2018 26th Mediterranean Conference on Control and Automation (MED)
Zadar, Croatia, June 19-22, 2018

Classification of Child Vocal Behavior for a Robot-Assisted Autism Diagnostic Protocol

Mirko Kokot[1], Frano Petric[1], Maja Cepanec[2], Damjan Miklić[1], Ivan Bejić[1] and Zdenko Kovačić[1]

Abstract— **Autism is a neurodevelopmental disorder affecting an increasing fraction of children, with severe social and economic consequences for affected persons and their families. Including robotic technologies in the diagnostic process could potentially increase its speed and reliability, opening the way towards earlier and more efficient therapy. The diagnostic process requires multimodal interaction, in which the vocal behavior of the child plays an important role. In this paper, we present a method for automatic classification of child vocal behavior, based on supervised learning, which is suitable for real-time execution on an autonomous robot with limited computational resources. The main contribution of the paper is an empirically determined minimal set of sound features, which allow efficient vocal behavior classification of preschool children, relevant in the context of autism diagnostics. The classifier is verified on a dataset combined from publicly accessible audio recordings and recordings collected during diagnostic and therapeutic sessions.**

I. Introduction

Autism spectrum disorder (ASD) is primarily defined by deficits in social communication and interaction, as well as the presence of restricted, repetitive patterns of behavior, interests, or activities [1]. It has become a commonly diagnosed neurodevelopmental disorder, with increasing prevalence rates. Although there are many robotic applications focusing on teaching and intervention [2], diagnostic applications are scarce. Since there are no medical markers of autism that could be used in a diagnostic process, the diagnosis relies on behavioral observations made by experienced clinicians, as there are specific behavioral markers of autism that are observable in children as young as 12 months [3]. These markers include eye contact, gesture use, joint attention capabilities, but also pre-verbal and verbal vocal behavior.

Characteristics of pre-verbal and verbal vocal behavior (both communicative and non-communicative) are considered an important diagnostic and prognostic factor in autism. Studies have shown that differences in vocalization patterns between children with ASD and typically developing children may be observed even before the age of 6 months [4]. Generally, infants and toddlers with ASD show lower rates of vocalization [5], [6], less complex modulated vocal productions [4], lower rates of canonical babbling [7], significantly higher proportion of distress vocalization [8] and significantly more atypical non-speech vocalizations [9], [10]. Stereotyped vocal patterns in the form of prolonged existence of jargon (babble with speech-like inflection) and echolalia (repetition of vocalization made by another person) are also common [11]. Given the early onset of vocal behavior in infants, and often delayed onset of speech production in children with ASD, the analysis of pre-verbal vocal production is clinically important even in chronologically older children with ASD, especially if they are pre-verbal [12]. During the verbal period, prosodic aspects of speech (rate, loudness, pitch) in children with ASD are from a very early age marked by specific characteristics - they tend to produce more atypical stress patterns [13] and are less inclined to use the prosodic frequency range normally used by typical children [14]. Altogether, analysis of various aspects of vocal patterns have shown that vocal behavior could be a useful component of early screening and diagnosis model of ASD.

When combining data about vocal behavior and other communicative behaviors (e.g. eye contact, use of gesture etc.), both diagnostic and prognostic values are rising. For example, it is known that children with ASD have higher prevalence of non-social babbling than typically developing children [5], that they rarely use co-speech gestures [15], that only communicative vocalization (over noncommunicative) predicts expressive language outcome [8] etc. All these data suggest that a multimodal model is necessary to describe and diagnose social-communication deficits in children with ASD, and vocal behavior should be a key component of that model.

Scientists have tried to implement automated vocal analysis to add a fully objective measure to the battery used to detect speech-related disorders and differentiate children with ASD from typically developing children based on their vocalization patterns [16]. Moreover, some attempts were also made to create automatic techniques for quantifying the amount of repetition in speech of children with ASD [17].

Motivated by the promising prospect of employing robotic assistants in ASD diagnostics and intervention [2] and by the importance of child vocal behavior in both contexts, we set out to implement a method for automatically classifying vocal behavior, suitable for real-time execution on an autonomous robot. We have implemented a software solution for extracting relevant audio features and classifying child vocal behavior, which can run on the NAO robot, an off the shelf robotic platform which is already widely used in ASD research. For classification, we use the Random forest classifier with supervised learning. Our contribution in this paper is twofold:

- an empirically determined minimal set of audio features,

This work has been fully supported by the Croatian Science Foundation through the ADORE project (HRZZ-93743-2014)

[1]LARICS Laboratory, Faculty of Electrical Engineering and Computing, University of Zagreb, Unska 3, Zagreb, Croatia larics@fer.hr

[2]Faculty of Education and Rehabilitation Sciences, University of Zagreb, Borongajska cesta 83f, Zagreb, Croatia maja.cepanec@erf.hr

978-1-5386-7891-6/18/$31.00 ©2018 IEEE

which yield a classifier which is fast enough for real-time, onboard execution, yet accurate enough to provide meaningful information for informing robot actions.

- experimental evaluation of random forest classifier using humanoid robot NAO

II. Methodology

Guided by the widely accepted Autism Diagnostic Observation Schedule (ADOS)[18], we envision the robot-assisted ASD diagnostic protocol which emulates ADOS through a set of tasks that are to be administered by a humanoid robot. Relying only on observations of multiple social cues within each task, such as eye contact, gestures and vocal behavior, the robot provides an assessment of the child's ASD-relevant functioning level within a particular task and provides human evaluators with readable information.

During the diagnostic procedure, along with noting all the other diagnostically relevant cues, the human examiner also performs an assessment of the child's vocal performance by classifying its audible cues into 4 categories: non-articulated vocalizations, babbling (the child is pronouncing syllables that are rhythmically separated), jargon (the child is pronouncing syllables that have the melody of speech but have no actual meaning) and speech.

Since our goal is for the robot to produce as much information as possible for the human examiner, we investigate whether the robot can differentiate the aforementioned classes of vocal activity. During preliminary sessions with children, we observed that the presence of the robot may be stressful for the children which occasionally results in crying. Crying is a specific form of non-articulated vocalization, and in terms of diagnostic protocol usually signals that the behavior of the examiner needs to be changed. The robot which can reliably detect crying will be able to autonomously alter its behavior or even stop the session, which we deem to be beneficial towards our goal of increasing the robot autonomy, therefore we include the class of crying into a set of classes we consider in this paper.

As the robot behavior should be interactive and adaptive, real-time execution is of paramount importance. Our targeted robotic platform is the NAO robot, because of its off the shelf availability, widespread use in child behavior research and characteristics suitable for interaction with ASD affected children [2]. We focus on on-board execution in order to eliminate the need for external infrastructure and environment customization, thus facilitating deployment in realistic scenarios.

A. Dataset

As we are unable to find an appropriate publicly accessible dataset for training and evaluating our classifier, we resort to creating one. Our primary source of data are recordings made during clinical ASD diagnostic and intervention sessions, as well as pilot sessions involving the NAO robot. Recordings from the robots microphone are filtered with a noise reduction filter, in order to alleviate the well-known problem with NAO hardware [19]. This dataset is augmented with recordings collected from publicly available sources, such as YouTube. A total of 107 recordings has been collected, with a cumulative duration of about 8 hours. To simulate different scenarios that may occur in diagnostic session, we do not use audio recordings recorded in an ideal setup (using the same equipment, similar volume levels, similar levels of background noise) but rather incorporate in the dataset recordings of various volume (simulates distance between a child and the robot) and with various levels of background noise.

TABLE I: Dataset summary. The dataset contains only pure child vocalizations, i.e., all the intervals of silence and other voices have been removed.

Class	Number of samples	Total duration [s]
Speech	296	765.02
Babbling	247	699.97
Nonarticulated	227	578.09
Crying	210	559.70
Jargon	211	618.56
Total	1191	3221.34

The original recordings are first split into samples at silence intervals. The split recordings are then cropped of silent periods and samples with excessive background noise, multiple speakers or adult voices are discarded. This results in a dataset of 1191 samples with a total duration of 3221.34 seconds of pure child vocalizations, as summarized in Table I. These samples are then categorized by a phonetically trained listener into the four standard categories of child vocalization, with the addition of crying as a category with special relevance for informing robot actions during interaction with children.

III. Child vocal behavior classification

The dataset for training the classifier is small for a machine learning application and slightly imbalanced in terms of number of samples and total duration of each class. Therefore we opt for a random trees classifier (also known as random forest), as it was proven to be able to cope with even imbalanced data [20]. An additional benefit of using the random trees classifier is that the training error is estimated internally during the training, so there is no need to split the training part of the dataset further into train and validation sets for the purpose of training, which would accentuate the problem of small dataset. We use the OpenCV implementation of random trees[1], which is available on the robot, but also use random forest implementation in Python module *sklearn* as it facilitates feature selection.

A. Feature selection

Another advantage of using the random trees classifier is the inherent calculation of variable (feature) importance, which corresponds to the importance of each feature in making the final decision on the class label. This allows for multiple feature selection approaches, from univariate

[1] http://docs.opencv.org/2.4/modules/ml/doc/random_trees.html

TABLE II: Audio features considered for the child vocal behavior classification problem. Numbers in brackets for vector features indicate the number of components. Features are extracted using *libXtract*. Features selected for the final classifier version are printed in bold.

Feature	Domain	Dimension
Zero crossing rate	Time	Scalar
High zero crossing rate ratio	Time	Scalar
Root mean square	Time	Scalar
Kurtosis	Time	Scalar
Skewness	Time	Scalar
Spectral mean	Frequency	Scalar
Spectral variance	Frequency	Scalar
Spectral deviation	Frequency	Scalar
Spectral kurtosis	Frequency	Scalar
Spectral skewness	Frequency	Scalar
Sharpness	Frequency	Scalar
Loudness	Frequency	Scalar
Linear predictive coding	Frequency	Vector[10]
Linear prediction cepstral coefficients	Frequency	Vector[10]
Mel frequency cepstral coefficients	Frequency	Vector[10]
Bark frequency cepstral coefficients	Frequency	Vector[24]

in which only those features that have importance above a certain threshold can be kept to more complex feature selection methods such as recursive or model-based feature selection.

Given the variety in the vocal behaviors which are assessed, we consider several scalar and vector features, both in time and frequency domain, in order to extract the most information from a given audio sample. Features, which are listed in Table II, are extracted using the open source libXtract[2] library.

To select the best features for our application, we use recursive feature elimination with cross evaluation method. Feature elimination is performed by ranking features with respect to the importance in the classifier. Cross evaluation is then used to stop the feature elimination to ensure optimal number of features. In each iteration of the selection algorithm, random forest classifier is trained, feature importance is extracted and the worst feature is removed. This procedure is performed using *sklearn* Python module. With this approach, all components of linear predictive coding vector and linear prediction cepstral coefficients vector are removed from the feature set. Also, all but one of the 24 Bark frequency cepstral coefficient are staged for removal, so we eliminate all of the Bark frequency cepstral coefficient without loss of classifier accuracy. Recursive feature elimination also removed high zero crossing rate ratio, loudness and sharpness, leaving the following feature set comprised of 19 elements in total:

- Zero crossing rate;
- Root mean square of the signal;
- Kurtosis;
- Skewness;
- Spectral mean;
- Spectral variance;
- Spectral deviation;
- Spectral kurtosis;
- Spectral skewness;
- Mel frequency cepstral coefficients (MFCC), which are among the most popular features in automated speech recognition, music information retrieval and audio similarity measures; We experimented with several MFCC vector lengths starting at 40 and found that there is no significant degradation in classifier performance if only 10 coefficients are used, resulting in shorter computation times;

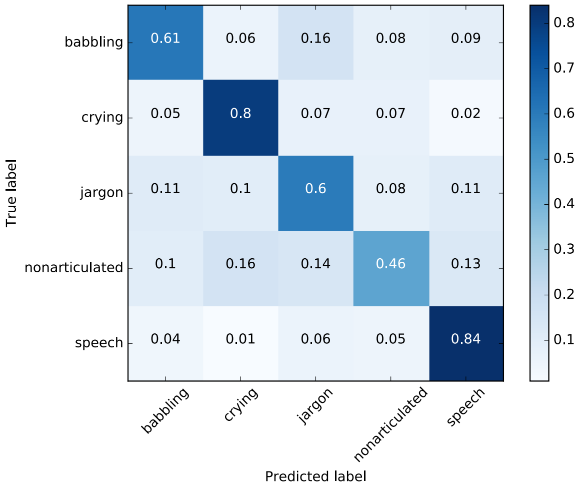


Fig. 1: Normalized confusion matrix for the test dataset.

IV. Classifier Cross-Validation

To verify the effectiveness of the proposed classification method, we perform repeated random subsampling validation. The dataset is randomly partitioned into training data (75% of samples in each class) and test data (25% of samples in each class), and training and validation are performed on these subsets. This is repeated until we achieve convergence of the average values of elements on the main diagonal of the test confusion matrix, which is usually achieved within 100 iterations.

Confusion matrix for test dataset averaged over 100 iterations is shown in Figure 1. Values on the main diagonal are dominant, confirming solid classifier performance.

Although the described feature selection method does not necessarily optimize the feature set for better generalization, we conclude that the random forest classifier successfully generalizes, with some issues in classifying nonarticulated utterances, which is to be expected since even humans can miss-classify such vocalizations. From the confusion matrix for the test dataset, we extract several classification metrics to perform a quantitative analysis of classifier performance. These metrics include but are not limited to true positives (TP), false positives (FP), false negatives (FN), true negatives (TN), true positive rate (recall, TPR), positive predictive value (precision, PPV), accuracy and diagnostic odd ratio

[2] https://github.com/jamiebullock/LibXtract

(DOR). The results for precision, recall, accuracy and DOR are summarized in table III.

TABLE III: Classification metrics extracted from the confusion matrix for the test dataset. Values for a class are obtained by looking at the classification in form of one-vs-all problem. Overall assessment in the last row is performed by micro-averaging the values across all classes.

Class	precision	recall	accuracy	DOR
Babbling	0.64	0.61	0.87	20.11
Crying	0.71	0.80	0.89	45.63
Jargon	0.62	0.60	0.83	13.30
Nonart.	0.59	0.46	0.85	11.98
Speech	0.75	0.84	0.89	53.78
Total	0.67	0.67	0.87	23.22

True positive rate (recall) for a given class describes the likelihood that the class label will be correctly identified, and attains low values for babbling and jargon, which is to be expected as those classes are easily misclassified as speech. Positive predictive value (precision) for a given class describes the likelihood that any of the other classes will not be misclassified into it. Diagnostic odd ratio is defined as $(TP \cdot TN)/(FP \cdot FN)$ and values greater than one indicate that the test is useful with greater values being indicative of better performance. The ideal classifier would have DOR tend to infinity, as FP and FN would be zero. However, if only one of FP and FN is zero, the DOR is not defined.

Using data from the confusion matrix in Figure 1 and Table III, we conclude that the proposed classifier is best suited for detection of crying and speech, while not being able to successfully classify jargon and nonarticulated vocalizations of a child. Values of accuracy in Table III suggest there are no systematic errors in the classifier.

V. Experimental evaluation

Given the fact that it is rather difficult to obtain data from sessions with children, evaluation on NAO is performed in the laboratory. Test dataset for experimental evaluation consists of 300 samples and is randomly drawn from the existing dataset. Each of the samples is played through speakers and recorded using NAO's front microphone. Experiment is performed twice, in one case the speakers are put at a distance of about one meter from the microphone and in the other at a distance of about 50 cm from the robot. In both cases, the speakers are slightly above the microphone level (see Figure 2) since the front microphone on NAO's head is directed at an upwards angle.

The confusion matrices for these two scenarios are shown in Figure 3a and Figure 3b, respectively.

By comparing confusion matrices in Figures 3a and 3b, we can conclude that the trained random forest classifier is not well suited for use with NAO. It is difficult to evaluate to which degree this is a result of a bad classifier design or feature selection, mostly due to large amount of noise caused by the fan of the computer which is located near the microphone in the head of NAO. This negative effect of NAO's fan noise on different kinds of audio processing is well-known and already reported in [19], where successful speech recognition is achieved only when a child is directly in front of the robot and very close. At a larger distance from the sound source, NAO is only capable of recognizing crying (Figure 3a). Picking up crying is even more pronounced when the sound source is closer to the robot, as most of the audio signals were classified as crying at a distance of 50 cm (see Figure 3b, which indicates that the classifier could be overtuned towards crying.

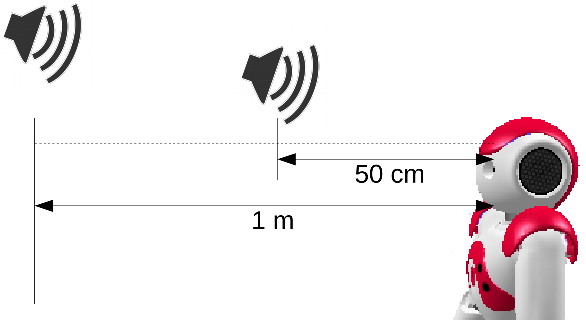


Fig. 2: Experimental setup for classification evaluation using NAO. For each of the experiments, speakers are positioned at a different distance from NAO. Depiction of NAO is not to scale.

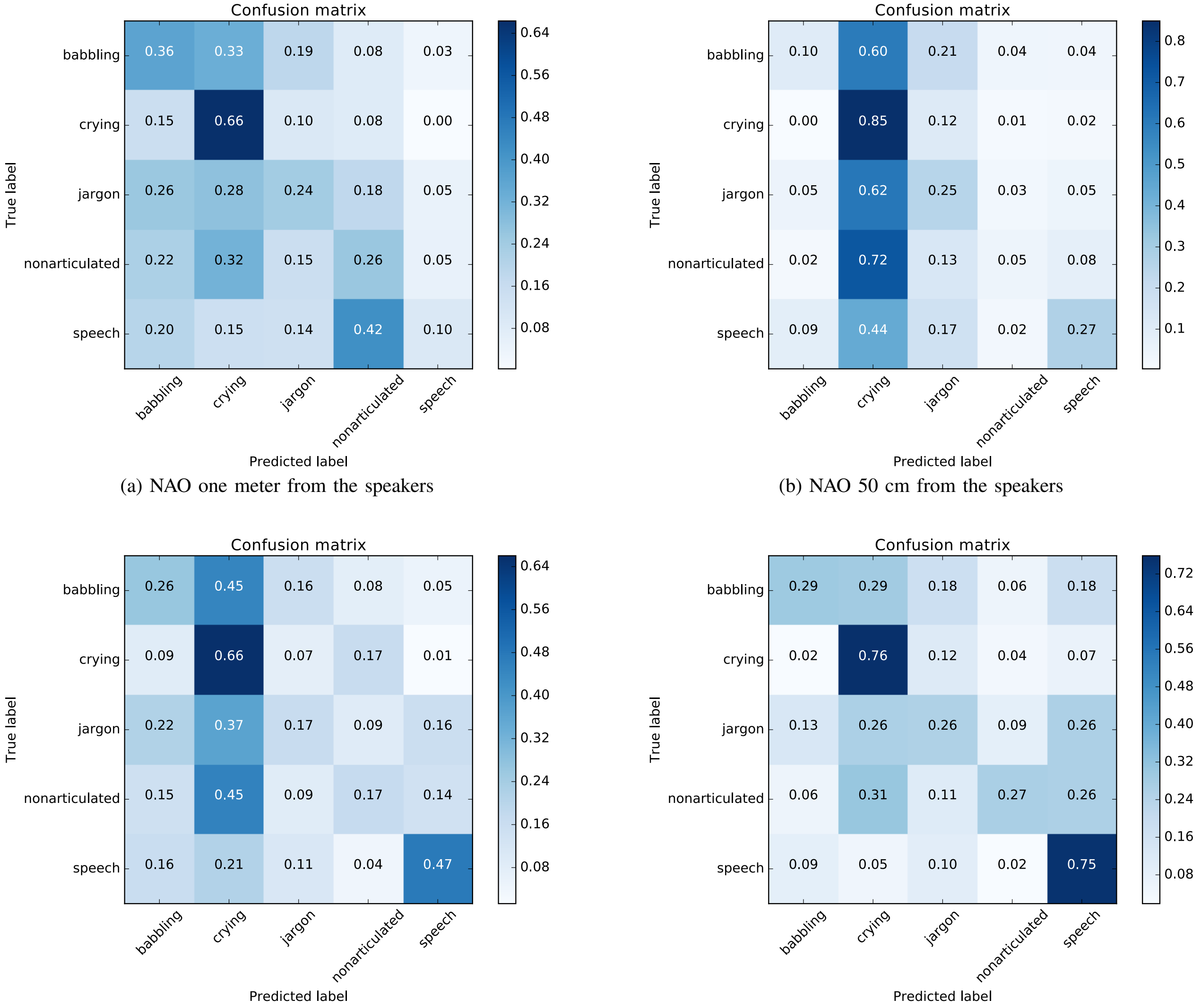
Next, we investigate whether the classifier performance improves when the noise is filtered. To filter the fan noise, we employ noise removal algorithm from Audacity[3]. First, several seconds of silence is recorded to estimate the spectral noise profile, then the noise is filtered from a sample by performing spectral subtraction of signal spectrum and estimated noise spectrum.

Classification results when the noise is removed, for both speaker distances, are shown in Figures 3c and 3d. Figure 3c shows that by removing the noise from a greater distance recordings, the classifier behavior is similar to that of a classifier for smaller distance but with noise (Figure 3b). That is, the classifier is mostly detecting crying, but one can also notice that the detection of speech is improved when noise is removed. Successful detection of speech is confirmed by examining the values in confusion matrix in Figure 3d.

For the last case, when the speaker is 50 cm away from the robot and the noise is removed, we also perform quantitative analysis by calculating precision, recall, accuracy and DOR. The results are summarized in Table IV.

From values in Table IV, but even more so by comparing Tables III and IV, we conclude that even when the noise is removed with the state-of-the-art algorithm, the performance of classifier in a realistic setting is severely deteriorated, especially for babbling, jargon and nonarticulated vocalizations

[3] Audacity(R) software is copyright (c) 1999-2018 Audacity Team. Web site: http://audacity.sourceforge.net/. It is free software distributed under the terms of GNU General Public License. The name Audacity(R) is a registered trademark of Dominic Mazzoni



(a) NAO one meter from the speakers

(b) NAO 50 cm from the speakers

(c) NAO one meter from the speakers, noise removed

(d) NAO 50 cm from the speakers, noise removed

Fig. 3: Classification results in form of normalized confusion matrices, obtained by classifying 300 samples recorded using NAO's microphones at different distances.

of a child. However, the robot is capable of successfully recognizing speech and crying, which are of critical importance in the robot-assisted ASD diagnostic protocol and indicate that the robot can be used to automatically evaluate some components of child vocalizations.

Software modules for NAO robot and tools that facilitated this analysis are available on Github:

- https://github.com/adore-hrzz/nao-sound-classification
- https://github.com/adore-hrzz/sound-classification-validation.

VI. Conclusions

We describe a classifier design for classifying child vocal behavior in the context of robot-assisted autism diagnostics. By using a stock open-source implementation of the random trees classifier and selecting only the most discriminating features for classification, we are able to meet the design goal of efficient real-time execution on the NAO robot. To the best of our knowledge, this is the first implementation of vocalization classification in the context of autism diagnostics, capable of running online on the NAO. The classifier is trained and tested on a custom dataset consisting of audio recordings collected during ASD diagnostics and intervention, recordings made by the NAO robot during pilot diagnostic sessions and publicly available recordings of child vocalizations. Classifier performance in terms of accuracy in laboratory setting is satisfactory, with an acceptable fraction of false positives and false negatives, making it a promising candidate for inclusion in a multi-modal robot-assisted autism diagnostics protocol. However, a bigger dataset is necessary for a more objective assessment of classifier performance.

Classification using front microphone of the NAO robot confirmed already reported negative effect of the computer

TABLE IV: Classification metrics obtained by classifying 300 samples recorded using NAO's microphones from distance of 50 cm and filtered using Audacity noise removal algorithm (see Figure 3d). Values for a class are obtained by looking at the classification in form of one-vs-all problem. Overall assessment in the last row is performed by micro-averaging the values across all classes.

Class	precision	recall	accuracy	DOR
Babbling	0.46	0.30	0.80	4.75
Crying	0.40	0.76	0.79	12.24
Jargon	0.37	0.26	0.74	2.47
Nonart.	0.5	0.27	0.85	7.28
Speech	0.6	0.75	0.79	12.10
Total	0.48	0.67	0.79	6.27

fan on audio signal processing. It was shown that the classifier can successfully recognize crying and speech, which provides encouragement for further investigations.

In our future work, we will focus on augmenting the dataset and evaluating additional audio features to improve classifier performance. We will also evaluate the classifier using Pepper robot, as it is deemed to be less prone to noise in the microphones while being compatible with NAO software. Furthermore, we are working towards fusing the results of vocalization classification with visual feedback to inform robot decisions during a pilot study with ASD and control groups.

Acknowledgment

The authors would like to thank Ines for immeasurable contribution in creating the dataset.

References

[1] A. P. Association, *Diagnostic and statistical manual of mental disorders (DSM-5®)*. American Psychiatric Publishing, 2013.

[2] P. Pennisi, A. Tonacci, G. Tartarisco, L. Billeci, L. Ruta, S. Gangemi, and G. Pioggia, "Autism and social robotics: A systematic review," *Autism Research*, vol. 9, no. 2, pp. 165–183, oct 2015. [Online]. Available: https://doi.org/10.1002/aur.1527

[3] L. Zwaigenbaum, S. Bryson, T. Rogers, W. Roberts, J. Brian, and P. Szatmari, "Behavioral manifestations of autism in the first year of life," *International Journal of Developmental Neuroscience*, vol. 23, no. 2-3, pp. 143–152, apr 2005. [Online]. Available: https://doi.org/10.1016/j.ijdevneu.2004.05.001

[4] J. Brisson, K. Martel, J. Serres, S. Sirois, and J.-L. Adrien, "Acoustic analysis of oral production of infants later diagnosed with autism and their mother," *Infant Mental Health Journal*, vol. 35, no. 3, pp. 285–295, 2014. [Online]. Available: http://dx.doi.org/10.1002/imhj.21442

[5] N. Chericoni, D. de Brito Wanderley, V. Costanzo, A. Diniz-Gonçalves, M. Leitgel Gille, E. Parlato, D. Cohen, F. Apicella, S. Calderoni, and F. Muratori, "Pre-linguistic vocal trajectories at 6 - 18 months of age as early markers of autism," *Frontiers in Psychology*, vol. 7, p. 1595, 2016. [Online]. Available: http://journal.frontiersin.org/article/10.3389/fpsyg.2016.01595

[6] K. Chenausky, C. Nelson, III, and H. Tager-Flusberg, "Vocalization rate and consonant production in toddlers at high and low risk for autism," *Journal of Speech, Language, and Hearing Research*, vol. 60, no. 4, pp. 865–876, 2017. [Online]. Available: http://dx.doi.org/10.1044/2016_JSLHR-S-15-0400

[7] E. Patten, K. Belardi, G. T. Baranek, L. R. Watson, J. D. Labban, and D. K. Oller, "Vocal patterns in infants with autism spectrum disorder: Canonical babbling status and vocalization frequency," *Journal of Autism and Developmental Disorders*, vol. 44, no. 10, pp. 2413–2428, 2014. [Online]. Available: http://dx.doi.org/10.1007/s10803-014-2047-4

[8] A. M. Plumb and A. M. Wetherby, "Vocalization development in toddlers with autism spectrum disorder," *Journal of Speech, Language, and Hearing Research*, vol. 56, no. 2, pp. 721–734, 2013. [Online]. Available: http://dx.doi.org/10.1044/1092-4388(2012/11-0104)

[9] E. Schoen, R. Paul, and K. Chawarska, "Phonology and vocal behavior in toddlers with autism spectrum disorders," *Autism Research*, vol. 4, no. 3, pp. 177–188, 2011.

[10] T. P. Gabrielsen, M. Farley, L. Speer, M. Villalobos, C. N. Baker, and J. Miller, "Identifying autism in a brief observation," *Pediatrics*, 2015. [Online]. Available: http://pediatrics.aappublications.org/content/early/2015/01/07/peds.2014-1428

[11] S. D. Mayes and S. L. Calboun, "Symptoms of autism in young children and correspondence with the dsm." *Infants & Young Children*, vol. 12, no. 2, pp. 90–97, 1999.

[12] S. J. Sheinkopf, P. Mundy, D. K. Oller, and M. Steffens, "Vocal atypicalities of preverbal autistic children," *Journal of Autism and Developmental Disorders*, vol. 30, no. 4, pp. 345–354, 2000. [Online]. Available: http://dx.doi.org/10.1023/A:1005531501155

[13] A. McAlpine, L. W. Plexico, A. M. Plumb, and J. Cleary, "Prosody in young verbal children with autism spectrum disorder," *Contemporary Issues in Communication Science & Disorders*, vol. 41, no. 41, pp. 120–132, 2014.

[14] C. Baltaxe, J. Q. Simmons, and E. Zee, "Intonation patterns in normal, autistic and aphasic children," in *Proceedings of the Tenth International Congress of Phonetic Sciences*. Foris Publications Dordrecht, The Netherlands, 1984, pp. 713–718.

[15] H. Sowden, J. Clegg, and M. Perkins, "The development of co-speech gesture in the communication of children with autism spectrum disorders," *Clinical Linguistics & Phonetics*, vol. 27, no. 12, pp. 922–939, 2013, pMID: 23944149. [Online]. Available: http://dx.doi.org/10.3109/02699206.2013.818715

[16] D. K. Oller, P. Niyogi, S. Gray, J. Richards, J. Gilkerson, D. Xu, U. Yapanel, and S. Warren, "Automated vocal analysis of naturalistic recordings from children with autism, language delay, and typical development," *Proceedings of the National Academy of Sciences*, vol. 107, no. 30, pp. 13 354–13 359, 2010.

[17] J. P. H. van Santen, R. W. Sproat, and A. P. Hill, "Quantifying repetitive speech in autism spectrum disorders and language impairment," *Autism Research*, vol. 6, no. 5, pp. 372–383, 2013. [Online]. Available: http://dx.doi.org/10.1002/aur.1301

[18] C. Lord, M. Rutter, P. Dilavore, and S. Risi, *Autism Diagnostic Observation Schedule*. Western Psychological Services, 2002.

[19] J. Kennedy, S. Lemaignan, C. Montassier, P. Lavalade, B. Irfan, F. Papadopoulos, E. Senft, and T. Belpaeme, "Child speech recognition in human-robot interaction," in *Proceedings of the 2017 ACM/IEEE International Conference on Human-Robot Interaction - HRI'17*. ACM Press, 2017. [Online]. Available: https://doi.org/10.1145/2909824.3020229

[20] C. Chen, A. Liaw, and L. Breiman, "Using Random Forest to Learn Imbalanced Data," Department of Statistics, University of Berkeley, Tech. Rep., 2004. [Online]. Available: http://www.stat.berkeley.edu/users/chenchao/666.pdf