Zero-Shot Transfer Learning to Enhance Communication for Minimally Verbal Individuals with Autism using Naturalistic Data

Jaya Narain* and Kristina T. Johnson* (*Equal Contribution)
Rosalind Picard
Pattie Maes
MIT Media Lab
Cambridge MA, 02139
{jnarain, ktj, picard, pattie}@media.mit.edu

Abstract

We applied zero-shot transfer learning to classify vocalizations from a nonverbal individual with autism using captured audio. Data were recorded in natural environments using a small wearable camera and sparsely labeled in real-time with a custom-built open-source app. We then trained LSTM models on VGGish audio embeddings from the generic AudioSet database for three categories of vocalizations: laughter, negative affect, and self-soothing sounds. We applied these models to the unique audio recordings of a young autistic boy with no spoken words. The models identified laughter and negative affect with 70% and 69% accuracy, respectively, but classification of the self-soothing sounds produced accuracies around chance. This work highlights both the need and potential for specialized, naturalistic databases and novel computational methods to enhance translational communication technologies in underserved populations.

1 Background and motivation

A major challenge in deep learning is generalization, particularly for small, noisy datasets from uncurated, naturalistic domains. To build machine learning-based technology for these realistic scenarios, we need strategies that leverage existing well-validated deep learning datasets. In this paper, we introduce a novel dataset of sparsely labeled naturalistic vocalizations from a minimally verbal (mv) individual with autism spectrum disorder (ASD) that includes over 13 hours of audio. To the authors' knowledge, this is the first dataset of its kind. We designed an approach to collect this dataset, including in-the-moment labeling to denote affective states and communicate intent using a custom open-source mobile app. We then employed a zero-shot transfer learning approach [27] to adapt an LSTM model trained on subclasses of a large, generic audio database to classify an autistic[1] child's non-speech vocalizations.

There are over 800,000 people in the United States with nonverbal or minimally verbal ASD, meaning they use zero or fewer than 20 words/word approximations, respectively [3, 2]. Previous applications of machine learning in ASD populations have focused on diagnosing ASD in naturalistic and laboratory settings [34, 23, 16, 8] or detecting a single emotional valence in laboratory or outpatient care settings using physiology [13, 14, 19, 18, 21]. Most approaches have relied on labels by professionals, like researchers and therapists. Prior work in classifying non-speech vocalizations has focused on classifying typically developing infant cries by need (e.g. hunger, pain) using both humans and machines [35, 20, 25, 31, 5, 11]. Infant cries have also been used to diagnose ASD

[1] Both "person-first"[1] and "identity-first" [6] language will be utilized interchangeably in this work.

33rd Conference on Neural Information Processing Systems (NeurIPS 2019), Vancouver, Canada.

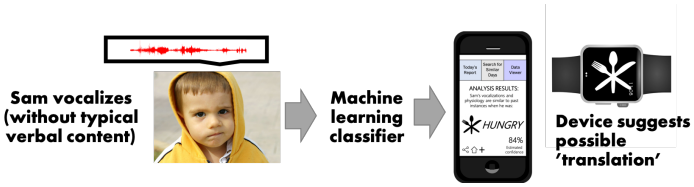

Figure 1: In the proposed platform, audio acquired in real-time is processed through a pre-trained LSTM model and "translated" into shareable communicative content.

[32]. There has been extensive prior work on affect detection in speech with typical verbal content [29, 30, 10, 28, 24, 7], including in ASD populations with verbal abilities in task-driven [22, 4] and natural settings [26]. However, no known work to date has attempted to classify communicative content in vocalizations from non-infant children or adults who are minimally or non-verbal.

The presented work was motivated by a real, vetted need. We conducted interviews and surveys with over 75 families and individuals with who had speech and language challenges and found that miscommunication was reported as a major source of stress. Respondents with ASD also noted that existing communication augmentation devices were difficult to use or did not sufficiently capture affect or complex communicative intent. For these individuals, affect or intent was often conveyed through non-traditional vocal communication such as hums, consonant utterances, babbling sounds. These utterances could have specific known meanings, such as "buhbuh" for "go to the playground," while others may have less clear mappings, like high-pitched yells to indicate general frustration. The parents of minimally verbal autistic children reported that they understood their offspring's communicative intent significantly better than others who interacted with their child, like teachers and babysitters. Hence, a machine that could learn from or with an expert caregiver could enable visiting caregivers to understand the individual more effectively, which is the motivation for this work (see Figure 1). Given the high heterogeneity of the mvASD population, it was necessary to begin with a deep case study.

2 Naturalistic dataset: collection and characterization

Data collection methods Data were collected with a non-speaking autistic boy of elementary school age. After providing informed consent/assent, the participant wore a t-shirt with an inexpensive mini-camera in the front pocket. Single channel audio was recorded at 32 kHz and 16 bits per second. The participant and his family were asked to continue their regular activities (playing, running errands, etc.) during recording sessions.

Early iterations of this study indicated that retrospective labeling of multi-hour videos by a caregiver was impractical. Since the same labeling by the researcher introduced a bias we were trying to avoid, a custom Android app[2] was developed to enable "live labeling" of affective states and communicative indicators by the caregiver in real time. The app included button-based labels like "request," "protest," "laughter," and other customizable labels, and the caregiver was instructed to label the child's vocalizations as soon as possible as they occurred. All labels from the app were timestamped and then synced to a server at the user's discretion.

As with any naturalistic or longitudinal dataset, the data acquired for this work posed a number of ethical and practical considerations that we have attempted to address. For example, the data collection methodology was developed over a 5-month iterative process with the pilot family and was designed to be inexpensive (i.e., easily replaced and deployable), comfortable, and unobtrusive. Paired with the open-source live-labeling app, this method enables a tenable solution for remote data collection – a critical feature for future participants in this underserved, geographically dispersed population – with minimal participant and caregiver burden. Privacy concerns for naturalistic audio and video remains an open and active field. In our study, participants had the ability to review and delete video/audio segments before sharing them with the research team. In the future, the final released dataset will include only a de-identified feature set (without raw video or audio) from participants who have chosen to opt-in.

[2]This app is open source and available by emailing the authors.

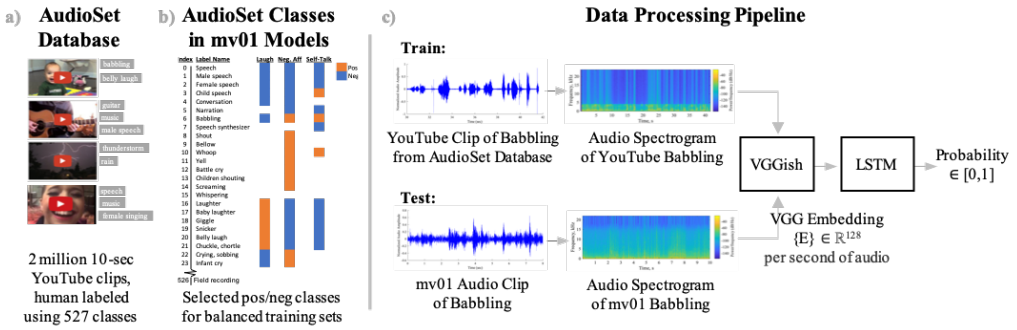

Figure 2: a) Each video in the AudioSet database [12, 15] was human-labeled using 527 possible event classes. b) We selected appropriate positive/negative classes from AudioSet to include in three different models of mv01 vocalizations. c) The LSTM model was trained on VGG-like (VGGish) embeddings of the selected subsets of AudioSet data. The inference task used the VGGish embeddings of the mv01_test data.

Dataset characterization and structure The acquired dataset, referred to in this paper as **mv01**, contains approximately 850 minutes of data spread across 5 non-consecutive days. Sessions ranged in length from 78 to 262 minutes, with an average session length of 168 minutes. The participant's caregiver live-labeled for a subset (approximately 15-30 minutes) of each session. Labels and vocalizations were non-uniformly distributed and regularly occurred in clusters – e.g., the participant often went 20-30 minutes without making a sound, and then made several vocalizations within a few minutes. Because of this, test segments for the models were selected to include clusters of live labels.

The first 3 days of data collection, containing 529 minutes of data, was used for validation (**mv01_validate**). The last 2 days of data, containing 312 minutes of data, was held out for testing (**mv01_test**) in an effort to mimic performance on true future data. The validation (test) set contained 177 (70) labels, covering 12 (7) label categories. For this work, we evaluated 3 frequent label categories: self-talk, laughter, and negative affect. *Self-talk* describes babbling, sing-song, or humming sounds not overtly associated with a communicative interaction and generally indicative of positive affect. *Laughter* was used as an easily recognizable indicator of model robustness. *Negative affect* is a meta grouping encompassing three live labels: frustration, protest, and dysregulation. "Dysregulation" tends to occur when the child is over- or under-stimulated. The labels in this meta-category produced similar vocal sounds, including high-pitched yells, whoops, and audible teeth grinding.

3 Machine learning approaches

Small, sparsely-labeled datasets pose a challenge in many naturalistic applications of deep learning, particularly when working with specialized populations as in mv01. In this dataset, the live labels do not point to a precise audio segment; rather they indicate a general time period where an event occurred ($\sim \pm 3$s). Hence, the mv01 dataset is not ideal for model training in its current state because of the lack of true, well-aligned labels for audio events and the small number of labels for each category. Note that future work will explore alignment methods, like dynamic time warping, to prepare the dataset for use in training semi-supervised models and other transfer learning approaches. Here, we took a zero-shot transfer learning approach to examine how models trained with a large, generic audio dataset performed with non-traditional vocalization data, and to inform how to augment training data for improved model performance with mvASD individuals.

In this work, the mv01 dataset was used to validate and test models trained on subsets of the AudioSet database [12, 15]. AudioSet is a large public database containing over 2 million 10-second segments from YouTube (see Figure 2a). Each segment has been human-labeled using 527 audio event labels, from speech to music to nature sounds. 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 pilot participant.

We first identified the three most frequently labeled categories from mv01: Self-Talk, Negative Affect (meta-category), and Laughter. We then manually selected AudioSet labels to create positive and

Table 1: Confusion matrices for the three mv01 models using held-out test data

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			

negative classes for each category based on our knowledge of the audio content in the mv01_validation data (see Figure 2b). 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 but that we did not want to model, like music. Note that using distinct models for each category (Self-Talk, Negative Affect, and Laughter) allowed us to explore overlapping identifications of ambiguous sounds. Feedback from the participant's caregiver was that the highest priority was identifying audio events of interest, even at the cost of a higher false positive rate, so the final positive/negative classes for each model were selected with this criterion in mind. The complete list of AudioSet labels used for each model is available in the Supplementary Material.

Using VGGish [9], each 0.96s of audio from the selected classes of AudioSet was converted into a high-level 128-dimensional feature vector (see Figure 2c). These embeddings were then fed into three-layer LSTM models that used batch normalization and Adam optimization [17] with balanced positive and negative classes, similar to [33]. The probability of a successful classification was determined using overlapping 9.6s segments with a window step size of 0.96s. Probability thresholds were selected for each model based on the performance with mv01_validation: 0.80 for Self-Talk, 0.70 for Negative Affect, and 0.95 for Laughter. To evaluate each model's efficacy, a 5-minute segment was selected from mv01_test for each category, and each second was hand-coded by the caregiver before the models were tested. The caregiver's live labels were used to identify a segment with many instances of that category before coding, addressing one of the challenges of sparse data discussed earlier. The performance of each model on the test data is summarized in Table 1.

4 Discussion and future work

The Laughter model produced an accuracy of approximately 70% and a strong true negative rate, correctly recognizing when audio segments were not laughter. However, its low true positive rate may be a consequence of the highly varied and noisy environment of real-world data. The Laughter and Negative Affect models performed similarly, with accuracies much higher than that of Self-Talk, which had an accuracy around chance. These results may be related to the availability – or lack thereof – of accurately representative event classes in the AudioSet dataset. For example, while AudioSet contained sounds similar to the participant's Negative Affect (e.g., yell, whoop, groan), the Self-Talk model was largely trained on infant babbling sounds or young children speaking with typical verbal content. Although we used the closest sounds available, it may be that neither of these event classes accurately represented the sing-song vocalizations of an elementary-aged child. In addition to any content differences between vocalizations in mv01 and AudioSet, the waveforms and spectral decompositions of naturalistic data are also distinct, as visible in Figure 2.

This dataset is one of the first of its kind and an important step in developing algorithms that can generalize to sparse, naturalistic data. The challenge of creating models using imprecise real-time labels in a small audio dataset with unique vocal characteristics may be best posed for a semi-supervised learning approach. In addition, future work will explore direct transfer learning between the VGGish embedding spaces of AudioSet and mv01. Contextual information from audio and video data could also be incorporated as features in future model iterations to improve model performance. The planned release of the de-identified dataset will include features representing contextual information that captures the environment surrounding the data, such as background noises and presence of other speakers. These results serve as a charge to build datasets and algorithms that enhance communication in minimally and non-verbal ASD populations. We invite other researchers to join us in exploring this unique and important challenge.

Acknowledgments

We are immensely grateful to the study participants for their time and energy, as well as to Craig Ferguson for the custom app development. We also thank Natasha Jaques, Sara Taylor, Ishwarya Ananthabhotla, Neo Mohsenvand, Doug Beeferman, and Dhaval Adjodah for helpful and insightful feedback. This research was partially funded by the Microsoft AI for Accessibility grant and was supported by the MIT Media Lab Consortium.

References

[1] Communicating with and about people with disabilities. URL https://www.cdc.gov/ncbddd/disabilityandhealth/pdf/disabilityposter_photos.pdf.

[2] Elizabeth C Bacon, Suzanna Osuna, Eric Courchesne, and Karen Pierce. Naturalistic language sampling to characterize the language abilities of 3-year-olds with autism spectrum disorder. *Autism*, 23(3):699–712, 2019.

[3] Jon Baio. Prevalence of autism spectrum disorder among children aged 8 years-autism and developmental disabilities monitoring network, 11 sites, united states, 2010. 2014.

[4] Alice Baird, Shahin Amiriparian, Nicholas Cummins, Alyssa M Alcorn, Anton Batliner, Sergey Pugachevskiy, Michael Freitag, Maurice Gerczuk, and Björn Schuller. Automatic classification of autistic child vocalisations: A novel database and results. International Speech Communication Association, 2017.

[5] Ioana-Alina Bănică, Horia Cucu, Andi Buzo, Dragoş Burileanu, and Corneliu Burileanu. Automatic methods for infant cry classification. In *2016 International Conference on Communications (COMM)*, pages 51–54. IEEE, 2016.

[6] Lydia Brown. Identity-first language, 2011. URL http://autisticadvocacy.org/about-asan/identity-first-language.

[7] Dragos Datcu and Leon JM Rothkrantz. Semantic audio-visual data fusion for automatic emotion recognition. *Emotion recognition: a pattern analysis approach*, pages 411–435, 2014.

[8] Helen L Egger, Geraldine Dawson, Jordan Hashemi, Kimberly LH Carpenter, Steven Espinosa, Kathleen Campbell, Samuel Brotkin, Jana Schaich-Borg, Qiang Qiu, Mariano Tepper, et al. Automatic emotion and attention analysis of young children at home: a researchkit autism feasibility study. *npj Digital Medicine*, 1(1):20, 2018.

[9] Dan Ellis, Shawn Hershey, Aren Jansen, and Manoj Plakal. Vggish. URL https://github.com/tensorflow/models/tree/master/research/audioset/vggish.

[10] Florian Eyben, Martin Wöllmer, and Björn Schuller. Openear—introducing the munich open-source emotion and affect recognition toolkit. In *2009 3rd international conference on affective computing and intelligent interaction and workshops*, pages 1–6. IEEE, 2009.

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

[12] Jort F Gemmeke, Daniel PW Ellis, Dylan Freedman, Aren Jansen, Wade Lawrence, R Channing Moore, Manoj Plakal, and Marvin Ritter. Audio set: An ontology and human-labeled dataset for audio events. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 776–780. IEEE, 2017.

[13] Matthew S Goodwin, Ozan Özdenizci, Catalina Cumpanasoiu, Peng Tian, Yuan Guo, Amy Stedman, Christine Peura, Carla Mazefsky, Matthew Siegel, Deniz Erdoğmuş, et al. Predicting imminent aggression onset in minimally-verbal youth with autism spectrum disorder using preceding physiological signals. In *Proceedings of the 12th EAI International Conference on Pervasive Computing Technologies for Healthcare*, pages 201–207. ACM, 2018.

[14] Elliott Hedman, Lucy Miller, Sarah Schoen, Darci Nielsen, Matthew Goodwin, and Rosalind Picard. Measuring autonomic arousal during therapy. In *Proc. of Design and Emotion*, pages 11–14. Citeseer, 2012.

[15] Shawn Hershey, Sourish Chaudhuri, Daniel PW Ellis, Jort F Gemmeke, Aren Jansen, R Channing Moore, Manoj Plakal, Devin Platt, Rif A Saurous, Bryan Seybold, et al. Cnn architectures for large-scale audio classification. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 131–135. IEEE, 2017.

[16] Kayleigh K Hyde, Marlena N Novack, Nicholas LaHaye, Chelsea Parlett-Pelleriti, Raymond Anden, Dennis R Dixon, and Erik Linstead. Applications of supervised machine learning in autism spectrum disorder research: a review. *Review Journal of Autism and Developmental Disorders*, 6(2):128–146, 2019.

[17] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[18] Azadeh Kushki, Ajmal Khan, Jessica Brian, and Evdokia Anagnostou. 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.

[19] Todd P Levine, Stephen J Sheinkopf, Matthew Pescosolido, Alison Rodino, Gregory Elia, and Barry Lester. Physiologic arousal to social stress in children with autism spectrum disorders: a pilot study. *Research in Autism Spectrum Disorders*, 6(1):177–183, 2012.

[20] Lichuan Liu, Wei Li, Xianwen Wu, and Benjamin X Zhou. Infant cry language analysis and recognition: an experimental approach. *IEEE/CAA Journal of Automatica Sinica*, 6(3):778–788, 2019.

[21] Sinéad Lydon, Olive Healy, and Martina Dwyer. An examination of heart rate during challenging behavior in autism spectrum disorder. *Journal of Developmental and Physical Disabilities*, 25 (1):149–170, 2013.

[22] Erik Marchi, Björn Schuller, Anton Batliner, Shimrit Fridenzon, Shahar Tal, and Ofer Golan. Emotion in the speech of children with autism spectrum conditions: Prosody and everything else. In *Proceedings 3rd Workshop on Child, Computer and Interaction (WOCCI 2012), Satellite Event of INTERSPEECH 2012*, 2012.

[23] D Kimbrough Oller, P Niyogi, S Gray, JA Richards, J Gilkerson, D Xu, U Yapanel, and SF Warren. 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.

[24] Chien Shing Ooi, Kah Phooi Seng, Li-Minn Ang, and Li Wern Chew. A new approach of audio emotion recognition. *Expert systems with applications*, 41(13):5858–5869, 2014.

[25] Lizbeth Peralta-Malváez, Omar López-Rincón, David Rojas-Velazquez, Luis Oswaldo Valencia-Rosado, Roberto Rosas-Romero, and Gibran Etcheverry. Newborn cry nonlinear features extraction and classification. *Journal of Intelligent & Fuzzy Systems*, 34(5):3281–3289, 2018.

[26] Fabien Ringeval, Erik Marchi, Charline Grossard, Jean Xavier, Mohamed Chetouani, David Cohen, and Björn Schuller. Automatic analysis of typical and atypical encoding of spontaneous emotion in the voice of children. In *Proceedings INTERSPEECH 2016, 17th Annual Conference of the International Speech Communication Association (ISCA)*, pages 1210–1214, 2016.

[27] Bernardino Romera-Paredes and Philip Torr. An embarrassingly simple approach to zero-shot learning. In *International Conference on Machine Learning*, pages 2152–2161, 2015.

[28] Björn Schuller, Michel Valstar, Florian Eyben, Gary McKeown, Roddy Cowie, and Maja Pantic. Avec 2011–the first international audio/visual emotion challenge. In *International Conference on Affective Computing and Intelligent Interaction*, pages 415–424. Springer, 2011.

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

[30] Mehmet Cenk Sezgin, Bilge Gunsel, and Gunes Karabulut Kurt. Perceptual audio features for emotion detection. *EURASIP Journal on Audio, Speech, and Music Processing*, 2012(1):16, 2012.

[31] Shivam Sharma and Vinay Kumar Mittal. Infant cry analysis of cry signal segments towards identifying the cry-cause factors. In *TENCON 2017-2017 IEEE Region 10 Conference*, pages 3105–3110. IEEE, 2017.

[32] Stephen J Sheinkopf, Jana M Iverson, Melissa L Rinaldi, and Barry M Lester. Atypical cry acoustics in 6-month-old infants at risk for autism spectrum disorder. *Autism Research*, 5(5): 331–339, 2012.

[33] Nat Steinsultz. Laugh detector, 2018. URL `https://github.com/ideo/LaughDetection`.

[34] Qandeel Tariq, Jena Daniels, Jessey Nicole Schwartz, Peter Washington, Haik Kalantarian, and Dennis Paul Wall. Mobile detection of autism through machine learning on home video: A development and prospective validation study. *PLoS medicine*, 15(11):e1002705, 2018.

[35] O Wasz-Höckert, TJ Partanen, V Vuorenkoski, K Michelsson, and E Valanne. The identification of some specific meanings in infant vocalization. *Experientia*, 20(3):154–154, 1964.

Zero-Shot Transfer Learning to Enhance Communication for Minimally Verbal Individuals with Autism using Naturalistic Data

Jaya Narain & Kristina T. Johnson*, Rosalind Picard, Pattie Maes (*Equal Contribution)*

AudioSet indices below match those in class_labels_indices.csv from the Features Dataset found here: https://research.google.com/audioset/download.html

As discussed in the text, the three model categories are Laughter, Negative Affect, and Self-Talk. Each positive class contains sounds that were similar (or as similar as possible) to the child's vocalizations for that category. Each negative class contains sounds that might confuse the model because they sounded similar to the child's vocalizations in a different category or appeared frequently in the dataset.

Index	Label Name	Laugh	Neg. Aff	Self-Talk
0	Speech	Negative class	Negative class	Negative class
1	Male speech	Negative class	Negative class	Negative class
2	Female speech	Negative class	Negative class	Negative class
3	Child speech	Negative class	Negative class	Positive class
4	Conversation	Negative class	Negative class	
5	Narration		Negative class	Negative class
6	Babbling	Negative class	Positive class	Positive class
7	Speech synthesizer			Negative class
8	Shout		Positive class	
9	Bellow		Positive class	
10	Whoop		Positive class	Positive class
11	Yell		Positive class	
12	Battle cry		Positive class	
13	Children shouting		Positive class	
14	Screaming		Positive class	
15	Whispering			
16	Laughter	Positive class	Negative class	Negative class
17	Baby laughter	Positive class	Negative class	Negative class
18	Giggle	Positive class	Negative class	Negative class
19	Snicker	Positive class	Negative class	Negative class
20	Belly laugh	Positive class	Negative class	Negative class
21	Chuckle, chortle	Positive class	Negative class	Negative class
22	Crying, sobbing	Negative class	Positive class	
23	Infant cry	Negative class	Positive class	
24	Whimper		Positive class	
25	Wail, moan		Positive class	

Positive class
Negative class

26	Sigh									
27	Singing									
28	Choir									
29	Yodeling									
30	Chant									
31	Mantra									
32	Male singing									
33	Female singing									
34	Child singing									
35	Synthetic singing									
36	Rapping									
37	Humming									
38	Groan									
39	Grunt									
	. . .									
45	Pant									
	. . .									
63	Clapping									
	. . .									
69	Crowd									
70	Speech babble									
	. . .									
137	Music									
	. . .									
161	Percussion									
	. . .									