Earlier identification of children with autism spectrum disorder: an automatic vocalisation-based approach

Florian B. Pokorny, Björn Schuller, Peter B. Marschik, Raymond Brueckner, Pär Nyström, Nicholas Cummins, Sven Bölte, Christa Einspieler, Terje Falck-Ytter

Angaben zur Veröffentlichung / Publication details:

Pokorny, Florian B., Björn Schuller, Peter B. Marschik, Raymond Brueckner, Pär Nyström, Nicholas Cummins, Sven Bölte, Christa Einspieler, and Terje Falck-Ytter. 2017. "Earlier identification of children with autism spectrum disorder: an automatic vocalisation-based approach." In *Situated interaction: 18th Annual Conference of the International Speech Communication Association (INTERSPEECH 2017), Stockholm, Sweden, 20-24 August 2017; Volume 1*, 309–13. Red Hook, NY: Curran Associates, Inc. https://doi.org/10.21437/interspeech.2017-1007.

INTERSPEECH 2017
August 20–24, 2017, Stockholm, Sweden

Earlier Identification of Children with Autism Spectrum Disorder: An Automatic Vocalisation-based Approach

Florian B. Pokorny[1;2;3], *Björn W. Schuller*[4;5], *Peter B. Marschik*[1;3;6], *Raymond Brueckner*[2;7], *Pär Nyström*[8], *Nicholas Cummins*[4], *Sven Bölte*[6;9], *Christa Einspieler*[1], *Terje Falck-Ytter*[6;8]

[1]Research Unit iDN – interdisciplinary Developmental Neuroscience, Department of Phoniatrics, Medical University of Graz, Austria
[2]Machine Intelligence & Signal Processing group (MISP), MMK, Technical University of Munich, Germany
[3]Brain, Ears & Eyes – Pattern Recognition Initiative (BEE-PRI), BioTechMed-Graz, Austria
[4]Chair of Complex & Intelligent Systems, University of Passau, Germany
[5]Machine Learning Group, Department of Computing, Imperial College London, UK
[6]Center of Neurodevelopmental Disorders (KIND), Department of Women's and Children's Health, Karolinska Institutet, Stockholm, Sweden
[7]Nuance Communications, Ulm, Germany
[8]Uppsala Child and Baby Lab, Department of Psychology, Uppsala University, Sweden
[9]Child and Adolescent Psychiatry, Center for Psychiatry Research, Stockholm County Council, Sweden

florian.pokorny@medunigraz.at

Abstract

Autism spectrum disorder (ASD) is a neurodevelopmental disorder usually diagnosed in or beyond toddlerhood. ASD is defined by repetitive and restricted behaviours, and deficits in social communication. The early speech-language development of individuals with ASD has been characterised as delayed. However, little is known about ASD-related characteristics of pre-linguistic vocalisations at the feature level. In this study, we examined pre-linguistic vocalisations of 10-month-old individuals later diagnosed with ASD and a matched control group of typically developing individuals (N = 20). We segmented 684 vocalisations from parent-child interaction recordings. All vocalisations were annotated and signal-analytically decomposed. We analysed ASD-related vocalisation specificities on the basis of a standardised set (eGeMAPS) of 88 acoustic features selected for clinical speech analysis applications. 54 features showed evidence for a differentiation between vocalisations of individuals later diagnosed with ASD and controls. In addition, we evaluated the feasibility of automated, vocalisation-based identification of individuals later diagnosed with ASD. We compared linear kernel support vector machines and a 1-layer bidirectional long short-term memory neural network. Both classification approaches achieved an accuracy of 75% for subject-wise identification in a subject-independent 3-fold cross-validation scheme. Our promising results may be an important contribution en-route to facilitate earlier identification of ASD.

Index Terms: autism spectrum disorder, early identification, infant vocalisation analysis, speech-language pathology

1. Introduction

Autism spectrum disorder (ASD) is a neurodevelopmental disorder defined by patterns of repetitive and restricted behaviours, and persistent deficits in the socio-communicative domain [1, 2]. According to recent estimates from the Autism and Developmental Disabilities Monitoring (ADDM) Network of the Centers for Disease Control and Prevention (CDC), ASD has a prevalence of about 1 in 68 children and occurs about 4.5 times more common in males than in females [3]. There is an increased recurrence risk of up to 18% for children with older siblings diagnosed with ASD [4, 5]. ASD is not curable, but there is increased evidence of early intervention being beneficial for affected individuals [6]. Early intervention requires early detection, and even though progress has been made in ASD screening, ASD is usually not diagnosed before toddlerhood [3].

Early speech-language development of children with ASD has repeatedly been characterised as delayed and deviant: Late onset canonical babbling, reduced volubility, and/or monotony in intonation are among the reported signs (e. g., [7, 8, 9, 10]). Only a few studies focussed on early ASD-related vocalisation specificities at the feature level (e. g., [11]). However, most of these studies assessed crying vocalisations by extracting single acoustic features, such as fundamental frequency or (cry) duration (e. g., [12, 13, 14]).

This study builds on our experience in depicting early speech-language phenomena in individuals with (neuro)developmental disorders, such as fragile X syndrome or Rett syndrome (e. g., [15, 16, 17]). In the present study, we aimed to gain better understanding of ASD-related characteristics of pre-linguistic vocalisations in order to identify potential early acoustic markers. Encouraged by our promising results for vocalisation-based early identification of infants with Rett syndrome [18], we aimed to evaluate the feasibility of automatic identification of infants later diagnosed with ASD.

To the best of our knowledge, this is one of the first attempts to explore ASD-related specificities in pre-linguistic vocalisations based on a wide range of acoustic parameters. Our ultimate goal is to initiate further research on signal-analytic approaches that might – one day – lead to an earlier identification of ASD, and thus, to an earlier entry into intervention.

http://dx.doi.org/10.21437/Interspeech.2017-1007

2. Methods

In this study, we retrospectively analysed data collected in the framework of the project EASE[1] (Early Autism Sweden). EASE is a longitudinal study following infants at increased risk for ASD, i. e., younger siblings of children with an ASD diagnosis.

2.1. Material

We reviewed audio-video recordings of 20 infants (10 males) in parent-child interaction settings[2]. Each recording had a length of 12 minutes resulting in a total audio-video length of 240 minutes. The parent-child interaction settings were recorded in a closed room. Parents were instructed to play with their children on a mat with toys as they would do at home. The study personnel left the room for the recordings. Apart from parental voice and sounds caused by playful manipulation of toys, no background noises were present. For audio-video recording, Panasonic HC-V700 cameras were mounted at the ceiling in the corners of the room in order to guarantee that a multitude of infant behaviours can be continuously assessed irrespective of the infant's orientation in the room. Audio information was extracted from one of the cameras and converted to single-channel 44.1 kHz AAC format for further analyses. At the time of recording, all infants included in this study were 10 months old. Ten infants (five males) stem from a Swedish population at heightened risk for ASD (younger siblings of children already diagnosed with ASD) and were diagnosed with ASD at 3 years of age using DSM-5 criteria [1]. Henceforth, we refer to this group of infants as "ASD group". The remaining ten infants (five males) stem from a Swedish low risk population with a normal outcome at 3 years of age. Henceforth, this group of typically developing (TD) infants is referred to as "TD group". Side note: One female subject from the ASD group and two male subjects from the TD group (temporarily) had pacifiers in their mouths during the recordings.

2.2. Segmentation

All 20 audio-video recordings were manually segmented for infant vocalisations using the video coding tool Noldus Observer XT[3]. Vocalisation boundaries were set on the basis of distinct vocal breathing groups [19]. Vegetative sounds, such as smacking sounds, breathing sounds, or hiccups, were excluded. We identified and segmented a total of 684 pre-linguistic vocalisations (Table 1) with a mean vocalisation length of 2.01s (median = 1.39s). A proportion of 37.9% (259) of the vocalisations were segmented from the material of the ASD group and a proportion of 62.1% (425) from the material of the TD group.

In order to estimate each subject's level in speech-language development, we annotated all types of vocalisations that were more complex than single canonical syllables (e. g., canonical babbling) usually well in place in TD infants at 10 months of age. Annotation was based on the Stark Assessment of Early Vocal Development-Revised (SAEVD-R) [20]. In Table 1 the occurrence of an annotated vocalisation type is specified per subject.

Segmentation and annotation was done by the first author, who was not informed about each subject's group membership (ASD group or TD group) during the segmentation and annotation process.

[1] http://www.earlyautism.se

[2] The ethical review boards of the various centres approved the study.

[3] http://www.noldus.com

Table 1: *Specification of gender, number of segmented vocalisations ($\#_{voc}$), vocalisation rate in vocalisations per minute ($\#_{voc}/min$), and assignment to partition in the 3-fold cross-validation scheme for ten ASD subjects and ten TD subjects. The vocalisation rate is rounded to two decimal points. ('*' indicates that at least one of the produced vocalisations was more complex than a single canonical syllable.)*

Subject	Gender	$\#_{\text{voc}}$	$\#_{\text{voc}}/min$	Partition
ASD01	f	10	0.83	3
ASD02	m	*38	3.17	1
ASD03	f	*42	3.50	2
ASD04	f	*26	2.17	1
ASD05	m	19	1.58	3
ASD06	f	*31	2.58	1
ASD07	m	28	2.33	2
ASD08	m	9	0.75	2
ASD09	m	17	1.42	3
ASD10	f	39	3.25	3
Σ		259		
TD01	m	18	1.50	3
TD02	m	45	3.75	3
TD03	m	35	2.92	2
TD04	f	15	1.25	1
TD05	f	*98	8.17	2
TD06	f	*29	2.42	3
TD07	m	*59	4.92	1
TD08	f	52	4.33	1
TD09	m	*37	3.08	2
TD10	f	37	3.08	3
Σ		425		

2.3. Analysis

2.3.1. Volubility

We calculated each subject's vocalisation rate in vocalisations per minute (see Table 1) and applied the Mann-Whitney U-test (group-specific data were not normally distributed; $\alpha = 0.05$) to analyse the difference in volubility between the ASD group and the TD group.

2.3.2. Feature extraction

To build the basis for acoustic vocalisation analysis and classification experiments, we extracted acoustic features from the vocalisations using the open-source tool kit openSMILE[4] [21, 22]. We used the recently defined extended Geneva Minimalistic Acoustic Parameter Set (eGeMAPS) [23] representing a basic standard feature set of 88 acoustic parameters selected for a wide range of automatic voice analysis applications, including applications of clinical speech analysis. The set comprises statistical functionals calculated for a compact composition of 25 frequency-related, energy-related, and spectral low-level descriptors that are extracted on a short-term basis and smoothed over time [23].

2.3.3. Feature analysis

Using the Mann-Whitney U-test (group-specific feature values were not normally distributed; $\alpha = 0.05$), for each feature we tested the null hypothesis that feature values extracted from vo-

[4] http://www.audeering.com

calisations of the ASD group and feature values extracted from vocalisations of the TD group were samples of continuous distributions with equal medians. The effect size r (z-value divided by the square root of the number of samples) was calculated for each significant group difference.

2.3.4. Classification

The binary classification paradigm ASD versus TD was tested on the basis of a subject-independent 3-fold cross-validation scheme. Therefore, we split our dataset into three partitions of subjects matched for gender and diagnosis, and containing an approximately equal number of vocalisations. We further tried to obtain a constant balance between the number of vocalisations of the ASD group and the TD group in each partition. As indicated in Table 1, we created two partitions (partitions 1 and 2) of six subjects each (three ASD and three TD) and one partition (partition 3) of eight subjects (four ASD and four TD). Partition 1 contained 221 vocalisations (95 ASD and 126 TD), partition 2 contained 249 vocalisations (79 ASD and 170 TD), and partition 3 contained 214 vocalisations (85 ASD and 129 TD).

For each of the three validation runs, we used one partition as training set, another as development set, and the remaining partition as test set. Throughout the 3-fold cross-validation procedure, each partition was used as training, development, and test partition exactly one time.

To evaluate the feasibility of vocalisation-based identification of class ASD versus class TD, we chose linear kernel support vector machines (SVMs) as baseline classification approach. SVMs are known to be robust, not sensitive to feature overfitting, and achieved good recognition performances in similar classification tasks (e. g., [18]). For SVM training, we applied the sequential minimal optimisation algorithm using the widely spread data mining tool kit Weka[5] [24]. For each of the three validation runs, SVMs were trained on the basis of vocalisations in the training partition. Next, the complexity parameter $C \in \{1, 10^{-1}, 10^{-2}, 10^{-3}, 10^{-4}, 10^{-5}\}$ was determined to achieve the best unweighted average recall (UAR) on vocalisations in the development set. Finally, the training and development partitions were merged to an ultimate training partition and validation was done on the basis of vocalisations in the test partition.

As a topical alternative to the baseline classification approach, we employed recurrent, bi-directional long short-term memory neural networks (BLSTM NNs). BLSTM NNs have been proven to be powerful models in many speech-related areas (e. g., [25, 26, 27]) and the incurred delay is of no concern in the task considered in this study. We used the vanilla BLSTM implementation described in [28] and trained our models with TensorFlow[6] [29] on 10 ms time steps utilising the first-order gradient-based Adam optimisation algorithm [30]. Interestingly, we found that cross-entropy loss worked best when averaging the posterior probabilities across the full utterance. Since Adam is an adaptive-learning rate algorithm, we followed a patience-based approach, where we stopped training, if there was no improvement of the UAR on the development set for more than five epochs and chose the best model. Finally, we did a grid search to determine the optimum number of cells and layers and found that, a single-layer BLSTM NN with eight cells performed best.

Building on vocalisation-wise classification decisions, we additionally generated subject-wise judgements. An ASD ratio was computed for each subject by counting all vocalisations of this subject classified as class ASD, divided by the total number of vocalisations of this subject. In case this ASD ratio exceeded a specific threshold, the subject was judged to be an infant from the ASD group. In a first approach, the threshold was set to 0.5 (majority voting). In a second approach, the threshold was optimised in the 3-fold cross-validation procedure on the basis of the respective merged training and development partitions of each validation run. As optimisation criterion for the threshold we defined a maximum distance between the mean ASD ratio for subjects from the ASD group versus the mean ASD ratio for subjects from the TD group.

[5] http://www.cs.waikato.ac.nz

[6] https://www.tensorflow.org

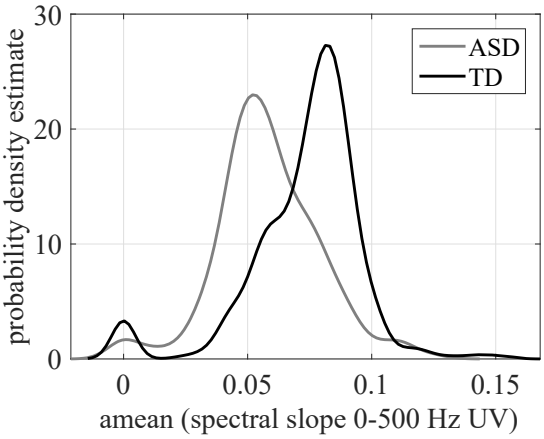

Figure 1: *Comparison between ASD group (grey) and TD group (black) by means of probability density estimates of feature values extracted from either group's vocalisations. (amean = arithmetic mean; UV = for unvoiced segments)*

3. Results

Eight subjects (four ASD and four TD) produced vocalisation types more complex than single canonical syllables. Moreover, there was no significant difference in volubility between the ASD group and the TD group ($p = 0.104$).

For 54 of 88 analysed features, significant differences between the feature value distributions related to vocalisations from the ASD group versus vocalisations from the TD group could be found. Table 2 lists the top ten features according to the effect size r. For five features, moderate effects ($|r| > 0.3$) were identified. For both the ASD group and the TD group, value distributions of the feature with the highest differentiation effect are shown in Figure 1.

For vocalisation-wise classification, the SVM approach and the BLSTM NN approach showed similar performances. Using SVMs, we achieved the highest UAR of $64.5\% \pm 3.3\%$ over the three validation runs. Detailed results including confusion matrices for both the SVM approach and the BLSTM NN approach are given in Table 3.

In our subject-wise judgement scenario, both the SVM approach and the BLSTM NN approach achieved an accuracy of 75% when using optimised decision thresholds. Eight of ten infants later diagnosed with ASD could be correctly assigned to the ASD group. Seven of ten TD infants could be correctly assigned to the TD group. A detailed overview of the subject-wise decisions is given in Table 4. Side note: One of the incorrectly assigned TD subjects was using a pacifier during the recording.

Table 2: *Top ten acoustic features for differentiation between the ASD group and the TD group according to the (magnitude of the) effect size r. r is rounded to four decimal places. (amean = arithmetic mean; f0 = fundamental frequency; MFCC = Mel-frequency cepstral coefficient; pctlR = percentile range; stddNorm = normalised standard deviation, UV = for unvoiced segments)*

Feature	r
$amean(spectral\ slope\ 0-500\ Hz\ UV)$	0.3966
$stddNorm(MFCC\ 4)$	0.3402
$mean\ unvoiced\ segment\ length$	−0.3371
$amean(Hammarberg\ index\ UV)$	0.3369
$amean(spectral\ slope\ 500-1500\ Hz\ UV)$	−0.3355
$amean(f0\ semitone\ from\ 27.5\ Hz)$	0.2932
$amean(alpha\ ratio\ UV)$	−0.2911
$voiced\ segments\ per\ second$	0.2551
$pctlR20-80(f0\ semitone\ from\ 27.5\ Hz)$	0.2540
$stddNorm(MFCC\ 1)$	−0.2502

Table 3: *Classification results of subject-independent 3-fold cross-validation in form of class-specific numbers of test vocalisations (in-)correctly classified as class ASD or TD (confusion matrix), and mean and standard deviation (SD) of weighted and unweighted average recall (WAR and UAR) for SVMs and the BLSTM NN. WAR and UAR are given in [%] and rounded to one decimal place.*

	SVM		BLSTM NN	
$classified\ as \rightarrow$	ASD	TD	ASD	TD
ASD	131	128	155	104
TD	83	342	141	284
	WAR	UAR	WAR	UAR
mean	69.0	64.5	70.7	62.9
SD	2.6	3.3	2.6	2.0

4. Discussion

In this study, we could not confirm a deviant volubility in 10-month old individuals later diagnosed with ASD. Nonetheless, we were able to provide evidence of acoustic parameters in pre-linguistic vocalisations as potential early markers for ASD and ASD-related neural mechanisms on the basis of our dataset. Our classification experiments could demonstrate basic feasibility of automated vocalisation-based identification of ASD. However, the impact of our results is limited due to the small number of subjects included in this first feasibility study. A significant performance difference between an SVM approach and a deep learning approach might only become evident on a dataset with a considerably higher number of vocalisations.

The number of incorrectly classified vocalisations of subjects from the ASD group indicates that, a certain proportion of vocalisations of an individual later diagnosed with ASD does not bear atypicalities in the acoustic signal domain. Therefore, compared to a vocalisation-wise classification approach, subject-wise judgements based on evaluating a set of vocalisations might be a more realistic scenario for practical applications. Imagine for example, a scenario in which early preventative judgements could be made on individuals on the basis of vocalisations produced during standard paediatric examinations in the first year of life.

Table 4: *Subject-wise judgements based on vocalisation-wise classification decisions of SVMs and the BLSTM NN by checking ASD ratios (R_{ASD}) against a 50% threshold ($Th_{0.5}$) compared to an optimised threshold (Th_{opt}). ASD ratios are rounded to two decimal places. ('✓' indicates that the subject was assigned to correct group. '✗' indicates that the subject was assigned to incorrect group. Th_{opt} for SVMs: 0.33 for partition 1; 0.42 for partition 2; 0.31 for partition 3. Th_{opt} for the BLSTM NN: 0.46 for partition 1; 0.44 for partition 2; 0.39 for partition 3.)*

	SVM			BLSTM NN		
Subject	R_{ASD}	$Th_{0.5}$	Th_{opt}	R_{ASD}	$Th_{0.5}$	Th_{opt}
ASD01	0.40	✗	✓	0.60	✓	✓
ASD02	0.68	✓	✓	0.71	✓	✓
ASD03	0.52	✓	✓	0.67	✓	✓
ASD04	0.77	✓	✓	0.81	✓	✓
ASD05	0.53	✓	✓	0.63	✓	✓
ASD06	0.45	✗	✓	0.45	✗	✗
ASD07	0.00	✗	✗	0.29	✗	✗
ASD08	0.33	✗	✗	0.78	✓	✓
ASD09	0.59	✓	✓	0.41	✗	✓
ASD10	0.56	✓	✓	0.64	✓	✓
TD01	0.83	✗	✗	0.72	✗	✗
TD02	0.04	✓	✓	0.31	✓	✓
TD03	0.06	✓	✓	0.23	✓	✓
TD04	0.00	✓	✓	0.00	✓	✓
TD05	0.06	✓	✓	0.17	✓	✓
TD06	0.38	✓	✗	0.52	✗	✗
TD07	0.54	✗	✗	0.54	✗	✗
TD08	0.06	✓	✓	0.25	✓	✓
TD09	0.22	✓	✓	0.35	✓	✓
TD10	0.11	✓	✓	0.16	✓	✓
Accuracy		70%	75%		70%	75%

5. Conclusions and outlook

In this study, we elaborated on the potential of an automated pre-linguistic vocalisation-based analysis approach for an early acoustic identification of individuals with ASD. Our examinations were based on a small but well gender- and family-language-balanced sample of subjects later diagnosed with ASD and TD controls recorded in semi-standardised parent-child interaction settings.

En-route to – potentially – enabling a reliable earlier identification of individuals with ASD, a number of more fine-grained studies on larger datasets are warranted. From a technological point of view, the capabilities of various acoustic feature sets, feature pre-processing strategies (e. g., bag-of-words processing [31]), and different classification approaches (e. g., different deep learning architectures) should be evaluated in detail.

6. Acknowledgements

The authors acknowledge funding from the Austrian National Bank (OeNB; P16430), the Austrian Science Fund (FWF; P25241), BioTechMed-Graz, the EU's H2020 Programme via RIA #688835 (DE-ENIGMA), the Stiftelsen Riksbankens Jubileumsfond, and the Swedish Research Council. We would like to thank Mathias Egger for assistance in linguistic matters. The authors express their gratitude to all the families participating in the longitudinal study EASE.

7. References

[1] A. P. Association, *Diagnostic and Statistical Manual of Mental Disorders (DSM-5®)*. American Psychiatric Pub, 2013.

[2] F. R. Volkmar, C. Lord, A. Bailey, R. T. Schultz, and A. Klin, "Autism and pervasive developmental disorders," *Journal of Child Psychology and Psychiatry*, vol. 45, no. 1, pp. 135–170, 2004.

[3] D. L. Christensen, J. Baio, K. V. N. Braun, D. Bilder, J. Charles, J. N. Constantino, J. Daniels, M. S. Durkin, R. T. Fitzgerald, M. Kurzius-Spencer, L.-C. Lee, S. Pettygrove, C. Robinson, E. Schulz, C. Wells, M. S. Wingate, W. Zahorodny, and M. Yeargin-Allsopp, "Prevalence and characteristics of autism spectrum disorder among children aged 8 years – Autism and developmental disabilities monitoring network," *MMWR. Surveillance Summaries*, vol. 65, no. 3, pp. 1–23, 2016.

[4] S. Ozonoff, G. S. Young, A. Carter, D. Messinger, N. Yirmiya, L. Zwaigenbaum, S. Bryson, L. J. Carver, J. N. Constantino, K. Karen Dobkins, T. Hutman, J. M. Iverson, R. Landa, S. J. Rogers, M. Sigman, and W. L. Stone, "Recurrence risk for autism spectrum disorders: A Baby Siblings Research Consortium study," *Pediatrics*, vol. 128, no. 3, pp. e488–e495, 2011.

[5] S. Sumi, H. Taniai, T. Miyachi, and M. Tanemura, "Sibling risk of pervasive developmental disorder estimated by means of an epidemiologic survey in Nagoya, Japan," *Journal of Human Genetics*, vol. 51, no. 6, pp. 518–522, 2006.

[6] S. Bölte, "Is autism curable?" *Developmental Medicine & Child Neurology*, vol. 56, no. 10, pp. 927–931, 2014.

[7] R. Paul, Y. Fuerst, G. Ramsay, K. Chawarska, and A. Klin, "Out of the mouths of babes: Vocal production in infant siblings of children with ASD," *Journal of Child Psychology and Psychiatry*, vol. 52, no. 5, pp. 588–598, 2011.

[8] L. Dilley, S. Cook, I. Stockman, and B. Ingersoll, "Prosodic characteristics in young children with autism spectrum disorder," *The Journal of the Acoustical Society of America*, vol. 136, no. 4, pp. 2312–2312, 2014.

[9] 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.

[10] Z. Azizi, "The acoustic survey of intonation in Autism Spectrum Disorder," *The Journal of the Acoustical Society of America*, vol. 137, no. 4, pp. 2207–2207, 2015.

[11] J. F. Santos, N. Brosh, T. H. Falk, L. Zwaigenbaum, S. E. Bryson, W. Roberts, I. M. Smith, P. Szatmari, and J. A. Brian, "Very early detection of autism spectrum disorders based on acoustic analysis of pre-verbal vocalizations of 18-month old toddlers," in *Proceedings of the 38th IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2013*. Vancouver, Canada: IEEE, May 2013, pp. 7567–7571.

[12] 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, 2010.

[13] G. Esposito, M. del Carmen Rostagno, P. Venuti, J. D. Haltigan, and D. S. Messinger, "Brief Report: Atypical expression of distress during the separation phase of the strange situation procedure in infant siblings at high risk for ASD," *Journal of Autism and Developmental Disorders*, vol. 44, no. 4, pp. 975–980, 2014.

[14] S. J. Sheinkopf, J. M. Iverson, M. L. Rinaldi, and B. M. Lester, "Atypical cry acoustics in 6-month-old infants at risk for autism spectrum disorder," *Autism Research*, vol. 5, no. 5, pp. 331–339, 2012.

[15] P. B. Marschik, K. D. Bartl-Pokorny, J. Sigafoos, L. Urlesberger, F. Pokorny, R. Didden, C. Einspieler, and W. E. Kaufmann, "Development of socio-communicative skills in 9-to 12-month-old individuals with fragile X syndrome," *Research in Developmental Disabilities*, vol. 35, no. 3, pp. 597–602, 2014.

[16] P. B. Marschik, W. E. Kaufmann, S. Bölte, J. Sigafoos, and C. Einspieler, "En route to disentangle the impact and neurobiological substrates of early vocalizations: Learning from Rett syndrome," *Behavioral and Brain Sciences*, vol. 37, no. 6, pp. 562–563, 2014.

[17] P. B. Marschik, C. Einspieler, J. Sigafoos, C. Enzinger, and S. Bölte, "The interdisciplinary quest for behavioral biomarkers pinpointing developmental disorders." *Developmental Neurorehabilitation*, vol. 19, no. 2, pp. 73–74, 2016.

[18] F. B. Pokorny, P. B. Marschik, C. Einspieler, and B. W. Schuller, "Does she speak RTT? Towards an earlier identification of Rett syndrome through intelligent pre-linguistic vocalisation analysis," in *Proceedings of the 17th Annual Conference of the International Speech Communication Association, Interspeech 2016*. San Francisco, CA, USA: ISCA, September 2016, pp. 1953–1957.

[19] M. P. Lynch, D. K. Oller, M. L. Steffens, and E. H. Buder, "Phrasing in prelinguistic vocalizations," *Developmental Psychobiology*, vol. 28, no. 1, pp. 3–25, 1995.

[20] S. Nathani, D. J. Ertmer, and R. E. Stark, "Assessing vocal development in infants and toddlers," *Clinical Linguistics & Phonetics*, vol. 20, no. 5, pp. 351–369, 2006.

[21] F. Eyben, M. Wöllmer, and B. Schuller, "openSMILE: The Munich versatile and fast open-source audio feature extractor," in *Proceedings of the 18th ACM International Conference on Multimedia, MM 2010*. Florence, Italy: ACM, October 2010, pp. 1459–1462.

[22] F. Eyben, F. Weninger, F. Groß, and B. Schuller, "Recent developments in openSMILE, the Munich open-source multimedia feature extractor," in *Proceedings of the 21st ACM International Conference on Multimedia, MM 2013*. Barcelona, Spain: ACM, October 2013, pp. 835–838.

[23] F. Eyben, K. R.Scherer, B. W. Schuller, J. Sundberg, E. André, C. Busso, L. Y. Devillers, J. Epps, P. Laukka, S. Narayanan, and K. P. Truong, "The Geneva Minimalistic Acoustic Parameter Set (GeMAPS) for voice research and affective computing," *IEEE Transactions on Affective Computing*, vol. 7, no. 2, pp. 190–202, 2016.

[24] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten, "The WEKA data mining software: An update," *ACM SIGKDD Explorations Newsletter*, vol. 11, no. 1, pp. 10–18, 2009.

[25] B. Schuller, *Intelligent Audio Analysis*. Springer, 2013.

[26] R. Brueckner and B. Schuller, "Hierarchical neural networks and enhanced class posteriors for social signal classification," in *Proceedings of the Automatic Speech Recognition and Understanding Workshop*, IEEE. Olomouc, Czech Republic: IEEE, December 2013, pp. 361–364.

[27] B. Schuller and A. Batliner, *Computational Paralinguistics: Emotion, Affect and Personality in Speech and Language Processing*. John Wiley & Sons, 2014.

[28] K. Greff, R. Srivastava, J. Koutník, B. Steunebrink, and J. Schmidhuber, "LSTM: A search space odyssey," *CoRR*, vol. abs/1503.04069, 2015.

[29] M. Abadi, A. Agarwal, P. Barham, E. Brevdo, Z. Chen, C. Citro, G. Corrado, A. Davis, J. Dean, M. Devin, S. Ghemawat, I. Goodfellow, A. Harp, G. Irving, M. Isard, Y. Jia, R. Jozefowicz, L. Kaiser, M. Kudlur, J. Levenberg, D. Mané, R. Monga, S. Moore, D. Murray, C. Olah, M. Schuster, J. Shlens, B. Steiner, I. Sutskever, K. Talwar, P. Tucker, V. Vanhoucke, V. Vasudevan, F. Viégas, O. Vinyals, P. Warden, M. Wattenberg, M. Wicke, Y. Yu, and X. Zheng, "TensorFlow: Large-scale machine learning on heterogeneous systems," *CoRR*, vol. abs/1603.04467, 2016.

[30] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[31] M. Schmitt and B. W. Schuller, "openXBOW – Introducing the Passau open-source crossmodal bag-of-words toolkit," *arXiv preprint arXiv:1605.06778*, 2016.