VERY EARLY DETECTION OF AUTISM SPECTRUM DISORDERS BASED ON ACOUSTIC ANALYSIS OF PRE-VERBAL VOCALIZATIONS OF 18-MONTH OLD TODDLERS

João F. Santos[1], *Nirit Brosh*[2], *Tiago H. Falk*[1], *Lonnie Zwaigenbaum*[3], *Susan E. Bryson*[4], *Wendy Roberts*[5], *Isabel M. Smith*[4], *Peter Szatmari*[6] *and Jessica A. Brian*[2]

[1]INRS-EMT, University of Quebec, [2]Holland Bloorview Kids Rehabilitation Hospital, [3]University of Alberta, [4]Dalhousie University, [5]University of Toronto, [6]McMaster University

ABSTRACT

With the increasing prevalence of Autism Spectrum Disorders (ASD), very early detection has become a key priority research topic, as early interventions can increase the chances of success. Since atypical communication is a hallmark of ASD, automated acoustic-prosodic analyses have received prominent attention. Existing studies, however, have focused on verbal children, typically over the age of three (when many children may be reliably diagnosed) and as high as early teens. Here, an acoustic-prosodic analysis of pre-verbal vocalizations (e.g., babbles, cries) of 18-month old toddlers is performed. Data was obtained from a prospective longitudinal study looking at high-risk siblings of children with ASD who were also diagnosed with ASD, as well as low-risk age-matched typically developing controls. Several acoustic-prosodic features were extracted and used to train support vector machine and probabilistic neural network classifiers; classification accuracy as high as 97% was obtained. Our findings suggest that markers of autism may be present in pre-verbal vocalizations of 18-month old toddlers, thus may be used to assist clinicians with very early detection of ASD.

Index Terms— Autism, biomarker, prosody, SVM, PNN.

1. INTRODUCTION

Recent statistics from the Centers for Disease Control and Prevention show that in the United States, 1 in 88 children (1 in 54 boys) are diagnosed with Autism Spectrum Disorders (ASD), thus translating to over 2 million individuals in the U.S. and tens of millions worldwide. In fact, these prevalence rates have increased 17% annually in recent years. Autism is a pervasive developmental disorder which is related to a triad of impairments: i) atypical development in reciprocal social interaction, ii) atypical communication, and iii) restricted, stereotyped and repetitive behaviours [1]. Diagnosis typically occurs by the age of three, when children start showing delays in developmental milestones. Very early detection of ASD, however, allows for interventions to be administered at an early age, thus improving their chances of success [2]. As such, very early detection has become a key research priority.

Over the last decade, a number of different tools have been developed and/or refined for assessment and diagnosis of ASD in toddlers (e.g, ADOS-Toddler module [3]), and increasing developments are being made in very early detection, through the use of screening questionnaires [4] as well as clinical observation tools (e.g., the Autism Observation Scale for Infants (AOSI) [5]). Neuroimaging [6] and physiological [7] indices are also being explored as possible avenues for early detection. Notwithstanding, one area that has received prominent attention recently is that of acoustic-prosodic analysis of children's speech [8, 9, 10]. Particular emphasis has been placed on prosody, as clinicians have reported "exaggerated" and "monotonous" vocalizations and/or cries from children diagnosed with ASD, as well as atypical stress and intonation patterns. The majority of these studies have been based on data collected from verbal children recently diagnosed with ASD, thus at 30-36 months of age or older.

There is growing evidence, however, that early markers of autism can be present before the age of two [11] and possibly before 12 months of age [12]. For example, parents of children with ASD have reported difficulty in identifying emotional content in *pre-verbal* vocalizations [13]. Being able to detect acoustic differences in *pre-verbal* vocalizations could assist clinicians with very early detection of autism, but such analysis is lacking in the literature. This retrospective study aims to fill this gap. To achieve this goal, we used data derived from an ongoing longitudinal prospective Canadian Infant Sibling Study [14], where controls (low-risk) and younger siblings of probands with ASD (high-risk) were followed throughout infancy and independently diagnosed with ASD at an age of three. Here, focus will be placed on acoustic-prosodic features extracted from audio recordings obtained during the children's 18-month assessment. Pattern recognition is used to sift out salient features and to discriminate typically developing toddlers from those with ASD.

The remainder of this paper is organized as follows. Section 2 describes the experimental setup, Section 3 introduces the investigated classifiers and their design, Section 4 presents the experimental results, and Section 5 compares the findings with existing studies. Lastly, Section 6 concludes the paper.

Table 1. Participant demographics and vocalization duration

Group	Male/ Female	Age (months)	Vocalization duration (seconds)
Control	13/7	18 ± 0.23	127.0
ASD	15/8	18 ± 0.42	194.5

Table 2. Summary of vocalization instances for each group

Group	Speech	Babble	Other	Total
Control	594	687	1311	2592
ASD	797	509	624	1930

2. EXPERIMENTAL DESIGN

2.1. Database of infant pre-verbal vocalizations

The database used for the experiment was extracted from a set of videotaped ADOS - Module 1 sessions, which are part of an ongoing longitudinal prospective study on behavioral manifestation of autism in the first years of life (also known as the Canadian "Infant Sibling Study") [14]. This ADOS module was designed for preverbal children, who utter no more than single words or simple phrases [15]. The study monitors siblings of children diagnosed with ASD, who are considered to be in the ASD high-risk group (approximately 19% of siblings of individuals with ASD also exhibit the disorder [16]), as well as low-risk age-matched "control" children of families without a history of ASD. Participants are followed from the age of 6 months and every 3-12 months undergo a series of (re)assessments, including the AOSI, ADOS, and standardized developmental and language tests. A final diagnosis for ASD is done at the age of three by a blinded experienced clinician, utilizing gold standard clinical tools, such as medical history, ADOS, and the Autism Diagnostic Interview - Revised (ADI-R) [17].

For this prospective study, a subset of the Infant Sibling Database is used. More specifically, recordings from 43 participants during their 18-month assessment were utilized; 23 of which were diagnosed with ASD at the 36-month assessment and 20 age-matched typically developing (TD) controls. All participants in the study came from English speaking homes. Participant demographics are presented in Table 1. The ADOS sessions had durations ranging from 24 to 52 minutes. Audio content was extracted from the video recordings and children's vocalizations were manually segmented and labelled using the following classes: speech, babble, cry, squeal, laugh, and other. Table 1 also lists the total duration of vocalizations extracted for each group. Instances of vocalizations with overlapping adult speech (parents, clinician) were also documented and discarded from further analysis.

2.2. Automated Segmentation

Despite the database being manually segmented and labeled, an additional automated segmentation step needed to be performed to avoid acoustic-prosodic analysis of silent or ambient-noise only audio intervals (due to manual segmentation limitations), as well as to separate long bursts of vocalizations into smaller segments for analysis. Here, an energy-thresholding strategy similar to that employed in [8] was used. More specifically, the signal energy was calculated for 10 ms frames and if the average signal energy was 90% above a pre-defined threshold for 50 ms, a vocalization "start" would be detected. Conversely, when the average signal energy went below 10% of the energy threshold for 50 ms, a vocalization "end" would be detected. For our simulations, an energy threshold was chosen empirically based on simulations with a subset of the audio recordings. After automated segmentation, a total of 4522 vocalization instances were obtained for the two groups, 57% of which belonged to the TD control group. Table 2 summarizes the number of instances (not overlapped with adult speech) of each vocalization type for both groups. The column labeled "other" included vocalizations labelled as "laugh," "cry," "squeal," "whine," and "yell." As can be seen, the control group had a larger prominence of babble and 'other' vocalization types, whereas the ASD group had a larger percentage of speech instances. This may seem counterintuitive, but the control group had, e.g., three fewer participants and several speech instances that were discarded due to overlap with adult speech.

2.3. Acoustic-Prosodic Feature Extraction

In order to develop a classifier to discriminate between ASD and TD, a number of acoustic-prosodic parameters were extracted for each of the vocalization instances described in Table 2. Parameters were extracted using the VoiceSauce software from the UCLA SPAP Laboratory [18] and MATLAB scripts developed in-house. Statistics such as mean, standard deviation and range were computed for each vocalization instance for all parameters across all participants and used as "features" to train our classifiers. A description of the extracted parameters and their motivation is described below:

- Fundamental frequency (F_0): related to speech intonation and different crying patterns [19];
- First four formant frequencies and bandwidths (F_1, F_2, F_3, F_4, B_1, B_2, B_3, B_4): linked to vocal tract control and maturation of speech [20];
- Harmonic spectra locations and magnitudes (H_1, H_2, H_4, A_1, A_2, A_3) and the differences between spectral harmonic magnitudes and spectrum magnitude at the formant frequencies, corrected for the effects of the vo-

cal tract ($H_1^* - H_2^*$, $H_2^* - H_4^*$, $H_1^* - A_1^*$, $H_1^* - A_2^*$, and $H_1^* - A_3^*$): linked to the so-called Open Quotient, which is a measure of vocal quality (creakiness and breathiness), as well as the closing velocity of the vocal folds or muscle tensions [21];
- Subharmonic-to-Harmonic Ratio (SHR): related to speech quality [22];
- Root mean square signal energy: associated to speech loudness;
- Cepstral Peak Prominence (CPP): related to breathy and modal voice quality [21];
- Harmonic-to-Noise Ratio (HNR) for three frequency bands (0-500 Hz, 0-1500 Hz and 0-2500 Hz): a measure of the spectral noise level, shown to be correlated with breathiness [23];
- Jitter and shimmer: both measurements were shown to be correlated with breathiness, jitter is also correlated with hoarseness and roughness [24];
- Voiced ratio (ratio of number of voiced frames to the total vocalization duration): shown to discriminate between cries of healthy and pathological infants.

It is important to emphasize that pitch and formant range calculation algorithms were adjusted to be in the appropriate range for children's vocalizations. More specifically, F_0 was measured in the range 60-1600 Hz, while formants were measured considering a nominal F_1 frequency of 1250 Hz (the nominal F_1 frequency for a 7 cm vocal tract) [8]. All harmonic measurements have also a corrected value based on the formant frequencies and bandwidths. Segments, originally digitalized in 44 kHz, were downsampled to 16 kHz, and features were calculated for 25 ms frames with 10 ms frame shifts. Spectral subtraction was used to reduce the uncorrelated ambient noise prior to feature extraction.

3. CLASSIFIER DESIGN

In order to develop an automated tool to assist clinicians with very early detection of ASD, a classification system needs to be developed. In this study, we explored two leading supervised pattern recognition models, namely support vector machines (SVMs) and probabilistic neural networks (PNNs), with inherent feature selection capabilities. While a complete description of the classifiers is beyond the scope of this paper, a brief overview is given for the sake of completeness. The interested reader is referred to [25, 26] for more details.

Support vector machines for classification are based on the idea that data points will be separable by a hyperplane in some higher-dimensional feature space obtained via a so-called kernel mapping. Commonly used kernels include linear, polynomial, and radial basis functions (RBF). Probabilistic neural networks, in turn, are feed-forward networks derived from Bayes decision networks. Their structure is similar to a feed-forward neural network, consisting of an input layer, a class layer, and a decision (or output) node. The input layer has a neuron for each instance in the training set; for a given input, this layer computes the dot product between the input and the neuron's center point and then applies the radial basis function kernel to the result. The sum of the results of all pattern units corresponding to training points of the class i is given by:

$$f_i(x) = \frac{1}{(2\pi)^{p/2}\sigma^p M_i} \sum_{j=1}^{M_i} \exp \frac{-(x - x_{ij})^T (x - x_{ij})}{2\sigma^2}$$

where j is the pattern unit number, x_{ij} is the j^{th} training vector from class i, x is the test vector, M_i is the number of training vectors in class i, p is the dimension of vector x, and σ is the smoothing factor. As such, $f_i(x)$ is the sum of the radial basis function applied to the dot product of the input with the training vectors. This is called the Parzen probability density function estimator. A decision is made to classify a vector x as belonging to class i based on the Bayes strategy rule, i.e., choose the class which gives the maximum likelihood value. The only input parameter for a PNN is the smoothing factor which can be the same for all features, different for each feature but shared between groups, or different for each group.

In our experiments, the DTREG software was used for classifier evaluation. For the SVM, 4-fold grid search was used to find optimal cost and RBF kernel parameters whilst guarding against over-fitting. For the PNN, a smoothing parameter was used for each input feature, but the same values were used for both groups. A leave-one-out cross-validation strategy was used to obtain the optimal parameters and to guard against over-fitting. Classifier performance is measured using 10-fold cross-validation with all the samples and average accuracy, sensitivity, specificity, and the area under the ROC curve (AUC) are used as performance metrics. Three different feature combinations were explored, namely the *mean* features (feature combination FC1), the *mean* and *standard deviation* features (FC2), and the *mean*, *standard deviation* and *range* features (FC3) described in Section 2.3.

4. EXPERIMENTAL RESULTS

Experimental results for the two classifiers are reported in Table 3 for the three different feature combinations described above. As can be seen, PNN classifiers obtained improved performance over SVM under all three feature combination categories. For feature combination FC1, a relative improvement in accuracy of 5% was obtained with PNN over SVM. The PNN gains increased as more features were explored and for feature combination FC3, a relative improvement in accuracy of 55.6% could be achieved over SVM. Similar gains were observed across the other performance metrics.

During training of the SVMs and PNNs, feature importance can be calculated via the DTREG software, as the two classifiers inherently perform feature selection. An in-depth

Table 3. Performance comparison of SVM and PNN classifiers. Columns labelled "Acc," "Sens," and "Spec" correspond to classifier accuracy, sensitivity, and specificity, respectively, averaged over ten cross-validation trials.

Classifier	FC	Acc (%)	Sens (%)	Spec (%)	AUC
PNN	1	83.1	87.0	85.0	0.90
	2	93.0	91.3	95.0	0.92
	3	97.7	95.6	100.0	0.97
SVM	1	79.1	82.6	75.0	0.87
	2	67.4	82.6	50.0	0.71
	3	62.8	69.6	55.0	0.66

investigation into the top-selected features showed that for the PNN classifier the *means* of $H_1^* - A_3^*$ and H_4, the *standard deviation* of H_1, and the *range* of F_0 and shimmer parameters obtained high importance. For the SVM, in turn, the *means* of the energy, H_4, $H_1^* - A_3^*$, jitter and A_2 parameters had the highest importance, which may account for the drop in performance for the SVM with feature combinations FC2 and FC3 (relative to FC1), as they included standard deviation and range-based features. While PNN and SVM classifiers inherently do feature selection, perhaps a dedicated feature selection step would lead to better results, especially in the SVM case. This, however, is not within the scope of this paper.

5. DISCUSSION AND RELATIONSHIP TO PREVIOUS FINDINGS

Results reported herein suggest that sufficient discriminability exists in pre-verbal vocalizations of toddlers to detect very early indicators of ASD, as early as 18 months of age. This is an important finding, as diagnostic behavioural traits may only be exhibited after two years of age and interventions have shown to be more effective when initiated at an early stage. As such, the developed tools may be used by clinicians to assist with very early diagnosis of ASD. Moreover, the majority of the discriminant features found by the classifiers were related to measures of vocal quality, particularly $H_1^* - A_3^*$ which was selected by both classifiers. This result may point to a difference in vocal quality between TD and ASD groups. The range of the fundamental frequency also stood out as an important feature. Statistics of F_0 have been shown in the previous literature to be a strong discriminant factor for verbal children over the age of three.

In [8], for example, the age range of the ASD group was 16-48 months, with an average of approximately 32 months. In their study, linear discriminant analysis and linear logistic regression were used to obtain sensitivity and specificity levels of 75% and 98%, respectively, using a 12-dimensional feature set comprised of pitch, formant, spectral tilt, harmonics and vocalization duration parameters. In [27], a Naive Bayes classifier trained only with pitch related statistics was used and achieved 74% accuracy when classifying between TD and ASD groups; the age range was $4 - 8.5$ years with a mean of 6.4 years. More recently, in [9], phonetic-level features were also considered and found to be discriminant; results were obtained, however, for children within an age range of $5.8 - 14.7$ years (mean= 9.8 years).

As can be seen, the age range of the participants in the present study is the lowest reported in the literature investigating acoustic-prosodic analysis for markers of autism, thus it truly represents an *early* marker that may assist clinicians with diagnosis. In fact, the tight age range of the participants may have helped with obtaining such high discrimination power. Children's vocalizations are known to change continuously during childhood and performing acoustic-prosodic analysis over a wide age range (e.g., $6 - 15$ years, as in [9]) may be a more challenging task, as natural age-related acoustic changes and variability need to be accounted for. By comparing data from only 18-month old toddlers, such variability was mitigated and allowed the classifiers to place focus on existing ASD discriminative features.

6. CONCLUSIONS

In this paper, the first steps towards the development of an assessment tool to assist in very early recognition of autism were taken. Acoustic-prosodic features encompassing pitch, formant, energy, harmonics, and vocal quality (e.g., breathiness) were extracted from audio recordings obtained from a prospective study that looked at high-risk siblings of children with ASD (and who were later also diagnosed with ASD) and typically-developing controls. Participant data was obtained during their routine 18-month assessment. Two classifiers, namely support vector machine and probabilistic neural network, were designed and tested using 10-fold cross validation. The PNN classifier achieved accuracy, sensitivity, and specificity above 95% and outperformed the SVM by as much as 50%. The salient features selected by the classifiers are inline with those previously reported in the literature, thus suggest that markers of autism may be present at early stages of life. Notwithstanding, further work is still needed to investigate if the obtained results will remain once toddlers with other non ASD-related language development disorders are included in the analysis. Lastly, our ongoing investigations include extending the present analysis to audio data collected during the children's 9- and 12-month routine assessments.

7. ACKNOWLEDGEMENTS

This work was supported by funding from the National Science and Engineering Research Council of Canada, the Canadian Institutes of Health Research, Autism Speaks, and NeuroDevNet. We wish to thank the children and their families for participating in the Canadian Infant Sibling Study.

8. REFERENCES

[1] American Psychiatric Association, *Diagnostic and Statistical Manual of Mental Disorders DSM-IV-TR Fourth Edition*, Amer Psychiatric Pub, 4th edition, June 2000.

[2] E Fenske et al., "Age at intervention and treatment outcome for autistic children in a comprehensive intervention program," *Analysis Interv. Developm. Disabil.*, vol. 5, no. 1-2, pp. 49–58, Jan. 1985.

[3] C Lord et al., *Autism Diagnostic Observation Schedule, 2nd Edition (ADOS-2) Manual (Part II): Toddler Module*, Western Psychological Services, CA, 2012.

[4] S. Swinkels et al., "Screening for autistic spectrum in children aged 14-15 months: Development of the early screening of autistic traits questionnaire," *J. Autism Dev. Disord.*, vol. 36, no. 6, pp. 723–732, 2006.

[5] S Bryson et al., "The Autism Observation Scale for Infants: scale development and reliability data.," *J. Autism Dev. Disord.*, vol. 38, no. 4, pp. 731–8, Apr. 2008.

[6] F. Duffy and H. Als, "A stable pattern of eeg spectral coherence distinguishes children with autism from neuro-typical controls-a large case control study," *BMC medicine*, vol. 10, no. 64, 2012.

[7] T. Chaspari, C.-C. Lee, and S. Narayanan, "Interplay between verbal response latency and physiology of children with autism during ECA interactions," in *InterSpeech*, 2012.

[8] D K Oller et al., "Automated vocal analysis of naturalistic recordings from children with autism, language delay, and typical development.," *Proc. Nat. Acad. Sci. U.S.A.*, vol. 107, no. 30, pp. 13354–9, July 2010.

[9] D Bone et al., "Spontaneous-Speech Acoustic-Prosodic Features of Children with Autism and the Interacting Psychologist," in *InterSpeech*, 2012, pp. 3–6.

[10] J. McCann and S Peppe, "Prosody in autism spectrum disorders: a critical review.," *Intl. Journal Lang. Commun. Disorders*, vol. 38, no. 4, pp. 325–50, 2003.

[11] S Jonsdottir et al., "Children diagnosed with autism spectrum disorder before or after the age of 6 years," *Research in Autism Spectrum Disorders*, vol. 5, no. 1, pp. 175–184, 2011.

[12] M.T. Kishore and A. Basu, "Early concerns of mothers of children later diagnosed with autism: Implications for early identification," *Research in Autism Spectrum Disorders*, vol. 5, no. 1, pp. 157–163, 2011.

[13] R Paul et al., "Perception and Production of Prosody by Speakers with Autism Spectrum Disorders," *Journal Autism Develop Disord*, vol. 35, no. 2, pp. 205–220, Apr. 2005.

[14] S. Georgiades et al., "A prospective study of autistic-like traits in unaffected siblings of probands with autism spectrum disorder," *Arch Gen Psych*, 2012.

[15] C Lord et al., "The autism diagnostic observation schedule-generic: a standard measure of social and communication deficits associated with the spectrum of autism.," *J. Autism Dev. Disord.*, vol. 30, no. 3, pp. 205–23, June 2000.

[16] S Ozonoff et al., "Recurrence risk for autism spectrum disorders: a Baby Siblings Research Consortium study.," *Pediatrics*, vol. 128, no. 3, pp. e488–95, Sept. 2011.

[17] C Lord, M Rutter, and A Le Couteur, "Autism Diagnostic Interview-Revised: a revised version of a diagnostic interview for caregivers of individuals with possible pervasive developmental disorders.," *J. Autism Dev. Disord.*, vol. 24, no. 5, pp. 659–85, Oct. 1994.

[18] Y. Shue, P. Keating, C. Vicenik, and K. Yu, "VoiceSauce: a program for voice analysis," in *Intl Congress of Phonetic Sciences*, 2010.

[19] G. Esposito and P. Venuti, "Developmental changes in the fundamental frequency (f0) of infants' cries: a study of children with Autism Spectrum Disorder," *Early Child Development and Care*, vol. 180, no. 8, pp. 1093–1102, Sept. 2010.

[20] H. Vorperian and R. Kent, "Vowel acoustic space development in children: a synthesis of acoustic and anatomic data.," *Journal of speech, language, and hearing research : JSLHR*, vol. 50, no. 6, pp. 1510–45, Dec. 2007.

[21] P. Keating and C. Esposito, "Linguistic voice quality," *UCLA Working Papers in Phonetics*, 2006.

[22] X. Sun, "Pitch determination and voice quality analysis using subharmonic-to-harmonic ratio," *Proc. ICASSP*, pp. 333–336, 2002.

[23] Y. Shue, G. Chen, and A. Alwan, "On the interdependencies between voice quality, glottal gaps, and voice-source related acoustic measures," *Interspeech*, 2010.

[24] A. McAllister, J. Sundberg, and S. Hibi, "Acoustic measurements and perceptual evaluation of hoarseness in children's voices," *Logopedics Phoniatrics Vocology*, vol. 23, no. 1, pp. 27–38, Jan. 1998.

[25] I. Steinwart and A. Christmann, *Support vector machines*, Springer, 2008.

[26] D Sprecht, "Probabilistic neural networks for classification, mapping and associative memory," *ICNN-88 Conf. Proc*, pp. 525–532, 1988.

[27] G. Kiss, J. van Santen, E. Prud, and L. Black, "Quantitative Analysis of Pitch in Speech of Children with Neurodevelopmental Disorders," in *InterSpeech*, 2012.