Transfer Learning with Real-World Nonverbal Vocalizations from Minimally Speaking Individuals

Jaya Narain [1] **Kristina Johnson** [1] **Rosalind Picard** [1] **Thomas Quatieri** [2 3] **Pattie Maes** [1]

Abstract

We applied transfer learning to classify the affect and communication intent of nonverbal vocalizations from eight minimally speaking individuals (mv*) with autism. Data was recorded in real-world settings with in-the-moment labels from a close family member. We trained deep neural nets (DNNs) on six audio datasets (including our dataset of nonverbal vocalizations) and then used the learned weights to initialize networks to train personalized models for each individual. We also evaluated a zero-shot approach for arousal and valence regression using an acted dataset of nonverbal vocalizations that occur amidst typical speech. Transfer learning improved model performance and there were weak groupings in arousal values inferred using zero-shot learning for two of the eight mv* communicators. The limited success of the evaluated approaches highlights the need for specialized datasets with mv* individuals.

1. Background and Motivation

[1] A major challenge in machine learning in health care is having sufficient data to train models that work for individuals in heterogeneous, specialized populations. Understanding if and when datasets created with the general population can be leveraged for model training can help target data collection efforts. We present an exploration of transfer learning with deep neural nets (DNNs) to classify the affect and communicative intent nonverbal vocalizations from individuals who are non- or minimally speaking with respect to verbal language. Here we focus on a sub-group of non- and minimally speaking individuals (designated by mv*) who have ten or fewer spoken words/word approximations and who have limited expressive language through speech, written word, or available AAC devices. There are over one million mv* individuals in the United States alone (CDC, 2020; Bacon et al., 2019), including individuals with Autism, Down Syndrome (DS), Cerebral Palsy (CP), and other diagnoses.

Mv* individuals communicate richly through many means, including nonverbal vocalizations which are used to express emotions, to participate in social exchanges, and to communicate wants and needs. Understanding and responding appropriately to this unique type of communication is important in both clinical and social settings. For example, a clinician who recognizes a nonverbal vocalization as expressing dysregulation (a state of general stress often related to under- or over-stimulation) could adjust their actions to prevent exacerbating stress and related negative health effects.

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. Additionally, transfer learning can be used to probe similarities in expressions among nonverbal vocalizations mv* communicators and between mv* communicators and other populations.

2. Related Work

A large body of previous work with mv* individuals has focused on using machine learning with physiological data for affect detection (Picard, 2009; Kushki et al., 2014; Hyde et al., 2019). While there is a large body of work characterizing nonverbal vocalizations that occur amidst typical

[1]MIT Media Lab, Massachusetts Institute of Technology, Cambridge, MA, USA [2]MIT Lincoln Laboratory, Lexington, MA, USA [3]Speech and Hearing Bioscience and Technology, Harvard University, Boston, MA, USA. Correspondence to: Jaya Narain <jnarain@mit.edu>.

Proceedings of the 38^{th} International Conference on Machine Learning, PMLR 139, 2021.

[1]DISTRIBUTION STATEMENT A. Approved for public release. Distribution is unlimited.

This material is based upon work supported by the Under Secretary of Defense for Research and Engineering under Air Force Contract No. FA8702-15-D-0001. Any opinions, findings, conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the Under Secretary of Defense for Research and Engineering.

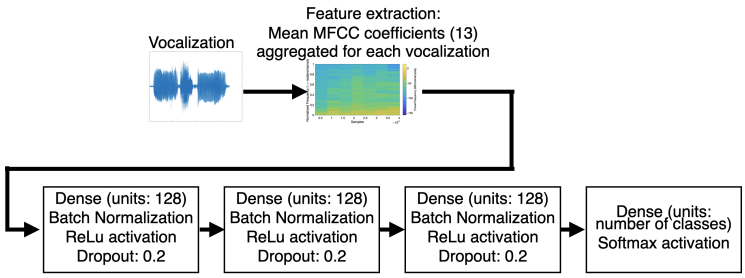


Figure 1. Model architecture used for transfer learning experiments.

verbal speech (e.g, sighs, laughter, grunts) (Sauter et al., 2010; Anikin & Persson, 2017; Holz et al., 2021) and on classifying infant vocalizations (Liu et al., 2019; Fuhr et al., 2015), prior work on nonverbal vocalizations with mv* individuals has focused on studying vocalizations as diagnostic markers for autism other developmental differences (Oller et al., 2010; Rescorla & Ratner, 1996; Martin et al., 2009) and not on vocalization function. There is extensive prior work in speech emotion recognition (SER) (Schuller, 2018) and on model development for clinical diagnostic tasks like depression (Cummins et al., 2011) with typical verbal speech, including transfer learning for affective and diagnostic classification tasks (Gerczuk et al., 2021; Laguarta et al., 2020). To our knowledge, no prior work has explored transfer learning for models of affect and intent in nonverbal vocalizations from mv* individuals.

3. Data

3.1. Data Collection and Pre-processing

The study was approved by an institutional review board (IRB). Data was collected with eight mv* communicators: P01 (M, age 18-25, diagnosis: autism, DS); P02 (M, age 18-25, diagnosis: autism); P03 (M, age 6-9, diagnosis: autism, genetic disorder); P05 (F, age 9-12, diagnosis: autism); P06 (M, age 9-12, diagnosis: autism, CP); P08 (F, age 6-9, diagnosis: autism); P11 (M, age 9-12, diagnosis: CP); P16 (M, age 6-9, diagnosis: autism). From a parent report, P01, P03, P05, and P08 had no spoken words/word approximations; P02 had 4: P05 had 3; P11 had 1; and P16 had 5-8.

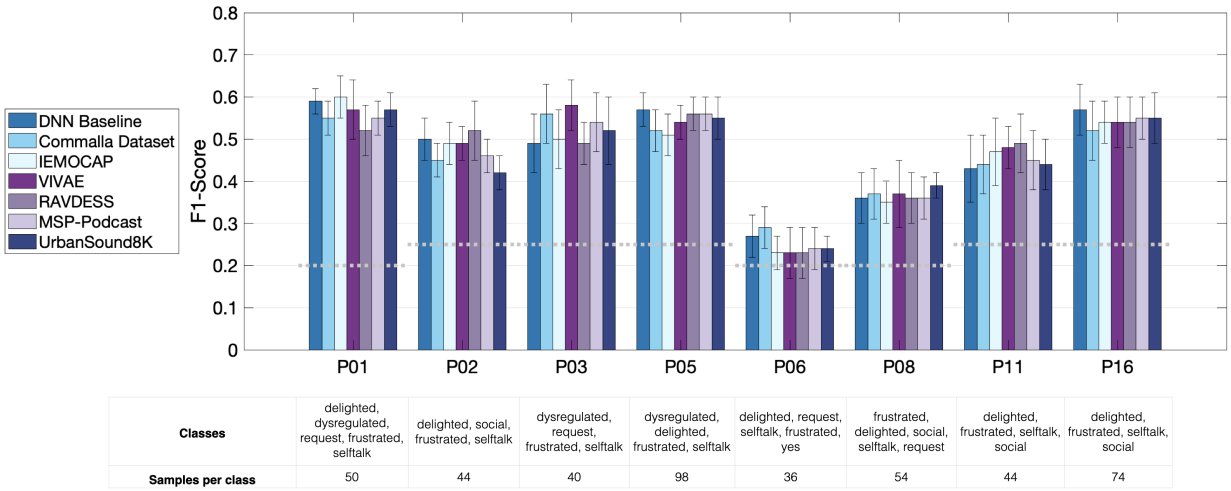
Families collected and labeled data during their day-to-day life in real-world settings, often in their homes. The study was designed to give participants flexibility in setting the pace, settings, and timeline for data collection, which was critical for working with an overburdened specialized population. The remote nature of the study allowed for data collection even during COVID-19. Audio was recorded using a Sony IDC-TX800 recorder in 16 bit, 44.1 kHz stereo. The lightweight recorder was attached with lapel magnets to the mv* communicator's clothing or placed nearby. A close family member (the "labeler") used a custom app (see Supplementary Material) to label vocalizations in real-time. The app had six labels common from all users (selftalk, dysregulated, frustrated, request, delighted, and social), selected based on interviews with families and a speech and language pathologist. Label descriptions were provided to families (see Supplementary Material). For example, selftalk is positive affect vocalizations like babbling and dysregulated is a state of general stress including over- and under-stimulation. Each family could also customize four labels from a provided list of 25 preset options spanning a range of affect and communicative functions.

The labels and audio recordings were timestamped and synced post-hoc. A volume-based filter was used to isolate audio segments of interest, enabled by the placement of the recorder close to the mv* communicator. Segments were matched with temporally adjacent labels, account for human delay and small clock shifts. A researcher listened to each segment to confirm it contained a vocalization from the mv* communicator and, if needed, trim any noise surrounding the vocalization.

3.2. Transfer Learning Datasets

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

- **Nonverbal vocalizations from mv* individuals**: the dataset collected as described in 3.1
- **IEMOCAP**, Interactive Emotional Dyadic Motion Capture (Busso et al., 2008): sentences from dyadic scripted and impromptu acted sessions
- **RAVDESS**, the Ryerson Audio-Visual Database of Emotional Speech and Sound (Livingstone & Russo, 2018): speech from scripted professional actors
- **VIVAE** full set, Variably Intense Vocalizations of of Affect and Emotion Corpus (Holz et al., 2021): acted



Classes	delighted, dysregulated, request, frustrated, selftalk	delighted, social, frustrated, selftalk	dysregulated, request, frustrated, selftalk	dysregulated, delighted, frustrated, selftalk	delighted, request, selftalk, frustrated, yes	frustrated, delighted, social, selftalk, request	delighted, frustrated, selftalk, social	delighted, frustrated, selftalk, social
Samples per class	50	44	40	98	36	54	44	74

Figure 2. Results of transfer learning experiments with personalized neural networks.

corpus of nonverbal vocalizations that occur amidst typical verbal speech (e.g., grunts, screams)

- **MSP-Podcast** Dataset (Lotfian & Busso, 2017): speech from podcasts annotated using crowdsourcing
- **Urban Sound 8K** (Salamon et al., 2014): annotated recordings of non-speech environmental sounds from Freesound

IEMOCAP and RAVDESS were selected because they are commonly used in SER 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.

4. Methods

4.1. Transfer Learning with Personalized Neural Networks

Personalized models were trained for each participant. Models were personalized because, to our knowledge, there is no prior work suggesting commonality in nonverbal vocal expressions among mv* individuals because the participants encompassed diverse ages, genders, and diagnoses. Classes were included in the model if there were least 30 samples spread across three distinct sessions (see Fig. 2). A session was a single recording file uploaded by participant; sessions generally were time-separated and had distinct background sounds. To prevent models fitting to background sounds for classes with many samples from a single session, the maximum number of training samples per class per participant was limited to 10. Classes were balanced to the minimum of the largest class size and twice the smallest class size using random undersampling and the synthetic minority oversampling technique (SMOTE) (Lemaître et al., 2017).

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 were trained for each participant (Fig. 1). Other model architectures, including long-short term memory (LSTM) recurrent neural nets (RNNs), the EmoNet architecture (based on ResNet) (Gerczuk et al., 2021), and smaller convolutional neural nets (CNNs) were also evaluated but tended to overfit the data. The selected architecture had the most robust performance across participants.

Models were trained on each dataset using an 80/20 training/validation data split (see Supplementary Material for base model class sizes and performances on validation data). With our 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, model weights were initialized using the learned weights of models trained on each dataset. The last layer was removed and replaced with a new softmax layer for the target participant's data. Baseline models without transfer learning were also trained for each participant. Because the amount of training data for each participant was small, models were evaluated using 5-fold cross validation with three distinct random seeds. The mean unweighted F1 score and the unweighted average recall (UAR) with 95% confidence intervals are reported.

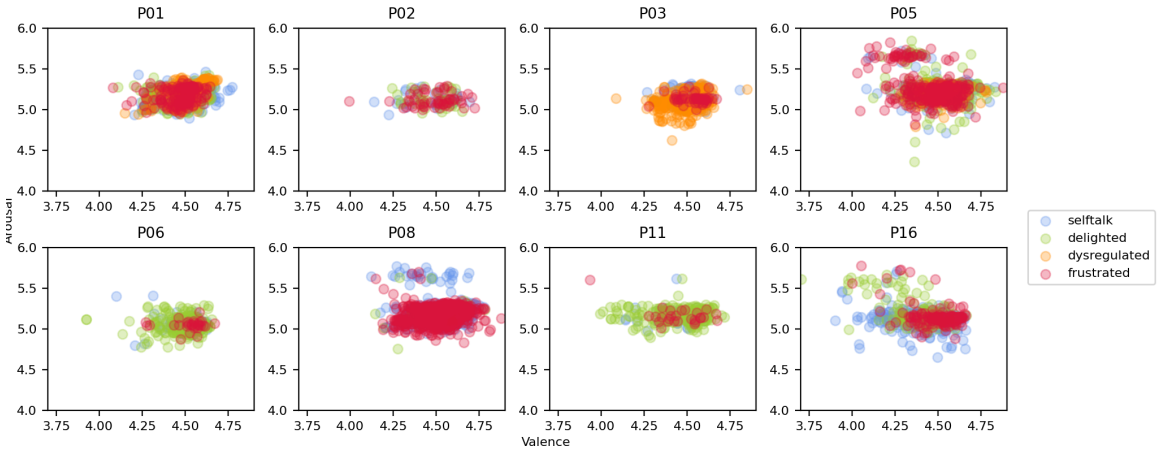


Figure 3. Valence and arousal ratings for each vocalization were inferred using zero shot transfer learning with a Random Forest regressor. Plot colors show the label of each vocalization, to visualize relations between predictions and labels.

4.2. Zero-Shot Learning for Valence and Arousal

The extended Geneva minimalistic acoustic parameter set (eGeMAPs) (Eyben et al., 2015) was extracted for each vocalization in the VIVAE core dataset (Holz et al., 2021). 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 value 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 vocalizations from mv* individuals with affective labels (Fig. 3).

5. Results and Discussion

5.1. Transfer Learning with Personalized Neural Networks

Results from the transfer learning experiments are shown in Fig. 2. Transfer learning consistently improved the F1 score models for P03 and P11, with the largest improvements with the VIVAE dataset. There were small improvements in the F1 score for P01, P02, P06, and P08 with the our dataset and the IEMOCAP, RAVDESS, and UrbanSound datasets. Transfer learning often did not change or slightly reduced the overall F1 score for other participants and base datasets.

5.2. Zero-Shot Learning for Valence and Arousal

Zero-shot learning was used to explore whether valence and arousal ratings inferred (Fig. 3) 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 and delighted than dysregulated and 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 did not clearly relate to the affective labels.

6. Conclusions and Future Work

Transfer learning with personalized neural networks had the greatest positive effect for P03 and P11, who had relatively small training class sizes. For other participants, effects were small or negative, particularly for participants with larger amounts of available training data like P05 and P16. There were weak visible relations between arousal predictions and affective label characteristics with the zero-shot modeling approach for two participants.

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 admist 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 that would be helpful in both clinical settings and for communication in day-to-day life.

References

Anikin, A. and Persson, T. Nonlinguistic vocalizations from online amateur videos for emotion research: A validated corpus. *Behavior research methods*, 49(2):758–771, 2017.

Bacon, E. C., Osuna, S., Courchesne, E., and Pierce, K. Naturalistic language sampling to characterize the language abilities of 3-year-olds with autism spectrum disorder. *Autism*, 23(3):699–712, 2019.

Busso, C., Bulut, M., Lee, C.-C., Kazemzadeh, A., Mower, E., Kim, S., Chang, J. N., Lee, S., and Narayanan, S. S. Iemocap: Interactive emotional dyadic motion capture database. *Language resources and evaluation*, 42(4): 335–359, 2008.

CDC. Key findings: CDC releases first estimates of the number of adults living with Autism Spectrum Disorder in the United States, 2020.

Cummins, N., Epps, J., Breakspear, M., and Goecke, R. An investigation of depressed speech detection: Features and normalization. In *Twelfth Annual Conference of the International Speech Communication Association*, 2011.

Eyben, F., Scherer, K. R., Schuller, B. W., Sundberg, J., André, E., Busso, C., Devillers, L. Y., Epps, J., Laukka, P., Narayanan, S. S., et al. The Geneva minimalistic acoustic parameter set (GeMAPS) for voice research and affective computing. *IEEE transactions on affective computing*, 7 (2):190–202, 2015.

Fuhr, T., Reetz, H., and Wegener, C. Comparison of supervised-learning models for infant cry classification/vergleich von klassifikationsmodellen zur säuglingsschreianalyse. *International Journal of Health Professions*, 2(1):4–15, 2015.

Gerczuk, M., Amiriparian, S., Ottl, S., and Schuller, B. Emonet: A transfer learning framework for multi-corpus speech emotion recognition. *arXiv preprint arXiv:2103.08310*, 2021.

Holz, N., Larrouy-Maestri, P., and Poeppel, D. The paradoxical role of emotional intensity in the perception of vocal affect. *Scientific Reports*, 11(1):1–10, 2021.

Hyde, K. K., Novack, M. N., LaHaye, N., Parlett-Pelleriti, C., Anden, R., Dixon, D. R., and Linstead, E. Applications of supervised machine learning in autism spectrum disorder research: a review. *Review Journal of Autism and Developmental Disorders*, 6(2):128–146, 2019.

Kushki, A., Khan, A., Brian, J., and Anagnostou, E. A kalman filtering framework for physiological detection of anxiety-related arousal in children with autism spectrum disorder. *IEEE Transactions on Biomedical Engineering*, 62(3):990–1000, 2014.

Laguarta, J., Hueto, F., and Subirana, B. Covid-19 artificial intelligence diagnosis using only cough recordings. *IEEE Open Journal of Engineering in Medicine and Biology*, 1: 275–281, 2020.

Langley, P. Crafting papers on machine learning. In Langley, P. (ed.), *Proceedings of the 17th International Conference on Machine Learning (ICML 2000)*, pp. 1207–1216, Stanford, CA, 2000. Morgan Kaufmann.

Lemaître, G., Nogueira, F., and Aridas, C. K. Imbalanced-learn: A python toolbox to tackle the curse of imbalanced datasets in machine learning. *The Journal of Machine Learning Research*, 18(1):559–563, 2017.

Liu, L., Li, W., Wu, X., and Zhou, B. X. Infant cry language analysis and recognition: an experimental approach. *IEEE/CAA Journal of Automatica Sinica*, 6(3): 778–788, 2019.

Livingstone, S. R. and Russo, F. A. The ryerson audio-visual database of emotional speech and song (ravdess): A dynamic, multimodal set of facial and vocal expressions in north american english. *PloS one*, 13(5):e0196391, 2018.

Lotfian, R. and Busso, C. Building naturalistic emotionally balanced speech corpus by retrieving emotional speech from existing podcast recordings. *IEEE Transactions on Affective Computing*, 10(4):471–483, 2017.

Martin, G. E., Klusek, J., Estigarribia, B., and Roberts, J. E. Language characteristics of individuals with Down syndrome. *Topics in language disorders*, 29(2):112, 2009.

Oller, D. K., Niyogi, P., Gray, S., Richards, J. A., Gilkerson, J., Xu, D., Yapanel, U., and Warren, S. F. Automated vocal analysis of naturalistic recordings from children with autism, language delay, and typical development. *Proceedings of the National Academy of Sciences*, 107 (30):13354–13359, 2010.

Picard, R. W. Future affective technology for autism and emotion communication. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 364(1535): 3575–3584, 2009.

Rescorla, L. and Ratner, N. B. Phonetic profiles of toddlers with specific expressive language impairment (SLI-E). *Journal of Speech, Language, and Hearing Research*, 39 (1):153–165, 1996.

Salamon, J., Jacoby, C., and Bello, J. P. A dataset and taxonomy for urban sound research. In *Proceedings of the 22nd ACM international conference on Multimedia*, pp. 1041–1044, 2014.

Sauter, D. A., Eisner, F., Calder, A. J., and Scott, S. K. Perceptual cues in nonverbal vocal expressions of emotion. *Quarterly Journal of Experimental Psychology*, 63(11): 2251–2272, 2010.

Schuller, B. W. Speech emotion recognition: Two decades in a nutshell, benchmarks, and ongoing trends. *Communications of the ACM*, 61(5):90–99, 2018.