

Automatic classification of autistic child vocalisations: a novel database and results

Alice Baird, Shahin Amiriparian, Nicholas Cummins, Alyssa M. Alcorn, Anton Batliner, Sergey Pugachevskiy, Michael Freitag, Maurice Gerczuk, Björn Schuller

Angaben zur Veröffentlichung / Publication details:

Baird, Alice, Shahin Amiriparian, Nicholas Cummins, Alyssa M. Alcorn, Anton Batliner, Sergey Pugachevskiy, Michael Freitag, Maurice Gerczuk, and Björn Schuller. 2017. "Automatic classification of autistic child vocalisations: a novel database and results." In *Proceedings of Interspeech 2017, 20-24 August 2017, Stockholm*, edited by Francisco Lacerda, 849–53. ISCA. <https://doi.org/10.21437/Interspeech.2017-730>.

Nutzungsbedingungen / Terms of use:

licgercopyright

Dieses Dokument wird unter folgenden Bedingungen zur Verfügung gestellt: / This document is made available under these conditions:

Deutsches Urheberrecht

Weitere Informationen finden Sie unter: / For more information see:

<https://www.uni-augsburg.de/de/organisation/bibliothek/publizieren-zitieren-archivieren/publiz/>





Automatic Classification of Autistic Child Vocalisations: A Novel Database and Results

Alice Baird¹, Shahin Amiriparian^{1,2}, Nicholas Cummins¹, Alyssa M. Alcorn³, Anton Batliner¹,
Sergey Pugachevskiy¹, Michael Freitag¹, Maurice Gerczuk¹, Björn Schuller^{1,4}

¹Chair of Complex & Intelligent Systems, Universität Passau, Germany

²Machine Intelligence & Signal Processing Group, Technische Universität München, Germany

³Centre for Research in Autism and Education, UCL Institute of Education, U.K.

⁴Machine Learning Group, Imperial College London, U.K.

alice.baird@uni-passau.de

Abstract

Humanoid robots have in recent years shown great promise for supporting the educational needs of children on the autism spectrum. To further improve the efficacy of such interactions, user-adaptation strategies based on the individual needs of a child are required. In this regard, the proposed study assesses the suitability of a range of speech-based classification approaches for automatic detection of autism severity according to the commonly used Social Responsiveness Scale™ second edition (SRS-2). Autism is characterised by socialisation limitations including child language and communication ability. When compared to neurotypical children of the same age these can be a strong indication of severity. This study introduces a novel dataset of 803 utterances recorded from 14 autistic children aged between 4 – 10 years, during Wizard-of-Oz interactions with a humanoid robot. Our results demonstrate the suitability of support vector machines (SVMs) which use acoustic feature sets from multiple Interspeech COMPARE challenges. We also evaluate deep spectrum features, extracted via an image classification convolutional neural network (CNN) from the spectrogram of autistic speech instances. At best, by using SVMs on the acoustic feature sets, we achieved a UAR of 73.7 % for the proposed 3-class task.

Index Terms: children, autism, vocal irregularities, speech classification, social responsiveness scale, SRS-2, spectral features, human-robot interaction, humanoid robotics

1. Introduction

The *Autism Spectrum Conditions* (ASC) are a group of neuro-developmental conditions which can be defined by difficulties in two core domains: social and communicative behaviours, and restricted and repetitive behaviours [1]. Often, ASC becomes noticeable in early childhood, as children begin to diverge from typical developmental trajectories. Currently, the diagnosis of ASC is based on direct behavioural observation or reports, e. g., [2, 3], which generally focus on these two main areas of difficulty. A number of the behaviours assessed or observed relate to comprehension and verbal language production ability.

It is estimated that 1 in 68 children are affected by an ASC [4]; although there is substantial variance in how and when children are diagnosed. The early and reliable diagnosis of ASC is crucial to enable access to appropriate services. These are supports and teaching programmes delivered at a young age, which have generally been shown to have more positive, long-term effects compared to intervention at a later age [5].

This paper compares the efficacy of various acoustic feature representations to classify autism severity of an ASC child using

vocalisations, based on *Social Responsiveness Scale™, Second Edition* (SRS-2) [2] scores; a widely-used measure for assessment of social and communicative behaviours. Audio-based severity categorisation can be used as ‘shorthand’ for a range of interaction variables, such as the complexity to which verbal instruction is likely to be understood. Considering this, and given that humanoid robots have been suggested for aiding ASC education since [6], integrating such a system in this manner could offer interaction and personalisation benefits, based on the needs of the ASC child.

Verbal irregularities are a useful evaluation criterion for ASC, and manifest in a variety of ways depending on severity. Children with ASC who are verbally able may show unusual tone, pacing, volume, and abnormal prosody [2, 7]. Supra-segmental acoustic features relating to articulation, loudness, pitch, and rhythm have been successfully used in speech-based interaction systems for improving the social skills of children with ASC [8, 9]. They have shown promising results when classifying vocalisations of ASC or typically developing children [10].

This study utilises a brand-new corpus of 803 speech instances, collected from 14 ASC children interacting with a humanoid robot during an emotion-recognition training programme. We investigate the suitability of three *Interspeech Computational Paralinguistics Challenge* features sets from 2009 (IS09-Emotion) [11], 2010 (IS10-Paraling) [12], and 2013 (COMPARE) [13]. These representations, COMPARE in particular, have been found suitable for similar classification tasks between the speech of typical or atypically developing children [13–16], and for recognising spontaneous emotional expressions in the vocalisations of ASC children [17].

We also investigate the suitability of the Hybrid ‘*end-to-evolution*’ (e2ev) classification approach [18]. This approach utilises a combination of *deep spectrum features* and *competitive swarm optimisation* (CSO) for feature selection. The deep spectrum features are derived from forwarding spectrograms through very deep *convolutional neural networks* (CNNs) pre-trained for image recognition. Specifically the deep spectrum features are activations from the second fully connected layer (fc7) of AlexNet [19]. We speculate that this approach will suit our task, as the supra-segmental acoustic features commonly associated with ASC can be thought of being inherent spectrogram representations.

The rest of this paper is structured as follows: SRS-2 is detailed in Section 2, and Section 3 describes ASC vocal behaviours, with corpus outline in Section 4. The experimental settings are presented in Section 5, and the results and discussion in Section 6, followed by final remarks in Section 7.

2. The Social Responsiveness Scale™-2

The *Social Responsiveness Scale*™-2 [2], is a widely-used, standardised measure of reciprocal social interaction difficulties based on ASC criteria as laid out by the Diagnostic and Statistical Manual of Mental Disorders (DSM-5) [1]. While not intended for diagnosis of autism, it is a common means for autism assessment, and measuring socio-communicative difficulties. It has a substantial number of questions relating to communication, including verbal and vocal/non-verbal productions. In this study, we use the SRS-2 *School Age* form, which is aimed at children from 4–18 years. The School Age SRS-2 consists of 65 statements used to assess a child’s behaviour and manner, completed by a parent, or teacher familiar with the child. For example, “53. Talks to people with an unusual tone of voice (for example, talks like a robot or like he or she is giving a lecture)”; the statements are rated on a 4-point Likert scale.

An accumulated item score produces a single raw SRS-2 score between 0–195, where higher scores are an indication of greater social interaction difficulties. Raw scores are normalised to T-scores (T) using a gender-specific formula. The T-scores categorise impairment, based on the reciprocal social interaction skills strongly associated with a clinical diagnosis of ASC [2]:

- **Below 59 T –within normal limits:** score is not clinically significant.
- **60 T –65 T –mild:** suggesting significant deficiencies in reciprocal social behaviours, which may impact everyday interactions.
- **66 T –75 T –moderate:** showing substantial deficiencies in reciprocal social behaviours leading to interference with social interactions.
- **Above 76 T –severe:** severely affected social behaviours of clinical significance which highly interfere with everyday socialisation.

3. Autistic Vocal Behaviours

Language and communication are a prominent part of many evaluation methods for ASC. In our case, SRS-2 has 12 of its 65 questions directed at spoken interaction, verbalisations, and auditory perception. ADOS-2 also takes into account the atypical vocal behaviours which are either unique to or commonly observed during interaction with a particular child [3].

Echolalia is a term first coined by Kanner in 1943, to describe ‘parrot-like repetition of heard word combinations’ [7]. Also known as ‘echoed speech’, this would be the immediate repetition of a series of statements made by the individual interacting with the autistic child. Some children also show substantial *delayed echolalia*, which would be vocalisations from much earlier interactions or observations. *Stereotyped or idiosyncratic phrases* are vocalisations which are often repeated by the child and seem specific to them as an individual [20]. For example, in Figure 1 we see three instances of a child in the DE-ENIGMA corpus exhibiting this vocalisation behaviour. A clear similarity between speech rhythm and a noticeable parallel in pitch declination can be seen. *Irregular intonation* is another common verbal trait of individuals with ASC; often described as robot-like, ASC speech can be narrow in pitch range and can show minimal variance in frequency and intonation intensity [21].

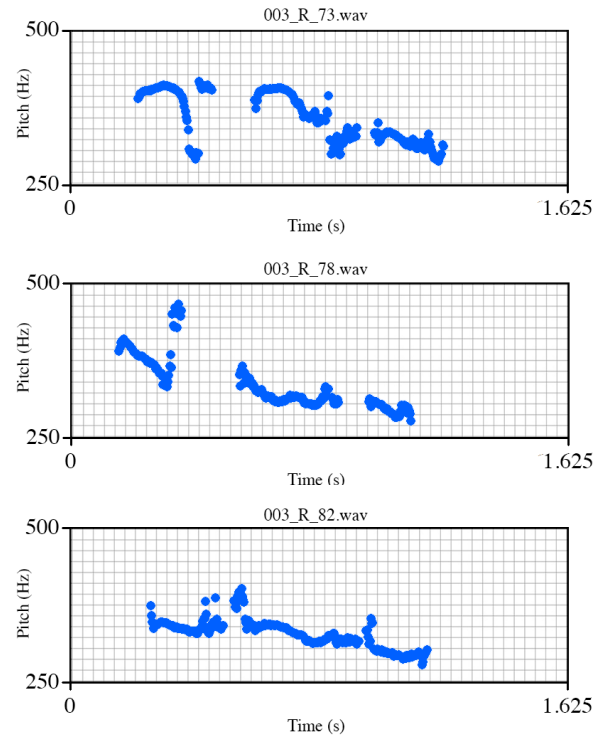


Figure 1: Examples of stereotyped speech. Instance name from corpus shown above; child ID (003), robot session (R) and instance number. The phrase “Ah o-ah ah” is repeated during the session. Pitch curve not corrected, irregular octave jumps due to transition phenomena (laryngealisations) [22].

4. DE-ENIGMA Corpus

The data collected for this study has been provided through the DE-ENIGMA, Horizon 2020 initiative, and is a forerunner to the full DE-ENIGMA database which is currently being collected and annotated. DE-ENIGMA is a research project, with the specific aim of advancing education for autistic children, through the use of humanoid robots [23]. Autistic children from differing cultural backgrounds (Serbia and the United Kingdom) are involved in the experiments; for our purpose, we have selected only instances collected in the United Kingdom¹.

A data set of 803 vocalisation instances has been gathered from 14 children who participated in the initial recordings. Ages ranging from 4–10 years; each child selected for this study has a prior diagnosis of ASC, and attends an autism-specific school in the UK. Additional diagnostic instruments and assessments (e. g., SRS-2 and ADOS-2) were administered as part of DE-ENIGMA in order to gain a clearer picture of each child. In our corpus, SRS-2 scores vary from mild to severe, with no child obtaining a T score below 59. A detailed display of instances per severity rating can be seen in Table 1.

Within our corpus the gender is split 2:12 (female:male) which gives us an (unavoidable) bias; however, this is representative of diagnosis rates in autism, which are currently 4.5 times higher in boys [4]. Previous studies [24] have also shown that gender has a less significant effect on the voice during childhood, thus, we do not expect this to impact our results.

¹Receiving full ethical approval under REC 796 DE-ENIGMA Multi-Modal HRI for Expanding Social Imagination in Autistic Children; as approved by the UCL IOE Research Ethics Committee.



Figure 2: Child participating in the U.K. DE-ENIGMA experiment. Figure shows child identifying the emotion of Zeno-R25 (DE-ENIGMA humanoid robot) with CRAE researcher. School staff member (in red), accompanying the child during the session. Circled is the Wizard-of-Oz keypad interface used to control the Zeno robot.

4.1. Data Collection and Annotation

Audio data was collected over 3–5 short daily sessions. The children participated in a *human-led* or *robot-led*, emotion-recognition training programme, based on the “Teaching Children with Autism to Mind-Read” workbook [25]. Robot-led sessions feature a Zeno-R25 [26], a humanoid-robot (controlled using a *Wizard-of-Oz* interface), with the aim of advancing social skills of ASC children.

For capturing audio, four microphones were placed in the room: 2 boundary microphones on the left and right of the child’s position; 1 overhead (approx. 1.5m above the child); and a close-talk microphone, placed on a *Centre for Research in Autism and Education* (CRAE) researcher guiding the child. The close-talk microphone was placed on the side of the child and generally picked up more of the child’s speech; thus it was the only recorded channel used for the segmented instances in our data set. Speaker diarisation of each child’s session was manually annotated by a native English speaker.

5. Key Experimental Settings

Our key experimental settings are as follows: we use a linear-kernel support vector machine (SVM) system trained on three different acoustic feature sets (cf. Section 5.1). We also implement a *e2ev* system, as proposed in [18], which combines feature extraction by deep convolutional neural networks (DCNN), with an evolutionary swarm algorithm for feature selection (cf. Section 5.2). The evaluation measure chosen for the tasks is the *Unweighted Average Recall* (UAR), i. e., the mean value of recognition accuracy for each class. As well as being a standard measure for the Interspeech COMPARE challenges, we use UAR as our corpora has an unbalanced class distributions (cf. Table 1).

5.1. Acoustic Feature Sets

Acoustic features were automatically extracted from speech through the use of our open-source openSMILE feature extractor [27]. Three different feature sets were investigated (cf. Table 3): 1) the *Interspeech 2009 Emotion Challenge* (IS09-emotion) set [11], 2) the *Interspeech 2010 Paralinguistic Challenge* (IS10-paraling) set [12], and 3) the *Interspeech 2013 Computational Paralinguistics Challenge* (COMPARE) set [13]. COMPARE includes features which have previously been used to classify speech corpora including social signals, conflict, emotion, and autism, achieving a baseline result for the autism diag-

Table 1: Number of instances per class in the train, development and test partitions, used for the SRS-2 classification task.

Classes	Train	Devel	Test
Mild	30	34	28
Moderate	45	35	21
Severe	205	211	194
Σ	280	280	243

Table 2: Distribution of the different acoustic feature sets which cover *Spec*(tral)/*energy*-related, *Sou*(rce)/*excitation*-related and *Dur*(ation)-related features with different levels of detail.

Feature set	Spec.	Sou.	Dur.	Total
IS09-emotion [11]	336	48	–	384
IS10-paraling [12]	1216	212	154	1582
COMPARE [13]	4366	397	1610	6373
Deep Spectrum [31]	4096	–	–	4096

nosis subset of 67.1 % [13]. An overview of distribution of the different feature sets regarding Spectral, Source and Duration related features is given in Table 2; a detailed description and implementation of these feature sets is given in [28].

In order to provide ‘baseline’ results, we use the open-source implementation provided by the WEKA data mining software [29] – version 3.8.1. Feature standardisation, i. e., subtracting the mean and dividing by the standard deviation, is applied. In particular, we use WEKA’s SVM implementation with the *Sequential Minimal Optimisation* (SMO; [30]) training algorithm, linear kernels for the classification tasks with epsilon-insensitive loss (known to be robust against overfitting). We optimised the complexity parameter on a logarithmic scale (10 values between 10^{-5} and 10^{-2}) and used a constant value (10^{-1}) for epsilon intensive – loss.

5.2. Hybrid ‘end-to-evolution’ System

As a spectrogram representation should inherently, contain all information relating to the linguistic and paralinguistic attributes associated with ASC, we expect that the hybrid *e2ev* approach will be suited to the task at hand. For the *e2ev* system, we first extract narrowband spectrograms from every instance in the corpus. These spectra are then passed on to pre-trained image classification CNNs, and the activations of a specific fully connected layer are extracted, resulting in a *deep spectrum* feature set containing 4 096 features [31] (cf. Table 2). These features can be interpreted as a high-level representation of the spectrograms as seen by the CNN.

A wrapper-based feature subset selection is performed on the deep spectrum features using *competitive swarm optimisation* (CSO). CSO is an evolutionary optimisation technique derived from particle swarm optimisation, which has recently been proposed for large-scale optimisation [32]. It evolves an optimised feature set by allowing candidate solutions, known as particles, to move through the search space over several generations in a self-organised way. It has been shown that, although the deep spectrum features alone can achieve high performance in paralinguistic recognition tasks, CSO feature selection can further boost classification accuracy [18].

We use a linear SVM in our hybrid *e2ev*. Since the optimal hyperparameter choice may be affected by subset selection, we have evaluated CSO for several combinations of SVM complexity $C \in [10^{-5}; 10^{-3}]$, number of generations $n_G \in [100; 400]$, and particle swarm size $n_P \in [100; 400]$.

Table 3: Classification results from the openSMILE acoustic feature based system, reporting UAR on development and test partitions. C : complexity parameter of the support vector machine. The chance level is 33.33 % UAR.

feature set	C	devel	test
IS09-emotion	10^{-5}	56.2	60.3
	10^{-4}	56.3	63.1
	10^{-3}	62.3	64.8
	10^{-2}	37.8	63.4
IS10-paraling	10^{-5}	56.5	57.4
	10^{-4}	62.2	73.7
	10^{-3}	48.0	65.6
	10^{-2}	37.8	63.4
COMPARE	10^{-5}	49.7	62.2
	10^{-4}	57.0	56.4
	10^{-3}	52.9	59.5
	10^{-2}	52.6	59.0

6. Results and Discussion

From our chosen acoustic feature systems, we see that the feature set IS10-paraling achieved the most promising results (cf. Table 3). This result indicates that IS10-paraling feature set contains potentially the most relevant features for the task at hand. While both IS09-emotion and IS10-paraling achieve comparable performance in the development set, IS10-paraling easily outperforms IS09-emotion on the test partition achieving the strongest test UAR of 73.7 % (cf. Table 3). We speculate this is due to the increased feature dimensionality and the addition of duration-related features (cf. Table 2).

Given the strong performance of COMPARE in similar tasks [17, 33], our results are weaker than expected. COMPARE feature set can be considered an omnibus feature set for paralinguistic tasks [34], and has been used successfully in the past for similar tasks of automatic diagnosis for ASC child vocalisations [17], as well as more recently for classifying typically developing children and children on the autism spectrum [33]. As our corpus size is relatively small (803 instances), and the dimensionality of COMPARE large (6373 features) we speculate the use of COMPARE introduced undesirable noise, which may have negatively impacted the result.

From the *e2ev* system, we see higher results in the development partition (77.8 %, 77.1 %, 72.2 %). However, given the weaker test partition results, we speculate that this is most likely a result of model overfitting. The best performing configuration in the *e2ev* system (200 n_G), achieved 66.9 % UAR on the development partition, and 61.9 % UAR on the test partition, having selected 3 137 of 4 096 deep spectrum features ($s_F = 76.6$ %).

On the development partition, stronger *e2ev* performances can be achieved with smaller feature sets, but this may result in a lower UAR on the test partition. Considering the *e2ev* systems promising combination of deep spectrum features and competitive swarm optimisation [18], for this task, the *e2ev* system consistently falls short of the acoustic feature systems on the test partition.

As has been mentioned in [35], swarm size (n_P) should not significantly impact the performance of the *e2ev* system. Our results would agree, the best UAR result on the test partition was 61.9 % with a n_P of 200, compared to 400 iterations, which shows no improvement (best 58.9 %, although the number of generations (n_G) may have had an effect.

Table 4: Classification results from the Hybrid ‘end-to-evolution’ system (*e2ev*) after differing feature selection configurations. We report UAR on development and test partition. n_G : different numbers of generation; n_P : swarm sizes; s_F : the % of deep spectrum features, and C : complexity parameter of the support vector machine. The chance level is 33.33 % UAR.

n_G	n_P	C	s_F	devel	test
100	400	$5 \cdot 10^{-5}$	76.5	65.5	58.9
		$1 \cdot 10^{-4}$	62.1	71.0	56.5
		$5 \cdot 10^{-4}$	49.9	71.3	58.4
		$1 \cdot 10^{-3}$	54.7	72.2	54.3
200	200	$5 \cdot 10^{-5}$	76.6	66.9	61.9
		$1 \cdot 10^{-4}$	66.1	71.3	59.7
		$5 \cdot 10^{-4}$	59.8	77.1	60.7
		$1 \cdot 10^{-3}$	54.1	73.5	57.8
300	134	$5 \cdot 10^{-5}$	71.1	68.9	58.2
		$1 \cdot 10^{-4}$	56.9	71.3	56.2
		$5 \cdot 10^{-4}$	54.9	77.8	58.3
		$1 \cdot 10^{-3}$	53.5	72.3	58.0

7. Summary and Conclusions

This study utilised a novel dataset of vocalisations from children with varying levels of autism severity, to explore the suitability of multiple classification approaches, based on the SRS-2 evaluation structure. We suggest, given the utilisation of humanoid robots within ASC education, that the integration of such an audio-based system could allow for discrete monitoring for improved human-robot interaction and personalisation.

Conventional feature extraction methods were executed using popular *Interspeech Computational Paralinguistics Challenge* feature sets, which previously have been shown to be suited to similar vocalisation classification tasks. Our results also show that a ‘default-standard’ acoustic features representation, in particular the IS10-paraling feature set, combined with a SVM backend, can achieve a high test set UAR of 73.7 %.

Results of the hybrid *e2ev* approach which combined competitive swarm optimisation and deep spectrum feature extraction, were not as strong as anticipated. However, the strong development partition results achieved by *e2ev* indicate the promise of this technique. Considering the early stage of the DE-ENIGMA project, we would anticipate the *e2ev* performance to improve with an increase in corpus size.

In future work we will consider improving our classification systems by exploring alternate feature selection methods, as well as exploring methods to fuse the more conventional openSMILE acoustic feature representations with deep spectrum features. We also plan to do an analysis using an increased dataset of children, to begin understanding the relationship between prosody and autism severity, specifically in instances showing typical ASC vocal behaviours.

8. Acknowledgements



This work was supported by the European Unions’ Seventh Framework and Horizon 2020 Programmes under grant agreements No. 338164 (ERC StG iHEARu) and No. 688835 (RIA DE-ENIGMA). The authors thank the staff and children involved in the DE-ENIGMA study, and to Professor Elizabeth Pellicano, Dr. Teresa Tavassoli and Rebecca Sealy from the CRAE research group.

9. References

- [1] American Psychiatric Association, *Diagnostic and statistical manual of mental disorders*, 5th ed., Washington, D.C., 2013.
- [2] J. M. Constantino and C. P. Gruber, *Social Responsiveness Scale, Second Edition (SRS-2)*. Western Psychological Services, 2012.
- [3] C. Lord and M. Rutter, *Autism Diagnosis Observation Schedule -2 (ADOS-2)*. Western Psychological Services, 2012.
- [4] D. L. Christensen, J. Baio, K. V. N. Braun, D. Bilder, J. Charles, J. N. C. and Julie 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," *Surveillance Summaries*, vol. 65, pp. 1–23, 2016.
- [5] S. J. Rogers, "Brief report: Early intervention in autism," *Journal of autism and developmental disorders*, vol. 26, pp. 243–246, 1996.
- [6] R. Emanuel and S. Weir, "Catalysing communication in an autistic child in a LOGO-like learning environment," in *Proc. AISB'76*, pp. 118–129, 1976.
- [7] L. Kanner, "Autistic Disturbances of Affective Contact," *Nervous Child*, vol. 2, pp. 217–250, 1943.
- [8] B. Schuller, E. Marchi, S. Baron-Cohen, A. Lassalle, H. O'Reilly, D. Pigat, P. Robinson, I. Davies, T. Baltrusaitis, M. Mahmoud, O. Golan, S. Friedenson, S. Tal, S. Newman, N. Meir, R. Shillo, A. Camurri, S. Piana, A. Stagliano, S. Bölte, D. Lundqvist, S. Berggren, A. Baranger, N. Sullings, M. Sezgin, N. Alyuz, A. Rynkiewicz, K. Ptaszek, and K. Ligmann, "Recent developments and results of ASC-Inclusion: An integrated internet-based environment for social inclusion of children with autism spectrum conditions," in *In Proc. IUI*, Georgia, USA, 2015, pp. 1–6.
- [9] E. Mower, M. Black, E. Flores, M. Williams, and S. Narayanan, "Rachel: Design of an emotionally targeted interactive agent for children with autism," in *Proc. ICME*, Barcelona, Spain, 2011, pp. 1–6.
- [10] D. Bone, C.-C. Lee, M. Black, M. Williams, S. Lee, P. Levitt, and S. Narayanan, "The psychologist as an interlocutor in autism spectrum disorder assessment: Insights from a study of spontaneous prosody," *Journal of Speech, Language, and Hearing Research*, vol. 57, pp. 1162–1177, 2014.
- [11] B. Schuller, S. Steidl, and A. Batliner, "The Interspeech 2009 Emotion Challenge," in *Proc. INTERSPEECH*, Brighton, UK, 2009, pp. 312–315.
- [12] B. Schuller, S. Steidl, A. Batliner, F. Burkhardt, L. Devillers, C. Müller, and S. Narayanan, "The INTERSPEECH 2010 Paralinguistic Challenge," in *Proc. INTERSPEECH*, Makuhari, Japan, 2010, pp. 2794–2797.
- [13] B. Schuller, S. Steidl, A. Batliner, A. Vinciarelli, K. Scherer, F. Ringeval, M. Chetouani, F. Weninger, F. Eyben, E. Marchi, M. Mortillaro, H. Salamin, A. Polychroniou, F. Valente, and S. Kim, "The INTERSPEECH 2013 Computational Paralinguistics Challenge: Social Signals, Conflict, Emotion, Autism," in *Proc. INTERSPEECH*, Lyon, France, 2013, pp. 148–152.
- [14] E. Marchi, B. Schuller, S. Baron-Cohen, O. Golan, S. Bölte, P. Arora, and R. Häb-Umbach, "Typicality and emotion in the voice of children with autism spectrum condition: Evidence across three languages," in *Proc. INTERSPEECH*, Dresden, Germany, 2015, pp. 115–119.
- [15] M. Schmitt, E. Marchi, F. Ringeval, and B. Schuller, "Towards cross-lingual automatic diagnosis of autism spectrum condition in children's voices," in *in Proc. ITG*, vol. 267, Paderborn, Germany, 2016, pp. 264–268.
- [16] D. K. Oller, P. Niyogi, S. Gray, J. A. Richards, J. Gilkerson, D. Xu, and S. F. Warren, "Automated vocal analysis of naturalistic recordings from children with autism, language delay, and typical development," in *Proc. of the National Academy of Sciences 107*. USA: PNAS, 2010, p. 13354–13359.
- [17] F. Ringeval, E. Marchi, C. Grossard, J. Xavier, M. Chetouani, D. Cohen, and B. Schuller, "Automatic Analysis of Typical and Atypical Encoding of Spontaneous Emotion in the Voice of Children," 2016, pp. 1210–1214.
- [18] M. Freitag, S. Amiriparian, M. Gerczuk, N. Cummins, and B. Schuller, "An 'End-to-Evolution' Hybrid Approach for Snore Sound Classification," in *Proc. of INTERSPEECH*. Stockholm, Sweden: ISCA, 2017, 5 pages.
- [19] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet Classification with Deep Convolutional Neural Networks," in *Advances in Neural Information Processing Systems*, F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger, Eds. Curran Associates, Inc., 2012, vol. 25, pp. 1097–1105.
- [20] N. Watt, A. Wetherby, A. Barber, and L. Morgan, "Repetitive and Stereotyped Behaviors in Children with Autism Spectrum Disorders in the Second Year of Life," *Journal of Autism and Developmental Disorders*, vol. 8, pp. 1518–1533, 2008.
- [21] L. Shriberg, R. Paul, J. McSweeney, A. Klin, D. Cohen, and F. Volkmar, "Speech and Prosody Characteristics of Adolescents and Adults With High-Functioning Autism and Asperger Syndrome," *Journal of Speech, Language, and Hearing Research*, vol. 44, pp. 1097–1115, 2001.
- [22] A. Batliner, S. Burger, B. Johne, and A. Kießling, "MÜSLI: A Classification Scheme For Laryngealizations," in *In Proc. ESCA Workshop on Prosody*, Lund, Sweden, 1993, pp. 176–179.
- [23] "DE-ENIGMA Playfully Empowering Autistic Children," Accessed: 2017-03-03. [Online]. Available: <http://de-enigma.eu/>
- [24] J. S. Hyde, "The Gender Similarities Hypothesis," *The American Psychologist Association*, vol. 60, pp. 581–592, 2005.
- [25] S. Baron-Cohen and J. Hadwin, *Teaching Children with Autism to Mind-read*. John Wiley and Sons Ltd., 1998.
- [26] "RoboKind Robots4Autism," Accessed: 2017-03-03. [Online]. Available: <http://www.robokindrobots.com>
- [27] F. Eyben, F. Weninger, F. Gross *et al.*, "Recent Developments in openSMILE, the Munich Open-Source Multimedia Feature Extractor," in *in Proc. ACM*, Barcelona, Spain, 2013, pp. 835–838.
- [28] F. Eyben, *Real-time Speech and Music Classification by Large Audio Feature Space Extraction*, ser. Springer Theses. Springer International Publishing, 2015.
- [29] M. Hall, E. Frank, G. Holmes *et al.*, "The WEKA Data Mining Software: An Update," *ACM SIGKDD Explorations Newsletter*, vol. 11, pp. 10–18, 2009.
- [30] J. Platt, "Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods," *Advances in large margin classifiers*, vol. 10, pp. 61–74, 1999.
- [31] S. Amiriparian, M. Gerczuk, S. Ottl, N. Cummins, M. Freitag, S. Pugachevskiy, and B. Schuller, "Snore Sound Classification Using Image-based Deep Spectrum Features," in *Proc. of INTERSPEECH*. Stockholm, Sweden: ISCA, 2017, 5 pages.
- [32] S. Gu, R. Cheng, and Y. Jin, "Feature selection for high-dimensional classification using a competitive swarm optimizer," *Soft Computing*, vol. 21, pp. 1–12, 2016.
- [33] J. Deng, N. Cummins, M. Schmitt, K. Qian, F. Ringeval, and B. Schuller, "Speech-based Diagnosis of Autism Spectrum Condition by Generative Adversarial Network Representations," in *Proc. of the ACM Digital Health*. London, U.K.: ACM, 2017, 5 pages.
- [34] F. Eyben, F. Weninger, F. Groß, and B. Schuller, "Recent Developments in openSMILE, the Munich Open-Source Multimedia Feature Extractor," in *Proc. of ACM MM*. Barcelona, Spain: ACM, 2013, pp. 835–838.
- [35] R. Cheng and Y. Jin, "A competitive swarm optimizer for large scale optimization," *IEEE Transactions on Cybernetics*, vol. 45, pp. 191–204, 2015.