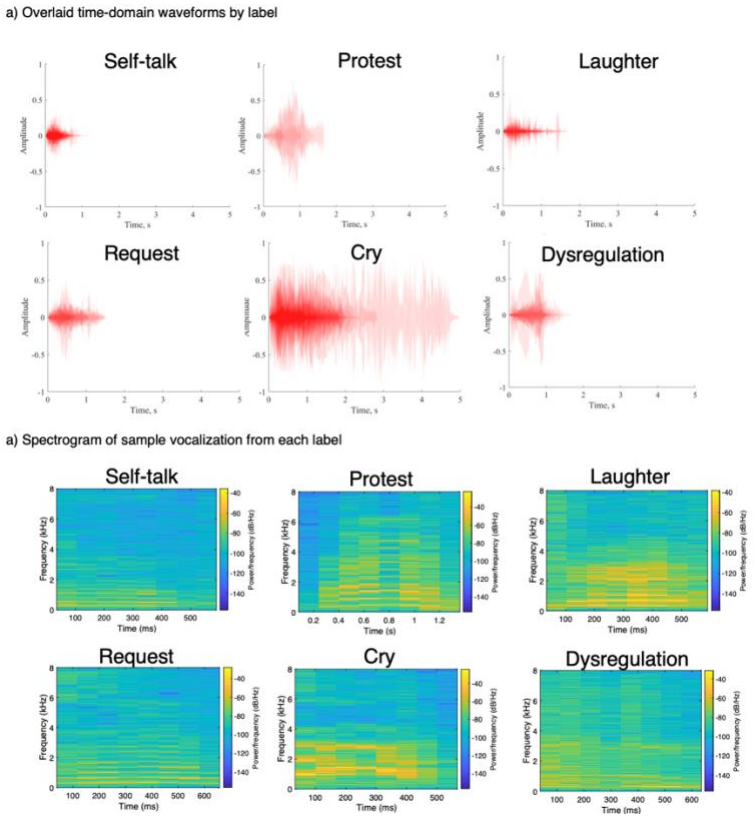


Figure 10: a) Time-domain visualization of researcher-labeled Phase 1 data. Each plot shows superimposed semi-transparent audio waveforms, suggesting distinct variations between sound types. b) Spectrograms of a selected sample of each vocalization type. The sample was identified by the author as being perceptually representative of a communicative function.

To explore the feasibility of machine-based classification of vocalizations without typical verbal content, a multi-class support vector machine (SVM) was trained and evaluated using 215 audio segments spanning six classes using the mean MFCC values for the first 13 coefficients. For this preliminary analysis, the clips were randomly split 80/20 (train/test). All data points had been seen by a researcher during the labeling phase.

4.4 Phase 1: Results & Discussion

The multi-class SVM produced a weighted F1 score of 0.67, suggesting that these researcher-labeled classes contained sufficient similarity to be partially clustered by a machine. These results were likely affected by the post-hoc researcher label assignment procedure.

The primary caregiver found it difficult to label the audio segments with high confidence without context, and watching the full video was too time intensive. It had taken the researcher over 4 hours to continuously label each hour of video and another 1-2 hours to label an hour's worth of extracted vocal segments. This process was an unrealistic load to place on the participant's family, and we realized it would limit the ability for a platform built on this labeling procedure to succeed with other families, who might have more limited resources or expertise. This experience prompted the development of an in-the-moment labeling app that was used in Phase 2 to incorporate real-time labels from a caregiver.

Qualitative analysis of the physiological signals indicated some post-hoc correlations between labeled context and physiology – e.g., high phasic EDA during challenging activities, low phasic EDA during sick days, and distinct skin conductance responses during stressful or exciting moments – but sensor comfort and long-term use was a concern. While the sensors themselves did

not appear to be irritating, the gel electrodes, particularly from the ECG Biopatch sensor, left a sticky residue on the skin that was difficult to remove. The family reported using soapy water, alcohol wipes, and other methods to remove the residue, which led to mild skin irritation. In response, a cloth-based Bioharness (Medtronic, USA) was trialed, but the data quality and fit was poor.

The light, non-gelled adhesives on the Byteflies sensors seemed comfortable and did not leave a residue, but they occasionally fell off the participant during physical play like swinging. From interviews with other families, we learned that many mv* individuals would not wear even wrist-worn sensors with dry electrodes (which would be significantly more comfortable than the wet electrode sensors used in the case study). In addition, the requirement of having the researcher present to attach the sensors – and the sensors being too expensive to leave with the participant – meant that this setup would not be able to reach the geographically distributed participants who could benefit from this technology.

The initial study design was built on the idea that high-quality data and scientific findings would precipitate complementary technology advancements. While this notion is still sensible, the naturalistic physiological data from this phase lacked sufficient context and data quality for reliable interpretations.

The data we collected were feature rich and the approach is worth further exploration (Figure 11), but the sensors and interpretability of the data did not fit the vision for a scalable communication platform that could integrate easily into day-to-day life. Additionally, recording physiological data interpreted as emotions and communication has complex considerations around consent for data

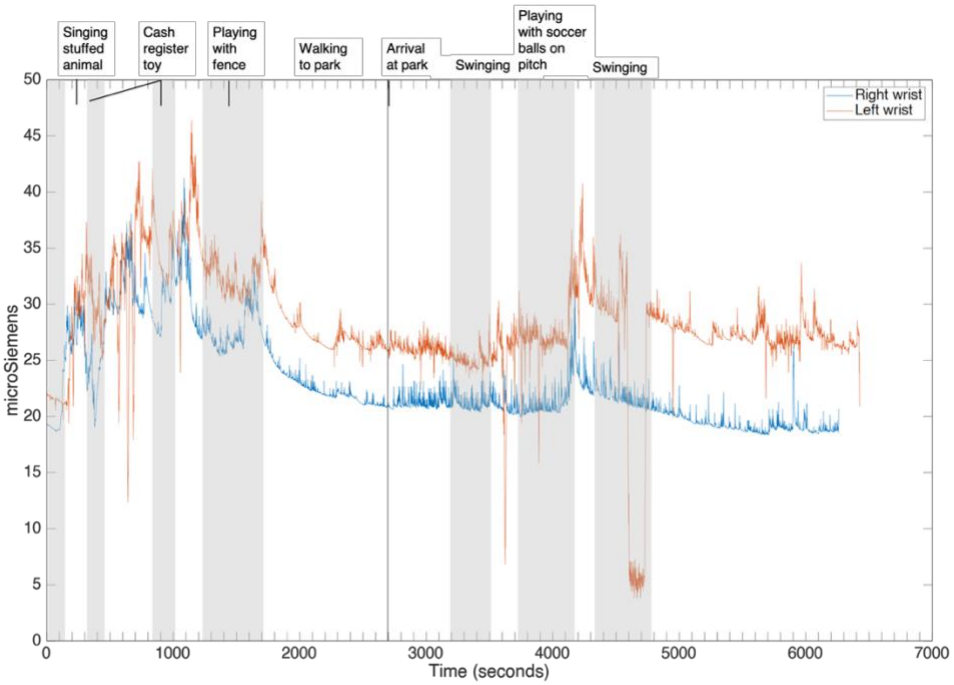


Figure 11: Wearable physiological sensors were not selected for continued study due to comfort and scalability issues but warrant future exploration. Qualitative analysis of the sensor data showed some correlations with contextual annotations of the recorded video. For instance, a peak in the EDA signal is visible as the child approached the front yard fence gate – the child found playing with swinging the fence gate highly enjoyable and exciting.

collection and sharing. Speech and vocalizations are volitional communication that an individual is actively choosing to share with those around them.

As a result, we chose to pivot our approach to explore how vocalizations, a more easily observable feature, could be used in a communication platform. Such an approach also has deployment advantages compared to a platform requiring costly sensors and attachment methods.

4.5 Phase 2: Methods

Phase 2 focused on first-person audio/video data capture for both affective and communicative expressions. This process was found to provide rich data in a comfortable, easy-to-use format.

Figure 12: One of the first iterations of the live-labeling app developed for Phase 2 of the longitudinal case study.

Pressing any button registers an event label, which can be turned into an event range by pressing "end" Because continuous, accurate caregiver labeling for long stretches of time was unrealistic, a progress bar was included at the top of the app. This bar tracks a 5-minute "Focus Mode" that was intended to encourage high-quality labeling for short periods of time to augment the number of high-quality ratings while minimizing caregiver demands

The in-the-moment labeling app (Figure 12) enabled over 300 labels across 13 hours of audio/video data.

4.5.1 Setup

A lightweight audio/video recorder ($40; BOBLOV, China) measuring 3.7" x 1.0" x 0.4" was placed in a small custom chest pocket on a t-shirt. Single channel audio was recorded at 32 kHz and 16 bits per second. Video was recorded at 30 frames per second and 1920 x 1080 resolution.

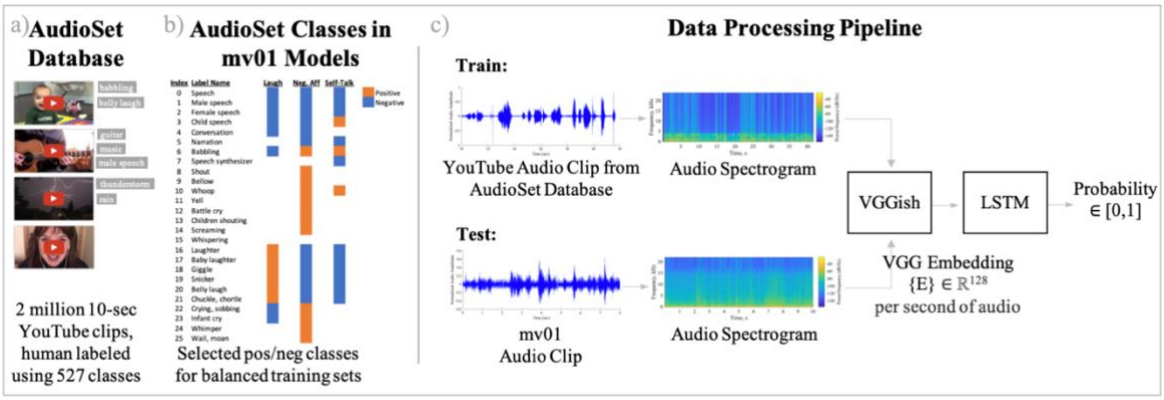


Figure 13 a) Each video in the AudioSet database [119], [120] *was human-labeled using 527 possible event classes. B) We selected appropriate positive/negative classes from AudioSet to include in three different models with Phase 2 vocalizations. C) The LSTM model was trained on VGGish embeddings of the selected subsets of AudioSet data.*

As before, after donning the shirt, the participant pursued his regular activities. The parents had control over the data and could delete files before sharing them with the research team.

4.5.2 Labeling

Using insights gleaned from Phase 1, we developed an Android app that enabled simple, intuitive, in-the-moment labeling (Figure 12). During Phase 2, the app went through five iterative design stages using feedback from every data session. Figure 12 shows the design of the app at the end of Phase 2. The labeling app later went through further design revisions that are detailed in Chapter 5. The app included icons and color gradients for quick visual scanning, customizable label presets, a contextual note option, easily adjustable timestamps, and the ability to "pin'" an event to label two things simultaneously. Each label could be registered as an instantaneous marker (one press) or as a range (press "End"). Users could review, edit, and delete past events through the hamburger menu button in the upper right. All labels were timestamped and synced to a server at the user's discretion. These labels were then integrated with the user audio/video files for further analysis.

To enhance label accuracy and synchronization with audio/video data, the caregiver was asked to label data in 5-minute focused chunks. A Focus Mode progress bar at the top of the screen served as a visual reminder, though the timer can be paused at any time.

4.5.3 Analysis Methods: Zero Shot Transfer Learning with AudioSet

To assess the applicability of existing general datasets for this platform, we implemented a zero-shot transfer learning (ZSL) classification approach using AudioSet. The goal of this analysis was to explore whether generic databases might be used to supplement our smaller specialized dataset and reduce the number of caregiver labels needed. This section contains excerpts from [124] which was co-written with Kristy Johnson. A deeper exploration of transfer learning with the full dataset that was collected later is presented in the Chapter 9.

We trained three models on sounds from the AudioSet database, a large public database of YouTube clips that have been human-labeled with over 500 audio event labels, including speech, music, nature sounds [119], [120]. Each model was built to classify one of three categories of audio from our participant: laughter, self-talk (similar to babbling), and general negative affect. The AudioSet database includes subtypes of vocalizations such as babble, giggle, wail, whimper, and sigh; however, no single category sufficiently captures the nuances of the vocalizations made by our case study participant. We manually selected AudioSet labels to create positive and negative classes for each category (Figure 13). For example, in the Laughter model, the positive class included AudioSet data for belly laugh, giggle, chuckle, and similar sounds. The negative class included sounds that might confuse the model, like speech, and other sounds that were common in the participant's environment that we did not want to model, like music.

Three-layer LSTM models were trained using batch normalization and Adam optimization with balanced positive and negative AudioSet training sets. The input to the LSTM model was a VGGish embedding of an audio waveform (Figure 13). A past/future split was then used to validate/test: the first 3 days of Phase 2 data was used for validation and optimization, while the last 2 days of data were held out for testing.

4.6 Phase 2: Results & Discussion

The ZSL Laughter and Negative Affect models built using the AudioSet database yielded 70% and 69% accuracy (Table 1), respectively, suggesting that including such data might be helpful in augmenting training for certain broad sound categories. However, the ZSL Self-Talk model performed just above chance, and all the models had high false positive rates (Table 1), highlighting the need for more targeted approaches and datasets that include unique data from this specialized population. In Chapter 6, I discuss developing a dataset of nonverbal vocalizations fully from mv* communicators to enable more targeted analysis, and present personalized modeling approaches in Chapter 8.

SELF-TALK		ACTUAL	
		Yes	No
PRED.	Yes	0.33	0.67
	No	0.41	0.59
ACCURACY: 0.511			

NEG. AFFECT		ACTUAL	
		Yes	No
PRED.	Yes	0.46	0.54
	No	0.27	0.73
ACCURACY: 0.690			

LAUGHTER		ACTUAL	
		Yes	No
PRED.	Yes	0.18	0.82
	No	0.06	0.94
ACCURACY: 0.703			

Table 1: Results from Phase 2 of the case study: confusion matrices for the three models evaluated using held-out test data

The hardware used in Phase 2 -- a recorder tucked into a t-shirt chest pocket -- appeared comfortable for the participant. Collecting data required minimal setup from the parents, and they

were able to label a large volume of data during a two-week period (Figure 9) with reported ease. Even so, the caregiver still had to wait for an utterance or interaction, interpret it, and press the appropriate button – this led to some unavoidable delays in the label timing which are analyzed in Chapter 7. Additionally, we realized that the recording device had significant clock drift, even between multiple recordings taken on the same day. The recorder did not have a sufficient dynamic range and a large amount of the recorded data was clipped.

Throughout the case study, we continued interviews and informal conversations with other families with mv* communicators. Feedback from many families was that recording video during day-to-day life was too invasive of privacy. Recording audio alone - with families having complete control on recording timing and content - was generally acceptable. Interviews also suggested that the approach would extend to other mv* communicators - parents of mv* individuals said that they understood their child's nonverbal vocalizations better than others and that it would be helpful to share their unique knowledge with the broader community.

Phase 2 of the case study illustrated the utility and feasibility of collecting real-world nonverbal vocalization data with in-the-moment labels from close family members. Based on the case study and continued interviews, we developed Phase 3 (Figure 8) of the study with a small, wearable audio recorder with a reliable internal clock and a large dynamic range along with an updated labeling interface. The data collection system for Phase 3 was scaled up for use with additional participants and is described in Chapter 5.3.

5 System Design

The interviews, surveys, and case study along with a review of prior work led to proposed designs for new AAC devices. We learned that nonverbal vocalizations from mv* individuals convey rich communicative and affective information that is not well understood by individuals who do not know the communicator well. We learned that there is an opportunity to help those who don't know a communicator well (e.g., teachers, new caregivers, extended family) better understand unique communication like nonverbal vocalizations. We also learned that integrating vocal controls into AAC devices could improve the usability of touch-to-chat programs for some mv* communicators. Currently, nonverbal vocalizations are understudied in the literature and are not integrated into available AAC devices.

Research and development related to augmentative and assistive communication (AAC) devices and autism has largely focused on tools to assist communicators in generating verbal speech via buttons and other motor-controlled inputs (e.g., apps like Proloquo2go). Such research is congruent but distinct from the focus of this work on improving the ability of the community-at-large to better understand a way in which mv* individuals already communicate - nonverbal vocalizations. Voice-based control of existing interfaces via nonverbal vocalizations lies at the intersection of these approaches and could also be facilitated by the presented research.

Here, I present the vision for novel AAC interfaces: (1) a classification interface to help someone who doesn't know the communicator well understand nonverbal vocalizations, (2) classification

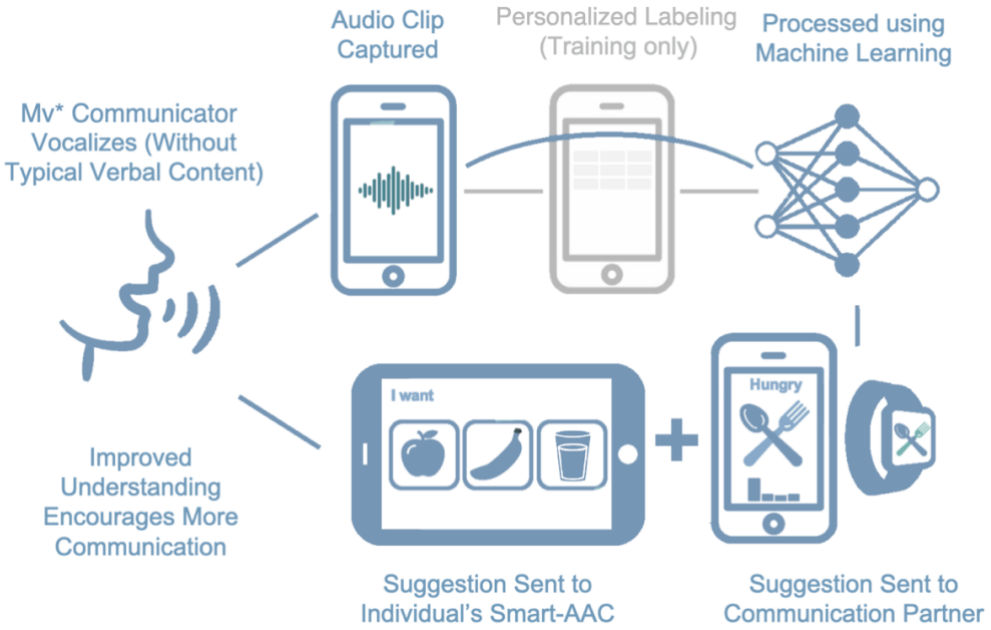


Figure 14: Vision for interface for real-time classification of nonverbal vocalizations. A personalized model would be used to classify a vocalization in real-time, and the result would be sent to the mv communicator's AAC or directly to a communication partner.*

integrated with available AAC technology, and (3) an educational application to help people who don't know the communicator well learn about the mv* individual's nonverbal communication style. Chapter 10 presents prototypes of a classification and educational interface. The work presented in this chapter was created collaboratively with Kristy Johnson and Craig Ferguson (app and UI developer) and this chapter includes excerpts from [125]–[127].

Because there have been no known systematic studies of the function of nonverbal vocalizations from mv* communicators, the first critical step towards improved communication technology and enhanced knowledge around nonverbal vocalizations was designing a novel data collection system for real-world audio with personalized in-the-moment labels. Ensuring that the collected data could inform the future development of interfaces, like the ones described in 5.1 and 5.2, was a guiding factor behind the design of the data collection system.

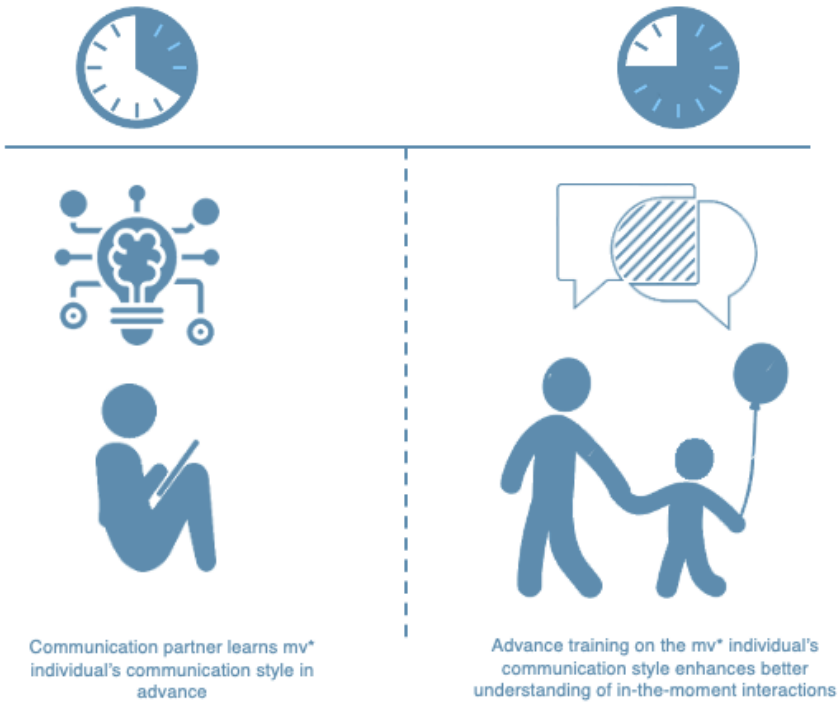


Figure 15: Vision for an educational interface to help the community-at-large better understand nonverbal vocalizations from mv communicators.*

5.1 Vision for Classification Application

In the classification interface, an app or device would capture an audio clip containing a vocalization and classify the vocalization using a pre-trained model. The interface would require a training stage where someone who knows the mv* communicator well (or the communicator themselves, if possible) provided personalized vocalization labels. The inferred classification could then be displayed as a suggestion to a communication partner. This feature would be used by communication partners who do not yet know the mv* communicator well (e.g., a new teacher or caregiver, extended family members). For mv* communicators who use AAC, the classification could also be sent to the AAC and could help the communicator parse AAC options more quickly. For example, if the vocalization was a request, the AAC interface could show items that the communicator might be requesting.

Figure 16: Data collection kits sent to participants. Data collection was conducted entirely remotely, allowing us to reach a small, geographically distributed population.

5.2 Vision for Educational Application

In the educational interface, a communication partner would listen to labeled vocalizations to learn about an mv* communicator's communication style. The interface would allow the user to listen to vocalizations grouped by class so that the user could learn what to listen for in vocalizations. After learning more about the mv* communicator's unique communication style, the communication partner would be better able to understand and respond to communication in-the-moment while interacting with the communicator.

5.3 A Novel System for Collecting Real-World Audio Data with In-the-Moment Labels

Developing new knowledge and communication technologies around nonverbal vocalizations from mv* individuals required designing a new data collection protocol for collecting real-world data with personalized labels.

Figure 17: Data was collected using a small audio recorder (Sony IDC-TX800) that could be attached magnetically to participants' clothing. The recorder was easy to use during day-to-day activities, like reading.

The study of nonverbal vocalizations by mv* individuals presents unique challenges. The population is relatively small and geographically distributed and the resource burden on the population is high. The study was designed to be conducted entirely remotely, to reach a small geographically distributed population and minimize time burden from extraneous activities like transportation to a study center. The remote nature of the study enabled data collection even during COVID-19. Data was 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.

5.3.1 Equipment

Participating families were sent a data collection kit (Figure 16) with all equipment, peripheral cables, spare components, and written instructions.

5.3.1.1 Audio recorder

Audio was recorded using a Sony IDC-TX800, recording in 16-bit 44.1 kHz stereo, attached magnetically to the communicator's clothing or placed near the communicator (Figure 17). The recorder was selected because it was small (38.0 x 38.0 x 13.7mm) and lightweight (22g) and could be worn comfortably during day-to-day life. The recorder also had a high dynamic range and could record loud vocalizations without clipping when worn by an mv* communicator. The recorder also had a reliable internal clock. These important parameters identified in Phase 2 of the case study.

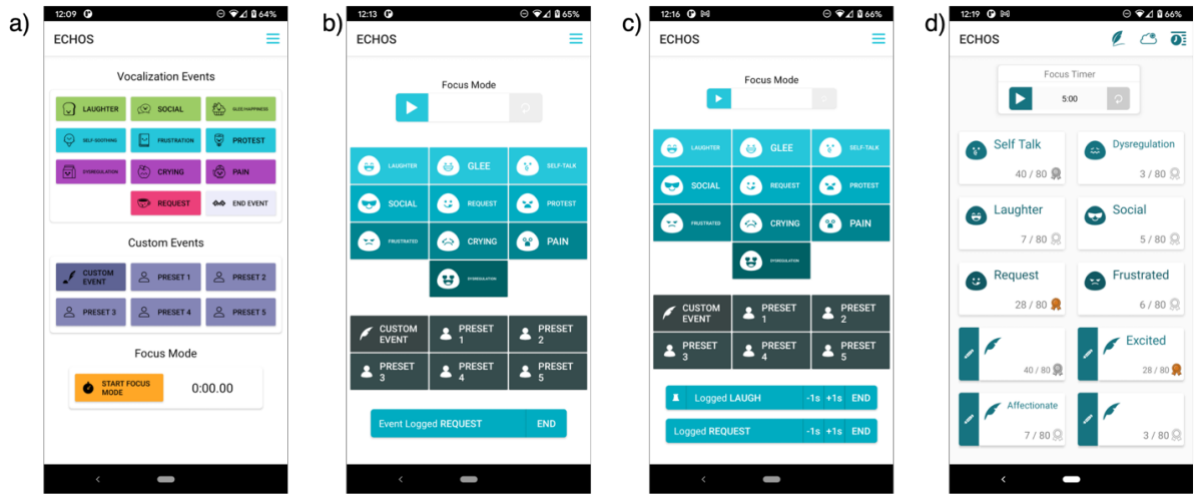


Figure 18: App design evolution showing major design changes. a) The first version of the app had smaller label buttons and a specific "End Event" button. b) The next version of the app had an updated UI design. After a label was pressed, a bar appeared at the bottom of the screen with a small button for the user to mark a label "end". c) In an intermediate version of the app, a labeler could 'pin' a label and then select another, allowing a user to specify multiple labels for a given vocalization. During testing, we found that allowing multiple simultaneous labels led to mislabels and confusion which reduced overall label fidelity and the feature was discontinued. d) In the next version, the number of labels was reduced, and label button sizes were increased. A user presses the label once to start a label, and then presses the label again to end it. The numbers on each label show the number of labels recorded by the user along with a target number. The labeling app was developed by Craig Ferguson.

The audio recorder also came with a remote control. This functionality allowed a parent to start and stop recording easily after and without having to physically access the recorder itself after the recorder had been attached to the participant's clothing or placed in the desired location.

5.3.1.2 In-the-moment labeling app

Vocalizations were labeled in-the-moment by a close family member (i.e., a "labeler") using a custom app. The app design resulted from iterating during the longitudinal case study described in Chapter 4 and feedback from an additional three families of mv* communicators who piloted the app. Figure 18 shows major design changes in the evolution of the labeling app. The first version

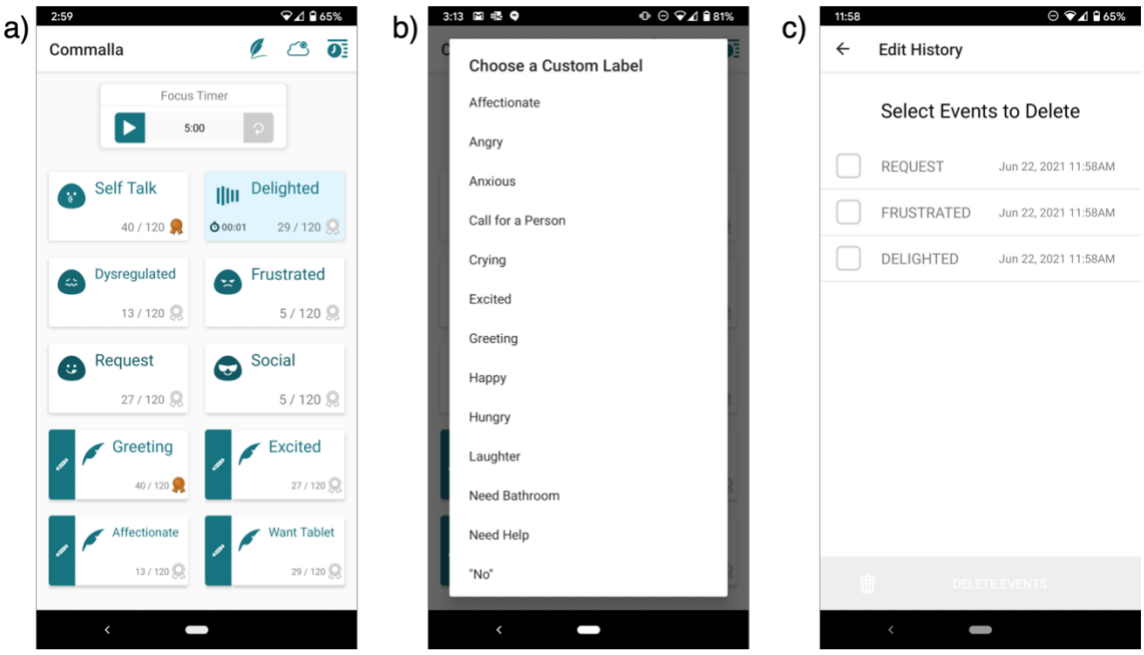


Figure 19: Final deployed version of custom in-the-moment labeling app, showing a) the main labeling interface, b) the list of preset labels, and c) the interface to edit the labeling history

the app (Figure 18a) had small, multi-color labeling buttons. A user could mark the label start time by pressing the button for that label and mark the label end time by pressing the "End Event" button that was in the labeling grid. Later versions of the app had an updated color scheme (Figure 18b-e). The first version of the app did not show the user which label was active and it was hard for a user to find the "End Event" when needed because it blended into the labeling grid. In the next version (Figure 18b), after a label was pressed by the user, a bar appeared at the buttom of the screen to show the user which label was active. The bar included a small "End" button. A version of the app that allowed for multiple simultaneous labels (Figure 18c) was tested. In pilot tests, we found that asking for multiple labels imposed too high of a cognitive load on the labeler and labelers forgot to "End" labels and often marked incorrect labels. The feature was removed for the next version of the app.

In the version shown in Figure 18d, a user presses the label once to start a label, triggering an animation on the button to visually remind the user that the label is active (as on the "delighted"

label in Figure 19a). The user then presses the label again to mark the end of the label. Each button has a counter that shows the number of labels the user has marked for that label and a target number of labels. Small medal icons are activated when labeling targets are reached to provide encouragement to the user. This version of the app is like the deployed version (Figure 19) but had smaller target label numbers.

The main labeling interface included six labels shared among all users: "selftalk", "delighted", "dysregulated", "frustrated", "request", and "social". These labels were chosen based on interviews with families of mv* communicators and conversations with a speech and language pathologist because they convey important communicative information are commonly associated with vocalizations. Families could customize four additional labels, by selecting from a dropdown list (Figure 19b) of twenty-five more specific preset options (e.g., "hungry", "greeting"). The "Focus Timer" shown on the app interface was designed to allow labelers to keep track of how long they had been engaged in labeling and encourage people to label for short, focused periods of time. The "Edit History" interface allowed labelers to delete labels that may have been incorrect or pushed accidentally (Figure 19c).

Participants were given a dedicated phone for labeling. A shortcut to the app was included on the home screen on all phones given to participants, to reduce the time required to access the app. Accessing the app required 3 button presses and screen touches: an initial press to turn the screen on, a swipe up to view main interface with installed app icons, and a touch on the app icon to select the app. These steps may add some delay when a labeler begins a new labeling session. As shown in Chapter 7, most participants assigned multiple labels in each labeling session. After the app is opened once to begin a labeling session, the labels can be pressed immediately without requiring any preceding button presses or screen touches.

5.3.2 Multi-Media Instructions

Because the study was conducted remotely, multi-media instructions were provided to families to ensure participants understood how to use the equipment and label data. The study activities and materials were approved by an institutional review board. At the beginning of the study, each participant had a personal meeting or call with the research team to review the study protocol, ask

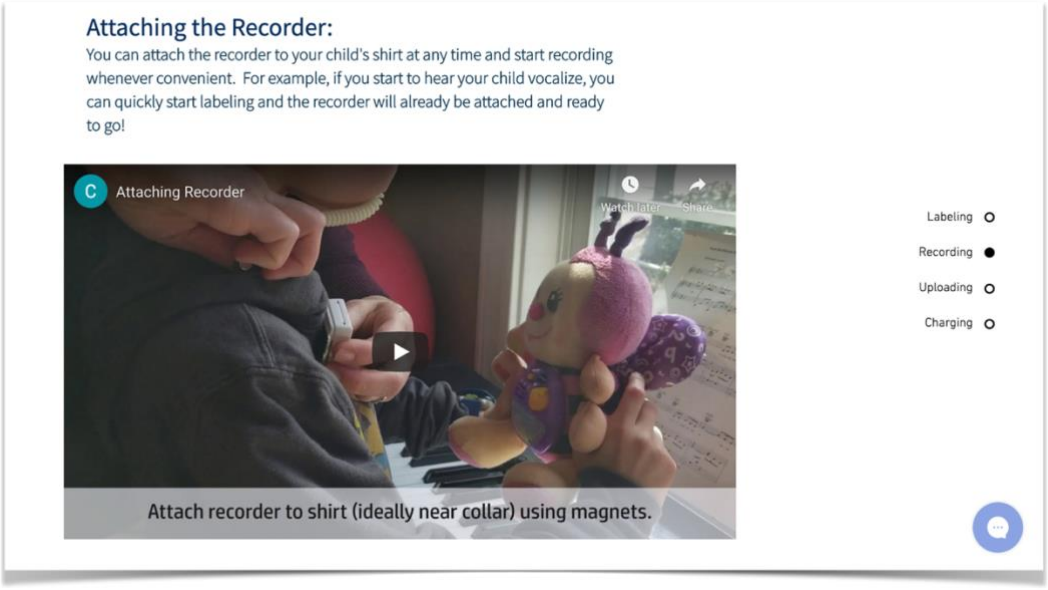


Figure 20: Screenshot of website (http://commallamit.wixsite.com) with instructions and embedded videos provided to participants. Participants were also provided with a YouTube playlist of instructional videos and step-by-step written instructions.

any questions, and set up the labeling app. Written instructions were included in each mailed data collection kit and provided as a PDF for each participant. Written instructions included both detailed step-by-step instructions as well as an abbreviated quick start guide for easy reference. Participants were also given access to a YouTube playlist of video instructions, as well as a website with written instructions and embedded videos (Figure 20). The YouTube playlist included an animated social story to help explain the study to mv* participants. Finally, participants were given the researcher's contact information and encouraged to reach out with any questions as the study progressed.

5.3.3 Conclusion

Real-world data collection, personalized in-the-moment labels, and remote participation were critical for collecting representative labeled vocalizations by mv* individuals. The presented data collection system and materials used to collect the ReCANVo dataset presented in Chapter 6. Chapter 6 also provides information on study participants and data processing. An analysis of usage of the data collection system and user feedback are presented in Chapter 7.

6 The ReCANVo Dataset of Nonverbal Vocalizations from Minimally Speaking Individuals

The ReCANVo dataset ("*Re*al-world *C*ommunicative and *A*ffective *N*onverbal *Vo*calizations") includes 7,077 vocalizations collected longitudinally with 8 mv* communicators. The ReCANVo dataset of nonverbal vocalizations from non-speaking individuals is, to our knowledge the only dataset with vocalizations from mv* individuals; the largest available dataset of nonverbal vocalizations, and one of the only datasets collected in real-world settings with personalized labels. Improved understanding of nonverbal vocalizations could contribute to the development of technology to augment communicative interactions, contribute to answering questions around language development, and expand awareness around this form of communication. We hope that the published dataset will engage other researchers in this critical field of study.

This chapter includes work created jointly with Kristy Johnson and excerpts from [126], a publication that will include the full dataset for use by other researchers.

6.1 Background

Often, speech emotion datasets are collected in lab-environments with actors [128], [129]. Naturalistic datasets have previously been compiled using data collected from specific activities like call center recordings, video conferencing, and podcasts [110], [130], [131]. The FAU-Aibo

dataset contains recordings of children interacting spontaneously with a pet robot [132] but is not publicly available. Lab and activity-based datasets provide valuable clearly annotated emotional speech sample but are limited by their settings. A dataset collected during day-to-day life can capture longitudinal expressions across contexts.

Several datasets exist that focus specifically on nonverbal vocalizations that occur amidst typical verbal speech. Lima et. al. published a dataset of 120 acted emotional nonverbal vocalizations containing achievement, amusement, pleasure, relief, anger, disgust, fear, and sadness [133]. Anikin et. al. compiled 260 naturalistic vocalizations from YouTube [100]. The OxVoc dataset includes 173 natural affective vocalizations from infants and adults [134]. The infant vocalizations were recorded in real-world settings for a research study and the adult vocalizations were compiled from YouTube. The ReCANVo dataset is the first to capture nonverbal vocalizations by mv* communicators used independently of typical verbal speech that are labeled by function.

This population of mv* communicators is highly heterogeneous, including diagnoses of autism, genetic disorders, Cerebral Palsy (CP), and global developmental delays. The specific etiologies of certain behaviors and symptoms are often not known. For example, a person may not speak due to motor planning difficulties, cognitive delays, differences in social motivation, some combination thereof, or alternative causes. In addition, abilities and/or behaviors affecting communication can evolve over time, further augmenting the heterogeneity of this group. Importantly, an individual may communicate in one way in their home or with their parents but completely differently or not at all in a laboratory setting or with examiners (e.g., [135], [136]), underscoring the need for in-situ environmental contexts, familiar people, and real-world data collection. Collecting a representative spread of vocalizations across labels required longitudinal real-world data collection.

Collecting ground-truth labels for nonverbal vocalizations is an open problem. Most mv* communicators cannot directly provide word-based labels. Understanding these vocalizations often requires familiarity and camaraderie with the communicator, so researcher or crowdsourced labels would have inaccuracies and bias. For this study, labels were provided in-the-moment by a close family member or caregiver. While any label from a person other than the communicator remains a proxy, labels from a person with a long-term relationship with the communicator are the closest obtainable ground truth. Because the labels were marked in-the-moment, the labeler had access to the full context (i.e., location and temporal information) as well as other communication from the mv* individual like gestures and body language. Personalized in-the-moment labels may have a positive impact on labeling fidelity, as labels based on acoustics alone may be inaccurate for ambiguous emotional expressions [101].

6.2 Participants

Participants were recruited for the IRB-approved study through conversations with community members and word of mouth. Sixteen participants enrolled in the study, though only eight collected enough data for analysis. Participants were included in the dataset and corresponding analyses if they collected data for at least ten recording sessions to ensure a sufficient number of captured vocalizations and diversity of audio soundscapes. Participating families were asked to fill out a background survey that included questions related to demographic information and communication practices. P02 did not fill out the intake survey but provided some background information via an interview with a researcher. Recorder placement was flexible to accommodator tactile sensitivities. The recorder was attached to the communicator's clothing (P01, P02, P03, P06, P11, P16), worn as a necklace (P08) or placed nearby (P05).

Participant ID	Gender	Age	Diagnoses	Time span of included data (weeks)	Number of spoken words and 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	16	0
P05	F	9-12	Autism	11	0
P06	M	9-12	Autism, Cerebral Palsy (CP)	4	3
P08	F	6-9	Autism	20	0
P11	M	9-12	CP	19	1
P16	M	6-9	Autism	10	5-8

Table 2: Demographic information for participants included in dataset and analysis

6.2.1 Demographics

Participants range in age from 6-23 years old and included diagnoses of autism spectrum disorder (ASD), cerebral palsy (CP), and genetic disorders. They all have fewer than 10 spoken words or word approximations, per parent report. No participants were excluded based on age, diagnosis, or other measures in order to capture the broadest possible sample of this unique and understudied population of communicators. Selected background and demographic information for participating mv* communicators are in Table 2.

6.2.2 Communication Profiles

Communication profiles were created for each participant using the survey data provided by the mv* communicator's parent. The communication profile includes the communication and word approximations used by the communicator, and feedback on if and how the communicator uses

speech (both verbal and nonverbal) across various communicative and affective categories. The parent-reported use of vocalizations and word/word approximations across provided categories are shown in Figure 21. Missing data for two categories for P11 was filled in by researchers using interview feedback.

The survey also asked respondents to note which modalities the communicator used when communicating via a multiple selection question that asked participants to select all relevant answers from a provided list and/or to submit a custom response. The responses are listed below:

- P01: Vocalizations that do not have typical speech content, gestures, hand leading, pulling, sign language (modified or official), picture cards, AAC device - "TouchChat"
- P03: Vocalizations that do not have typical speech content, gestures, hand leading, pulling, sign language (modified or official), picture cards, AAC device - "TouchChat"
- P05: Gestures, hand leading, pulling, sign language (modified or official)
- P06: Words or word approximations, vocalizations that do not have typical speech content, gestures, hand leading, pulling, AAC device - "Proloquo2go on iPad"
- P08: AAC device – “Novachat 10”, other - "Pointing or bringing what she wants"
- P11: Vocalizations that do not have typical speech content, gestures
- P16: Words or word approximations, gestures, hand leading, pulling, AAC device - "TouchChat"

6.3 Labels

While recording, the communication partner labeled vocalizations as they were produced. For example, a communicator might request a drink by vocalizing and gesturing toward a cup. The communication partner would then tap the "Request" label on the smartphone labeling app (Figure 19).

- **Selftalk**: Vocalizations that appear to be associated with being content, happy, or relaxed, and are generally made without other 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.

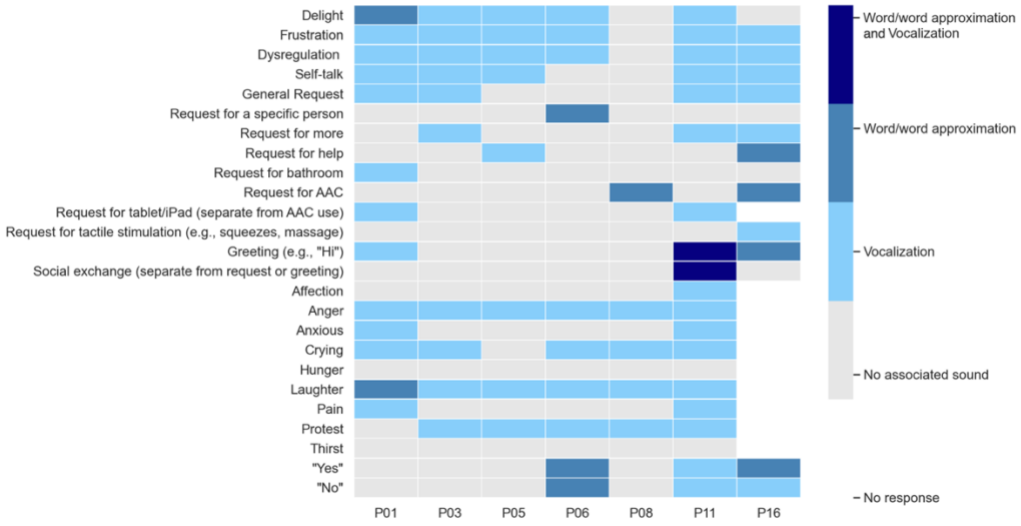


Figure 21: Parent-reported use of word and word approximations on survey.

- 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.

These six labels and descriptions were selected based on conversations with families of mv* communicators and a speech and language pathologist (SLP). Label descriptions were shared with families. Particularly, the distinctions between "dysregulated" and "frustrated" and between "selftalk" and "delighted" were discussed with families. As described above, "frustrated" was defined as an acute or active emotion while "dysregulated" described vocalizations associated with a state of general stress. Similarly, "delighted" was an active or acute state where "selftalk" described associated with a state of general contentment. Participants were instructed to use only those labels for which the communicator had a vocalization.

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"). The "social" label was used broadly for interactive vocalizations (e.g., a back-and-forth exchange with a family member) and the "request" label was used broadly for general request vocalizations. More specific "social" categories like "greeting" and "request" categories like "want more" could be selected from the preset options if applicable. 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, 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 a label that they felt would help others best understand how to respond to a vocalization.

In addition, families could customize four preset options from a drop-down menu. The labels included in the drop-down menu were: "affectionate", "angry", "anxious", "call for a person", "crying", "excited", "greeting", "happy", "hungry", "laughter", "need bathroom", "need help", "'no'", "pain", "protest", "singing", "social fun", "teeth grinding", "thirsty", "want more", "want tablet", "'yes'". Families could also provide custom feedback/labels via an "Add notes" button on the app (the feather icon in Figure 19) or by contacting the research team. Not all participants used the preset options; they were only used if the family indicated that the communicator had additional specific vocalizations that they wanted to capture. Hence, the presets were specific to each participant. Participants were guided through selecting presets during app setup, after which these labels were not changed. Thus, each participant had a fixed set of 6-10 labels to use throughout their study. Figure 24 shows the presets and labels used by each family.

Labelers were asked to achieve as close to a 1:1 mapping between a vocalization and a label as possible. However, not all participants followed this instruction closely; some participants designated long periods of time containing multiple vocalizations as a single label. The app required labelers to indicate a ‘start’ and ‘end’ time for a vocalization by tapping the corresponding label. A color change and animation appeared on a label that had been ‘started’ to visually indicate which label was active (shown on the “Delighted” label in Figure 19).

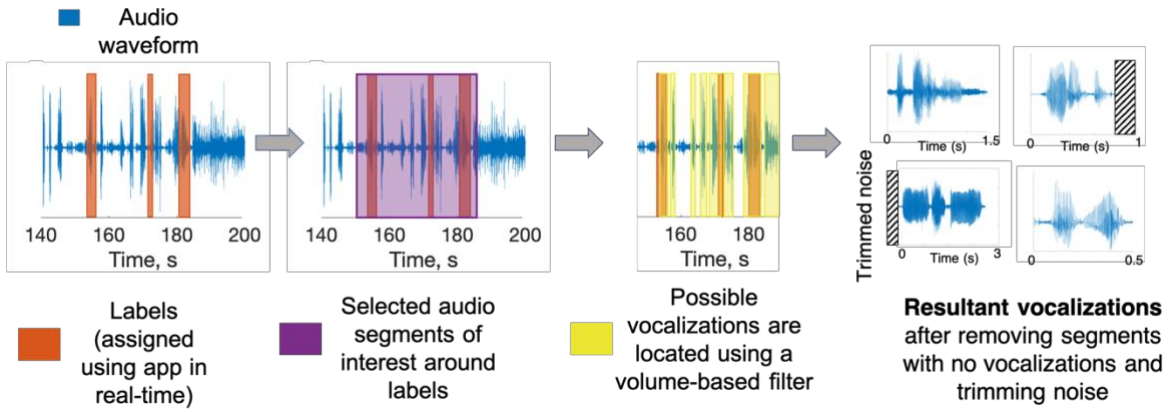


Figure 22: The audio data and real-time labels from the app were processed to align labels with vocalizations. A volume-based filter was used to isolate audio segments of interest. Segments temporally near a label were assigned to that label. A researcher listened to each segment to ensure it contained a vocalization and, if necessary, trim excess noise around the vocalization.

6.4 Dataset Preparation

The data collection system is described in detail in section 5.3. 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.

6.4.1 Alignment of Audio and Labels

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

The audio recordings and label information were then processed to isolate vocalizations of interest and align them with the assigned vocalization labels. Because participants were instructed to

record and label at their discretion, we first isolated regions of audio that were temporally near labels (Figure 22 and Figure 23). Then, because the recorder was attached to the communicator's clothing or placed nearby, a volume-based filter was used to isolate smaller audio segments within these regions that were likely to be vocalizations (Figure 22). The volume filter thresholds were selected for each session based on the recording levels during that session; they ranged between -20 and -45 dB. Vocalizations were considered distinct (separate vocalizations) if they were separated by approximately 250-450 ms of silence, determined based on the volume levels and background noise of that session's recording.

Isolated segments were assigned a label based on the following rules (Figure 23):

1. The audio segment was within the label bounds.
2. The audio segment ended during a label. This timing occurred naturally when a labeler pressed the label after hearing and recognizing a vocalization.
3. The audio segment started before the label started, and the label began within 15s of the segment start. This threshold was determined after listening to many files to account for human labeling delay.
4. The label began 3 seconds or less after the segment ended, even if the audio segment ended before the label started. This threshold was determined after listening to many files to account for small clock drifts and human delay.
5. The segment started within a label and ended within 3s of the label end. This timing threshold accounted for changing vocalizations during real-world data collection. For example, a labeler might have ended a label because the vocalization had changed, and the current label was no longer accurate. This rule is necessary because some labels

encompassed multiple vocalizations and so some vocalizations began after the label had been pressed.

These rules were determined heuristically by comparing the timings of labels to the full-length audio files. The background audio and conversational exchange in the audio files provided context to determine if a label matched a given vocalization. If multiple distinct labels satisfied the rules above, a single label was selected, prioritizing the label with the rule with the lowest number in the list above. Note that not every vocalization in a recording was assigned a label. Unlabeled vocalizations were not included in this dataset.

After labels had been assigned to audio segments, a researcher listened to each labeled audio segment. Segments that did not contain vocalizations were discarded. Segments that contained additional noise or voices surrounding a vocalization were manually trimmed.

6.4.2 Technical Validation

We identified three expected sources of labeling error:

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

To mitigate the first two sources of error, a researcher listened to the audio surrounding each labeled vocalization. A researcher also listened to full-length audio recordings for each participant at least every two weeks of collected data. The surrounding context from the audio recording, such as spoken dialogue that confirmed an emotional label (E.g., "Oh, I can see that you're frustrated.") or answered a communicator's request (E.g., "You would like a snack?"), was used to confirm that the assigned labels matched the audio context near the label.

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

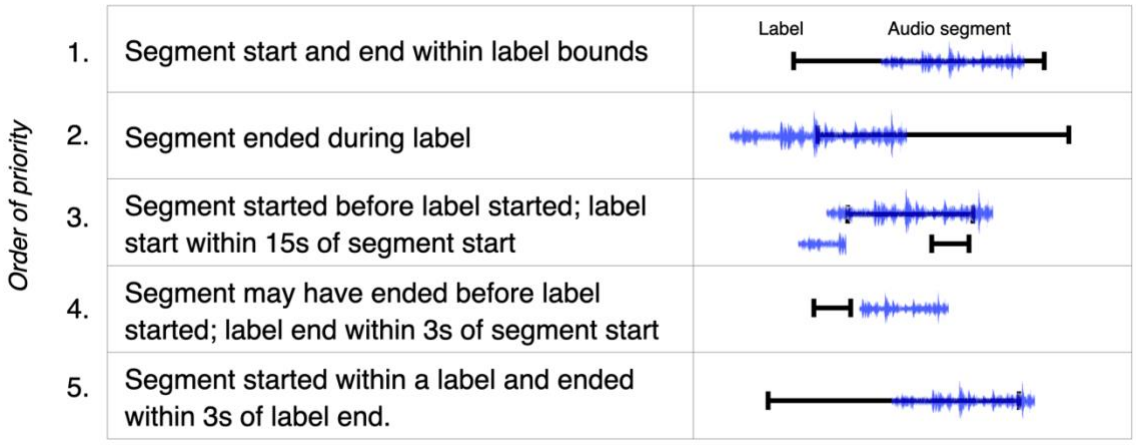


Figure 23: Illustration of rules for assigning labels to segments. The rule numbers in the figure correspond to the descriptions in the body of the thesis

Because of the longitudinal nature of the study, some clock drift (~10 seconds or less) was observed for some participants. This drift was accounted for manually when aligning the labels with vocalizations. In addition, there were expected sources of noise associated with real-world data, including environmental noise (e.g., wind, movement, background toys and electronics), overlapping voices, intensity changes due to variable location of the recorder. Many extraneous sources of noise were removed during the segmentation process or through manual trimming, however, vocalization segments of all qualities were included here to ensure naturalistic, real-world data transfer.

6.4.3 Data Privacy

Families were given full control over sharing any audio recordings. Only text-based labels were synced automatically to give participants control over the continuous audio recording files. The provided instructional materials included instructions on how to delete audio files that families did not wish to share. Families could delete any recordings directly on the recorder or on their computer before uploading data.

6.5 Dataset Overview

The published dataset [126] contains audio recordings of segmented vocalizations, labeled by vocalization meaning or function. A data folder containing 16-bit, 44.1 kHz .wav files for each participant is provided. The vocalizations are organized by label for each participant. A .csv file is also provided that has the name of each vocalization file, its path relative to the dataset, and the corresponding participant ID and label. In addition, demographic and communication profile data are provided in a .csv files.

The ReCANVo dataset includes 7,077 vocalizations collected longitudinally with 8 mv* communicators. Figure 24 shows the number of vocalizations in the dataset for each class and participant. To our knowledge, the ReCANVo dataset is the first dataset of nonverbal communication that occurs independent of typical verbal speech, the largest existing dataset of nonverbal vocalizations, and the first public dataset of affective speech collected longitudinally during day-to-day life across settings. Access instructions for the dataset will be available at https://www.media.mit.edu/projects/commalla/overview/.

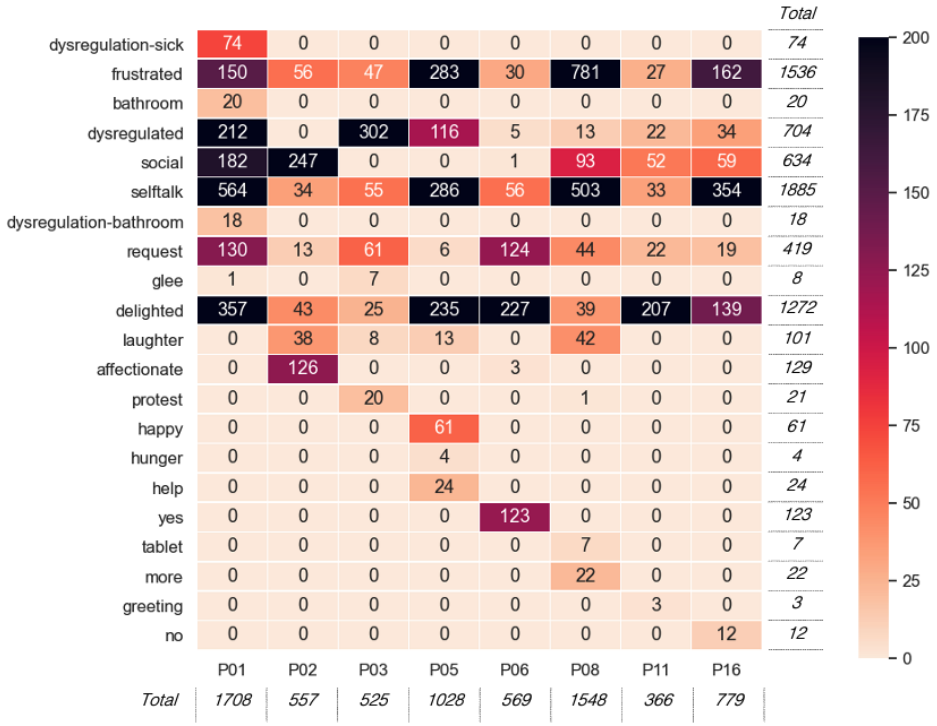


Figure 24: The number of vocalizations included in the dataset, organized by participant and labeled affective or communicative function

7 Data Collection System Evaluation

The data collection system was evaluated by compiling usage statistics for each participating family and by conducting exit interviews with labelers. The analysis of labeling and data collection practices provides insight into how our novel in-the-moment labeling system was used by each labeler, which can inform the design of in-the-moment tagging systems more generally. Additionally, the analysis of labeling and data collection practices provides valuable context for interpreting the modeling results presented later in Chapter 8.

7.1 Labeling Statistics

7.1.1 Methods

Labeling and data collection practices were tabulated for each participant including labeling delay, average recording session length, number of uploaded sessions, average label duration, average number of vocalizations per label, average number of labels per session, and median number of unique labels per session. The labeling delay was defined as 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.

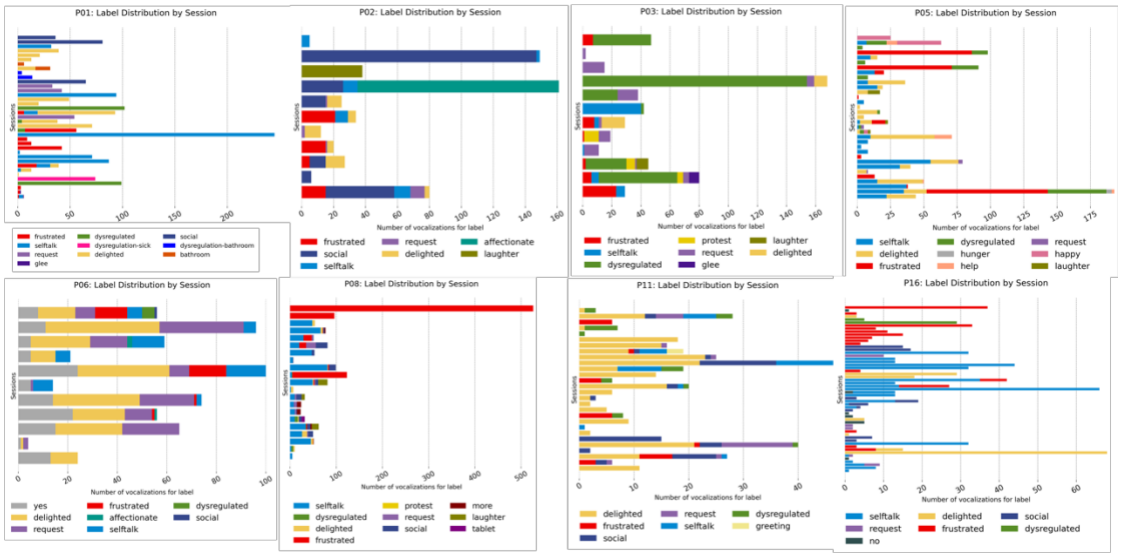


Figure 25: Distribution of labels by session for each participant. Each horizontal bar shows the distribution of labels in a particular session for a participant.

7.1.2 Results

Figure 25 shows the distribution of labels by session for each participant. For each participant, Figure 26 and Figure 27 show box-and-whisker plots of the labeling delay and label duration for each participating labeler, respectively. The presented statistics can inform the design of other in-the-moment tagging systems. Table 3 shows the average session length, number of sessions, average label duration, average e number of vocalizations per label, the average number of labels per session, and the median number of unique labels per session.

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 (Figure 25) and were not included in the delay analyses. Some of the tabulated statistics could also not be calculated for P01 and P16 for this reason. The other participants more closely followed the preferred collection protocol, designating labels for vocalizations that occurred within a recording.

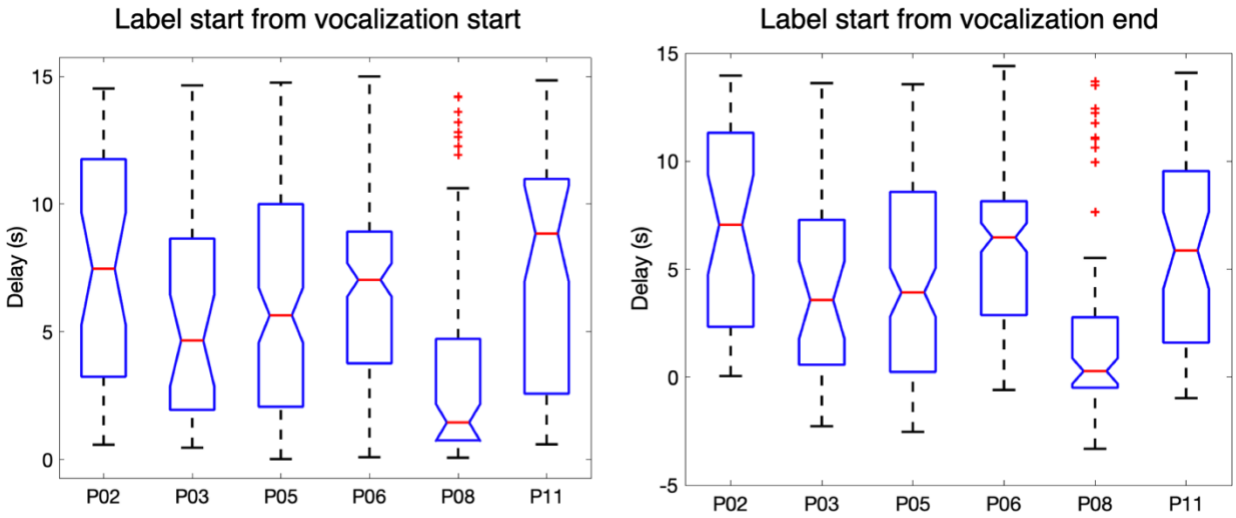


Figure 26: Distribution of labeling delays for each participant. The labeling delay was defined as the difference in time between the start (left) or end (right) of a vocalization in the audio recording and the timestamp of the button press on the labeling app for the corresponding label.

P01 and P16 were not included in the delay analysis because they deviated from the preferred labeling protocol by assigning labels to longer recordings instead of individual vocalizations

The average labeling delays ranged from 3.5-7 seconds and were different between labelers (Figure 26). 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. For example, P08 had significantly longer average label durations than other participants (Figure 27). Very long labels risk encompassing vocalizations of multiple categories and could be associated with mislabeled data. P02, P03, P05, P06, P08, and P11 all tended to have multiple unique labels

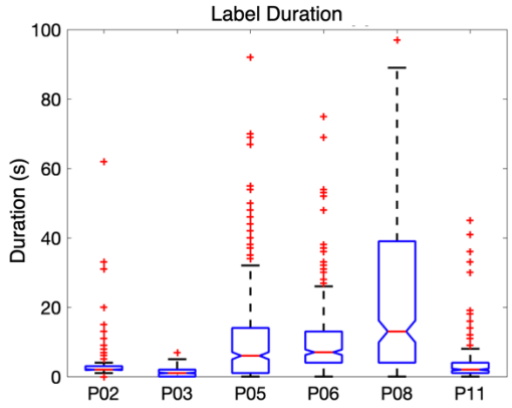


Figure 27: Distribution of label duration for each participant

in each session. Having multiple unique labels per session may reduce the model's likelihood of learning to associate a label with a particular background soundscape.

	P01	P02	P03	P05	P06	P08	P11	P16
Average 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
Average label duration (s)	N/A	9.7	0.9*	13.5	10.7	50.3	3.3	N/A
Average vocs. per label	47.4	8.0	2.5	4.7	2.4	7.2	2.1	13.7
Average number of labels per session	2.6	11.3	26.1	11.7	30.0	16.7	11.7	1.1
Unique labels/session (median)	1	3	3.5	2	5	4	3	1

*Table 3: Statistics describing labeling and data collection practices. *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*

7.2 Interviews

7.2.1 Methods

Conversational interviews were conducted with parents of the mv* communicators in the study to gather feedback on the study design and proposed interface designs, and better understand labeling and data collection practices among study participants. Interviewed parents had all collected and labeled data for at least one month prior to the interview. An outline of questions used to guide interviews is provided below:

Interview Questions

1. Did you run into any difficulties getting started with data collection?
2. Was the recorder easy to use?
3. Did anything on the labeling app confuse you?
4. When was it easiest to collect data (e.g., time of day, activity?)
5. Do you feel you were able to label vocalizations confidently?
6. Do you feel you were able to label vocalizations with accurate timing? About how long do you think normally passed between when you heard a vocalization and when you pressed the button on the app to label it?
7. Did you use any context to help choose a label (e.g., gestures, facial expressions)
8. Did participating in this study change anything in your vocal exchanges and interpretations in your family?
9. How long did a particular type of vocalization tend to last when you were labeling?
10. Did the labels on the app include what you wanted to label? Was anything missing or unclear?
11. Are there things that would have made it easier to collect data?
12. How did you feel about labeling on a mobile phone? Do you think you might have preferred something else (e.g., a smartwatch or tablet)?
13. Are there other ways that your child communicates that you would want to help other people understand?
14. Would a classification interface for vocalizations have utility for your family? (An interface like the one in 5.1 to classify vocalizations in real-time and integrate with AAC devices was described to the interviewee)

15. Would an educational interface for vocalizations have utility for your family? (An interface like the one in 5.2 to teach new communication partners what to listen for in vocalizations of different types was described to the interviewee)

Interviews were conducted over the phone and lasted 20-40 minutes each. Interviewees were asked follow-up questions based on their responses. Interview audio was recorded and transcribed. The data was qualitatively analyzed to extract themes and commonalities between interviewees.

7.2.2 Results

Interviews were conducted months after data collection with P02 was complete, and P02's family was not available for an interview at the time of inquiry. Interviews were conducted with parents of all other participating mv* communicators. Feedback obtained during interviews was grouped by category based on whether the feedback was related to the study design, study equipment, labeling accuracy, labeling practices, and label interface design. Feedback on communication interfaces is discussed later along with the prototype interface in Chapter 10. The feedback for each category is summarized below, where a parenthetical number designates the number of interviewees who made a particular comment. Because the interviews were conversational, sometimes participants provided feedback that related to multiple questions without prompting. For the reader's reference, the question most related to the cited response is shown in italic. Feedback on interface design (Q13-Q15) is presented in Chapter 10. Given the personal nature of the interviews, quotes were edited to remove gender to preserve anonymity.

7.2.2.1 Study Design

Q1. Interviewees reported that the combination of 1:1 instructions via video calls, video tutorials, and written instructions were clear. Interviewees did not frequently use the website but did watch the instructional videos on YouTube (5). One interviewee reported some confusion as to when to start and stop a recording and said that they tried to specifically record around short vocalizations at the beginning of the study but realized they should make long recordings with shorter labels after a couple weeks (1). *Q11.* There was an ebb and flow in collecting data based on life events, with some difficulty finding time to collect data when things were particularly busy (3). Two interviewees mentioned that sometimes they did not have the recorder on hand when they wanted to collect data (2).

Q8. Interviewees generally said that participating in the study did not change their vocal exchanges because they already listed to their child's vocalizations (4). Families who paid attention to vocal exchanges were likely more willing to and interested in participating in the study. One interviewee mentioned that their family had been recording their child's vocalization for years prior to the study and were excited that they could now record vocalizations as part of a larger study (1). One interviewee mentioned that they paid attention to vocalizations more during the study but didn't notice any specific patterns other than a clear differentiation between happy and mad screams (1).

Quotes from interviewees on the overall study design:

- *"Unlike other studies, it isn't intrusive at all…it doesn't affect our day-to-day life."*
- *"The main problem recently has been when we go into these modes when life seems too hectic. Life gets hectic and then you get less disciplined about these things"*

Quotes from interviewees on the study instructions:

- *" When I first started I just read the directions, then I went back and watched them too. It was more clear watching it then from reading. The two together (video and written) helped. I got it and was so excited I just wanted to get started and do it "*
- *" Definitely needed the Zoom tutorial. Just reading it didn't make sense to me"*

7.2.2.2 Study Equipment

Q2. Two participating families found an alternative to attaching the recorder directly to the clothes because attaching to the clothes was uncomfortable because of the mv* individual's tactile sensitivities (2). One participant placed the recorder near the mv* communicator when collecting data (1), and one participant attached the recorder to a necklace worn by the communicator (1). Two parents emphasized that they thought the recorder with magnets was very comfortable to wear, but that the mv* communicator had high tactile sensitivities and would not like to wear anything attached to their clothing (2). This feedback suggests that in the long-run non-wearable communication devices might be best suited to meet the needs of mv* communicators.

Q12. Interviewees reported that they would prefer labeling on a phone compared to other devices like tablets and smartwatches (5), though this feedback may have been biased because labelers were only given a labeling app on a mobile device and so they did not have a chance to try other modalities. One interviewee mentioned some issues with labels being pressed accidentally when taking the phone out of their pocket (1). One interviewee mentioned that the phone could be distracting because when the mv* communicator saw it they wanted to play on it (1), and one parent mentioned the communicator's siblings wanted to play with the phone (1). Two interviewees said that they might be interested in labeling on a smartwatch because of the portability of the watch but were concerned the buttons on a smaller screen would not be large enough to label

quickly and accurately (2). Two interviewees mentioned that they would have liked to collect data in the car because the communicator often vocalized in the car, but that they couldn't do so because the labeling process required using an app (2). The feedback from interviewees suggests that labeling on a phone app is tenable, but that there is a potential latent need for a portable hands-free labeling device.

Quotes from interviewees on the recorder:

- *"There were a couple times where the recorder was not charged. That was annoying. Once I got the hang of using the remote that was much easier. That was definitely an asset to the set up to have the remote, so I didn't have to go finagle with the recorder to turn things on and off"*
- *" [They] would not keep the recorder on at all. No matter where it was on their clothes they would pull it off. We were able to just set it next to them. If they got up and moved, we would just follow them around moving it...I can't think of anything that would have made it easier to wear. I think it was just the size and weight....They are not drawn to a character or a color. It's just this thing that's on them and they want it off. We tried to hide it – on their pants, on their back. It's not a big deal but they are so highly sensitive and so aware of anything on their body that they are not having it. It's just them, it was really small and light and easy to wear"*
- *"They [the communicator] would usually pull on it or mess around on it when we first put it on but then then they didn't seem particularly bothered by it after so that was good."*

Quotes from interviewees on labeling using the phone:

- *"I like the phone. I think the phone is nice especially because everyone has phones anyway. The one complaint I did have is that if you put it into a pocket to carry it sometimes it will label if you didn't put it to black. Sometimes it will label when you are pulling it out."*
- *"It's hard because I would like to do it in the car but I'm driving. If you could tell it you were in a car ride and just press to record...You [a researcher] would have to listen to see what's going on as opposed to the button pressing."*

7.2.2.3 Labeling practices and interface

Q3. Interviewees reported that the labeling interface was intuitive and easy to use (7). Two interviewees reported some confusion about the relation between the phone and recorder and mistakenly thought that the phone was recording audio (2). One interviewee had initial difficulty using the preset menu options (1), which was resolved via consultation with a researcher.

Q5. Interviewees generally felt confident that they were able to label vocalizations soon after they occurred but mentioned that there was an added delay when they were multi-tasking (2) or had to reach for the phone if they had just started labeling or had put the phone down (3). *Q6.* Interviewees estimated it took about 5 seconds (1), 10 seconds (1), and 3 seconds (1) to label a vocalization on the phone after identifying it. One interviewee estimated that approximately 10% of their labels were not accurate because they hit the wrong button while trying to label in-the-moment (1).

Q4. Interviewees said that it was easier to label when another person was present (4), like another parent or an at-home therapist like an ABA provider. Interviewees reported collecting data during playtime or down-time (5), during evenings/nighttime when things were less hectic (2), while

watching videos (1) and while on walks (1). Interviewees mentioned that there were certain times of the day during which they were not able to record data during the study (e.g., getting dressed) which affected the breadth of vocalizations recorded (5). One interviewee mentioned that the communicator's vocalizations were more varied when they were outside of the home, and it was not possible to collect data in settings that might elicit those vocalizations (e.g., theme parks) because of COVID-19 (1).

Q10. Interviewees reported that some vocalizations had multiple or ambiguous meanings, like a "frustrated request" where the communicator was both requesting something but also expressing frustration that their request had not been previously fulfilled (3). One interviewee mentioned that more fine-grained categories would be useful for their family, like "requesting a fruit" (1). *Q9.* Interviewees reported that how quickly the vocalization changed depended on the type of vocalization (3). One interviewee said that selftalk and dysregulation tend to happen over longer time scales while frustrated and request are pointed occurrences that happen for short time scales. The interviewee said that vocalizations that occurred on shorter time scales were more difficult to capture because the communicator may have stopped vocalizing by the time that they were able to get the equipment to record (1). Another interviewee said that for their child, frustration, selftalk, and dysregulated tended to occur on longer time scales than delighted (1).

Q7. All interviewees reported using other context in-the-moment to choose labels for vocalizations. Interviewees mentioned using gestures (5), facial expressions (2), hand pulling (2), AAC output (2), general context like the activity (2), posture (1), and physical objects brought by the communicator (1). In future work, capturing additional communication modalities like facial expressions and gestures could be informative. However, there is a privacy trade-off involved with collecting

additional data streams. In the study design phase (described in Chapter 4), we found that collecting video data that would have captured gestures and facial expressions was too invasive of privacy. Labeling data in-the-moment gives the labeler access to the full context so that they can label data more accurately without requiring extra data to be recorded. The presented data collection process may provide a privacy-sensitive way of including synthesized contextual information in studies.

Quote from interviewee on data collection timing:

- *"[It was easiest to collect data] if there was another person around, often during ABA therapy when someone else was playing with them [the communicator]. That was a good opportunity for me to get data label while they were occupied doing the activity. The activities themselves do facilitate certain types of vocalizations that may not be occurring as much at other times - things like request or frustration, especially when they are being asked to do something as part of a semi-structured activity. Things like selftalk I was more likely to be able to capture in the middle of the weekend on a Saturday morning or something like that where they are relaxed and doing their own thing."*

Quotes from interviewees on overlapping labels:

- *"Sometimes it is hard to tell the difference between social and selftalk. It's hard to tell if [communicator] is talking to get someone's attention or if they are talking to themselves." – Parent*
- *" Things that were complicated were states that were overlapping like 'frustrated request' or 'delighted dysregulation' which are very close states"*

- *Sometimes it is like hard to differentiate because sometimes they are frustrated but also requesting so it's like a frustrated request but I'm not sure which when to do so I use my best judgement."*

Quotes from interviewees on context used when labeling:

- *" I have to be at the scene and know what is happening. I know what label it is just because they are my child. I know when they [the communicator] are upset or happy or frustrated."*
- *"It's hard when they [the communicator] says something to press a button as opposed to leaving it on through what is going on. I started just leaving it on like for like the whole selftalk period as opposed to pressing it each time."*
- *"If they [the communicator] are really happy they'll laugh and smile. They'll cry if upset, and make grumpy face if mad. There's a lot of facial expressions but you really have to know them to recognize the expression. Not everyone does.*
- *"[The communicator] will use facial expressions. It's easy to tell if he is happy or sad or upset. The facial expressions definitely help with expressing those emotions. For them the facial expression is very important."*

7.3 Conclusion

Understanding how participants used the data collection system can provide valuable background when analyzing the collected data and interpreting results. In Chapter 8, the machine learning modeling results are discussed in the context of how participants used the data collection system. The presented interview feedback on the data collection system guided the design of the new app-based data collection process presented in Chapter 10.

8 Personalized Modeling of Real-World Affective and Communicative Nonverbal Vocalizations from Minimally Speaking Individuals

This chapter presents the results of the largest multi-class real-world nonverbal vocalization classification experiments to date with eight mv* communicators. The experiments show that nonverbal vocalizations can be classified using audio alone for each individual. This chapter presents evaluation and sampling strategies to work with messy, real-world data with uneven class distributions and varying background noise. A custom feature set was designed end evaluated for nonverbal vocalizations from mv* individuals. The model performance results are discussed in the context of how data was collected by each participant. In the presentation and discussion of models and results, the term "sample" is used to refer to a labeled vocalization used for model training or evaluation.

8.1 Methods

8.1.1 Machine Learning Evaluation and Sampling Strategies

Because of the heterogeneity of the participants (Table 2) 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

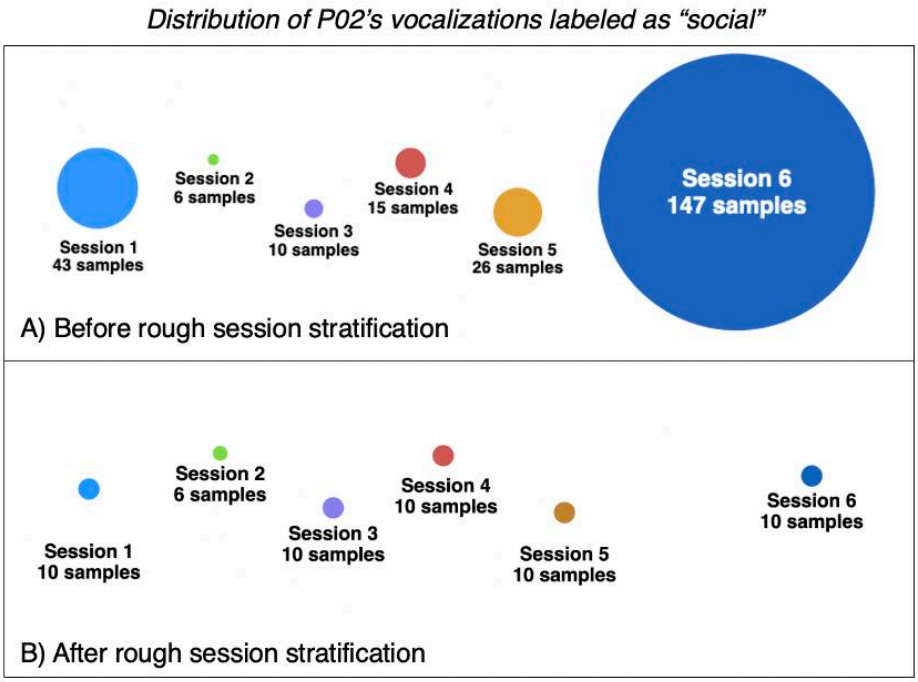


Figure 28: Distribution of P02's "social" vocalizations by session before and after rough session stratification

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. To illustrate this, Figure 28 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

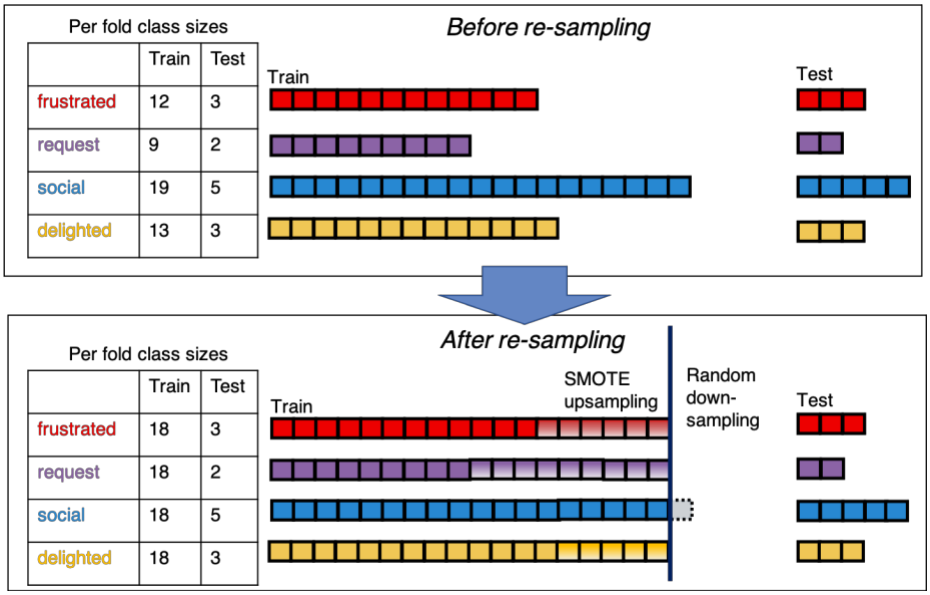


Figure 29: The class sizes were balanced in each fold using random downsampling and the synthetic minority oversampling technique (SMOTE). In each fold, the number of training samples per class was balanced to the minimum of twice the smallest class size and the largest class size. The figure illustrates the balancing strategy within a fold, using pseudo class sizes. The test data was not balanced, so metrics are reported using macro-averages to give each class equal importance.

vocalization recordings. The 5-fold cross validation evaluations were run with 3 distinct random seeds. The average macro-F1 score 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") [137] (Figure 29). SMOTE create synthetic data points by selecting randomly along lines in the feature space between nearest neighbor points in the minority classes [137]. 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 applicability across loops with varying class distributions.

Table 4 and Table 5 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 test data 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.

8.1.2 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 Table 4 and Table 5. For a participant, a label was included in the analysis if there were at least 30 vocalization samples spread across at least 3 sessions. If, for a given participant, there were more than five labels with sufficient data, only one of any closely related labels was included in the model. Closely related labels had similar valence and arousal characteristics (like "delighted" and "happy") or were directly overlapping (e.g., "dysregulation-bathroom" and "dysregulation-sick"). The presented analysis is the first of its kind and is intended to explore the feasibility of distinguishing well differentiated vocalizations - differentiating very closely related categories with large multi-class models was beyond the scope of this analysis. 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.

Multi-class with session stratification								
	P01	**P02**	**P03**	**P05**	**P06**	**P08**	**P11**	**P16**
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 valence with session stratification								
positive 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

Table 4: Number of samples for each class for each participant with rough session stratification. 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.

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

Table 5: Number of samples for each class for each participant without rough session stratification. 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

8.1.3 Binary Valence Classification

Binary 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 4. 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". 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.

8.1.4 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 [138].

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

- the extended Geneva minimalistic acoustic parameter set extracted using openSMILE [113], [114], size 88
- 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-based 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

Feature	Applied Functionals
Audio amplitude	Duration; Mean autocorrelation
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

Table 6: Features and applied functionals used in the custom feature set. Additional implementation details are provided in Appendix B.1.1.

Table 6 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. The custom feature set includes features that capture characteristics of the pitch contour like number of peaks and polynomial fit coefficients. The custom feature set applies a functional that extracts the value of the longest

constant first and second format, which is 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 [104], the audio amplitude duration and mean auto-correlation, and functionals applied to calculated power, cepstral peak prominence (related to voice quality), and harmonics. Implementation details for the custom feature set are provided in Appendix B.1.1. Details on the extraction of the cepstral coefficients and filter-bank coefficients are provided in Appendices B.1.2 and B.1.3 respectively.

The eGeMAPs feature sets and cepstral coefficient features have been used extensively in prior work in speech emotion recognition and other speech modeling tasks. Filter bank-based features have been used recently in speech emotion recognition [139]. 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 [140]. The bag-of-phones included the 50 phones that appeared most frequently in the dataset. In a similar manner to bag-of-words approaches, for each vocalization the bag-of-phones encoded the number of each phone in that vocalization. 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 parameters defining the DNN architecture were optimized for each participant. The selected parameters are provided in Appendix B.2. Because of the computationally intensive nature of training the autoencoder and selecting hyperparameters for a given fold split, the evaluation metrics with auDeep features are provided using one set of folds.

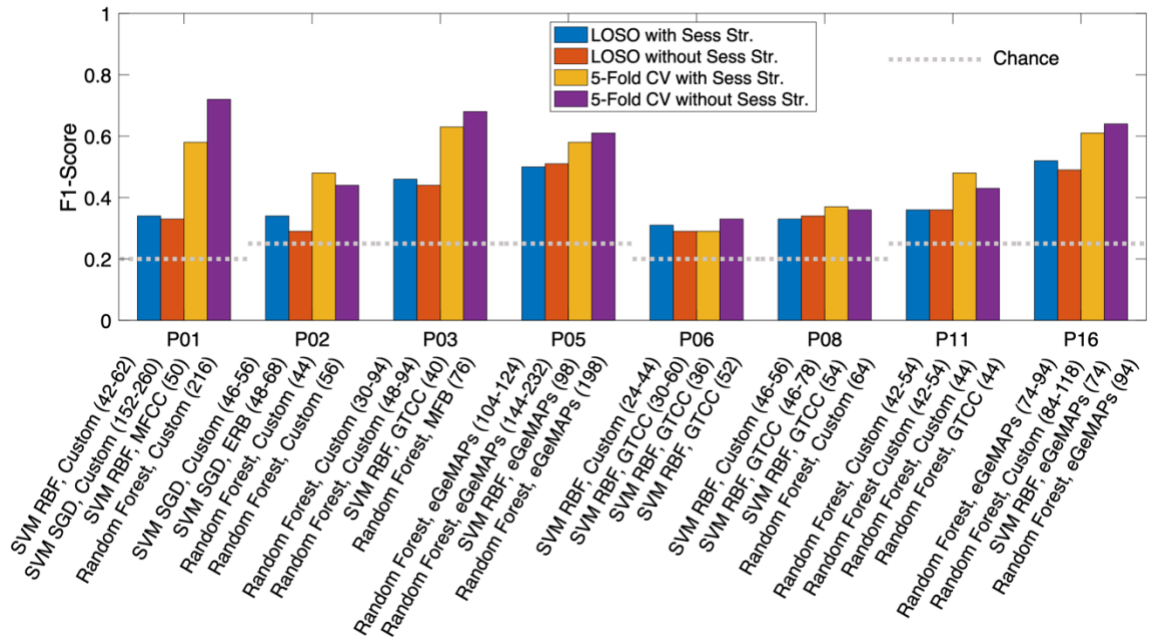
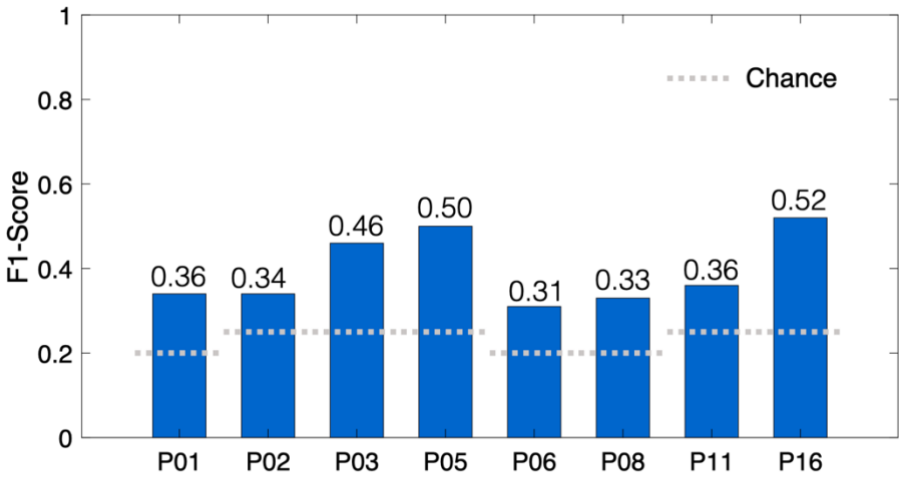


Figure 30: F1 score for best performing model and feature set for each participant, evaluated using leave one session out ("LOSO") and 5-fold cross-validation ("CV") with and without session stratification ("Sess Str.") The labels under each bar indicates the model and feature. The number of training samples per class is shown in parentheses. A range is provided for leave one session out evaluations, where the number of training samples varied between folds. The confidence intervals for the 5-fold CV evaluation are provided in Table 9 and Table 10.

8.2 Results and Discussion

Evaluation metrics are reported using micro-averages across folds, and 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.



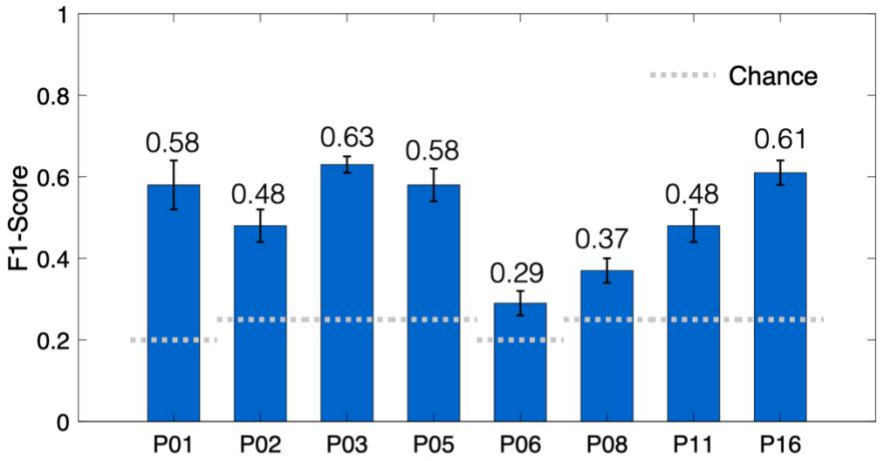
Feature Set	Custom	Custom	Custom	eGeMAPs	Custom	Custom	Custom	eGeMAPs
Model	SVM RBF	SVM SGD	Random Forest	Random Forest	SVM RBF	SVM RBF	SVM RBF	Random Forest
Samples per class	42-62	46-56	30-50	104-124	24-44	46-56	42-54	74-94

Figure 31: F1 score for best performing model and feature set for each participant, evaluated with LOSO and session stratification

8.2.1 Multi-Class Classification

Multi-class classification results for each model and evaluation strategy (LOSO and 5-fold cross validation) with each aggregated feature set with and without session stratification are provided at the end of this chapter in section 8.2.9.1 in Table 7, Table 8, Table 9, and Table 10. Results with the feature-learned auDeep features and the bag-of-phones feature set are provided in section 8.2.9.1 in Table 11 and Table 12, respectively.

Figure 30 shows the highest F1 score across aggregate feature sets and model types for each evaluation strategy. Figure 31 shows model performance for only the LOSO evaluation with session stratification, the strategy least likely to fit to background sounds. Figure 32 shows model performance for only the 5-fold cross-validation with session stratification, with 95% confidence



Feature Set	MFCC	Custom	GTCC	eGeMAPs	GTCC	Custom	Custom	eGeMAPs
Model	SVM RBF	Random Forest	SVM RBF	SVM RBF	SVM RBF	SVM RBF	Random Forest	SVM RBF
Samples per class	50	44	40	98	36	54	44	74

Figure 32: F1 score with 95% confidence intervals for best performing model and feature set for each participant, evaluated with 5-fold cross-validation and session stratification. Confidence intervals were calculated across folds with three random seeds.

intervals calculated across folds with three random seeds. The best performing models for each evaluation strategy had multi-class F1 scores higher than chance for all participants (Figure 30).

8.2.2 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, P06), models with session stratification had better performance than models without session stratification even though session stratification reduced the number of available training samples. For LOSO evaluations, the test data for each split has distinct background soundscapes from the training data and so fitting to background noise can actually reduce model performance. In these cases, session stratification can improve model performance. The P01 data tended to have a single label for an entire session (Figure 25) which may have made models for P01 particularly susceptible to fitting to background soundscapes.

8.2.3 Differences in Model Performance Between Participants

All F1 scores are above chance, but there are large variations in performance between participants (Figure 30, Figure 31, and Figure 32). These 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 (Figure 31 and Figure 32). Still, the variations are not due to training sizes alone -- P03 also had a relatively high model performance but had one of the smallest numbers 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 3). 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 3).

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 2). 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 (Figure 30).

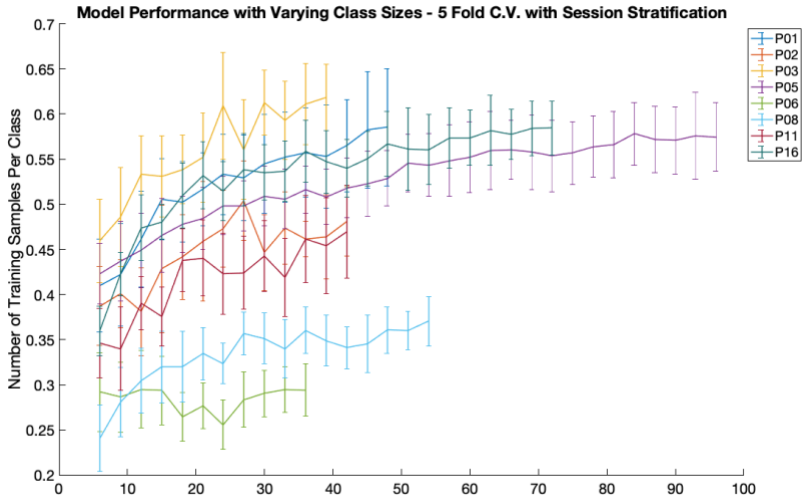


Figure 33: F Model performance with varying numbers of training samples per class. The error bars show the 95% confidence intervals for each result, calculated using the F1 score for each fold in the 5-fold cross-validation scheme with three random seeds.

8.2.4 Effect of Training Set Size

To explore the effect of training set size on model performance, the performance of the best performing aggregate feature set and model (Figure 31), was evaluated while varying the balanced training class sizes for each participant. Because the balanced training class size changes for each fold with the LOSO evaluation strategy depending on how samples are distributed among labels in the held-out session and in the remaining sessions used for training, the experiments to investigate the effect of the training set size were conducted with 5-fold cross-validation evaluation with session stratification. The maximum number of samples per class per session per participant, or the session stratification bound *b*, was defined as a function of the training set size to ensure that each class included samples from at least three sessions in training (as was the case for prior experiments with session stratification). The session stratification bound *b* as a function of the number of per class training samples *n* was set as

$$b(n) = \begin{cases} n \bmod 3, & n < 30 \\ 10, & x \geq 30. \end{cases}$$

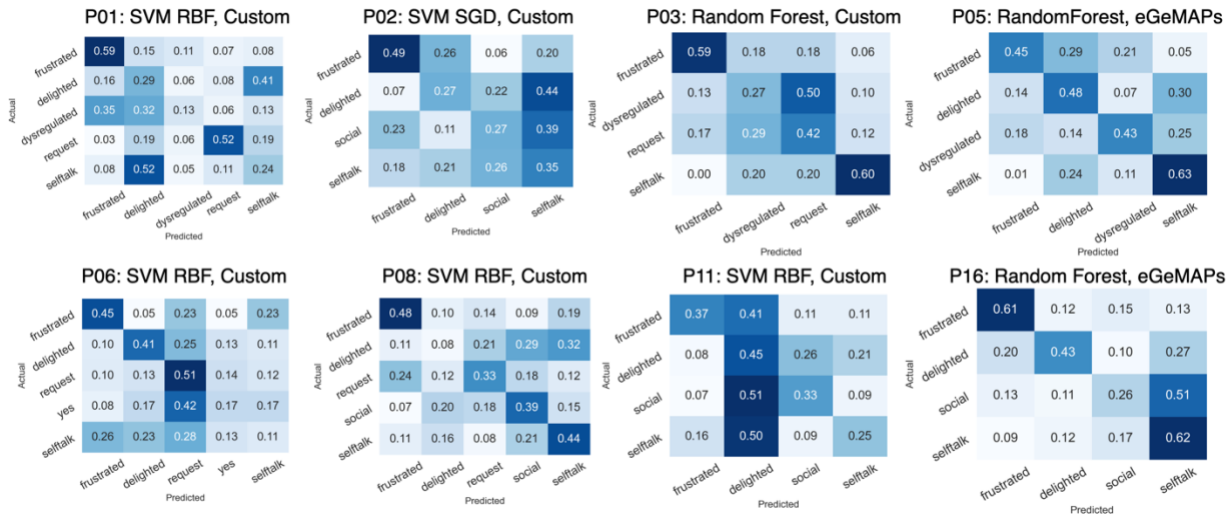


Figure 34: Multi-class confusion matrices for best performing model and feature set for each participant with leave one session out with session stratification evaluation. The diagonal entries of the matrix are the recall for each class.

For each participant, *n* ranged from 6 to the maximum per class training size selected for prior 5-fold cross validation experiments based on total training data availability (Figure 30). For each *n*, the model performance was evaluated with three random seeds. The results of the experiments are shown in Figure 33. As expected, the F1 score for the evaluated model for each participant trends upwards as the per class training size increases. There are some groupings in performance at a given *n* - for instance, P01, P05, and P16 have closely overlaid performance curves. Still, even for a given value of *n*, there are marked variations in model performance between participants. These performance differences could be due to the factors discussed in 8.2.3 like differences in labeling quality, audio quality, and vocalization practices.

8.2.5 Labels

Figure 34 shows multi-class confusion matrices for the leave one session out with session stratification evaluation. Differences in how well a particular label could be classified could be

characteristic of the vocalization itself or 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 4). 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 (i.e., the proportion of "delighted" samples that were correctly classified by the model as "delighted") and the largest number of training samples (Table 4).

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 ground truth label. 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.

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.

8.2.6 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 six of the eight participants with LOSO and session stratification (Figure 31), suggesting that the distinct

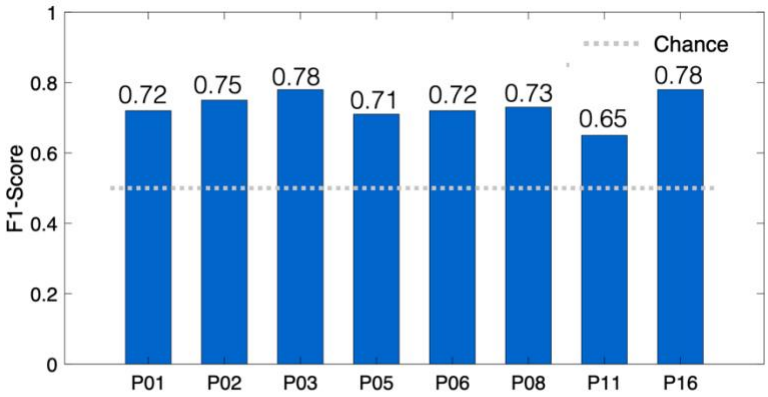
features and applied functionals chosen for the custom feature set capture the unique differences between nonverbal vocalizations of different types.

The 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 [140], 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.

8.2.7 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 noise in Section 8.2.9.2 in Table 13. Figure 35 shows the best performing model and feature set for binary valence classification.

Variations in model performance between participants for the binary classification task were less pronounced (Figure 35). 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



	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

Figure 35: F1 score for best performing binary valence model and feature set for each participant, evaluated with LOSO and session stratification.

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.

8.2.8 Conclusions

This chapter presented results from the largest study of nonverbal vocalizations by 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 from mv* individuals using audio alone. The overall classification performance is in the same general range as real-world multi-class affect classification for typical verbal speech [117] (though the cited approaches focus on speaker independent classification).

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 be used to explore whether there are subgroups of communicators with similar vocalization practices

8.2.9 Results Tables

8.2.9.1 Multi-Class Classification

	P01		*P02*		*P03*		*P05*		*P06*		*P08*		*P11*		*P16*	
Chance	0.20		0.25		0.25		0.25		0.20		0.20		0.25		0.25	
Random Forest	*F1*	*UAR*	*F1*	*UAR*	*F1*	*UAR*	*F1*	*UAR*	*F1*	*UAR*	*F1*	*UAR*	*F1*	*UAR*	*F1*	*UAR*
eGeMAPs	0.34	0.35	0.27	0.28	0.37	0.37	**0.50**	0.50	0.26	0.26	0.24	0.26	0.30	0.31	0.48	0.48
ERB	0.27	0.29	0.25	0.27	0.41	0.42	0.35	0.35	0.25	0.26	0.31	0.33	0.26	0.27	0.46	0.47
GTCC	0.28	0.30	0.30	0.31	0.40	0.42	0.40	0.40	0.28	0.30	0.29	0.30	0.34	0.35	0.52	0.52
Custom	0.29	0.30	0.23	0.23	**0.46**	0.47	0.47	0.47	0.23	0.24	0.30	0.31	**0.36**	0.38	0.51	0.52
MFB	0.31	0.33	0.33	0.35	0.39	0.41	0.38	0.38	0.22	0.24	0.32	0.33	0.24	0.25	0.47	0.48
MFCC	0.29	0.30	0.27	0.27	0.33	0.33	0.43	0.43	0.25	0.27	0.25	0.26	0.26	0.27	0.48	0.49
SVM RBF																
eGeMAPs	0.30	0.33	0.24	0.25	0.39	0.39	0.46	0.47	0.27	0.27	0.24	0.28	0.23	0.24	**0.52**	0.53
ERB	0.30	0.34	0.27	0.26	0.40	0.41	0.38	0.39	0.28	0.31	0.29	0.31	0.23	0.23	0.49	0.49
GTCC	0.26	0.27	0.23	0.23	0.34	0.34	0.43	0.44	0.31	0.35	0.32	0.33	0.29	0.30	0.50	0.51
Custom	**0.34**	0.35	0.28	0.26	0.43	0.43	0.49	0.50	0.29	0.33	**0.33**	0.35	0.34	0.35	0.45	0.46
MFB	0.32	0.35	0.30	0.30	0.44	0.44	0.39	0.41	0.23	0.24	0.29	0.31	0.24	0.24	0.50	0.50
MFCC	0.30	0.30	0.30	0.30	0.29	0.30	0.46	0.48	0.26	0.28	0.29	0.30	0.32	0.34	0.49	0.50
SVM SGD																
eGeMAPs	0.20	0.22	0.27	0.28	0.45	0.46	0.36	0.37	0.22	0.23	0.26	0.28	0.24	0.28	0.43	0.45
ERB	0.31	0.33	0.30	0.32	0.42	0.43	0.34	0.36	0.21	0.23	0.31	0.34	0.22	0.26	0.42	0.46
GTCC	0.20	0.22	0.24	0.26	0.39	0.41	0.43	0.43	0.26	0.28	0.30	0.32	0.24	0.31	0.37	0.42
Custom	0.26	0.28	**0.34**	0.34	0.36	0.38	0.41	0.43	**0.31**	0.34	0.29	0.29	0.22	0.26	0.36	0.37
MFB	0.31	0.34	0.29	0.29	0.41	0.42	0.38	0.41	0.21	0.22	0.27	0.30	0.24	0.27	0.42	0.44
MFCC	0.24	0.28	0.23	0.27	0.36	0.38	0.42	0.44	0.26	0.32	0.29	0.31	0.22	0.27	0.30	0.34

Table 7: Macro-F1 and UAR for leave one session out evaluation with session stratification

	P01		*P02*		*P03*		*P05*		*P06*		*P08*		*P11*		*P16*	
Chance	0.20		0.25		0.25		0.25		0.20		0.20		0.25		0.25	
Random Forest	*F1*	*UAR*	*F1*	*UAR*	*F1*	*UAR*	*F1*	*UAR*	*F1*	*UAR*	*F1*	*UAR*	*F1*	*UAR*	*F1*	*UAR*
eGeMAPs	0.27	0.31	0.22	0.23	0.37	0.45	**0.51**	0.51	0.25	0.27	0.29	0.33	0.33	0.35	0.49	0.51
ERB	0.23	0.27	0.28	0.31	0.31	0.39	0.35	0.35	0.27	0.29	0.31	0.35	0.24	0.25	0.48	0.49
GTCC	0.23	0.26	0.24	0.26	0.35	0.42	0.44	0.44	0.24	0.27	0.33	0.37	0.33	0.36	0.50	0.51
Custom	0.26	0.30	0.22	0.25	**0.44**	0.52	0.46	0.46	0.26	0.29	0.33	0.37	**0.36**	0.37	**0.59**	0.60
MFB	0.26	0.31	0.23	0.26	0.34	0.40	0.39	0.39	0.23	0.23	0.31	0.35	0.26	0.27	0.49	0.50
MFCC	0.23	0.26	0.23	0.26	0.33	0.39	0.42	0.42	0.24	0.26	0.30	0.32	0.24	0.26	0.49	0.51
SVM RBF																
eGeMAPs	0.23	0.29	0.20	0.24	0.39	0.42	0.49	0.49	0.23	0.24	0.30	0.35	0.30	0.31	0.53	0.58
ERB	0.24	0.28	0.27	0.28	0.37	0.44	0.36	0.37	0.28	0.32	0.29	0.33	0.21	0.22	0.51	0.52
GTCC	0.24	0.28	0.22	0.26	0.36	0.44	0.42	0.41	**0.29**	0.33	**0.34**	0.38	0.26	0.27	0.47	0.48
Custom	0.28	0.31	0.25	0.27	0.44	0.51	0.45	0.45	0.29	0.32	0.32	0.36	0.32	0.33	0.55	0.59
MFB	0.24	0.29	0.26	0.29	0.38	0.44	0.37	0.38	0.28	0.31	0.31	0.36	0.22	0.23	0.49	0.51
MFCC	0.23	0.26	0.26	0.33	0.32	0.39	0.40	0.40	0.25	0.26	0.30	0.34	0.26	0.29	0.47	0.49
SVM SGD																
eGeMAPs	0.21	0.27	0.24	0.25	0.30	0.37	0.49	0.49	0.20	0.23	0.29	0.37	0.23	0.25	0.41	0.44
ERB	0.20	0.25	**0.29**	0.31	0.31	0.42	0.40	0.41	0.21	0.25	0.31	0.38	0.23	0.26	0.54	0.57
GTCC	0.27	0.36	0.21	0.24	0.34	0.43	0.35	0.36	0.23	0.27	0.32	0.35	0.21	0.24	0.35	0.43
Custom	**0.33**	0.36	0.16	0.22	0.39	0.49	0.43	0.44	0.23	0.27	0.30	0.34	0.24	0.26	0.46	0.52
MFB	0.23	0.28	0.28	0.30	0.33	0.40	0.39	0.41	0.18	0.22	0.33	0.40	0.22	0.26	0.42	0.45
MFCC	0.21	0.26	0.27	0.31	0.29	0.33	0.35	0.35	0.16	0.21	0.32	0.35	0.27	0.31	0.39	0.46

Table 8: Macro-F1 and UAR for leave one session out evaluation with session stratification

	P01		P02		P03		P05		P06		P08		P11		P16	
Chance	**0.20**		**0.25**		**0.25**		**0.25**		**0.20**		**0.20**		**0.25**		**0.25**	
Random Forest	*F1*	*UAR*	*F1*	*UAR*	*F1*	*UAR*	*F1*	*UAR*	*F1*	*UAR*	*F1*	*UAR*	*F1*	*UAR*	*F1*	*UAR*
eGeMAPs	0.52 ± 0.03	0.54 ± 0.03	0.43 ± 0.04	0.44 ± 0.04	0.56 ± 0.05	0.57 ± 0.05	0.56 ± 0.03	0.57 ± 0.03	0.28 ± 0.03	0.31 ± 0.04	0.31 ± 0.03	0.35 ± 0.03	0.37 ± 0.03	0.40 ± 0.04	0.59 ± 0.04	0.60 ± 0.04
ERB	0.51 ± 0.03	0.55 ± 0.03	0.46 ± 0.06	0.47 ± 0.05	0.58 ± 0.05	0.58 ± 0.05	0.51 ± 0.03	0.52 ± 0.04	0.26 ± 0.03	0.29 ± 0.03	0.31 ± 0.03	0.34 ± 0.03	0.41 ± 0.05	0.44 ± 0.05	0.55 ± 0.04	0.57 ± 0.05
GTCC	0.55 ± 0.04	0.57 ± 0.04	0.47 ± 0.04	0.47 ± 0.04	0.63 ± 0.04	0.63 ± 0.04	0.56 ± 0.04	0.56 ± 0.04	0.27 ± 0.03	0.29 ± 0.03	0.34 ± 0.03	0.36 ± 0.04	0.43 ± 0.02	0.46 ± 0.02	0.55 ± 0.03	0.57 ± 0.03
Custom	0.54 ± 0.05	0.57 ± 0.05	**0.48 ± 0.04**	0.49 ± 0.05	0.61 ± 0.03	0.62 ± 0.03	0.56 ± 0.04	0.56 ± 0.04	0.27 ± 0.03	0.29 ± 0.03	0.35 ± 0.03	0.38 ± 0.03	**0.48 ± 0.04**	0.50 ± 0.03	0.59 ± 0.04	0.60 ± 0.04
MFB	0.55 ± 0.04	0.58 ± 0.04	0.41 ± 0.05	0.42 ± 0.06	0.57 ± 0.06	0.57 ± 0.06	0.52 ± 0.03	0.53 ± 0.03	0.26 ± 0.02	0.28 ± 0.02	0.32 ± 0.03	0.35 ± 0.04	0.40 ± 0.04	0.43 ± 0.04	0.54 ± 0.04	0.56 ± 0.05
MFCC	0.56 ± 0.04	0.58 ± 0.05	0.48 ± 0.05	0.49 ± 0.05	0.54 ± 0.05	0.55 ± 0.05	0.52 ± 0.04	0.52 ± 0.04	0.25 ± 0.03	0.28 ± 0.03	0.32 ± 0.02	0.34 ± 0.03	0.44 ± 0.06	0.48 ± 0.06	0.54 ± 0.04	0.56 ± 0.04
SVM RBF																
eGeMAPs	0.48 ± 0.03	0.52 ± 0.03	0.43 ± 0.05	0.43 ± 0.05	0.54 ± 0.03	0.55 ± 0.04	**0.58 ± 0.04**	0.59 ± 0.04	0.28 ± 0.02	0.31 ± 0.03	0.31 ± 0.04	0.35 ± 0.04	0.35 ± 0.04	0.40 ± 0.05	**0.61 ± 0.03**	0.61 ± 0.03
ERB	0.50 ± 0.03	0.57 ± 0.04	0.40 ± 0.05	0.40 ± 0.05	0.54 ± 0.06	0.55 ± 0.06	0.44 ± 0.03	0.47 ± 0.04	0.27 ± 0.02	0.31 ± 0.02	0.32 ± 0.02	0.36 ± 0.03	0.4 ± 0.04	0.44 ± 0.04	0.58 ± 0.03	0.58 ± 0.03
GTCC	0.53 ± 0.03	0.56 ± 0.04	0.46 ± 0.03	0.47 ± 0.03	**0.63 ± 0.02**	0.63 ± 0.02	0.57 ± 0.03	0.58 ± 0.03	**0.29 ± 0.03**	0.33 ± 0.03	0.35 ± 0.03	0.37 ± 0.04	0.44 ± 0.05	0.47 ± 0.05	0.58 ± 0.03	0.58 ± 0.03
Custom	0.50 ± 0.03	0.52 ± 0.04	0.44 ± 0.05	0.45 ± 0.06	0.57 ± 0.04	0.57 ± 0.04	0.56 ± 0.04	0.57 ± 0.04	0.28 ± 0.04	0.31 ± 0.04	**0.37 ± 0.03**	0.39 ± 0.03	0.46 ± 0.05	0.47 ± 0.05	0.58 ± 0.04	0.58 ± 0.04
MFB	0.51 ± 0.03	0.58 ± 0.04	0.40 ± 0.03	0.41 ± 0.04	0.53 ± 0.06	0.55 ± 0.05	0.43 ± 0.03	0.46 ± 0.04	0.26 ± 0.03	0.28 ± 0.03	0.32 ± 0.03	0.35 ± 0.04	0.39 ± 0.05	0.43 ± 0.05	0.58 ± 0.04	0.58 ± 0.04
MFCC	**0.58 ± 0.06**	0.60-± 0.06	0.45 ± 0.03	0.46 ± 0.03	0.54 ± 0.03	0.55 ± 0.04	0.55 ± 0.05	0.57 ± 0.05	0.27 ± 0.02	0.30 ± 0.03	0.30 ± 0.02	0.32 ± 0.03	0.45 ± 0.06	0.48 ± 0.06	0.59 ± 0.02	0.59 ± 0.02
SVM SGD																
eGeMAPs	0.42 ± 0.04	0.48 ± 0.03	0.42 ± 0.05	0.44 ± 0.05	0.58 ± 0.06	0.58 ± 0.06	0.48 ± 0.03	0.51 ± 0.03	0.21 ± 0.03	0.28 ± 0.03	0.26 ± 0.04	0.32 ± 0.04	0.28 ± 0.03	0.34 ± 0.03	0.55 ± 0.03	0.55 ± 0.03
ERB	0.43 ± 0.02	0.51 ± 0.04	0.34 ± 0.05	0.38 ± 0.05	0.46 ± 0.03	0.50 ± 0.04	0.46 ± 0.04	0.49 ± 0.04	0.22 ± 0.02	0.29 ± 0.03	0.32 ± 0.03	0.35 ± 0.03	0.33 ± 0.04	0.38 ± 0.03	0.55 ± 0.05	0.55 ± 0.05
GTCC	0.38 ± 0.03	0.46 ± 0.04	0.36 ± 0.03	0.41 ± 0.03	0.51 ± 0.04	0.53 ± 0.04	0.47 ± 0.04	0.51 ± 0.03	0.22 ± 0.04	0.3 ± 0.04	0.31 ± 0.03	0.36 ± 0.03	0.33 ± 0.05	0.41 ± 0.04	0.48 ± 0.05	0.48 ± 0.05
Custom	0.43 ± 0.04	0.49 ± 0.04	0.38 ± 0.04	0.42 ± 0.03	0.46 ± 0.07	0.50 ± 0.06	0.52 ± 0.03	0.54 ± 0.03	0.22 ± 0.04	0.28 ± 0.04	0.30 ± 0.05	0.34 ± 0.04	0.33 ± 0.03	0.4 ± 0.04	0.52 ± 0.03	0.52 ± 0.03
MFB	0.41 ± 0.04	0.48 ± 0.04	0.38 ± 0.04	0.40 ± 0.05	0.53 ± 0.05	0.54 ± 0.04	0.45 ± 0.04	0.49 ± 0.04	0.18 ± 0.02	0.25 ± 0.03	0.32 ± 0.03	0.36 ± 0.03	0.36 ± 0.06	0.41 ± 0.06	0.52 ± 0.04	0.52 ± 0.04
MFCC	0.40 ± 0.03	0.48 ± 0.03	0.42 ± 0.04	0.45 ± 0.04	0.47 ± 0.03	0.49 ± 0.03	0.44 ± 0.03	0.48 ± 0.03	0.18 ± 0.03	0.28 ± 0.04	0.24 ± 0.03	0.31 ± 0.03	0.32 ± 0.04	0.43 ± 0.05	0.45 ± 0.04	0.45 ± 0.04

Table 9: Macro-F1 and UAR with 95% confidence intervals for 5-fold cross validation with session stratification

	P01		P02		P03		P05		P06		P08		P11		P16	
Chance	0.20		0.25		0.25		0.25		0.20		0.20		0.25		0.25	
Random Forest	F1	UAR	F1	UAR	F1	UAR	F1	UAR	F1	UAR	F1	UAR	F1	UAR	F1	UAR
eGeMAPs	0.63 ± 0.02	0.68 ± 0.01	0.39 ± 0.03	0.46 ± 0.04	0.64 ± 0.04	0.68 ± 0.04	**0.61 ± 0.02**	0.61 ± 0.02	0.32 ± 0.02	0.35 ± 0.03	0.33 ± 0.02	0.39 ± 0.03	0.42 ± 0.03	0.49 ± 0.05	**0.64 ± 0.02**	0.67 ± 0.02
ERB	0.66 ± 0.01	0.69 ± 0.01	0.43 ± 0.03	0.49 ± 0.04	0.65 ± 0.03	0.71 ± 0.03	0.55 ± 0.02	0.56 ± 0.02	0.27 ± 0.02	0.31 ± 0.03	0.34 ± 0.01	0.40 ± 0.02	0.39 ± 0.05	0.45 ± 0.05	0.59 ± 0.02	0.62 ± 0.02
GTCC	0.67 ± 0.01	0.71 ± 0.01	0.39 ± 0.03	0.44 ± 0.03	0.68 ± 0.03	0.72 ± 0.03	0.59 ± 0.02	0.60 ± 0.02	0.31 ± 0.02	0.35 ± 0.03	0.36 ± 0.02	0.43 ± 0.02	**0.43 ± 0.02**	0.49 ± 0.04	0.58 ± 0.03	0.60 ± 0.03
Custom	**0.72 ± 0.02**	0.74 ± 0.02	**0.44 ± 0.04**	0.49 ± 0.05	0.66 ± 0.04	0.71 ± 0.04	0.59 ± 0.01	0.59 ± 0.01	0.32 ± 0.02	0.36 ± 0.03	0.36 ± 0.01	0.42 ± 0.01	0.41 ± 0.04	0.45 ± 0.05	0.64 ± 0.02	0.66 ± 0.02
MFB	0.70 ± 0.01	0.73 ± 0.01	0.42 ± 0.04	0.48 ± 0.05	0.68 ± 0.03	0.72 ± 0.03	0.58 ± 0.02	0.58 ± 0.02	0.27 ± 0.02	0.29 ± 0.02	0.35 ± 0.02	0.41 ± 0.03	0.40 ± 0.03	0.48 ± 0.03	0.60 ± 0.03	0.62 ± 0.03
MFCC	0.64 ± 0.02	0.67 ± 0.02	0.41 ± 0.03	0.47 ± 0.03	0.58 ± 0.03	0.63 ± 0.03	0.54 ± 0.02	0.54 ± 0.02	0.28 ± 0.03	0.31 ± 0.04	0.35 ± 0.01	0.41 ± 0.02	0.41 ± 0.04	0.48 ± 0.05	0.58 ± 0.02	0.60 ± 0.02
SVM RBF																
eGeMAPs	0.60 ± 0.01	0.66 ± 0.01	0.38 ± 0.02	0.45 ± 0.02	0.64 ± 0.02	0.68 ± 0.02	0.59 ± 0.02	0.60 ± 0.02	0.31 ± 0.02	0.35 ± 0.02	0.32 ± 0.01	0.40 ± 0.02	0.37 ± 0.02	0.43 ± 0.03	0.63 ± 0.02	0.66 ± 0.02
ERB	0.60 ± 0.01	0.66 ± 0.01	0.34 ± 0.03	0.40 ± 0.04	0.59 ± 0.03	0.66 ± 0.03	0.48 ± 0.02	0.50 ± 0.02	0.30 ± 0.02	0.35 ± 0.02	0.34 ± 0.01	0.40 ± 0.02	0.39 ± 0.02	0.46 ± 0.04	0.60 ± 0.02	0.60 ± 0.02
GTCC	0.69 ± 0.02	0.73 ± 0.02	0.40 ± 0.03	0.47 ± 0.03	**0.68 ± 0.02**	0.73 ± 0.03	0.60 ± 0.02	0.61 ± 0.02	**0.33 ± 0.02**	0.39 ± 0.03	**0.36 ± 0.01**	0.43 ± 0.01	0.41 ± 0.04	0.47 ± 0.05	0.63 ± 0.02	0.68 ± 0.03
Custom	0.69 ± 0.01	0.72 ± 0.01	0.42 ± 0.04	0.45 ± 0.05	0.67 ± 0.03	0.71 ± 0.03	0.58 ± 0.02	0.58 ± 0.02	0.33 ± 0.03	0.38 ± 0.04	0.36 ± 0.02	0.42 ± 0.03	0.43 ± 0.03	0.48 ± 0.04	0.63 ± 0.03	0.66 ± 0.03
MFB	0.64 ± 0.01	0.69 ± 0.01	0.36 ± 0.03	0.41 ± 0.03	0.59 ± 0.04	0.65 ± 0.04	0.48 ± 0.02	0.50 ± 0.02	0.29 ± 0.03	0.32 ± 0.03	0.34 ± 0.02	0.41 ± 0.04	0.38 ± 0.04	0.45 ± 0.05	0.59 ± 0.02	0.61 ± 0.02
MFCC	0.67 ± 0.01	0.71 ± 0.01	0.41 ± 0.03	0.48 ± 0.04	0.61 ± 0.02	0.67 ± 0.03	0.56 ± 0.02	0.56 ± 0.02	0.31 ± 0.02	0.35 ± 0.03	0.33 ± 0.02	0.38 ± 0.03	0.40 ± 0.03	0.45 ± 0.05	0.62 ± 0.03	0.66 ± 0.04
SVM SGD																
eGeMAPs	0.50 ± 0.02	0.60 ± 0.01	0.33 ± 0.03	0.39 ± 0.03	0.55 ± 0.04	0.65 ± 0.03	0.52 ± 0.03	0.54 ± 0.02	0.23 ± 0.02	0.28 ± 0.03	0.28 ± 0.02	0.35 ± 0.03	0.31 ± 0.03	0.40 ± 0.04	0.51 ± 0.07	0.56 ± 0.04
ERB	0.48 ± 0.04	0.57 ± 0.04	0.33 ± 0.05	0.41 ± 0.04	0.55 ± 0.04	0.63 ± 0.04	0.45 ± 0.02	0.48 ± 0.01	0.23 ± 0.02	0.32 ± 0.03	0.34 ± 0.02	0.41 ± 0.03	0.33 ± 0.04	0.44 ± 0.04	0.56 ± 0.04	0.59 ± 0.03
GTCC	0.45 ± 0.02	0.57 ± 0.02	0.34 ± 0.06	0.42 ± 0.04	0.56 ± 0.02	0.64 ± 0.03	0.49 ± 0.04	0.51 ± 0.03	0.26 ± 0.02	0.35 ± 0.03	0.29 ± 0.02	0.38 ± 0.03	0.31 ± 0.03	0.45 ± 0.03	0.52 ± 0.03	0.59 ± 0.02
Custom	0.54 ± 0.03	0.63 ± 0.02	0.35 ± 0.04	0.42 ± 0.04	0.56 ± 0.04	0.62 ± 0.03	0.50 ± 0.02	0.53 ± 0.02	0.25 ± 0.02	0.33 ± 0.02	0.31 ± 0.04	0.38 ± 0.02	0.34 ± 0.03	0.42 ± 0.04	0.54 ± 0.04	0.60 ± 0.03
MFB	0.51 ± 0.02	0.59 ± 0.02	0.35 ± 0.04	0.41 ± 0.04	0.54 ± 0.03	0.61 ± 0.04	0.44 ± 0.02	0.47 ± 0.02	0.23 ± 0.03	0.32 ± 0.03	0.33 ± 0.02	0.42 ± 0.03	0.34 ± 0.03	0.44 ± 0.03	0.55 ± 0.03	0.58 ± 0.03
MFCC	0.44 ± 0.02	0.57 ± 0.02	0.27 ± 0.04	0.40 ± 0.04	0.50 ± 0.05	0.56 ± 0.04	0.42 ± 0.03	0.46 ± 0.02	0.23 ± 0.03	0.31 ± 0.03	0.31 ± 0.02	0.39 ± 0.02	0.35 ± 0.03	0.43 ± 0.03	0.48 ± 0.04	0.55 ± 0.04

Table 10: Macro-F1 and UAR with 95% confidence intervals for 5-fold cross validation without session stratification

	P01		*P02*		*P03*		*P05*		*P06*		*P08*		*P11*		*P16*	
Chance	**0.20**		**0.25**		**0.25**		**0.25**		**0.20**		**0.20**		**0.25**		**0.25**	
LOSO (with Sess. Str.)	*F1*	*UAR*	*F1*	*UAR*	*F1*	*UAR*	*F1*	*UAR*	*F1*	*UAR*	*F1*	*UAR*	*F1*	*UAR*	*F1*	*UAR*
Random Forest	**0.23**	0.26	0.25	0.26	**0.33**	0.35	0.28	0.29	0.16	0.18	0.19	0.24	0.20	0.24	0.28	0.32
SVM RBF	0.22	0.26	0.24	0.24	0.33	0.34	**0.29**	0.33	0.15	0.16	0.17	0.21	**0.24**	0.25	0.27	0.35
SVM SGD	0.22	0.25	**0.30**	0.30	0.31	0.32	0.25	0.27	**0.20**	0.24	**0.21**	0.23	0.22	0.26	**0.34**	0.39
LOSO (without Sess. Str.)																
Random Forest	**0.21**	0.26	0.18	0.24	**0.26**	0.31	**0.28**	0.31	0.18	0.18	0.11	0.16	**0.23**	0.28	0.18	0.27
SVM RBF	0.21	0.24	**0.20**	0.25	0.25	0.29	0.27	0.31	0.19	0.19	**0.14**	0.19	0.20	0.23	0.17	0.26
SVM SGD	0.17	0.21	0.17	0.23	0.24	0.27	0.27	0.27	**0.20**	0.21	0.14	0.17	0.20	0.23	**0.20**	0.23
5-Fold CV (with Sess. Str.)																
Random Forest	**0.21 ± 0.03**	0.26 ± 0.02	0.28 ± 0.02	0.30 ± 0.03	0.34 ± 0.05	0.35 ± 0.05	0.24 ± 0.03	0.25 ± 0.04	**0.20 ± 0.03**	0.23 ± 0.03	**0.17 ± 0.02**	0.20 ± 0.02	**0.17 ± 0.02**	0.20 ± 0.02	**0.25 ± 0.03**	0.31 ± 0.03
SVM RBF	0.18 ± 0.03	0.23 ± 0.04	0.26 ± 0.04	0.29 ± 0.04	0.30 ± 0.05	0.29 ± 0.06	0.25 ± 0.03	0.28 ± 0.03	0.19 ± 0.03	0.22 ± 0.04	0.14 ± 0.02	0.16 ± 0.03	0.14 ± 0.02	0.16 ± 0.03	0.23 ± 0.02	0.30 ± 0.03
SVM SGD	0.21 ± 0.04	0.25 ± 0.04	**0.29 ± 0.04**	0.31 ± 0.03	**0.32 ± 0.05**	0.34 ± 0.05	**0.25 ± 0.02**	0.28 ± 0.02	0.14 ± 0.03	0.17 ± 0.03	0.13 ± 0.02	0.16 ± 0.03	0.13 ± 0.02	0.16 ± 0.03	0.23 ± 0.04	0.30 ± 0.03
5-Fold CV (without Sess. Str.)																
Random Forest	**0.21 ± 0.01**	0.26 ± 0.01	**0.23 ± 0.01**	0.30 ± 0.03	**0.27 ± 0.03**	0.33 ± 0.04	0.28 ± 0.01	0.30 ± 0.02	0.18 ± 0.02	0.20 ± 0.02	**0.16 ± 0.01**	0.21 ± 0.02	**0.22 ± 0.02**	0.28 ± 0.04	**0.23 ± 0.02**	0.31 ± 0.02
SVM RBF	0.21 ± 0.01	0.26 ± 0.01	0.23 ± 0.02	0.28 ± 0.03	0.23 ± 0.02	0.27 ± 0.03	**0.28 ± 0.01**	0.31 ± 0.02	**0.19 ± 0.03**	0.21 ± 0.03	0.16 ± 0.01	0.19 ± 0.02	0.22 ± 0.02	0.27 ± 0.04	0.20 ± 0.02	0.30 ± 0.02
SVM SGD	0.18 ± 0.02	0.24 ± 0.01	0.23 ± 0.03	0.28 ± 0.04	0.21 ± 0.02	0.26 ± 0.02	0.26 ± 0.02	0.29 ± 0.02	0.16 ± 0.03	0.2 ± 0.02	0.15 ± 0.02	0.20 ± 0.02	0.20 ± 0.03	0.27 ± 0.03	0.23 ± 0.02	0.30 ± 0.03

Table 11: Macro-F1 and UAR for models with bag-of-phones feature set

	P01		*P02*		*P03*		*P05*		*P06*		*P08*		*P11*		*P16*	
Chance	**0.20**		**0.25**		**0.25**		**0.25**		**0.20**		**0.20**		**0.25**		**0.25**	
5-Fold CV (with Sess. Str.)	*F1*	*UAR*	*F1*	*UAR*	*F1*	*UAR*	*F1*	*UAR*	*F1*	*UAR*	*F1*	*UAR*	*F1*	*UAR*	*F1*	*UAR*
Random Forest	0.58 ± 0.12	0.60 ± 0.14	**0.52 ± 0.07**	0.53 ± 0.07	0.54 ± 0.12	0.58 ± 0.12	0.54 ± 0.04	0.54± 0.04	0.27 ± 0.11	0.29 ± 0.12	**0.32 ± 0.05**	0.32 ± 0.07	0.41 ± 0.09	0.43 ± 0.11	0.62 ± 0.06	0.63 ± 0.07
SVM RBF	**0.61± 0.03**	0.62 ± 0.04	0.36 ± 0.05	0.40 ± 0.05	0.49 ± 0.09	0.52 ± 0.12	**0.60 ± 0.07**	0.64 ± 0.08	**0.28 ± 0.08**	0.29 ± 0.08	0.31 ± 0.06	0.34 ± 0.06	**0.41 ± 0.06**	0.43 ± 0.06	**0.62 ± 0.06**	0.64 ± 0.06
SVM SGD	0.41 ± 0.09	0.47 ± 0.09	0.30 ± 0.05	0.38 ± 0.06	**0.55 ± 0.13**	0.59 ± 0.13	0.48 ± 0.06	0.54 ± 0.05	0.22 ± 0.04	0.28 ± 0.05	0.28 ± 0.02	0.36 ± 0.03	0.33 ± 0.05	0.42 ± 0.07	0.53 ± 0.03	0.58 ± 0.04
5-Fold CV (without Sess. Str.)																
Random Forest	0.72 ± 0.01	0.76 ± 0.02	**0.42 ± 0.08**	0.46 ± 0.11	**0.64 ± 0.07**	0.70 ± 0.07	0.59 ± 0.02	0.59 ± 0.02	0.27 ± 0.07	0.29 ± 0.08	**0.35 ± 0.02**	0.40 ± 0.05	**0.43 ± 0.04**	0.46 ± 0.06	0.64 ± 0.09	0.66 ± 0.09
SVM RBF	**0.80 ± 0.02**	0.80 ± 0.02	0.34 ± 0.04	0.43 ± 0.02	0.63 ± 0.03	0.70 ± 0.05	**0.59 ± 0.02**	0.60 ± 0.02	**0.27 ± 0.04**	0.27 ± 0.05	0.35 ± 0.02	0.39 ± 0.02	0.39 ± 0.06	0.43 ± 0.07	**0.65 ± 0.02**	0.69 ± 0.02
SVM SGD	0.66 ± 0.02	0.71 ± 0.03	0.29 ± 0.05	0.42 ± 0.04	0.56 ± 0.07	0.64 ± 0.03	0.48 ± 0.06	0.53 ± 0.05	0.19 ± 0.03	0.27 ± 0.02	0.31 ± 0.02	0.36 ± 0.02	0.25 ± 0.03	0.38 ± 0.04	0.60 ± 0.01	0.65 ± 0.02

Table 12: Macro-F1 and UAR for models with feature set learned from data using auDeep

8.2.9.2 Binary Classification

	P01		P02		P03		P05		P06		P08		P11		P16	
Chance	**0.50**		**0.50**		**0.50**		**0.50**		**0.50**		**0.50**		**0.50**		**0.50**	
Random Forest	*F1*	*UAR*	*F1*	*UAR*	*F1*	*UAR*	*F1*	*UAR*	*F1*	*UAR*	*F1*	*UAR*	*F1*	*UAR*	*F1*	*UAR*
eGeMAPs	0.67	0.66	0.64	0.63	0.60	0.60	0.67	0.67	0.67	0.72	0.62	0.63	0.63	0.64	0.73	0.72
ERB	0.68	0.68	0.66	0.66	0.62	0.62	0.56	0.56	0.69	0.78	0.65	0.66	0.62	0.65	0.75	0.74
GTCC	0.69	0.69	0.64	0.63	0.77	0.76	0.65	0.64	**0.72**	0.79	0.70	0.70	0.59	0.61	**0.78**	0.77
Custom	0.72	0.71	0.58	0.58	0.74	0.74	0.65	0.65	0.69	0.75	0.69	0.70	**0.65**	0.66	0.77	0.76
MFB	0.72	0.72	**0.75**	0.74	0.66	0.67	0.55	0.55	0.60	0.64	0.65	0.66	0.59	0.61	0.73	0.72
MFCC	0.62	0.62	0.64	0.63	0.57	0.58	0.67	0.67	0.71	0.79	0.65	0.65	0.58	0.60	0.72	0.72
SVM RBF																
eGeMAPs	0.65	0.65	0.57	0.57	0.57	0.57	0.67	0.67	0.66	0.68	0.62	0.64	0.63	0.63	0.71	0.71
ERB	0.68	0.68	0.71	0.71	0.59	0.61	0.55	0.55	0.69	0.77	0.70	0.73	0.54	0.55	0.75	0.73
GTCC	0.68	0.68	0.59	0.58	0.59	0.60	0.63	0.63	0.68	0.75	0.64	0.64	0.57	0.58	0.77	0.75
Custom	0.68	0.68	0.66	0.65	0.71	0.71	0.66	0.66	0.69	0.74	**0.73**	0.74	0.58	0.59	0.73	0.73
MFB	**0.72**	0.72	0.66	0.65	0.57	0.59	0.54	0.54	0.68	0.77	0.70	0.72	0.54	0.55	0.77	0.76
MFCC	0.66	0.66	0.65	0.64	0.60	0.61	**0.71**	0.71	0.66	0.74	0.65	0.66	0.59	0.61	0.71	0.71
SVM SGD																
eGeMAPs	0.68	0.69	0.60	0.60	0.56	0.56	0.69	0.69	0.52	0.56	0.67	0.72	0.55	0.61	0.72	0.73
ERB	0.62	0.62	0.67	0.68	0.61	0.62	0.58	0.59	0.68	0.75	0.66	0.67	0.58	0.63	0.63	0.63
GTCC	0.55	0.56	0.69	0.69	0.59	0.62	0.54	0.56	0.70	0.76	0.66	0.66	0.61	0.64	0.65	0.66
Custom	0.69	0.69	0.66	0.66	**0.78**	0.77	0.62	0.64	0.70	0.78	0.68	0.69	0.48	0.52	0.67	0.68
MFB	0.63	0.63	0.66	0.68	0.61	0.63	0.59	0.62	0.65	0.73	0.66	0.67	0.52	0.55	0.71	0.71
MFCC	0.54	0.54	0.74	0.74	0.59	0.61	0.53	0.54	0.62	0.71	0.63	0.64	0.64	0.69	0.61	0.61

Table 13: Macro-F1 and UAR for binary valence classification models with aggregate features with LOSO evaluation and session stratification

9 Boosting and Transfer Learning

Having sufficient data to train models that work for individuals in heterogeneous, specialized populations can be challenging. Understanding if and when datasets created with the general population can be leveraged for model training can help target data collection efforts. This chapter presents several types of transfer learning experiments to classify the affect and communication intent of nonverbal vocalizations from mv* individuals.

For our work, collecting data in the real-world with personalized labels was critical to capturing representative, motivation-driven communication. However, building the data set took over a year and required significant time and effort from researchers and participating families. Because collecting data is time-intensive, it is important to understand if and how existing datasets can be leveraged towards interpreting nonverbal vocalizations from mv* individuals. Transfer learning can be used to probe similarities in expressions among nonverbal vocalizations from mv* communicators and between mv* communicators and other populations. Boosting can be used to leverage particularly informative samples within a small, messy dataset by re-weighting data to prioritize hard to classify samples. This chapter also explores boosting methods for classification and boosting combined with transfer learning.

Transfer learning has been applied for affective and diagnostic classification tasks [117], [141]. To my knowledge, no prior work has explored transfer learning for models of affect and intent in nonverbal vocalizations from mv* individuals. Parts of this chapter are published in [142].

9.1 Transfer Learning Datasets

Datasets were selected for transfer learning experiments based on their relevance and availability:

- ReCANVo, Real World Communicative and Affective Nonverbal Vocalizations: communicative and affective vocalizations from mv* individuals (presented in Chapter 6)
- IEMOCAP, Interactive Emotional Dyadic Motion Capture [129]: verbal speech from dyadic scripted and impromptu acted sessions
- RAVDESS, Ryerson Audio-Visual Database of Emotional Speech and Sound [128]: verbal speech from scripted professional actors
- VIVAE, Variably Intense Vocalizations of Affect and Emotion [101]: acted corpus of nonverbal vocalizations that occur amidst typical verbal speech (e.g., grunts, screams)
- MSP-Podcast Dataset [110]: speech from podcasts perceptually annotated using crowdsourcing
- Urban Sound 8K [143]: recordings of non-speech environmental sounds scraped from Freesound

IEMOCAP and RAVDESS were selected because they are commonly used in speech emotion recognition studies. VIVAE was selected because it contained nonverbal vocalizations. The MSP-Podcast dataset was selected because of its naturalistic nature. The Urban Sound 8K dataset was selected because it has real-world sounds with varying recording environments.

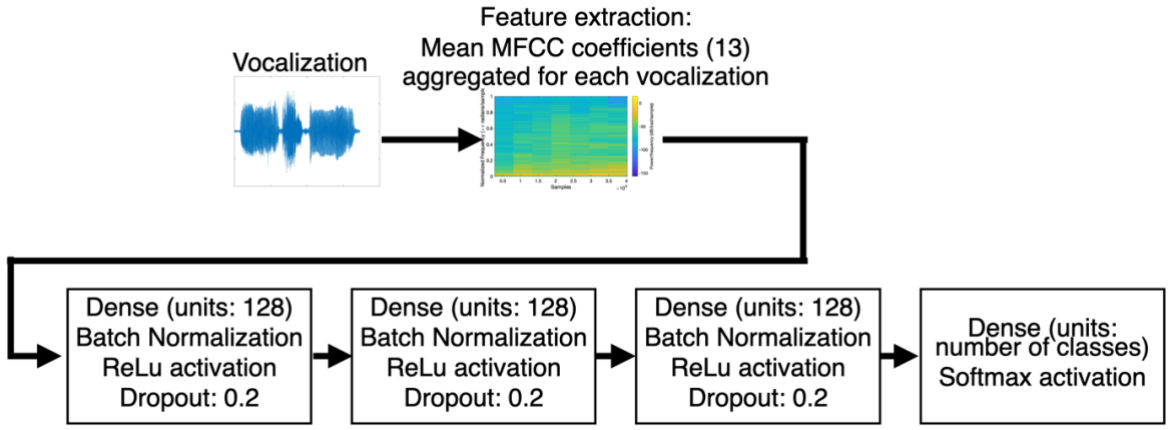


Figure 36: DNN model architecture used for transfer learning experiments described in 9.2.19.2.1 . Models were trained with each dataset listed in 9.1 and then fine-tuned for each participant

9.2 Methods

9.2.1 Transfer Learning with Personalized Neural Networks

The mean of each of thirteen mel frequency cepstral coefficients (MFCC) was extracted for each vocalization. Deep neural net (DNN) models with two hidden layers (Figure 36) were trained for each participant as a baseline. Baseline models were trained with a learning rate of 0.001. Other model architectures, including long-short term memory (LSTM) recurrent neural nets (RNNs), the EmoNet architecture (based on ResNet) [117], and smaller convolutional neural nets (CNNs) were also evaluated but tended to overfit the data. The selected DNN architecture (Figure 36) had the most robust performance across participants.

Models were trained on each dataset listed in 9.1 using an 80/20 training/validation data split. Table 14 shows the base model class sizes and performance on validation data. With the ReCANVo dataset, unique base models were trained for each participant to include only the other mv* communicators, to prevent prior exposure to any samples used in model evaluation. Then, models trained on the base dataset were fine-tuned for each participant. The network was fine-tuned with an adaptive learning rate with an exponential decay schedule with initial learning rate

0.001, a decay rate of 0.1, and 10,000 decay steps. The last layer was removed and replaced with a new softmax layer for the target participant's data. Models were evaluated using leave one session out (LOSO) evaluation with session stratification. The total/test class sizes were as reported in Table 4, with the per class training sizes as shown in Figure 30.

Base Model	F1 (validation)	UAR (validation)	Chance	Training Data Counts
ReCANVo, excluding P01	0.41	0.41	0.167	selftalk: 472, frustrated: 273, delighted: 385, social: 174, dysregulated: 94, request: 123
ReCANVo, excluding P02	0.45	0.43	0.167	selftalk: 508, frustrated: 294, delighted: 439, social: 129, dysregulated: 119, request: 148
ReCANVo, excluding P03	0.48	0.47	0.167	selftalk: 515, frustrated: 295, delighted: 472, social: 174, dysregulated: 77, request: 106
ReCANVo, excluding P05	0.43	0.41	0.167	selftalk: 421, frustrated: 272, delighted: 376, social: 174, dysregulated: 67, request: 148
ReCANVo, excluding P06	0.45	0.45	0.167	selftalk: 497, frustrated:304, delighted: 399, social: 174, dysregulated: 119, request: 93
ReCANVo, excluding P08	0.43	0.41	0.167	selftalk: 442, frustrated:276, delighted: 442, social: 117, dysregulated: 119, request: 122
ReCANVo, excluding P11	0.41	0.40	0.167	selftalk: 509, frustrated: 300, delighted: 358, social: 140, dysregulated: 119, request: 148
ReCANVo, excluding P16	0.47	0.46	0.167	selftalk: 414, frustrated: 240, delighted: 433, social: 136, dysregulated: 119, request: 148
IEMOCAP	0.43	0.44	0.25	neutral: 614, frustrated: 549, excited: 400, sad: 285
VIVAE	0.56	0.56	0.167	pleasure: 162, surprise: 150, pain: 148, fear: 141, anger: 139, achievement: 129
RAVDESS	0.59	0.59	0.143	angry: 154, fearful: 154, disgust: 154, sad: 154: surprised: 154, happy: 154, calm: 154
MSP-Podcast	0.21	0.16	0.125	neutral: 16518, happy: 8564, surprise: 2451, contempt: 2179, angry: 1839, sad: 1681, disgust: 1363, fear: 969
UrbanSound8K	0.42	0.42	0.125	engine idling: 805, jackhammer: 802, street music: 800, dog bark: 800, children playing: 800, air conditioner: 800, drilling: 800, siren: 730

Table 14: Performance of base models on validation data and training class sizes. A base model using our dataset of vocalizations from mv individuals was trained for each participant without the target participant's data to prevent prior exposure to any samples used in model evaluation.*

9.2.2 Zero Shot Learning for Valence and Arousal Modeling

The extended Geneva minimalistic acoustic parameter set (eGeMAPs) [113] was extracted for each vocalization in the VIVAE core dataset [101]. The VIVAE dataset was used in this experiment because its content (acted nonverbal vocalizations) was most similar to the ReCANVo dataset although the former were acted and the latter natural. Random forest regressors (150 estimators, 4 minimum samples/split) were trained to infer valence and arousal with an 80/20 training/validation data split. Training labels were mean valence (1-7 for negative to positive) and arousal (1-7 for minimal to maximal) ratings from 30 raters included with the published dataset for training. The R^2 values on validation data were 0.33 and 0.76 for the valence and arousal models, respectively. The trained model was used to infer valence and arousal ratings for nonverbal vocalization from mv* individuals. Only vocalizations with affective labels were included in this analysis.

9.2.3 AdaBoost and TrAdaBoost

Adaboost and multi-class TrAdaBoost were implemented to explore models with boosting and boosting combined with transfer learning. The target data was the participant for whom the model was being evaluated, and the source data was data from other participants that was used to augment training. The algorithm is from [144] for multi-class TrAdaBoost using forward stagewise modeling of SAMME as in the multi-class AdaBoost algorithm by [145]. The weights, w, are updated according to

$$w_i^{t+1} = f(x) = \begin{cases} w_i^t \cdot K(1-\varepsilon_t) \cdot e^{-\frac{K-1}{K}\alpha_t} & if\ h_{t(x_i)} = y(x_i),\ 1 \le i \le m \\ w_i^t \cdot K(1-\varepsilon_t) \cdot e^{-\alpha} \cdot e^{-\frac{K-1}{K}\alpha_t} & if\ h_{t(x_i)} \ne y(x_i),\ 1 \le i \le m \\ w_i^t \cdot K(1-\varepsilon_t) \cdot e^{-\frac{K-1}{K}\alpha_t} & if\ h_{t(x_i)} = y(x_i),\ m+1 \le i \le n+m \\ w_i^t \cdot e^{\frac{1}{K}\alpha_t} & f\ h_{t(x_i)} \ne y(x_i),\ m+1 \le i \le n+m, \end{cases}$$

where i indicates a sample in the training set, t indicates the boosting iteration, w_i^t denotes the weight of sample i at iteration t, K is the number of classes, n is the number of samples in the target dataset, m is the number of samples in the source (augmenting) dataset, N is the total number of boosting iterations, $h_{t(x_i)}$ is the prediction for training sample x_i at iteration t, $y(x_i)$ is the true label of training sample x_i, ε_t is the error on the target dataset at iteration t given by

$$\varepsilon_t = \sum_{i=1}^{n} \frac{w_i^t \cdot (I(h_{t(x_i)} \neq y(x_i))}{\sum_{i=1}^{n} w_i^t}$$

where I is an indicator function defined as

$$I(h_{t(x_i)} \neq y(x_i)) = \begin{cases} 1, & h_{t(x_i)} \neq y(x_i) \\ 0, & h_{t(x_i)} = y(x_i) \end{cases},$$

and α and α_t are scaling factors defined as

$$\alpha_t = log(\frac{1-\varepsilon_t}{\varepsilon_t}) + log(K-1) \text{ and}$$

$$\alpha = log(\frac{1}{1+\sqrt{2ln(m/N)}}).$$

The model prediction is given by $H(x) = \underset{k}{argmax} \sum_{t=1}^{N} \alpha_t \cdot I\left(h_{t(x)} = k\right)$

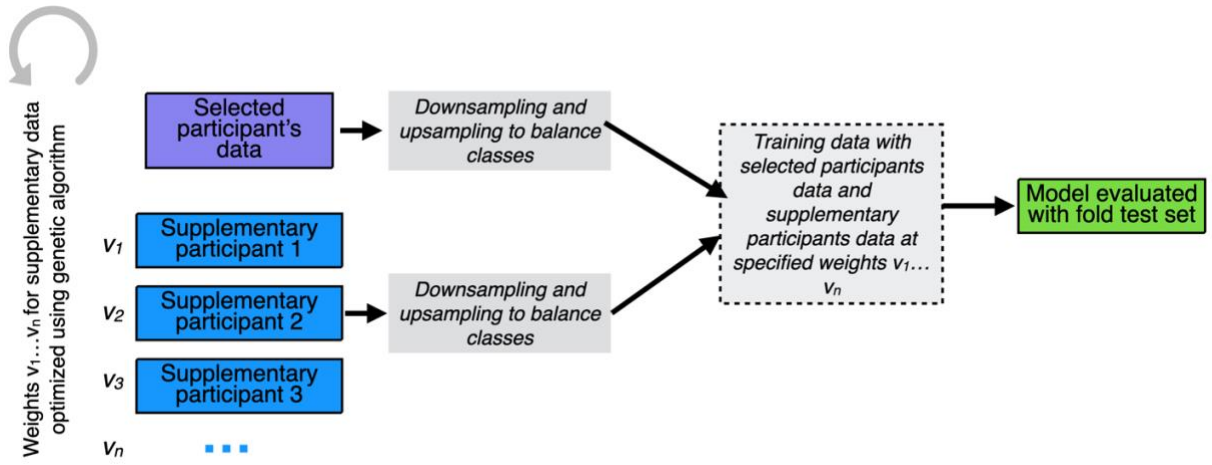


Figure 37: A genetic algorithm was used to select sample weights for models trained with all participants' data.

Models were trained separately for each participant, where the selected participant's data had weight 1 and the supplementary data had weights <1. A single weight was assigned to all data from a given participant.

Experiments were conducted with the leave one session out evaluation with session stratification and the aggregate feature sets described in 8.1.4. In the experiments, the base estimator was a decision tree with a maximum depth of 2. The above formulation was used for both Adaboost and TrAdaBoost experiments. For the AdaBoost experiments, $m = 0$ and no source (augmenting) data was included in training. Because TrAdaBoost experiments required overlap between the target participants' labels and the labels for the other participants, the "yes" class was not included for P06 because no other participants had samples of vocalizations for that label.

9.2.4 Cross-Training Among Participants

For cross-training experiments, a personalized model was trained for each participant where other participants' data was included in training with lower weight (Figure 37). Experiments were conducted with the leave one session out evaluation with session stratification. As in the TrAdaBoost experiments, the "yes" class was not included for P06 because no other participants had samples of vocalizations for that label. A genetic algorithm was used to optimize the weights of

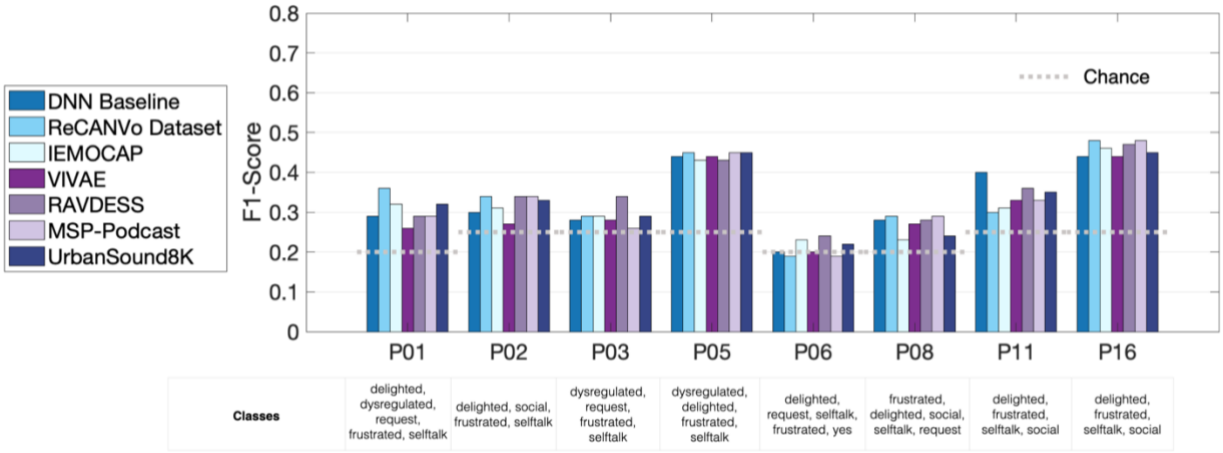


Figure 38: Results of transfer learning experiments described in 9.2.1 with personalized DNNs. For each participant, the reported F1-scores correspond to the model performance in predicting the labels listed below the graph.

the augmenting data. A single weight was assigned to all data from a given participant. The genetic algorithm was run for 50 generations, with a population size of 6 and a cross-over rate of 0.9. The optimization was run with the best performing aggregate feature set for each participant.

9.3 Results and Discussion

9.3.1 Transfer Learning with Personalized Neural Networks

Results from the transfer learning experiments classifying valence and arousal with DNNs are shown in Figure 38. Transfer learning within the ReCANVo dataset improved model performance compared to the baseline DNN for six participants, though some improvements were small and may not be significant. There was a larger improvement with the RAVDESS dataset for P03. Transfer learning often did not change, slightly improved, or slightly reduced the overall F1 score for other participants and base datasets. The performance changes suggest that there may be some transferability in learning between participants and in some cases between nonverbal vocalizations from mv* individuals and more general affective speech datasets.

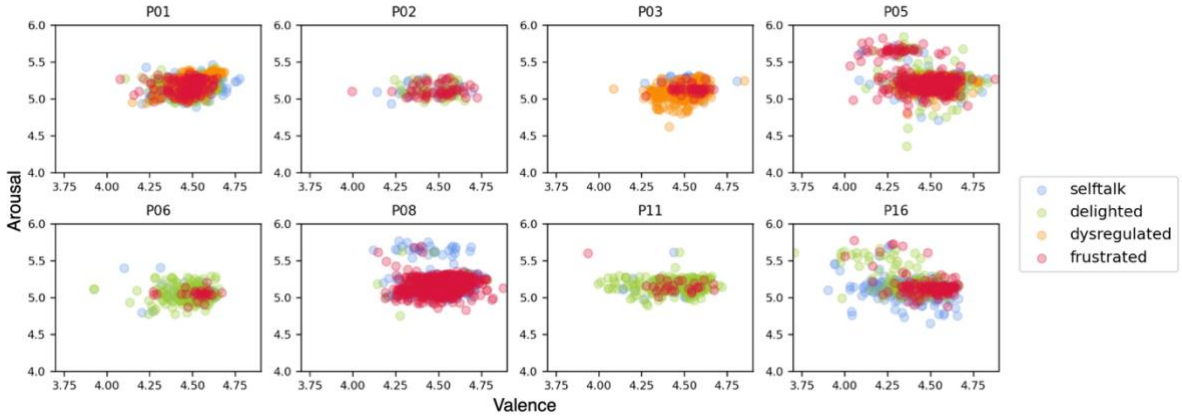


Figure 39: Valence and arousal ratings for each vocalization were inferred using zero shot transfer learning with a Random Forest regressor (as described in 9.2.2). Plot colors show the label of each vocalization, to visualize relations between predictions and labels

9.3.2 Zero Shot Learning for Valence and Arousal Modeling

Zero-shot learning was used to explore whether valence and arousal ratings inferred (Figure 39) by a model trained on the VIVAE dataset (nonverbal vocalizations that occur amidst typical speech) captured expected relative valence and arousal characteristics between labels for each participant - i.e., higher relative arousal for "frustrated" than "dysregulated and for "delighted" than "selftalk" and more positive valence for "delighted" and "selftalk" than "frustrated" and "dysregulated". For some participants (P03, P08, P11, and P16), vocalizations of the same label appear in clusters on the valence-arousal plots; for others like P01 and P02 there are no distinct groupings. Frustrated" had a high relative average arousal within P03's vocalizations, and "delighted" had a high relative average arousal within P16's vocalizations but generally the predicted valence and arousal rating estimated using VIVAE did not clearly relate to characteristics of the known affective labels.

9.3.3 AdaBoost and TrAdaBoost

The model performance of the AdaBoost and TrAdaBoost models are shown in Table 15. AdaBoost models had higher F1 scores than the modeling approaches described in 8.2.1 for P06. TrAdaBoost for P06 were compared to the best performing 4-class model (ERB features with SVM

RBF; F1 score: 0.40, ERB; UAR: 0.42) because the "yes" class could not be used with the described transfer learning approaches. Multi-class TrAdaBoost did not improve model F1 scores for any participants.

	P01		P02		P03		*P05*		*P06*		*P08*		*P11*		*P16*	
Chance	**0.20**		**0.25**		**0.25**		**0.25**		**0.25**		**0.20**		**0.25**		**0.25**	
Adaboost	*F1*	*UAR*	*F1*	*UAR*	*F1*	*UAR*	*F1*	*UAR*	*F1*	*UAR*	*F1*	*UAR*	*F1*	*UAR*	*F1*	*UAR*
eGeMAPs	0.25	0.26	0.23	0.24	0.40	0.38	0.37	0.38	0.30	0.32	0.28	0.30	0.30	0.34	0.41	0.42
ERB	0.30	0.35	0.13	0.14	0.35	0.35	0.34	0.34	0.33	0.35	0.33	0.37	0.21	0.24	0.32	0.33
GTCC	0.20	0.21	0.29	0.30	0.36	0.36	0.38	0.38	0.28	0.29	0.27	0.31	0.26	0.31	0.36	0.38
Custom	0.30	0.32	0.20	0.20	0.38	0.38	0.36	0.37	0.22	0.23	0.27	0.27	0.35	0.40	0.38	0.41
MFB	0.20	0.20	0.27	0.27	0.29	0.30	0.33	0.34	**0.34**	0.34	0.27	0.31	0.29	0.31	0.39	0.40
MFCC	0.26	0.25	0.26	0.27	0.33	0.35	0.35	0.36	0.22	0.23	0.24	0.27	0.26	0.32	0.39	0.44
TrAdaboost																
eGeMAPs	0.23	0.27	0.21	0.26	0.33	0.36	0.35	0.38	0.33	0.45	0.13	0.17	0.34	0.36	0.24	0.25
ERB	0.26	0.34	0.31	0.38	0.32	0.32	0.36	0.47	0.23	0.26	0.17	0.23	0.22	0.23	0.30	0.29
GTCC	0.21	0.23	0.18	0.22	0.22	0.25	0.26	0.30	0.27	0.35	0.26	0.29	0.26	0.26	0.32	0.36
Custom	0.26	0.30	0.25	0.29	0.31	0.34	0.25	0.28	0.17	0.19	0.28	0.30	0.23	0.29	0.27	0.30
MFB	0.28	0.30	0.24	0.29	0.21	0.24	0.21	0.28	0.27	0.29	0.14	0.18	0.31	0.32	0.39	0.38
MFCC	0.19	0.28	0.17	0.25	0.30	0.34	0.32	0.34	0.28	0.40	0.19	0.25	0.21	0.23	0.31	0.43

Table 15: AdaBoost and TrAdaBoost model performance with leave one session out evaluation with session stratification. F1 scores higher than the best performance for the models described in 8.2.1 are shown in green.

Model performances for P06 - the participant for whom AdaBoost improved model performance - were lower than that model performance for other participants across modeling approaches. The lower relative performances and the relative success of AdaBoost could be related to properties of the training data, like labeling quality and number of samples. The improvements from AdaBoost were small but suggests that there may be cases where boosting can help improve model performance.

9.3.4 Cross-Training Among Participants

Cross-training consistently had the best results with Random Forest Models. Model results with the optimal weights selected by the genetic algorithm are presented in Table 16.

Participant	Chance	Feature Set	F1	F1 Δ	Data weight for best performing model							
					P01	*P02*	*P03*	*P05*	*P06*	*P08*	*P11*	*P16*
P01	0.20	Custom	0.409	+0.070	1	0.73	0.77	0.48	0.49	0.28	0.41	0.70
P02	0.25	Custom	0.315	-0.024	0.96	1	0.96	0.60	0.14	0.35	0.15	0.34
P03	0.25	Custom	0.549	+0.089	0.66	0.30	1	0.62	0.31	0.63	0.54	0.67
P05	0.25	eGeMAPs	0.542	+0.041	0.48	0.93	0.043	1	0.53	0.58	0.89	0.20
P06*	0.25	ERB	0.387	-0.013	0.30	0.11	0.69	0.75	1	0.50	0.54	0.73
P08	0.20	Custom	0.293	-0.021	0.29	0.83	0.70	0.41	0.42	1	0.60	0.33
P11	0.25	Custom	0.376	-0.014	0.29	0.54	0.48	0.10	0.18	0.30	1	0.68
P16	0.25	eGeMAPs	0.515	-0.004	0.98	0.40	0.66	0.069	0.32	0.43	0.37	1

Table 16: F1 Score for cross-training with optimal weights for each participant, evaluated with leave one session out and session stratification.

A four-class model was used for P05, since the "yes" category was unique to P06 and could not be used in cross-training. The performance change for P06 is reported with respect to the best performing 4-class model and feature set with the methods described in 8.1.

Cross-training improved the model F1 score for P01, P03, and P05. Improvements ranged from 4.1-8.9%. For other participants, cross-training decreased model performance by 0.9-2.4%. These preliminary results suggest that there may be some overlap in how affect and intent are communicated between participants.

9.4 Conclusions

This is the first exploration of transfer learning applied to nonverbal vocalizations from mv* individuals. The results suggest that there may be some overlap in how affect is expressed in nonverbal vocalizations from mv* individuals and nonverbal vocalizations that occur amidst typical verbal speech, and that there may be some cases where existing speech datasets can be used in

modeling. Future work with larger datasets from more mv* individuals could further explore these hypotheses. The limited success of the presented approaches, particularly with the zero-shot model for valence and arousal inference, highlights the need for creating datasets directly with the specialized population of mv* individuals. Such data collection efforts and appropriate targeted models could enable improved understanding of mv* communicators.

10 Commalla: Communication Interface Design and Evaluation

The presented offline classification results are a step towards a real-world, real-time, personalized AAC system that enhances dyadic interaction and understanding between mv* individuals and the world. To that end, we have developed a prototype app, Commalla, for AAC systems based on the concepts described in 5.1 and 5.2. The name "Commalla" is derived from the phrase "Communication for all". The work presented in this chapter was created collaboratively with Kristy Johnson and Craig Ferguson (app and UI developer), and excerpts from [125].

10.1 Alpha Prototype

10.1.1 Design

We created a prototype smartphone app that records a vocalization and employs a personalized, pre-trained model to classify that individual's vocalizations in real-time (Figure 40) and show the classification result to the user. Pressing and recording the microphone button records and sends a vocalization to a cloud-based server for real-time classification (Figure 40a). Users are instructed to start and stop recording as close to the start and end of a single vocalization as possible. A 2 second buffer is included with each recording to minimize clipping at the start of a vocalization. The

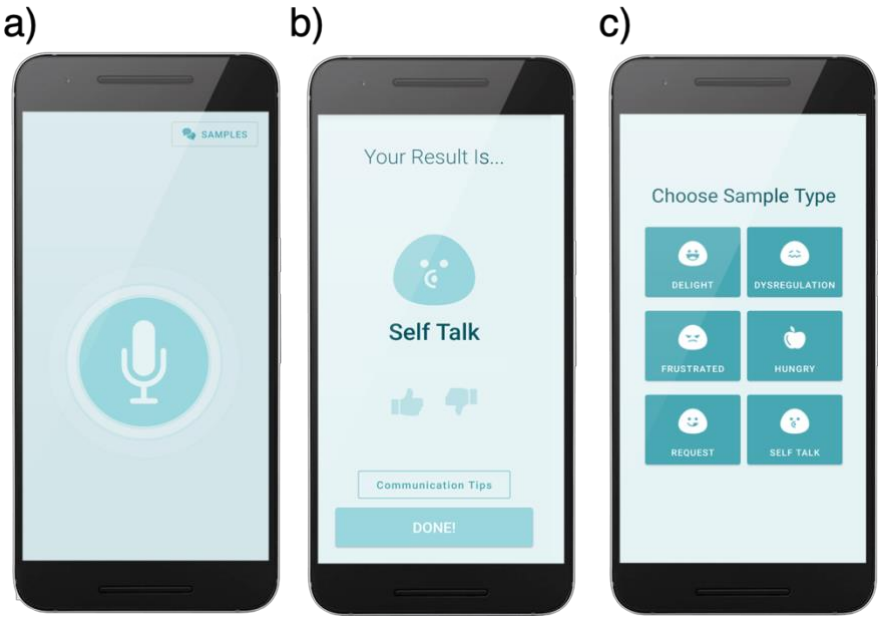


Figure 40: Alpha prototype for a personalized real-time classification application and sample sound library.

a) Pressing the microphone button records and sends a vocalization to a cloud-based server for real-time classification. b) In < 1 sec, the result is shown to the user, who can provide feedback as to whether the prediction was correct or incorrect. c) The user can also listen to personalized examples of each vocalization type to share with others to expand awareness and dyadic communication.

buffer is recorded continuously whenever the app is open but is continuously re-written to prevent privacy. Data is only uploaded to the cloud when the user initiates a classification.

10.1.2 Evaluation

The app was pilot tested with the family of P03 (Table 2), who provided qualitative feedback. One of P03's parents was a researcher involved in the project. The family reported that the app was generally easy to use and navigate. The family said that recording data using the app was much easier than the recorder and app system used for collecting the dataset.

A binary classification model was trained to distinguish between two labeled states -- selftalk (positive affect) and dysregulation (negative affect). To more closely resemble audio recorded via a smartphone, each original audio segment from P03's wearable recorder was passed through a 6th-order lowpass Butterworth filter with a cutoff frequency of 8 kHz, and the amplitude was multiplied

by 0.22. These altered segments were combined with the original segments to create an augmented training set twice the original set size (n=48 selftalk; n=244 dysregulation). Then the binary model was trained and validated using downsampled, balanced sets of 48 samples from the augmented dataset, and tested using a held-out, future test set of smartphone-recorded audio files. The resultant model had a 90.4% classification accuracy, with 14 self-talk true positives (TP), 1 self-talk false positive (FP), 5 dysregulation TP, and 1 dysregulation FP.

The audio data recorded during the app had more variable amplitude compared to data using the recorder, because the app was being held by another person in the room instead of being worn or placed nearby. Because vocalizations were recorded via a press and hold mechanism on the app, vocalization segments recorded using the app had start and endpoints defined by the user in real-time as compared to post-processing.

10.2 Latest Prototype

10.2.1 Iterative Design and Evaluation

Recording data directly on the phone was easier for the family pilot testing the app compared to using the recorder because they did not have to set-up multiple pieces of equipment to collect data. Pilot testing with Prototype I suggested that it was feasible to train models using the app-recorded data, which had lower quality and higher variability than audio captured with the dedicated recorder. Because of the press and hold recording modality, data recorded on the app were already reasonably segmented around vocalizations and could be analyzed without the time-intensive processing to align and segment vocalizations that was required with the recorder data. For these reasons, the next prototype of the app integrated data collection and labeling along with the classification features and sample sound player.

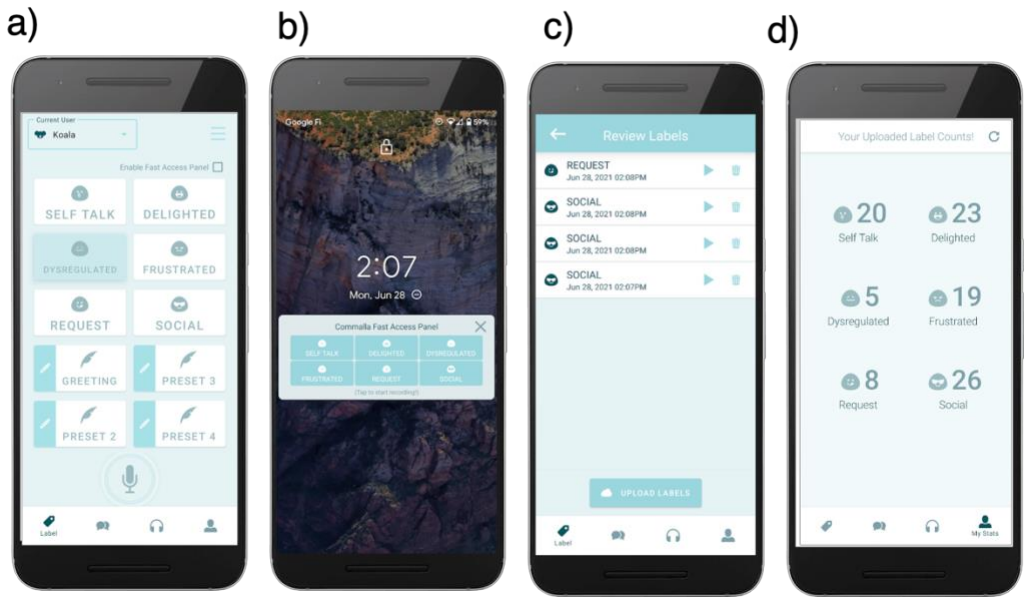


Figure 41: Data collection features on the latest app version.

a) The main labeling and recording screen. A user presses and holds a label to record a vocalization. An animation appears to mark the selected label and clearly show the user that sound is being recorded. A checkbox for "Enable Fast Access Panel" activates a widget to label on the lock screen. The "Current User" dropdown menu keeps track of who is labeling data, in case there are multiple distinct labelers. The hamburger menu allows users to access the "Review Labels" screen. b) The "Fast Access Panel" can be used to quickly label data without opening the full app. The panel allows users to record vocalizations for the six shared labels directly from the lock screen. c) Users can listen to data they have recorded and delete any samples they do not want to upload. d) The "My Stats" screen is accessed via the menu bar at the bottom of the app and is designed to help users keep track of how many samples of each vocalization type they have recorded.

The integrated app was continuously evaluated with P03 and his family. The app includes data recording and labeling, classification, listening, and tracking features. The data collection features in the latest version of the app are shown in Figure 41. Like the previous version, the app includes a 2s pre-buffer to minimize clipping of vocalization in the recording and classification features. In pilot testing, we found that there was sometimes slight clipping at the end of a vocalization and so the latest prototype also includes a 0.25 second post-buffer. The buffer is continuously re-written to prevent privacy.

The design was updated iteratively based on feedback from pilot testers. For example, the "Review Labels" panel was added based on feedback from the family emphasized that having the chance to

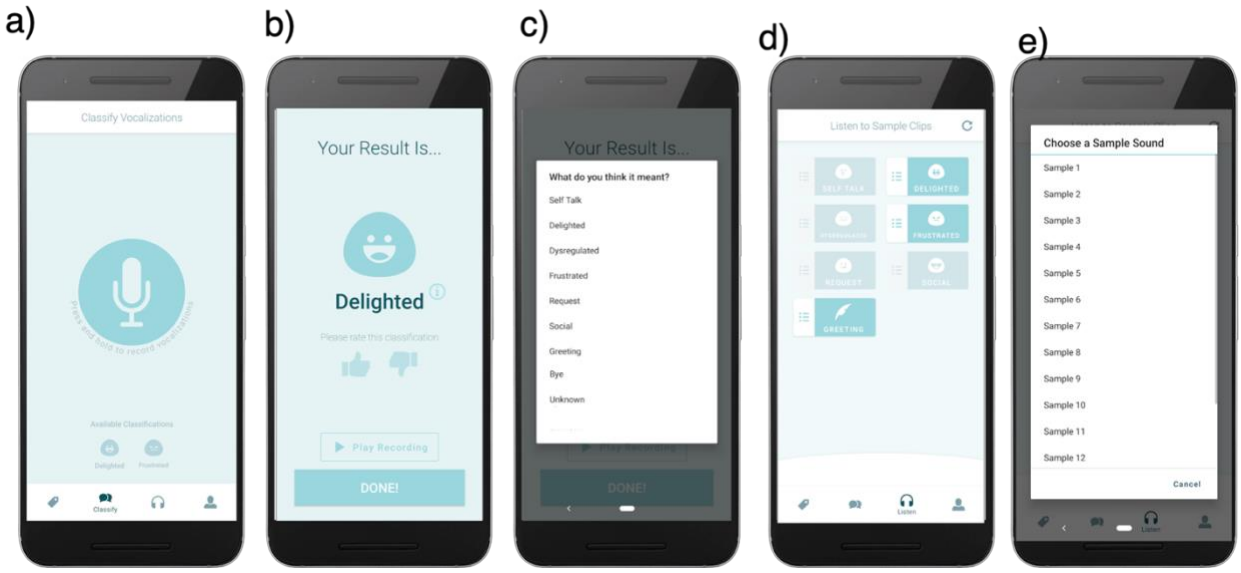


Figure 42: The classification and sample sound library in the latest app design.

b) To use the real-time classification feature, first the user records a vocalization by pressing and holding the microphone. The screen shows the classes that the trained model can infer to ensure the user understands that the model outputs are limited. b) The recording is sent to a server, where it is processed using a personalized machine learning model and the predicted class is shown to the user. The user can upvote or downvote the classification and listen to the file that they recorded. c) If a user downvotes a classification, they can provide feedback as to what they think the vocalization was expressing. d) The app includes a sample sound library where users can listen to recorded vocalization samples grouped by label. Pressing a label plays a vocalization associated with that label at random. e) Pressing the bullet list symbol icon next to a label allows the user to select from all the samples available for that label.

review and delete inaccurately labeled data was important. The "Review Labels" panel allows a user to listen to recordings before they are uploaded and delete any recordings they do not wish to upload (Figure 41c). Dark button shading (as shown on the "Dysregulated" label in Figure 41a) along with a microphone animation to highlight to a user which label they are pressing was included based on user feedback that the more subtle background color shift in previous versions was not marked enough. Additionally, the "Fast Access Panel" (Figure 41b) was added based on user feedback that by the time a labeler out the phone and started the app to collect data, the vocalization they were hoping to record was over. The "Fast Access Panel" allows a user to label and record data from the lock screen which can be accessed more quickly than the full app.

The app design was also informed by insights from the longitudinal study with the combined recorder and app labeling system. While collecting data with the system described in 5.3, we found that updating labelers on how many vocalizations of each label they had collected helped users collect more balanced datasets. Based on this experience, we added a "My Stats" summary page to the app to show users how many samples they have labeled (Figure 41d).

The classification and sample sound library features are shown in Figure 42 (based on the classification and education interface concepts described in Chapter 5). These features are accessed via a menu bar at the bottom of the screen. In the "Classify" screen, a user can initiate a real-time classification by recording a vocalization by pressing and holding the microphone icon. The labels at the bottom of the screen (Figure 42a) highlight to the user that the model can only output limited options and that those options are pre-specified. Once the user submits a vocalization, the audio is sent to a server where a personalized model classifies the vocalization. The result is then shown to the user (Figure 42b). The user can upvote or downvote the classification. If the user downvotes the classification, they can provide feedback as to what they thought the vocalization meant (Figure 42c).

The app also includes a sample sound library (Figure 42d), where users can listen to vocalizations grouped by label. Someone who doesn't know the mv* communicator well could use this feature to learn about the communicator's vocalizations, and how different sounds convey affect and intent. Pressing a label in the library plays a vocalization of that category. The user can select a specific vocalization in a category via a menu that appears (Figure 42e) when the bullet list icon next to a label is selected. During the pilot testing, one of P03's parents said that they found the sample

sound library particularly useful and had already used the library to teach someone about their child's vocalizations.

10.3 Concept Evaluation: Interviews

Participants in the larger data collection study (Table 2) were asked for feedback on the classification and sample sound library educational interface concepts during exit interviews. At least one of the proposed interface functionalities would have utility for 6/7 of the interviewees (6).

Five interviewees reported that they would find an app that could classify vocalizations in real-time useful (5). Four said they would give the app to help babysitters, new teachers, and others who don't know the communicator well better understand communication in real-time (4). One mentioned that they would particularly like a classification app to help people understand different types of frustrated vocalizations from the communicator, particularly frustration due to boredom (1). One interviewee mentioned that they would like to use the app to communicate to others in the general vicinity in public that the communicator's vocalizations express happy sounds because the communicator's happy sounds are often misinterpreted (1). One interviewee said that a classification app that the communicator could train for safety and comfort related communication (e.g., "yes", "no", to share a phone number) would be particularly helpful (1). One interviewee said that when they began the study they thought the classification app would be helpful, but they are now more interested in the sample sound library (1). Five interviewees said that they would find an educational app (e.g., a sound library where people can listen and learn about the communicator's vocalizations) useful (5). Interviewees also mentioned interest in a tool to track language development (2) and pinpoint sources of pain (3). While our study did not capture specific information about pain, studying how people communicate about pain nonverbally could be an

impactful area for future research and development. One interviewee mentioned that it was important that the communication interface did not require the communicator to wear something, due to tactile sensitivities (1).

Quotes from interviewees on communication interfaces:

- *"I don't believe the people who work with [my child] pay attention much to his vocalizations. I don't think they necessarily try to interpret his vocalizations. So, in that setting it [a tool] could be very useful."*
- *"I do think the greater utility is going to be in the sample library and just hearing these ideas. I think the biggest leap is in seeing that there are different vocalization types and that they are discrete and discernable. Really helping them [people who don't know the communicator well] discern the sounds and know to be aware of those sounds and then teaching them how to respond."*
- *"[Communicator] will scream happily in a store and I'll turn around and go 'Oh you're really happy to be in Target right now?'. It would be nice if their device said 'I'm happy' to communicate that to people"*
- *"Anything that is going to help communication would be helpful. Something that can translate and help facilitate communication – it would be helpful. Ultimately the dream is that we can get this higher-level information from them [the communicator], not just this…I know that's not what you're doing, but I want to understand not just what this sound means but why are they making this sound? Why are they feeling joyful? What is bringing them joy so we can bring it to them again?"*
- *"I think that [a classification app for vocalizations] would be useful for new teachers or even new helpers at our house or even family members that might not visit very often. That*

would be helpful. If we are around then we can help because we know them [the communicator] well enough but we are not always around so that would be helpful.

10.4 Conclusions

The model results presented in Chapters 8 and 9 along with interface designs and feedback presented in this chapter suggest that vocalization-based communication interfaces are feasible and useful. Additional work with broad user testing and deployment could help further refine the proposed interfaces. Directions for further development are discussed in Chapter 11.

11 Future Work and Conclusions

This thesis included longitudinal data collection and analysis of nonverbal vocalizations from mv* individuals and human-centered design applied to new AAC interfaces that incorporate vocalizations. This thesis contributed a study of user needs, novel data collection systems and communication interfaces, and evaluation strategies and analysis with real-world data. The presented work provides a foundation for further work analyzing nonverbal vocalizations and developing new models and communication interfaces.

11.1 Future Work

11.1.1 Data Analysis and Modeling

The presented ReCANVo dataset captures motivation-driven communication in day-to-day life. The dataset can be used in analysis towards better understanding speech and language development trajectories. Data analyses with the collected data could also study relations between conversational turns, context, and vocalizations.

Creating the ReCANVo dataset involved some manual audio trimming and verification steps. Expanded data collection and analysis would be enabled by training a vocalization onset and offset detector using the segmented vocalizations. Then, the trained model could be used to find unlabeled vocalization segments in the dataset, with label hypotheses developed using contextual

audio information. Semi-supervised and weak label learning approaches could be explored with the expanded dataset.

As part of exploratory analyses, additional features were evaluated including features related to perceptual loudness (as in [146]) but were not included in the final custom feature set. As more data with more participants is collected, such approaches (along with the approaches presented in detail in Chapter 8) should be re-assessed as they may be more informative with a larger dataset.

11.1.2 Expanded Data Collection

Additional data collection with more mv* communicators could enable new analysis approaches. A larger participant-base could help identify whether there are grouping of individuals with overlapping nonverbal communication practices. Identifying groups with shared practices could better enable transfer learning and domain adaptation approaches in modeling. Expanded data collection could also help in identifying whether there are individuals for whom vocalization-based communication is particularly interpretable via modeling approaches. Collecting data over an even longer time span (i.e., multiple years) could enable studying the trajectory of vocalization use and content within an individual, in a manner similar to the Speechome project which was a multi-year data collection undertaking with a neurotypical toddler [147]. Further data collection could incorporate additional modalities like gestures by asking labelers to record short videos that capture additional information, though such a system would have heightened privacy concerns.

11.1.3 Communication Interfaces

Additional work on the development and evaluation of the proposed interfaces includes expanded user testing with diverse communicators and the development of new features. In the current prototype, models must be manually updated for each participant as new data becomes available.

Incorporating active learning via automatic evaluation and deployment of models could be an impactful next step in the technical development of the classification interface.

The utility of the classification feature could be improved by integrating the feature with AAC devices. Vocalization-based control of AAC apps (e.g., apps like Prolouquo2Go) could help users navigate through options more quickly, while interacting with communication partners. Allowing users to freely customize all labels could expand the population for whom the interface is useful. Additional features could also include more general voice-based control of tablets and cell phones, which could include integration with speech interfaces like Alexa and Siri.

The sample sound library feature is an educational interface that can be used to help new communication partners learn about an mv* individual's communication style. Additional development could include creating an interactive educational game to enhance learning and including other communication modalities into the library like videos of gestural communication.

Future development could also include an interface to allow users fine-grained control over the uploaded audio. BirdNET [148] is an app that allows users to record bird songs and then attempts to identify the bird species using the recording, shows users a spectrogram of a recorded clip. Users can then clip the recording to trim excess noise. Including similar features in future revisions could improve data quality. Presenting visual representations of recordings could also help users learn about different vocalization types and identify patterns in vocalizations. However, asking users to complete additional cleaning tasks could increase cognitive load on users and reduce data quantity. Future studies could carefully evaluate the trade-off between collecting additional information directly from users and having a minimal interface with low time burden.

11.2 Conclusions

This thesis presented the development of interfaces and models towards improved understanding of real-world communicative and affective nonverbal vocalizations from minimally speaking individuals. The contributions of this thesis include:

- The identification of user needs and research opportunities for AAC via interviews, survey, and a case study along with the compilation and publication of survey results for use by other researchers
- The development and evaluation of a novel data collection protocol for real-world audio with personalized in-the-moment labels
- The creation of ReCANVo, the first database of nonverbal vocalizations with affective and communicative labels from mv* individuals
- The development of machine learning evaluation strategies and algorithms suitable for real-world data that can classify nonverbal vocalizations from mv* individuals with F1 scores well above chance
- The design and evaluation of novel communication interfaces towards enhanced understanding of nonverbal vocalizations

While this work was driven by the potential for impact with a small, underserved population, the work can more broadly inform designing unobtrusive remote studies and data collection protocols, and model development and evaluation for real-world in-the-moment labeled data.

References

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

[2] "Key Findings: CDC Releases First Estimates of the Number of Adults Living with Autism Spectrum Disorder in the United States | Autism | NCBDDD | CDC." [Online]. Available: https://www.cdc.gov/ncbddd/autism/features/adults-living-with-autism-spectrum-disorder.html. [Accessed: 09-Jun-2021].

[3] G. E. Martin, J. Klusek, B. Estigarribia, and J. E. Roberts, "Language characteristics of individuals with down syndrome," *Top. Lang. Disord.*, vol. 29, no. 2, pp. 112–132, Apr. 2009.

[4] R. Didden *et al.*, "Communication in individuals with Rett syndrome: An assessment of forms and functions," *J. Dev. Phys. Disabil.*, vol. 22, no. 2, pp. 105–118, 2010.

[5] "Mowat-Wilson syndrome: MedlinePlus Genetics." [Online]. Available: https://medlineplus.gov/genetics/condition/mowat-wilson-syndrome/. [Accessed: 18-May-2021].

[6] "Rubinstein-Taybi Syndrome - American Association for Pediatric Ophthalmology and Strabismus." [Online]. Available: https://aapos.org/glossary/rubinstein-taybi-syndrome. [Accessed: 18-May-2021].

[7] "Pitt-Hopkins syndrome: MedlinePlus Genetics." [Online]. Available: https://medlineplus.gov/genetics/condition/pitt-hopkins-syndrome/. [Accessed: 18-May-2021].

[8] American Speech-Language-Hearing Association, "Augmentative and Alternative Communication (AAC)," 2021. [Online]. Available: https://www.asha.org/public/speech/disorders/aac/. [Accessed: 04-May-2021].

[9] J. Narain and P. Maes, "Understanding AAC Usage and Needs through a Web Survey with Device Users and Families," in *Proceedings of the Annual International Conference of the IEEE Engineering in Medicine and Biology Society, EMBS*, 2020, vol. 2020-July, pp. 3864–3868.

[10] M. Alper, *Giving Voice: Mobile Communication, Disability, and Inequality*. Cambridge, MA: MIT Press, 2017.

[11] D. McNaughton, J. Light, D. R. Beukelman, C. Klein, D. Nieder, and G. Nazareth, "Building capacity in AAC: A person-centred approach to supporting participation by people with complex communication needs," *AAC Augment. Altern. Commun.*, vol. 35, no. 1, pp. 56–68, Jan. 2019.

[12] B. Batorowicz, S. Mcdougall, and T. A. Shepherd, "AAC and community partnerships: The participation path to community inclusion," *AAC Augment. Altern. Commun.*, vol. 22, no. 3, pp. 178–195, Sep. 2006.

[13] C. Binger *et al.*, "Personnel roles in the AAC assessment process," *AAC Augment. Altern. Commun.*, vol. 28, no. 4, pp. 278–288, Dec. 2012.

[14] “Book Review: ‘Nobody’s Normal’ chronicles the intertwined history of mental illness and stigma | Spectrum | Autism Research News.” [Online]. Available: https://www.spectrumnews.org/opinion/book-review-nobodys-normal-chronicles-the-intertwined-history-of-mental-illness-and-stigma/. [Accessed: 24-May-2021].
[15] M. Alper and G. Goggin, “Digital technology and rights in the lives of children with disabilities,” *New Media Soc.*, vol. 19, no. 5, pp. 726–740, May 2017.
[16] P. Hess, “Autism behind bars | Spectrum | Autism Research News.” [Online]. Available: https://www.spectrumnews.org/features/deep-dive/autism-behind-bars/. [Accessed: 24-May-2021].
[17] “Identity-First Language | Autistic Self Advocacy Network.” [Online]. Available: https://autisticadvocacy.org/about- asan/identity-first-language/. [Accessed: 24-May-2021].
[18] “Communicating With and About People with Disabilities | CDC.” [Online]. Available: https://www.cdc.gov/ncbddd/disabilityandhealth/materials/factsheets/fs-communicating-with-people.html. [Accessed: 24-May-2021].
[19] S. Lydon, O. Healy, P. Reed, T. Mulhern, B. M. Hughes, and M. S. Goodwin, “A systematic review of physiological reactivity to stimuli in autism,” *Developmental Neurorehabilitation*, vol. 19, no. 6. Taylor and Francis Ltd., pp. 335–355, 01-Nov-2016.
[20] E. Hedman, L. Miller, S. Schoen, D. Nielsen, M. Goodwin, and R. Picard, “Measuring Autonomic Arousal During Therapy.,” in *Proceedings of Design and Emotion*, 2012, pp. 11–14.
[21] K. K. Hyde *et al.*, “Applications of Supervised Machine Learning in Autism Spectrum Disorder Research: a Review,” *Review Journal of Autism and Developmental Disorders*, vol. 6, no. 2. Springer New York LLC, pp. 128–146, 15-Jun-2019.
[22] R. W. Picard, “Future affective technology for autism and emotion communication,” *Philos. Trans. R. Soc. B Biol. Sci.*, vol. 364, no. 1535, pp. 3575–3584, Dec. 2009.
[23] M. S. Goodwin *et al.*, “Predicting imminent aggression onset in minimally-verbal youth with autism spectrum disorder using preceding physiological signals,” in *ACM International Conference Proceeding Series*, 2018, pp. 201–207.
[24] 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.
[25] M. De Leon, “Total Communication Approach | Strategies for Non-Verbal Children | NAPA.” [Online]. Available: https://napacenter.org/nonverbal-communication/. [Accessed: 19-May-2021].
[26] R. Paul, *Language disorders from infancy through adolescence : assessment & intervention*. Mosby, 2001.
[27] *Speech and Language Disorders in Children: Implications for the Social Security Administration’s Supplemental Security Income Program*. National Academies Press, 2016.
[28] AssistiveWare, “Speak up with symbol-based AAC,” 2021. .
[29] TouchChat App, “TouchChat - Communication Apps for iPad, iPhone, and iPod Touch,” 2021. [Online]. Available: https://touchchatapp.com/. [Accessed: 10-May-2021].
[30] “Products: NOVA chat, Chat Fusion, TouchChat Express.” [Online]. Available: https://saltillo.com/products. [Accessed: 18-May-2021].
[31] “Picture Exchange Communication System (PECS)® | Pyramid Educational Consultants.” [Online]. Available: https://pecsusa.com/pecs/. [Accessed: 18-May-2021].
[32] “Eye trackers and software for research | View our products.” [Online]. Available: https://www.tobiipro.com/product-listing/. [Accessed: 18-May-2021].
[33] “Big Talk One Message Communicator - Big Mack Switch | Enabling Devices.” [Online].

Available: https://enablingdevices.com/product/big-talk/. [Accessed: 18-May-2021].
[34] “Disability Adapted Switches for People with Disabilities.” [Online]. Available: https://enablingdevices.com/product-category/switches/. [Accessed: 18-May-2021].
[35] “Products » Control Bionics.” [Online]. Available: https://www.controlbionics.com/products/. [Accessed: 18-May-2021].
[36] “PISON.” [Online]. Available: https://pisontechnology.com/. [Accessed: 18-May-2021].
[37] Y. Elsahar, S. Hu, K. Bouazza-Marouf, D. Kerr, and A. Mansor, “Augmentative and alternative communication (AAC) advances: A review of configurations for individuals with a speech disability,” *Sensors (Switzerland)*, vol. 19, no. 8. MDPI AG, 02-Apr-2019.
[38] A. Moorcroft, N. Scarinci, and C. Meyer, “A systematic review of the barriers and facilitators to the provision and use of low-tech and unaided AAC systems for people with complex communication needs and their families.,” *Disabil. Rehabil. Assist. Technol.*, vol. 14, no. 7, pp. 710–731, Oct. 2019.
[39] S. Baxter, P. Enderby, P. Evans, and S. Judge, “Barriers and facilitators to the use of high-technology augmentative and alternative communication devices: a systematic review and qualitative synthesis.,” *Int. J. Lang. Commun. Disord.*, vol. 47, no. 2, pp. 115–29.
[40] J. Caron and J. Light, “‘Social Media has Opened a World of “Open communication:”’ experiences of Adults with Cerebral Palsy who use Augmentative and Alternative Communication and Social Media.,” *Augment. Altern. Commun.*, vol. 32, no. 1, pp. 25–40, 2016.
[41] D. H. Angelo, “Impact of augmentative and alternative communication devices on families,” *AAC Augment. Altern. Commun.*, vol. 16, no. 1, pp. 37–47, 2000.
[42] J. Marshall and J. Goldbart, “‘Communication is everything I think.’ Parenting a child who needs Augmentative and Alternative Communication (AAC).,” *Int. J. Lang. Commun. Disord.*, vol. 43, no. 1, pp. 77–98.
[43] J. M. Johnson, E. Inglebret, C. Jones, and J. Ray, “Perspectives of speech language pathologists regarding success versus abandonment of AAC.,” *Augment. Altern. Commun.*, vol. 22, no. 2, pp. 85–99, Jun. 2006.
[44] N. Franke and E. A. von Hippel, “Satisfying Heterogeneous User Needs via Innovation Toolkits: The Case of Apache Security Software,” *SSRN Electron. J.*, 2002.
[45] J. Light *et al.*, “Challenges and opportunities in augmentative and alternative communication: Research and technology development to enhance communication and participation for individuals with complex communication needs,” *AAC: Augmentative and Alternative Communication*, vol. 35, no. 1. Taylor and Francis Ltd, pp. 1–12, 02-Jan-2019.
[46] W.-D. Chang, H.-S. Cha, D. Y. Kim, S. H. Kim, and C.-H. Im, “Development of an electrooculogram-based eye-computer interface for communication of individuals with amyotrophic lateral sclerosis.,” *J. Neuroeng. Rehabil.*, vol. 14, no. 1, p. 89, 2017.
[47] H. Collaguazo, P. Cordova, and C. Gordon, “Communication and Daily Activities Assistant System for Patient with Amyotrophic Lateral Sclerosis,” in *2018 5th International Conference on eDemocracy and eGovernment, ICEDEG 2018*, 2018, pp. 218–222.
[48] A. Larson, J. Herrera, K. George, and A. Matthews, “Electrooculography based electronic communication device for individuals with ALS,” in *SAS 2017 - 2017 IEEE Sensors Applications Symposium, Proceedings*, 2017.
[49] G. S. Meltzner, J. T. Heaton, Y. Deng, G. De Luca, S. H. Roy, and J. C. Kline, “Silent speech recognition as an alternative communication device for persons with laryngectomy,” *IEEE/ACM Trans. Audio Speech Lang. Process.*, vol. 25, no. 12, pp. 2386–2398, Dec. 2017.
[50] X. Zhang, L. Yao, Q. Z. Sheng, S. S. Kanhere, T. Gu, and D. Zhang, “Converting Your Thoughts to Texts: Enabling Brain Typing via Deep Feature Learning of EEG Signals,” Sep.

2017.
[51] K. Trnka, J. Mccaw, D. Yarrington, K. F. Mccoy, and C. Pennington, “User interaction with word prediction:The effects of prediction quality,” *ACM Trans. Access. Comput.*, vol. 1, no. 3, Feb. 2009.
[52] D. G. Park, S. Song, and D. H. Lee, “Smart phone-based context-aware augmentative and alternative communications system,” *J. Cent. South Univ.*, vol. 21, no. 9, pp. 3551–3558, Sep. 2014.
[53] R. Y.-Y. Chan, E. Sato-Shimokawara, X. Bai, M. Yukiharu, S.-W. Kuo, and A. Chung, “A Context-Aware Augmentative and Alternative Communication System for School Children With Intellectual Disabilities,” *IEEE Syst. J.*, pp. 1–12, May 2019.
[54] S. Hossain, M. Takanokura, H. Sakai, and Hi. Katagiri, “Using Context History and Location in Context-aware AAC Systems for Speech-language Impairments,” in *Proceedings of the International MultiConference of Engineers and Computer Scientists* , 2018.
[55] B. Wisenburn and D. J. Higginbotham, “An AAC application using speaking partner speech recognition to automatically produce contextually relevant utterances: objective results.,” *Augment. Altern. Commun.*, vol. 24, no. 2, pp. 100–9, 2008.
[56] S. Kafle, C. O. Alm, and M. Huenerfauth, “Modeling Acoustic-Prosodic Cues for Word Importance Prediction in Spoken Dialogues,” Mar. 2019.
[57] J. Light, R. Page, J. Curran, and L. Pitkin, “Children’s ideas for the design of AAC assistive technologies for young children with complex comminucation needs,” *AAC Augment. Altern. Commun.*, vol. 23, no. 4, pp. 274–287, Dec. 2007.
[58] T. Huijbregts and J. R. Wallace, “Talkingtiles: Supporting personalization and customization in an AAC app for individuals with aphasia,” in *Proceedings of the 2015 ACM International Conference on Interactive Tabletops and Surfaces, ITS 2015*, 2015, pp. 63–72.
[59] K. O’Leary *et al.*, “Design goals for a system for enhancing aac with personalized video,” in *ASSETS’12 - Proceedings of the 14th International ACM SIGACCESS Conference on Computers and Accessibility*, 2012, pp. 223–224.
[60] M. Abadi *et al.*, “TensorFlow: Large-Scale Machine Learning on Heterogeneous Distributed Systems,” Mar. 2016.
[61] J. Konečný, H. B. McMahan, D. Ramage, and P. Richtárik, “Federated Optimization: Distributed Machine Learning for On-Device Intelligence,” *arXiv Prepr.*, Oct. 2016.
[62] C. Mims, “GPS Receivers Now Small Enough to Attach to Almost Anything,” *MIT Technology Review*, 2011.
[63] M. J. Heron, V. L. Hanson, and I. Ricketts, “Open Source and Accessibility: Advantages and Limitations,” *J. Interact. Sci.*, vol. 1, no. 1, p. 2, 2013.
[64] P. Korn, E. Bekiaris, and M. Gemou, “Towards open access accessibility everywhere: The ÆgIS concept,” in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 2009, vol. 5614 LNCS, no. PART 1, pp. 535–543.
[65] Coral, “Build intelligent ideas with our platform for local AI,” 2019. [Online]. Available: https://coral.withgoogle.com/. [Accessed: 15-Sep-2019].
[66] OpenBCI, “Open Source Brain-Computer Interfaces,” 2019. [Online]. Available: https://openbci.com/. [Accessed: 15-Sep-2019].
[67] S. Boucenna *et al.*, “Interactive Technologies for Autistic Children: A Review,” *Cognit. Comput.*, vol. 6, no. 4, pp. 722–740, Dec. 2014.
[68] J. A. Kientz, M. S. Goodwin, G. R. Hayes, and G. D. Abowd, “Interactive Technologies for Autism,” *Synth. Lect. Assist. Rehabil. Heal. Technol.*, vol. 2, no. 2, pp. 1–177, Nov. 2013.
[69] J. Daniels *et al.*, “Exploratory study examining the at-home feasibility of a wearable tool for

social-affective learning in children with autism," *npj Digit. Med.*, vol. 1, no. 1, Dec. 2018.
[70] M. J. Smith *et al.*, "Virtual reality job interview training in adults with autism spectrum disorder," *J. Autism Dev. Disord.*, vol. 44, no. 10, pp. 2450–2463, Oct. 2014.
[71] A. Begel *et al.*, "Lessons Learned in Designing AI for Autistic Adults," in *ASSETS 2020 - 22nd International ACM SIGACCESS Conference on Computers and Accessibility*, 2020.
[72] I. Cha, H. Hong, H. Yoo, and Y.-K. Lim, "Exploring the Use of a Voice-based Conversational Agent to Empower Adolescents with Autism Spectrum," in *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, 2021, pp. 1–15.
[73] 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 *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, 2020.
[74] 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 *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, p. 15.
[75] C. H. Chiang, W. T. Soong, T. L. Lin, and S. J. Rogers, "Nonverbal communication skills in young children with autism," *J. Autism Dev. Disord.*, vol. 38, no. 10, pp. 1898–1906, Nov. 2008.
[76] 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, Dec. 2010.
[77] H. Sowden, J. Clegg, and M. Perkins, "The development of co-speech gesture in the communication of children with autism spectrum disorders," *Clin. Linguist. Phonetics*, vol. 27, no. 12, pp. 922–939, 2013.
[78] 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 Dev. Disord.*, vol. 27, no. 6, 1997.
[79] R. G. Gordon and L. R. Watson, "Brief Report: Gestures in Children at Risk for Autism Spectrum Disorders," *J. Autism Dev. Disord.*, vol. 45, no. 7, pp. 2267–2273, Jul. 2015.
[80] 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, Oct. 2006.
[81] M. Gratier and E. Devouche, "Imitation and Repetition of Prosodic Contour in Vocal Interaction at 3 Months," *Dev. Psychol.*, vol. 47, no. 1, pp. 67–76, Jan. 2011.
[82] 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," *Dev. Sci.*, vol. 23, no. 1, p. e12843, Jan. 2020.
[83] J. McDaniel, K. D'Ambrose Slaboch, and P. Yoder, "A meta-analysis of the association between vocalizations and expressive language in children with autism spectrum disorder," *Research in Developmental Disabilities*, vol. 72. Elsevier Inc., pp. 202–213, 01-Jan-2018.
[84] L. Morgan and Y. E. Wren, "A Systematic Review of the Literature on Early Vocalizations and Babbling Patterns in Young Children," *Commun. Disord. Q.*, vol. 40, no. 1, pp. 3–14, Nov. 2018.
[85] S. R. Morris, "Clinical application of the mean babbling level and syllable structure level," *Lang. Speech. Hear. Serv. Sch.*, vol. 41, no. 2, pp. 223–230, Apr. 2010.
[86] 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, Apr. 2019.

[87] 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, Oct. 2012.
[88] D. K. Oller *et al.*, “Automated vocal analysis of naturalistic recordings from children with autism, language delay, and typical development,” *Proc. Natl. Acad. Sci.*, vol. 107, no. 30, pp. 13354–13359, Jul. 2010.
[89] 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, Aug. 2020.
[90] 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, Nov. 2019.
[91] L. Rescoria 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.
[92] 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, Jan. 2009.
[93] 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, Jul. 2010.
[94] D. A. Sauter, F. Eisner, A. J. Calder, and S. K. Scott, “Perceptual cues in nonverbal vocal expressions of emotion,” *Q. J. Exp. Psychol.*, vol. 63, no. 11, pp. 2251–2272, Nov. 2010.
[95] I. Poggi, A. Ansani, and C. Cecconi, “Sighs in everyday and political communication,” in *Laughter Workshop*, 2018.
[96] M. L. Knapp, *Essentials of nonverbal communication*. Holt, Rinehart and Winston, 1980.
[97] D. A. Sauter, F. Eisner, P. Ekman, and S. K. Scott, “Cross-cultural recognition of basic emotions through nonverbal emotional vocalizations,” *Proc. Natl. Acad. Sci. U. S. A.*, vol. 107, no. 6, pp. 2408–2412, Feb. 2010.
[98] A. Anikin and C. F. Lima, “Perceptual and acoustic differences between authentic and acted nonverbal emotional vocalizations,” *Q. J. Exp. Psychol.*, vol. 71, no. 3, pp. 622–641, Jan. 2018.
[99] D. A. Sauter, F. Eisner, A. J. Calder, and S. K. Scott, “Perceptual cues in nonverbal vocal expressions of emotion,” *Q. J. Exp. Psychol.*, vol. 63, no. 11, pp. 2251–2272, Nov. 2010.
[100] A. Anikin and T. Persson, “Nonlinguistic vocalizations from online amateur videos for emotion research: A validated corpus,” *Behav. Res. Methods*, vol. 49, no. 2, pp. 758–771, Apr. 2017.
[101] 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, p. 9663, Dec. 2021.
[102] A. Anikin, “A Moan of Pleasure Should Be Breathy: The Effect of Voice Quality on the Meaning of Human Nonverbal Vocalizations,” *Res. Artic. Phonetica*, vol. 77, pp. 327–349, 2020.
[103] O. Wasz-Höckert, T. J. Partanen, V. Vuorenkoski, K. Michelsson, and E. Valanne, “The identification of some specific meanings in infant vocalization,” *Experientia*, vol. 20, no. 3, p. 154, Mar. 1964.
[104] L. Liu, W. Li, X. Wu, and B. X. Zhou, “Infant cry language analysis and recognition: An experimental approach,” *IEEE/CAA J. Autom. Sin.*, 2019.
[105] I.-A. Banica, H. Cucu, A. Buzo, D. Burileanu, and C. Burileanu, “Automatic methods for infant cry classification,” in *2016 International Conference on Communications (COMM)*, 2016, pp. 51–54.
[106] S. Sharma and V. K. Mittal, “Infant cry analysis of cry signal segments towards identifying the cry-cause factors,” in *IEEE Region 10 Annual International Conference, Proceedings/TENCON*, 2017, vol. 2017-Decem, pp. 3105–3110.

[107] T. Fuhr, H. Reetz, and C. Wegener, “Comparison of Supervised-learning Models for Infant Cry Classification / Vergleich von Klassifikationsmodellen zur Säuglingsschreianalyse in: International Journal of Health Professions Volume 2 Issue 1 (2015),” *Int. J. Heal. Prof.*, vol. 2, no. 1, pp. 4–15, 2015.
[108] D. R. Beukelman and P. Mirenda, *Augmentative and alternative communication: Supporting children and adults with complex communication needs*, 4th ed. Baltimore, 2013.
[109] J. Prinz, “Which Emotions Are Basic?,” *Emot. Evol. Ration.*, vol. 69, 2004.
[110] R. Lotfian, S. Member, C. Busso, and S. Member, “Building Naturalistic Emotionally Balanced Speech Corpus by Retrieving Emotional Speech From Existing Podcast Recordings,” *J. IEEE Trans. Affect. Comput.*, vol. XX, 2016.
[111] B. W. Schuller, “Speech emotion recognition: Two decades in a nutshell, benchmarks, and ongoing trends,” *Communications of the ACM*, vol. 61, no. 5. Association for Computing Machinery, pp. 90–99, 01-May-2018.
[112] B. Schuller *et al.*, “The INTERSPEECH 2013 Computational Paralinguistics Challenge: Social Signals, Conflict, Emotion, Autism,” in *Proceedings INTERSPEECH 2013, 14th Annual Conference of the International Speech Communication Association*, 2013.
[113] F. Eyben *et al.*, “The Geneva Minimalistic Acoustic Parameter Set (GeMAPS) for Voice Research and Affective Computing,” *Trans. Affect. Comput.*, vol. 7, no. 2, pp. 190–202, 2016.
[114] F. Eyben, M. Wöllmer, and B. Schuller, “openSMILE-The Munich Versatile and Fast Open-Source Audio Feature Extractor,” in *Proceedings of the international conference on Multimedia - MM ’10*.
[115] F. Weninger, Y. Zhang, and R. W. Picard, “openXDATA: A Tool for Multi-Target Data Generation and Missing Label Completion,” *arXiv Prepr.*, 2020.
[116] S. Amiriparian, M. Freitag, N. Cummins, and B. Schuller, “Sequence to Sequence Autoencoders for Unsupervised Representation Learning from Audio,” in *Workshop on Detection and Classification of Acoustic Scenes and Events*, 2017.
[117] M. Gerczuk, S. Amiriparian, S. Ottl, B. Bj¨, and B. W. Schuller, “EmoNet: A Transfer Learning Framework for Multi-Corpus Speech Emotion Recognition,” *arXiv Prepr.*, 2021.
[118] A. Van Den Oord *et al.*, “WaveNet: A Generative Model for Raw Audio,” in *arXiv preprint*, 2016.
[119] S. Hershey *et al.*, “CNN Architectures for Large-Scale Audio Classification,” in *2017 IEEE Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2017, pp. 131–135.
[120] J. F. Gemmeke *et al.*, “Audio Set: An ontology and human-labeled dataset for audio events,” in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2017, pp. 776–780.
[121] S. Amiriparian *et al.*, “Snore Sound Classification Using Image-based Deep Spectrum Features,” in *INTERSPEECH 2017*, 2017, pp. 3512–3516.
[122] Y. Zhang, F. Weninger, S. Bjorn, and R. Picard, “Holistic Affect Recognition Using PaNDA: Paralinguistic Non-metric Dimensional Analysis,” *IEEE Trans. Affect. Comput.*, 2019.
[123] 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 *Extended Abstracts of the 2020 CHI Conference on Human Factors in Computing Systems*, p. *Equal contribution.
[124] 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 *NeurIPS 2019 Joint Workshop on AI for Social Good*, 2019, p. *Equal Contribution.
[125] J. Narain* and K. T. Johnson* *et al.*, “Personalized Modeling of Real-World Vocalizations

from Nonverbal Individuals," in *ICMI 2020 - Proceedings of the 2020 International Conference on Multimodal Interaction*, 2020, vol. 20, p. *Equal Contribution, 665-669.
[126] J. Narain*, K. Johnson*, T. Quatieri, R. Picard, and P. Maes, "ReCANVo: A Database of Real-World Communicative and Affective Nonverbal Vocalizations," *Under Rev.*, p. *Equal Contribution.
[127] J. Narain, K. Johnson, T. Quatieri, R. Picard, and P. Maes, "Modeling Real-World Affective and Communicative Nonverbal Vocalizations from Minimally Speaking Individuals," *Under Rev.*
[128] S. R. Livingstone and F. A. Russo, "The Ryerson audio-visual database of emotional speech and song (ravdess): A dynamic, multimodal set of facial and vocal expressions in north American english," *PLoS One*, vol. 13, no. 5, p. e0196391, May 2018.
[129] C. Busso *et al.*, "IEMOCAP: Interactive emotional dyadic motion capture database," *Lang. Resour. Eval.*, vol. 42, no. 4, pp. 335–359, Dec. 2008.
[130] A. Canavan, D. Graff, and G. Zipperlen, "CALLHOME American English Speech LDC97S42," *Philadelphia: Linguistic Data Consortium*, 1997. .
[131] F. Ringeval *et al.*, "Automatic analysis of typical and atypical encoding of spontaneous emotion in the voice of children," in *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, 2016, vol. 08-12-September-2016, pp. 1210–1214.
[132] S. Steidl, "Automatic classification of emotion related user states in spontaneous children's speech," University of Erlangen-Nuremberg, 2009.
[133] C. F. Lima, S. L. Castro, and S. K. Scott, "When voices get emotional: A corpus of nonverbal vocalizations for research on emotion processing," *Behav. Res. Methods*, vol. 45, no. 4, pp. 1234–1245, Dec. 2013.
[134] C. E. Parsons, K. S. Young, M. G. Craske, A. L. Stein, and M. L. Kringelbach, "Introducing the oxford vocal (OxVoc) sounds database: A validated set of non-acted affective sounds from human infants, adults, and domestic animals," *Front. Psychol.*, vol. 5, no. JUN, p. 562, Jun. 2014.
[135] M. D. Barokova, S. Hassan, C. Lee, M. Xu, and H. Tager-Flusberg, "A comparison of natural language samples collected from minimally and low-verbal children and adolescents with autism by parents and examiners," *J. Speech, Lang. Hear. Res.*, vol. 63, no. 12, pp. 4018–4028, Dec. 2020.
[136] V. Bernard-Opitz, "Pragmatic analysis of the communicative behavior of an autistic child," *J. Speech Hear. Disord.*, vol. 47, no. 1, pp. 99–109, 1982.
[137] G. Lemaître, F. Nogueira, and C. K. Aridas char, "Imbalanced-learn: A Python Toolbox to Tackle the Curse of Imbalanced Datasets in Machine Learning," *J. Mach. Learn. Res.*, vol. 18, pp. 559–563, 2017.
[138] 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. 173, pp. 1–5, 2018.
[139] M. Jaiswal and E. M. Provost, "Best Practices for Noise-Based Augmentation to Improve the Performance of Emotion Recognition 'In the Wild,'" *arXiv Prepr.*, Apr. 2021.
[140] X. Li *et al.*, "Universal Phone Recognition with a Multilingual Allophone System," *ICASSP, IEEE Int. Conf. Acoust. Speech Signal Process. - Proc.*, vol. 2020-May, pp. 8249–8253, Feb. 2020.
[141] J. Laguarta, F. Hueto, and B. Subirana, "COVID-19 Artificial Intelligence Diagnosis Using Only Cough Recordings," *IEEE Open J. Eng. Med. Biol.*, vol. 1, pp. 275–281, Sep. 2020.
[142] J. Narain, K. Johnson, T. Quatieri, R. Picard, and P. Maes, "Transfer Learning with Real-

World Nonverbal Vocalizations from Minimally Speaking Individuals," in *Interpretable Machine Learning in Health Care Workshop at ICML 2021*, 2021.
[143] J. Salamon, C. Jacoby, and J. P. Bello, "A Dataset and Taxonomy for Urban Sound Research," in *Proceedings of the 22nd ACM international conference on Multimedia*.
[144] H. He, K. Khoshelham, and C. Fraser, "A multiclass TrAdaBoost transfer learning algorithm for the classification of mobile lidar data," *ISPRS J. Photogramm. Remote Sens.*, vol. 166, pp. 118–127, Aug. 2020.
[145] J. Zhu, S. Rosset, H. Zou, and T. Hastie, "Multi-class AdaBoost," *Stat. Its Inference*, vol. 2, no. 3, pp. 349–360, 2006.
[146] R. Fernandez, "A Computational Model for the Automatic Recognition of Affect in Speech," Massachusetts Institute of Technology, 2004.
[147] D. Roy *et al.*, "The Human Speechome Project Stepping into the Shoes of Children," in *International workshop on emergence and evolution of linguistic communication*, 2006, pp. 192–196.
[148] "BirdNET – The easiest way to identify birds by sound." [Online]. Available: https://birdnet.cornell.edu/. [Accessed: 18-Jul-2021].
[149] D. D. Mehta, D. Rudoy, and P. J. Wolfe, "Kalman-based autoregressive moving average modeling and inference for formant and antiformant tracking," *J. Acoust. Soc. Am.*, vol. 132, no. 3, pp. 1732–1746, Sep. 2012.
[150] Y.-L. Shue, "The Voice Source in Speech Production: Data, Analysis and Models," University of California, Los Angeles, 2010.

Appendix A

1. User Needs Survey

1. How old are you in years? (Please enter your age here, even if you are filling out this form on behalf of someone else)

If the person is a minor over 13, verify that the respondent has permission from a parent/guardian to participate. If the person is a minor under 13, the survey exits

2. Are you filling out this form on behalf of someone else?
- Yes
- No

NOTE: If the person answered "Yes" to Q2, then questions were displayed with the text "your loved one" instead of "you" (e.g., for Q4:
"What condition does your loved one have"?). =

If response to Q2 was "Yes"
3. What is your relation to the person you are filling out this form on behalf of?
- Parent
- Child
- Sibling
- Professional caregiver
- Other _____

4. What condition do you have?
- Amyotrophic lateral sclerosis (ALS)
- Cerebral Palsy (CP)
- Stroke
- Autism spectrum disorder (ASD)
- Multiple sclerosis (MS)
- Spinal cord injury
- Other_____

If response to Q2 was "Yes"
5. How old is your loved one in years?

6. Do you have a intellectual disability?
- Yes

- No
- Other response _____

If response to Q4 was "Autism spectrum disorder (ASD)"
7. What activities do you find most challenging as a person living with ASD?

If response to Q4 was "Amyotrophic lateral sclerosis (ALS)"
8. When were you diagnosed with ALS?

If response to Q4 was "Multiple sclerosis (MS)"
9. When were you diagnosed with MS?

If response to Q4 was "Stroke"
10. When did you have a stroke?

11. Is there any other information about your condition you would like to share? (e.g., sub-condition type, what activities are affected most by the condition)

12. Do you experience spasticity, tics, or sudden jerking motions?
- Yes
- No

If response to Q12 was "Yes"
13. How often do you experience spasticity, tics, or sudden jerking motions?
- Every few minutes
- Every hour or so
- Daily
- Weekly or monthly
- Other _____

If response to Q12 was "Yes"
14. Where do you experience spasticity, tics, or sudden jerking motions? You can select multiple options.
- Eye gaze
- Facial muscles
- Head and/or neck
- Hands
- Arms
- Legs
- Trunk
- Other_____

15. Do your speech and/or motor abilities change with time?
- Abilities improve
- No change
- Abilities worsen

If response to Q15 is not "No change"

16. How frequently do you notice changes in your speech and/or motor abilities?
    - Daily
    - Weekly
    - Monthly
    - Every few months
    - Yearly
    - Other _____

If response to Q15 is not "No change"

What types of changes in speech/motor abilities do you have?

17. Which of the following better describes your speech abilities?
    - Used to be able to speak, but lost speech ability
    - Have never been able to speak

18. Which of the following applies to your verbal speech? You can select multiple options.
    - No auditory speech
    - Can speak in short phrases, not complete sentences (e.g., "milk: instead of "I want some milk")
    - Can speak in sentences with full clarity
    - Can speak in sentences, with reduced clarity (still understandable by most people)
    - Can speak in sentences, with reduced clarity (only understandable by close friends and family)
    - Other _____

19. Which of the following applies to your communication practices? You can select multiple options.
    - Has more difficulty communicating in unfamiliar settings
    - Hand gestures have consistent meanings when communicating
    - Non-word vocalized sounds have consistent meanings when communicating
    - Primarily communicates through crying/yelling when unhappy
    - Assistive device used to aid communication
    - Communication speed is significantly reduced
    - Communication is often misunderstood
    - Other _____

20. Would you like to provide any more information on how you communicate (optional)?

21. Can you do the following tasks?
    - Walk
    - Talk (to friends and family)
    - Talk (to strangers)
    - Write
    - Type (physical keyboard)
    - Mouth words (without producing sound)

Answer yes or no questions

- Make hand gestures
- Make facial expressions
- Express emotions and/or feelings
- Push physical buttons
- Push touch screen buttons
- Use a computer or tablet to send text or email
- Make a purchase at a store
- Ask for food when hungry
- Let people know when in pain
- Express need for sleep

Select from:
- Can do independently with no difficulty
- Can do independently with some difficulty
- Can fully do with help or prompting (e.g., from a device or person)
- Can somewhat do with help or prompting (e.g., from a person or device)
- Cannot complete task

22 If you would like, please use this space to provide more information about how you do the tasks above (optional).

23. What communication related tasks do you find most difficult or frustrating?

24. What, if any, devices do you use to help with communication? You can select more than option. If you do not use a device, please select "Other" and write that you don't use a device in the associated text box.

If you use a specific product, please list the product name (if applicable).
- Tobii
- Other gaze tracking device _____
- Head motion tracking device _____
- Proloquo2Go app
- Other tablet app_____
- Non-electronic word board
- Physical keyboard _____
- Stylus/pen + writing surface_____
- Buttons that produce audio messages (e.g., GoTalk 9+, BigTalk) _____
- Other_____
- Other_____
- Other_____

For each item selected in Q24:
[Q25-Q49]
What difficulties do you have with [device name from response]?
What do you like about [device name for response]?

E.g., if "Tobii" was selected the questions displayed would be

What difficulties do you have with Tobii?
What do you like about Tobii?

50. Use this space to provide any other information you would like to share about the assistive devices you use (optional).

51. Would you be interested in using a device with the following outputs?

- Computer generated speech
- Written text
- Private auditory prompting to device user (e.g., device suggests a specific word or phrase)
- Controlling a tablet or computer
- Controlling a mobile phone
- Calling a caregiver
- Controlling the lights in a room
- Controlling a television
- Information about user's physical state (e.g., pain, emotions, hunger, tiredness)

Select from:

- Extremely interested
- Very interested
- Moderately interested
- Slightly Interested
- Not at all interested

If selection for interested in information about a user's physical state was "Extremely interested", "Very interested", or "Moderately interested"
52. If a device had the capability to send physical state information to designated people, what information output would be useful for you?

- Hunger
- Presence of pain
- Location of pain
- Tiredness
- Need to urinate
- Need to defecate
- Emotions (e.g., sadness)
- Excitedness
- Stress

53. Are there any device outputs not listed above you would find useful?

54. How would you rate the difficulty of the following input mechanisms for an assistive device for communication?

- Speech
- Mouthing words (without producing sound)
- Pushing tactile buttons
- Pushing touch screen buttons
- Moving facial muscles
- Moving fingers

- Moving eye muscles
- Moving wrist
- Moving foot

Select from:
- Not difficult at all
- Slightly difficult
- Moderately difficult
- Very difficult
- Cannot do

55. Are there any device input mechanisms not listed above you?

56. How important are the following characteristics for an assistive device to augment communication for you?
- Input speed
- Output accuracy
- Device is wearable
- No tablet needed
- Can be used outside
- Does not take long to train device
- Aesthetics
- Device is robust to being dropped and treated roughly
- Device does not take long to set up
- Data privacy

Select from:
- Extremely important
- Very important
- Moderately important
- Slightly important
- Not at all important

57. Please rate the likelihood you would be willing to wear an assistive device with the displayed form factors:

(Worn under clothes)

Select from:
- Very likely
- Moderately likely
- Neither likely nor unlikely
- Moderately unlike
- Very unlikely

58. Consider the following potential device architectures:

Trained Speech Generation: Imagine a wearable headset-like device that can generate speech and text. The user trains the device to recognize the intent of facial muscle movements. When he/she makes a facial motion the device recognizes, it outputs the corresponding word as auditory speech, or as text input to a device (e.g., a phone or tablet). The user can choose to only have the device output messages after they have been checked and approved. The device can only be used to output messages it has been trained for.

Context-based Speech Generation: Imagine a goggle-like device. The user trains the device to recognize certain facial muscles, and the device has a camera, microphone, and other sensors to help interpret context. The device can process the speech of other people in a conversation, and understand basic information about the user's setting (e.g., restaurant, home). The device uses the context to suggest words or phrases to the user, who can accept, deny, or edit the suggestions. When the user approves a message, it is output as auditory speech or as text input to a device (e.g., a phone or tablet).

Speech Prompting: Imagine a headset-like device that can prompt the wearer in conversations. The device can recognize simple speech phrases, and also has a camera and microphone to interpret the context. For instance if he/she says the word "sandwich" and they are in a kitchen, the device will ask them easy yes/no questions to better understand their goals (e.g., Do you want a peanut butter sandwich? Do you want a sandwich now?). The device will process the given information and will prompt them to say a more complex thought, like "Can you please make me a ham sandwich now?")

Physical State Messaging: Imagine a glove or armband that records basic physiological data that can be correlated to physical states like pain, tiredness, excitement, or hunger. When the glove a change in physical state, it can help the wearer communicate how they may be feeling (e.g., excitement, sadness) and/or provide information on the wearer's physical state (e.g., pain or hunger). The device can be programmed to communicate with pre-designated family members or carers and/or with individuals in the wearer's vicinity. If desired, the device can be set-up to share information only with the wearer's permission.

59. How helpful would each of the devices described above be for you?
- Trained Speech Generation
- Context-based Speech Generation
- Speech Prompting
- Physical State Messaging

Select from:
- Extremely helpful
- Very helpful
- Moderately helpful
- Slightly helpful
- Not at all helpful

60. Do you have any other comments or suggestions?

Appendix B

1. Feature Extraction Implementation Details

1. Custom Feature Set

MATLAB 2020b was used to extract the custom feature set. The signal autocorrelation was calculated using the built-in *xcorr* function. The power was calculated as the normalized square of the discrete Fourier transform (computed using the built-in function *fft* applied to the signal.

Formants were extracted using KARMA [149]. Sustained values were defined as regions of sufficient length and amplitude with derivatives under a specified value. The thresholds were determined heuristically; to register as a constant formant section a segment had a minimum length of 0.2s, a minimum amplitude of 0.008, and a maximum numerical derivative of 4000 Hz/s. The large threshold for the maximum derivative was due to amplified noisiness in the numerical derivative with a small step size. The amplitude threshold was included to prevent erroneously tracking formants in periods of silence.

The pitch was calculated using the built-in *pitch* function with a 0.05s Hamming window and 0.025s overlap. The built-in functions *findpeaks* and *polyfit* were used to extract the number of peaks and polynomial fit values, respectively. The overall rise/fall was calculated as the difference in pitch between the first 0.05s of the vocalization and the last 0.05s of the vocalization.

Gammatone cepstral coefficients (GTCC) were extracted using the built-in *gtcc* function with a 0.05s periodic Hanning window with 0.025s overlap. Bark frequency cepstral coefficients were extracted by first creating a bark frequency scale filter bank (using the built-in *designAuditoryFilterBank* function) and then applying the filter bank to a short-time Fourier transform (using the built-in *spectrogram* function) applied to the vocalization with a 0.05s Hamming window and 0.025s overlap.

The cepstral peak prominence and harmonics were calculated using MATLAB functions provided in VoiceSauce [150].

2. Cepstral Coefficients

Cepstral coefficients were extracted in MATLAB 2020b. The GTCC and MFCC were extracted using the built-in *gtcc* and *mfcc* functions respectively with a 0.05s periodic Hanning window with 0.025s overlap.

3. Filter Bank Features

Filter bank derived features were extracted in MATLAB 2020b. The filter banks were defined using the built-in *designAuditoryFilterBank* function with mel and ERB frequency scales. The filter banks were then applied to the absolute value of the STFT of the signal (extracted using the built-in *spectrogram* function with a 0.05s Hanning window with 0.025s overlap).

2. Selected Model Architectures for auDeep Feature Extraction

	Segment length (s)	Batch size	Learning rate	Number of layers	Number of units
P01	9.04	64	0.00098	5	256
P02	1.86	128	0.00063	2	256
P03	3.23	64	0.00052	4	256
P05	8.08	64	0.00094	4	128
P06	0.77	64	0.00077	3	256
P08	4.84	32	0.00042	5	256
P11	2.7	64	0.00064	1	256
P16	7.11	64	0.00055	3	256