# VERBAL PROTEST RECOGNITION IN CHILDREN WITH AUTISM

*Jonah Casebeer, Hillol Sarker, Murtaza Dhuliawala, Nicholas Fay, Mary Pietrowicz, Amar Das*

IBM Research, USA

## ABSTRACT

Real-time detection of verbal protest (sensory overload-induced crying) in children with autism is a first step towards understanding the precursors of challenging behaviors associated with autism. Detection of verbal protest is useful for both autism researchers interested in exploring just-in-time intervention techniques and researchers interested in audio event detection in routine living environments.In this paper, we examine, adapt, and improve upon two techniques for verbal protest recognition and tailor them for children with autism spectrum disorder (ASD). The first technique investigated is a Gaussian Mixture Model (GMM) with stacking. The second technique uses Convolutional Neural Networks (CNN) trained on log Mel-filter banks (LMFB). We proceed to examine accuracy with a focus on real-world false positive rates and minimization of dataset biases through the introduction of noise and input perturbation.

***Index Terms*—** Audio Event Detection, Convolutional Neural Networks, Gaussian Mixture Model, Ubiquitous Computing

## 1. INTRODUCTION

According to the Centers for Disease Control and Prevention (CDC), the cases of autism have increased by a factor of 10 since 1980. Children with autism display challenging behaviors, such as verbal protest which we define as sensory overload-induced crying, screaming, shouting, and yelling. Detection of verbal protest will help us understand the frequency and context in which these challenging behaviors occur, allowing for better prediction and prevention, thereby reducing the burden on both the children and the parents involved.

In this paper, we propose methods for exploring and detecting sensory overload-induced verbal protest in children with autism in the presence of confounding sounds, such as noise, and speech. We demonstrate GMM and CNN based machine learning and feature extraction techniques capable of detecting verbal protest with low false positive rates. We also demonstrate the ability of the resulting models to generalize across children from different age groups, in spite of the differences in gender, vocal capabilities, and vocal properties. Finally, we discuss the capability of the system to function in a variety of environments, from wearable devices and smartphones to cloud ecosystems. In the case of verbal protest, this flexibility will allow for real-time assessment and reactive Just-In-Time Intervention (JITI) by a caregiver. The primary contributions of this paper are 1) the curation of a dataset for studying verbal protest in autism, 2) a technique for leveraging observer-reported behavioral data via voluntary contributions on YouTube, 3) development of efficient detection models which can run in both embedded and cloud environments, 4) techniques for enhancing noise tolerance and reducing false positives, and 5) a model which detects verbal protests. To the best of our knowledge, this is the first work which presents a model to detect verbal protests in children with autism.

Section 2 explores prior related works on audio event recognition. Section 3 explains the datasets used for training and validation. Section 4 examines the feature extraction techniques used. Section 5 focuses on the GMM based approach, while Section 6 focuses on the CNN based approach. Section 7 reports findings, and Section 8 concludes the work.

## 2. RELATED WORK

Audio event detection has become increasingly popular due to the recent progress in ubiquitous computing with new techniques for data acquisition, model development, and interesting applications. Techniques for the detection of shouting, crying, and audio events are explored in [1, 2, 3, 4, 5, 6]. Prior work by Takahashi et al., [2], and by Salamon et al., [3] focused on the more general audio event or environment classification. Both of these works use deep learning techniques and combat the issue of data scarcity with data augmentation. As reported in [3], audio data augmentation helps generalize the resulting models, such that performance improves across unobserved data. Multiple instance learning was also shown to be an effective technique on these augmented datasets [2]. Specialized mel frequency cepstral coefficient (MFCC) feature-based Gaussian mixture models (GMMs) have been proven to perform well in shout detection despite the moderately noisy environments [4], and other representations such as Log Mel-filter bank responses have shown that deep learning applied to audio can distinguish between a baby's cry and confounding domestic sounds [1]. These prior approaches, however, were evaluated individually and used speech samples from children who were not on the autism spectrum. We leverage these techniques from the prior work in audio event detection, data augmentation, and machine learning, and propose a new model that is tuned to detect verbal protest for children on the autism spectrum.

## 3. DATA DESCRIPTION

To the best of our knowledge, no publicly-available dataset exists, featuring children with autism having meltdowns (behavioral outbursts). In order to study verbal protest in children with autism, we curated a dataset by collecting suitable YouTube recordings found via search strings related to autism challenging behaviors (e.g., "autism children shouting"). We download the relevant videos which contain episodes of children with autism having a verbal protest, as tagged by the uploader. This set contains videos from seven children aged between two and twelve years. Next, we extract the audio and employ an annotator to code the sound with the tags "verbal protest," "noise," or "speech." Here, noise and speech represented the events of daily life in contrast to the sounds of children with autism during a verbal protest. Then, we set aside approximately 15% of the 10 minutes of resulting annotated speech to test the generalizability of the model. Given the small dataset size, we also apply data augmentation techniques [2, 3] which have been successfully applied to acoustic event detection and sound classification problems. These techniques, which includes pitch shifting, time stretching, and dynamic range compression, increased the dataset size to 51 minutes and improved model robustness. To train the model, we also use the RML dataset [7] that contains confounding emotional speech (e.g., fear) from eight subjects in six different languages. Lastly, we use the *Urban Sound* dataset [8] to introduce noise in the training dataset to limit bias in the audio samples.

Two independent datasets are also used to validate the verbal protest detection model. First, we use a smartwatch to collect 105 minutes of audio in the noisy natural environment, such as riding a subway and walking around the city. This dataset provides us with a baseline for false positive rates in the free-living (natural environment) condition. Second, we use an independent "Audio Set" dataset [9] containing 527 ontology class labels. We select 18 classes under four high-level categories – child, adult, vehicle, and environment (see Figure 5 for a detailed breakdown). Audio segments are 10-second-long each containing one or more ontology-based class labels. This dataset contains many examples of audio events, which simulates a wide range of acoustic confounds present in our day-to-day life.

## 4. FEATURES

The audio streams, sampled at 44.1 kHz, are segmented into 25 ms windows with a 10 ms step. These 25 ms frames are then used as stationary representations of the incoming signal. We extract 34 features from each frame, including 24 *Mel-Frequency Cepstrum Coefficients* (MFCC), *Zero Crossing Rate (ZCR)*, *spectral roll-off*, *spectral centroid*, and *spectral contrast*. MFCCs are a staple of speech processing, well-suited to tasks of vocal differentiation as they imitate human perception. MFCC-based features also have been proven very useful in shouted speech detection in the presence of ambient noise [4]. Because of the variety of sonic backgrounds that this system is likely to encounter in actual use, we also extract features likely to differentiate confounding audio events, such as music and other noise. ZCR, roll off, centroid, and contrast proved useful in making these distinctions.

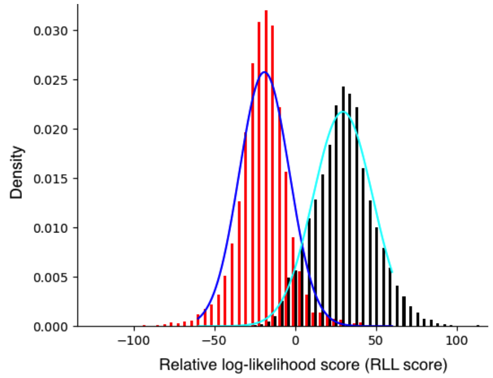
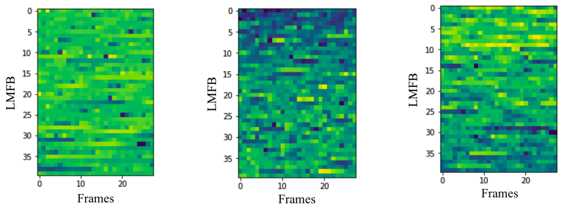


**Fig. 1**. Histogram of relative log-likelihood scores obtained from the GMM model. The Left distribution represents non-verbal-protest and the right represents verbal-protest.

## 5. MODELING APPROACH 1: GAUSSIAN MIXTURE MODEL (GMM)

Gaussian mixture model (GMM) techniques traditionally operate in an unsupervised manner under the assumption that data has been generated from one or more normal distributions. In a closely-related prior work [4], the "shout" audio event had been detected based on the relative likelihood metric computed from $L_{Shout} - max(L_{Speech}, L_{Noise})$, where *shout*, *speech*, and *noise* each had a separate, eight-component GMM. We extend this work by making three changes. First, we use two additional confounding classes, i.e., the *music* and the *emotional speech* samples obtained from the RML database. In addition, we fit two component GMMs for each of the five classes. The Relative Log Likelihood (RLL) score of verbal protest is computed as $L_{VerbalProtest} - max(L_{Speech}, L_{Noise}, L_{Music}, L_{EmotionalSpeech})$. Third, we compute the RLL scores for all five classes and make a union of them. These RLL scores are fitted into a one-dimensional space and used to train a two-component GMM. This allows us to make predictions by setting an interpretable probability-like threshold instead of setting a



Verbal Protest Noise Speech

**Fig. 2**. The spectral images passed to our CNN. Differences in verbal protest and speech are most apparent when examining closeness and continuity of formants. Noise does not have the spectral lines created by formants, thus easily distinguishable.

domain-specific threshold, effectively making the GMMs into high-level feature extractors (see Figure 1). Figure 1 shows that two-component GMM models can distinguish the verbal protest and the non-verbal protest classes.

## 6. MODELING APPROACH 2: CONVOLUTIONAL NEURAL NETWORK(CNN)

CNNs have been shown to perform well on classification and recognition tasks. In this work, we apply CNNs to "spectrogram images". To prepare an audio stream for the CNN, we again segment data into 25 ms frames with a 10 ms slide. From each of these frames, a set of 40 LMFB activations were extracted. 28 consecutive frames are then concatenated to form a 40 by 28 "picture". The difference between a typical MFCC and LMFB is that no DCT or other decorrelation step is performed. Correlation is quite helpful for CNNs [1].

Figure 2 shows typical verbal protest, speech, and noise LMFBs pictures. The short spectral lines indicate the most interesting information. In the case of vocal sounds, these lines are the formants. Verbal protest instances do not show changes in formant location or grouping as often as speech because speech requires constant vocal tract modification to produce new sounds. We leverage this quality by using relatively wide filters. Pooling is avoided in the shallow layers to prevent information loss. This is made up for by having two fully connected hidden layers. These architectural modifications build upon [1] by allowing the model to develop a higher-level representation of the learned features. We use an Adam optimizer [10] with standard categorical cross-entropy as the loss function. Additionally, dropouts are employed to prevent overfitting and dead zones within the network. The architecture is shown graphically in Figure 3.

**Model Interpretation**: Neural network-based techniques are promising and have high accuracy across multiple domains. However, unlike previous statistical methods, they do not provide visibility into the hidden structure within the network; and analysis of information inside the hidden layers is difficult. To overcome this to an extent, Local Interpretable Model-Agnostic Explanations (LIME) from [11] is used. LIME perturbs the input data with the intent to reveal what sections of the input caused the given classification (and uncover the inner workings of the classifier). LIME lends itself to image, tabular, and text data very well; however, we tailored it for spectral images. Figure 4 shows three LMFB images of verbal protest from the same child where red zones contribute positively to the verbal protest class. The determining regions are not temporally or spectrally constant, suggesting the model did not learn potential bias in the dataset.

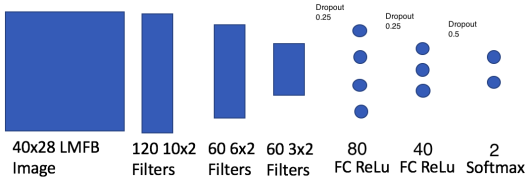


**Fig. 3**. CNN architecture for Verbal Protest detection. The data flows through the architecture from left to right.

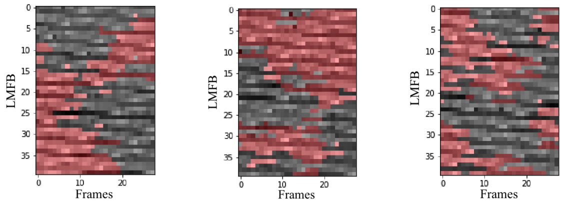


**Fig. 4**. Three example spectrograms with LIME overlay showing the regions of the image contributed to the final class label. Note that the regions of the spectrogram examined are unique across all three pictures.

**Training Perturbation**: Finally, to limit the model from learning potential bias within the dataset, we select 20 noises from the Urban Sound dataset and randomly mix them into our training sets. Table 1 shows accuracy, precision, and recall for this model.

## 7. RESULTS

Table 2 compares accuracy among the naive GMM (GMM-N), the exposed GMM (GMM-E), the naive CNN (CNN-N), and the exposed CNN (CNN-E) models. Exposed models are exposed to a small proportion (2.5%) of external noise data prior to testing (e.g., noise captured from the subway). The evaluation does not use any data from the training dataset. As discussed earlier "*Intra-Group*" means both training and testing is performed on children from same age group. "*Inter-Group*" refers to training on the younger children ($\leq$6) and testing on the older. As expected, *Intra-Group* outperforms *Inter-Group* for both GMM and CNN based models. Surprisingly, GMM performs better for the *Intra-Group* setting, and CNN performs better for the *Inter-Group* setting.

**Table 1**. Performance of the model on the perturbed dataset. *Intra-Group* referred to as training and testing both in same age group (≤6)

| | Accuracy | Precision | Recall |
|---|---|---|---|
| Intra-Group | .868 | .881 | .832 |
| Inter-Group | .713 | .706 | .760 |

**Table 2**. Comparison of performance using GMM and CNN-based models. N stands for the naive model and E stands for the exposed model. During the training phase, the exposed model includes a small fraction of additional confounding data, such as noise captured in the subway. A, P, and R are accuracy, precision, and recall, respectively.

| | Intra-Group | | | Inter-Group | | |
|---|---|---|---|---|---|---|
| | A | P | R | A | P | R |
| GMM-N | .933 | .916 | .967 | .690 | .741 | .780 |
| GMM-E | .930 | .913 | .950 | .643 | .726 | .699 |
| CNN-N | .901 | .929 | .834 | .722 | .817 | .617 |
| CNN-E | .880 | .927 | .808 | .710 | .896 | .515 |

The model is designed to run on ubiquitous devices which listen to audio in the environment and provide just-in-time intervention by suggesting an advisory message to the caregiver. Frequent false positive inferences would unnecessarily burden the caregiver. We apply four models (see Table 2) on 105 minutes of audio captured from a smartwatch in a noisy environment (e.g., subway). A wearable computing device such as a smartwatch allows for auditory processing within close proximity to the user. This is important to ensure that the foreground auditory signal is that of the primary user, while also creating the potential for continuous reinforcement learning by utilizing the constant stream of audio data. GMM-N and CNN-N provide false positive rates of 0.100 and 0.367, respectively. Exposed GMM-E and CNN-E provided lower false positives rates of 0.003 and 0.007, respectively. Next, we consider an ensemble model of GMM-E and CNN-E. We observe that in the majority of the cases, false positives were one or two windows long. Hand-curated training data suggest that the minimum duration of a child's verbal protest is 0.3 seconds. Therefore, we define an ensemble verbal protest as an episode of 0.3 second, where both models inferred verbal protest in 90% of the frames. As expected, the ensemble model reduces the false positive rate to 0.001.

Next, we evaluate the ensemble model on Audio Set. Table 5 shows the proportion of videos classified as verbal protest in each class. We further categorize the labels into logical sections – *Child*, *Adult*, *Vehicle*, and *Environment*. Most of the classes in the *Child* and *Adult* sections closely resemble the verbal protest category and we observe a higher proportion of videos labeled as containing verbal protest episodes. In the *Vehicle* and *Environment* sections, we observe a lower percentage of videos labeled as having verbal protest.

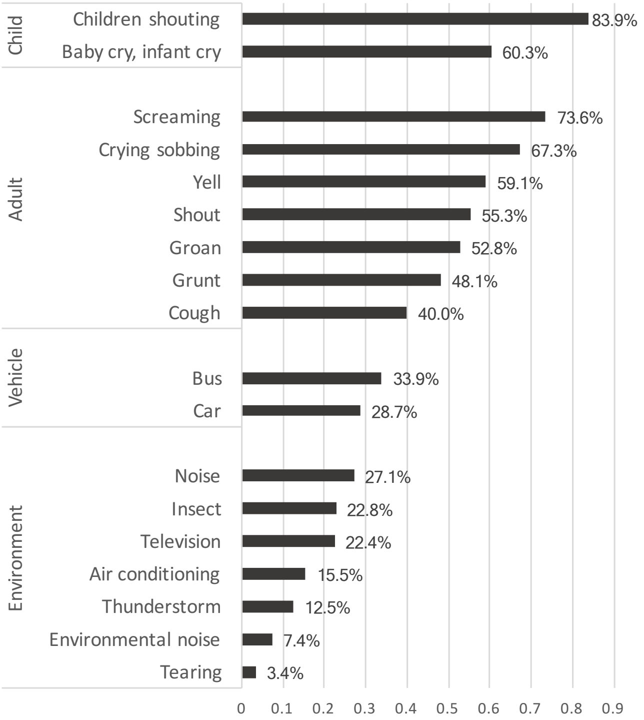


**Fig. 5**. Ensemble model-based predictions on 18 classes of Audio Set dataset. 83.9% of children shouting videos are inferred as containing verbal protest.

## 8. CONCLUSION AND FUTURE WORK

Statistical and neural network based approaches for verbal protest recognition present a set of trade-offs. GMMs can be particularly lightweight, because in the case of a diagonal covariance matrix, they will only require limited memory to be stored. However, CNNs provide significantly improved performance at the cost of resources. Despite the larger storage and computational burden of CNNs, we believe that ensemble models are the superior choice for solving this class of problems. This proposed system can eventually help the healthcare providers to understand the frequency and patterns of challenging behaviors. In the future, multiple instance learning on weakly-labeled data (e.g., tags) can reduce the dependency on manual annotation and improve the scalability of the system. In addition, we would like to add more confounding examples, such as non-autistic child crying or shouting in the training data.

## 9. ACKNOWLEDGEMENTS

We would like to thank Soumya Ghosh from IBM Research for the useful comments and discussions on the methodology.

## 10. REFERENCES

[1] Yizhar Lavner, Rami Cohen, Dima Ruinskiy, and Hans IJzerman, "Baby cry detection in domestic environment using deep learning," in *Science of Electrical Engineering (ICSEE), IEEE International Conference on the*. IEEE, 2016, pp. 1–5.

[2] Naoya Takahashi, Michael Gygli, Beat Pfister, and Luc Van Gool, "Deep convolutional neural networks and data augmentation for acoustic event detection," *arXiv preprint arXiv:1604.07160*, 2016.

[3] Justin Salamon and Juan Pablo Bello, "Deep convolutional neural networks and data augmentation for environmental sound classification," *IEEE Signal Processing Letters*, vol. 24, no. 3, pp. 279–283, 2017.

[4] Jouni Pohjalainen, Tuomo Raitio, and Paavo Alku, "Detection of shouted speech in the presence of ambient noise," in *Twelfth Annual Conference of the International Speech Communication Association*, 2011.

[5] Mahesh Kumar Nandwana and John HL Hansen, "Analysis and identification of human scream: Implications for speaker recognition," in *Fifteenth Annual Conference of the International Speech Communication Association*, 2014.

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

[7] Yongjin Wang and Ling Guan, "Recognizing human emotional state from audiovisual signals," *IEEE transactions on multimedia*, vol. 10, no. 5, pp. 936–946, 2008.

[8] Justin Salamon, Christopher Jacoby, and Juan Pablo Bello, "A dataset and taxonomy for urban sound research," in *Proceedings of the 22nd ACM international conference on Multimedia*. ACM, 2014, pp. 1041–1044.

[9] 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 *IEEE ICASSP*, 2017.

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

[11] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin, "Why should i trust you?: Explaining the predictions of any classifier," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, 2016, pp. 1135–1144.