Towards Cross-lingual Automatic Diagnosis of Autism Spectrum Condition in Children's Voices

Maximilian Schmitt[1]*, Erik Marchi*[1]*, Fabien Ringeval*[1]*, Björn Schuller*[1,2]

[1]Chair of Complex & Intelligent Systems, University of Passau, Germany
[2]Machine Learning Group, Department of Computing, Imperial College London, UK
Email: maximilian.schmitt@uni-passau.de

Abstract

Automatic diagnosis of Autism Spectrum Conditions (ASC) from the voice is still in its infancy. The comparably few studies up to now focus mostly on the relevance of acoustic features and optimal learning algorithms. However, cross-lingual studies with a higher number of speakers are a white spot in the literature. The present contribution thus focusses on extensive cross-lingual evaluations based on four databases collected in English, French, Hebrew, and Swedish. The datasets contain speech of children with ASC and typically developing (TD) children matched in both age and gender. Their speech further intentionally varies in emotion. This introduces an additional challenge besides the change of languages. We demonstrate automatic ASC vs TD classification to be feasible despite such variation with a remaining error.

Keywords: Autism Spectrum Condition, Automatic diagnosis, Cross-lingual analysis

1 Introduction

Autism is a pervasive developmental disorder influencing especially social and communicative skills of the affected person. It may refer to different level of severities, usually summarised under one umbrella term: "*autism spectrum disorders*" (ASD) or, preferably, "*autism spectrum conditions*" (ASC) [1]. DSM-4 distinguished ASC subtypes, such as the *Asperger syndrom*, which reveals later in life than ASC [2]. Though ASC are generally reported to be lifelong, there are some reported cases where children suffering from ASC regain social and communication skills, and reach the range of cognitive, adaptive and social skills as met for typically developing (TD) children [3].

Diagnose of people suffering from ASC is usually triggered by behavioural interventions [4]. Therefore, it is necessary to have a simple and universal method to care and monitor the different types of ASC. Autism usually manifests itself especially in the prosody of children's or adults' speech, such as pitch and rhythm [5]. A typical trait of people under ASC is, e. g., a monotonous voice. Research in the field has proved that individuals on the spectrum very often experience significant difficulties in both recognising and expressing emotions from different modalities, such as facial expressions, speech, and gestures.

In recent years, much research work has been done in the field of automatic recognition and monitoring of emotions from speech and helping children under ASC to improve their social skills. Examples for already existing systems are the Rachel Embodied Conversational Agent (ECA) [6] and the Mind-Reading software [7]. Those aim to provoke a specific emotion through an interactive agent and to teach people under ASC to recognise complex emotions of their conversational partner.

An internet-based platform assisting children with ASC to improve their social and emotional communication skills was developed within the scope of the ASC-Inclusion project [8–10]. This platform provides an interactive game to give scores on the typicality and on the naturalness of the presented emotions. The ASC-Inclusion platform combines several recent technologies in one virtual environment, combining multimodal emotion analysis and giving feedback on the appropriateness of the expressions, while also providing ideas for improvement of the vocal, facial, and gestural behaviour.

However, the automatic analysis of children's speech is highly challenging, as both acoustic and linguistic characteristics are correlated with their age and gender [11]. Performance of automatic recognition systems is additionally impaired by background noise, present in the children's rooms or at their doctor's offices. In [12, 13], the relevance of selected prosodic features with respect to their significance in emotion classification of children under ASC was investigated.

For computational research purposes, only few databases of children under ASC – and a corresponding control group – exist. One corpus in French language is the CPSD database, which has been proposed to automatically quantify the differences in the imitation of prosodic contours between groups of ASC and TD children [14, 15]. Another version of this corpus, CPESD, has been recently proposed for the study of spontaneous emotions [16]. A corpus in English language, the USC CARE Corpus [17], was used to investigate how cues in prosodic speech of children under ASC can be quantified during spontaneous interaction [18].

In this article, we would like to evaluate the performance of data-based ASC recognition systems across four different languages: English, French, Hebrew, and Swedish. A similar study based on three languages has been performed in [19], however, they did not consider cross-lingual recognition of emotions.

1.1 Contribution of this work

The present study focuses on the speech-based recognition of children with ASC vs typically developing children. The classification performances are investigated across languages, by training on one or several languages and evaluating on the remaining languages present in our data. The goal of this research work is, first of all, to learn about how ASC manifests in different languages, and secondly, to create a more robust ASC classifier by pooling data across languages.

The article is structured as follows: first, a detailed description of the databases are given (Section 2); then we define the experimental tasks, features and setup (Section 3). We then comment on the evaluation results (Section 4) before concluding the paper (Section 5).

ISBN 978-3-8007-4275-2

2 Databases

In this section, we describe the datasets in four different languages used throughout our experiments. The datasets in English, Hebrew, and Swedish have been recorded within the scope of the ASC-Inclusion project [8], the French dataset is the CPESD corpus [16].

Compared to state-of-the-art databases used in speech processing tasks, such as, e. g., automatic speech recognition (ASR), the presented databases are relatively small. It must be stated, however, that the recruitment of subjects from the target group (ASC) is quite challenging, alike the successful realisation of the experiments and the recordings. Compared to other studies within the field of ASC, those datasets can be considered as fairly representative. The specific characteristic is that, all datasets contain TD children and children with ASC under the same recording conditions, which is exploited in this work measuring the performance of recognition of ASC across different languages.

2.1 ASC-Inclusion children's emotional speech database

A database for the recognition of emotions and for the thorough analysis of speech features relevant for the task at hand was created by Marchi et al. [12, 13, 20]. It contains recordings of children with ASC and TD children, speaking out a set of prototypical emotional utterances in English, Swedish, and Hebrew.

All children with ASC were diagnosed by trained clinicians, based on established criteria (DSM IV/ICD 10) [2]. In order to limit the effort of the children, the experimental task was designed to focus on the six 'basic' emotions except for *disgust*: *happy, sad, angry, surprised, afraid,* and *neutral*, plus other three mental states: *ashamed, calm,* and *proud*. During two hour meetings with each child and her/his parents, a semi-structured observation was conducted, which included free-play in a virtual environment, followed by a directed play in pre-selected games, and by an interview with the child. Only then, the recording session was held, since it requires a good rapport with the child. The recordings took place at the childrens' homes according to the following setup: the child and the examiner sat at a table in front of a laptop. The examiner read out a sequence of short stories to the child, who was asked to imagine that he/she was the main character in the story. The microphone stood next to the laptop, about 20 cm in front of the child. The data was then annotated by two expert clinicians per site. This recording protocol was used to collect the three following datasets:

English dataset – The English dataset contains recordings of 18 children from England (cf. Table 1); all of them are native English speakers. The ASC group has 8 children (3 female, 5 male) at the age of 7 to 11 (mean=8.8, standard dev.=1.5). The control group (TD) consists of 10 children (5 female, 5 male) at the age of 5 to 10 (mean=7.9, standard dev.=1.6). For the audio recordings, a *Zoom H1 Handy Recorder* was used, with a sampling rate of 96 kHz and a quantisation of 16 bits. Details on the number of utterances per group are given in Table 1.

Hebrew dataset – The Hebrew dataset consists of 7 children under ASC (1 female, 6 male) at the age of 6 to 10 (mean=8.1, standard dev.=1.6), all diagnosed by trained clinicians. The control group is formed by 10 typically developing children (5 female, 5 male) at the age of 5 to 9 (mean=7.2, standard dev.=1.8). For the audio recordings, again a *Zoom H1 Handy Recorder* was used, with a sampling rate of 96 kHz and a quantisation of 16 bits. Details on the number of utterances per group are given in Table 1.

Table 1: Number of utterances per group for the four languages. Diagnosis categories: Typically developing children (TD) and children with Autism Spectrum Condition (ASC). Gender: number of female (f) and male (m) subjects.

Language	Diagnosis	# Subjects		# Utterances
		f	m	
English	TD	5	5	847
	ASC	3	5	658
French	TD	6	10	4052
	ASC	3	10	1191
Hebrew	TD	5	5	350
	ASC	1	6	178
Swedish	TD	5	6	397
	ASC	0	9	331

Swedish dataset – For the Swedish dataset, a total number of 20 children took part. All children were native speakers and the language is Swedish throughout recordings. The group under ASC consists of 9 children (all male) at the age of 7 to 11 (mean=9.1, standard dev.=1.2). The control group (TD) consists of 11 children (5 female, 6 male) at the age of 5 to 9 (mean=6.8, standard dev.=1.7). For the audio recordings, a *Zoom H4* device with a *RØDE NTG-2* microphone was used, with a sampling rate of 96 kHz and a quantisation of 16 bits. Details on the number of utterances per group are given in Table 1.

2.2 Child Pathological & Emotional Speech Database

French dataset – The French dataset we used consists of recordings taken from 29 children overall; 13 children under ASC (3 female, 10 male) and 16 TD children (6 female, 10 male).

The exploited dataset was extracted from a larger dataset recorded in France [16]. The Ethical Committee of the Pitié-Salpétrière Hospital gave approval to conduct recruitment and speech recording of children. In total, 35 monolingual participants were recruited in two university departments of child and adolescent psychiatry located in Paris, France. All children were equipped with communicative verbal skills. For each child, one of the following diseases or impairments had been diagnosed: autism disorders (AD), pervasive developmental disorders not-otherwise specified (PDD-NOS), or specific language impairment (SLI), according to DSM IV criteria [2]. The patients were matched for age, sex, academic grades and lexical abilities. A deeper insight of the socio-demographic and clinical characteristics of the participating children is found in [14].

To have a control group, 70 TD children from elementary schools were recruited. Participants were also matched for age and sex (2 TD for 1 patient). Their teachers were asked to fill in a questionnaire to exclude children with learning disorders.

The main goal of this study was to compare children's abilities to use prosody to encode emotion and affect in speech. The first task, the children were asked to complete, was based on the reproduction of an intonation contour [14]. The second task was based on storytelling of the pictured book 'Frog where are you?' by Mayer [21]. In this story, a little boy tries to find his escaped frog. This task was originally developed as a standard for the assessment of language production. However, in CPESD, the children were supposed to produce prosodic cues during the storytelling [16]. Those cues should encode different levels of the emotional valence. Valence was categorised here into three nominal levels by a psychologist: Negative/Neutral/Positive. In total, the book includes 15 emotionally negative, 6 emotionally neutral and 5 emotionally positive pictures. Three further pictures which could not be interpreted unambiguously with respect to emotion, were excluded from the experiment.

In total, almost 10 hours of audio were recorded: 7 h 38 min for TD children, 1 h 35 min for children with AD, 1 h 12 min for children with PDD-NOS, and 1 h 56 min for children with SLI.

Recordings were segmented automatically based on the energy contour of the speech signal, pursuing a division into groups of breath. Due to perturbation during the recordings, the obtained speech segments were manually processed in order to obtain a dataset of utterances with a complete prosodic contour only. This already provided some interesting findings: utterances produced by TD children were significantly longer than those of AD, PDD-NOS, and SLI children ($p < 0.5$, two-tailed t test) [16]; the opposite was observed on the task of intonation contour imitation [14].

For our proposed experiments, we used recordings of both groups AD and PDD-NOS as instances of the ASC group, which is well-founded, yet all recordings from patients with SLI were omitted. All subjects were between 6 and 18 years old at the time of the recording. The average age was 9.8 years for all groups (TD, AD, PDD-NOS), with the following standard deviations: TD: 3.3, AD: 3.5, PDD-NOS: 2.5.

As described above, we only used a subset of this whole dataset in our experiments. Details on the number of utterances per group are given in Table 1.

3 Experiments

The goal is now to train and evaluate a language-independent binary classifier to distinguish children with ASC from typically developing children.

3.1 Acoustic feature sets

Acoustic low-level descriptors (LLDs) were extracted from the speech waveform on frame level using our open-source feature extraction tool openSMILE [22]. Two different feature sets were applied: a large brute-forced feature set (ComParE) and a smaller, expert-knowledge based feature set (eGeMAPS). A detailed description and implementation of the latter is given in [23].

With the INTERSPEECH 2013 ComParE Challenge [15], a large acoustic feature set was provided (**ComParE**). It is the outcome of a continuous refinement of acoustic descriptors optimised for automatic analysis tasks in paralinguistics. It has been successfully employed for the recognition of various speaker traits and states, e. g., personality [24], pathology [15], cognitive and physical load [25], and eating condition [26].

The ComParE set consists of functionals of LLDs of various types, such as energy, spectral, cepstral (MFCC) and voicing related LLDs, as well as logarithmic harmonic-to-noise ratio (HNR), spectral harmonicity, and psychoacoustic spectral sharpness; altogether these are, 65 LLDs and their first order derivatives, i. e., 130 LLDs in total. Functionals such as, e. g., mean, standard deviation, higher-order moments, and percentiles are then applied to the LLDs of each utterance resulting in 6 373 features.

Table 2: eGeMAPS acoustic feature set: 25 low-level descriptors (LLDs).

6(8) frequency related LLD	**Group**
F_0 (linear & semi-tone)	Prosodic
Jitter (local), Formant 1 (bandwidth)	Voice quality
Formants 1, 2, 3 (frequency)	Vowel quality
Formant 2, 3 (bandwidth)	Voice quality
3 energy/amplitude related LLD	**Group**
Sum of auditory spectrum (loudness)	Prosodic
log. HNR, shimmer (local)	Voice qual.
9(14) spectral LLD	**Group**
Alpha ratio (50–1000 Hz / 1–5 kHz)	Spectral
Hammarberg index	Spectral
Spectral slope (0–500 Hz, 0–1 kHz)	Spectral
Formants 1, 2, 3 (rel. energy)	Voice qual.
Harmonic difference H1–H2, H1–A3	Voice qual.
Spectral flux	Spectral
MFCC 1–4	Cepstral

As a comparison to this large scale brute-forced feature set, a smaller, expert-knowledge based acoustic feature set has been tried, the so-called *Geneva Minimalistic Acoustic Parameter Set* in its *extended* version (**eGeMAPS**). It has been proposed for the analysis of speaker states and traits in [27] and already proven to be suitable and robust for modelling of short-term paralinguistic states, such as emotion [28–30]. In principle, a minimalistic feature set has the advantage of reducing the risk of overfitting a classifier as compared to larger feature sets.

The selection of features for **eGeMAPS** was mainly done in consideration of the potential to describe affective physiological changes in voice production. It contains supplemental 7 LLDs compared to the 18 LLDs in the original minimalistic feature set *GeMAPS*. A detailed list of all LLDs is provided in Table 2. Selected functionals are applied to the LLDs to obtain the final features.

3.2 Setup and evaluation

As classifier, support vector machines (SVMs) with linear kernel were used, where a fast implementation exists with LIBLINEAR [31]. All features have been standardised with an *on-line* approach, i. e., the *mean* and the *standard deviation* of each feature have been derived only from the training set and then used for standardisation on both training and test sets.

To optimise the complexity parameter of the SVM, the data sets of each language, were split into a training and a validation partition. The partitioning was done manually, with the goal of disparate speakers in each partition. Gender and class (ASC / TD) distribution in both partitions were chosen to be as equal as possible. For training, between 65 % and 70 % were taken from the data sets of each language, respectively. Complexity was optimised between 10^{-10} and 1, with a step factor of 10.

For training, we chose iteratively the datasets of each language and all their possible combinations. Training of the final classifier was based on the whole datasets (training+validation) with the complexity where the largest un-

Table 3: Results in terms of UAR (%) for cross-lingual ASC vs TD classification with eGeMAPS, balanced classes & languages. Results in parentheses are evaluated on the speaker-independent training / validation split for each language.

Trained on → Tested on ↓	E	F	H	S	E+F	E+H	E+S	F+H	F+S	H+S	E+F+H	E+F+S	E+H+S	F+H+S
English	(81.8)	51.3	56.7	46.0	(76.1)	(73.5)	(72.9)	54.9	43.7	52.1	(72.7)	(73.1)	(70.0)	**60.6**
French	49.9	(82.9)	**57.5**	40.2	(73.1)	51.3	48.7	(71.6)	(73.0)	47.9	(66.6)	(61.6)	51.6	(64.6)
Hebrew	60.9	50.6	(62.2)	55.5	60.6	(68.2)	61.8	(68.2)	59.7	(60.8)	(72.3)	**64.9**	(63.5)	(57.4)
Swedish	51.2	46.0	52.7	(73.2)	46.4	**63.5**	(59.1)	50.4	(65.0)	(57.3)	55.2	(58.6)	(55.0)	(61.4)

Table 4: Results in terms of UAR (%) for cross-lingual ASC vs TD classification with IS13, balanced classes & languages. Results in parentheses are evaluated on the speaker-independent training / validation split for each language.

Trained on → Tested on ↓	E	F	H	S	E+F	E+H	E+S	F+H	F+S	H+S	E+F+H	E+F+S	E+H+S	F+H+S
English	(86.9)	50.1	54.9	38.9	(84.5)	(86.5)	(81.6)	45.3	47.9	54.9	(83.7)	(77.6)	(75.3)	**56.9**
French	48.6	(87.1)	46.8	47.7	(83.4)	**55.4**	44.9	(85.4)	(81.4)	46.4	(82.0)	(76.7)	44.4	(83.0)
Hebrew	54.7	55.2	(71.6)	55.4	62.7	(70.3)	62.7	(58.1)	59.6	(58.6)	(72.1)	**67.0**	(72.1)	(59.9)
Swedish	48.5	49.8	53.2	(78.6)	42.5	49.4	(69.5)	**56.6**	(74.5)	(78.6)	51.9	(70.9)	(75.0)	(75.5)

weighted average recall (UAR) was achieved on the respective validation partition.

4 Results

Complexity was found to be optimum between 10^{-7} and 1, where it tends to be lower for the larger feature vectors of the ComParE feature set.

In Tables 3 and 4, the results of the evaluations are displayed for each language separately. Each column describes the datasets which have been used for training (E: English, F: French, H: Hebrew, S: Swedish). Numbers in parentheses represent results where the test language is present in the respective training data. In these cases, the evaluation was done based on the split into training and validation partition, where each speaker is only present in one of the two partitions.

For all shown results, the classes and languages have been balanced in the training sets by upsampling of the utterances of the minority class(es), as this practice usually improves the results.

From the results, it is evident that, in some configurations, not even the chance level (50 %) in binary classification is reached. The main difficulty in recognising ASC from speech in the given setting is that the recording conditions, especially the room and the background noises, were different between databases, and between recordings of children with ASC and typically developing children. This might be a reason for the better results in within-language recognition of ASC, reported also in previous publications.

For English and Hebrew, a classifier trained on all three other languages performed best for both feature sets. For French and Swedish, the most interesting finding is that in the optimum training set, Hebrew is always present. Interestingly, the Hebrew corpus provides the worst results in case of evaluation within-language. On average, better results are achieved with the minimalistic eGeMAPS feature set than with the ComParE feature set.
This is in contrast to the recognition performance within each language. Here, the UAR achieved with the ComParE features is higher in almost every case.

5 Conclusions and outlook

We have assessed the performance of automatic detection of ASC in children's speech with two state-of-the-art acoustic feature sets, one large brute-force feature set and one smaller set designed by experts in the field of computational paralinguistics. The evaluations were done based on training data from four different languages, considering especially recognition across languages. Results show that, transferring knowledge from other languages for classification of ASC vs TD is highly challenging, because of high variability in the recording conditions across languages.

As the performance of cross-lingual recognition of ASC is yet not very good, we will use transfer-learning [32, 33] to exploit the knowledge from the training data in a better way to predict in recordings with previously unseen languages. Speech denoising techniques, either based on pure signal processing [34], or machine learning methods [35], will be used as well as a front end to decrease the impact of variable background noises in the recordings.

Acknowledgements

The research leading to these results has received funding from the European Unions's Seventh Framework and Horizon 2020 Programmes under grant agreements No. 338164 (ERC StG iHEARu), No. 645094 (IA SEWA), No. 645378 (RIA ARIA-VALUSPA), and No. 688835 (RIA DE-ENIGMA).

References

[1] "Diagnostic and statistical manual of mental disorders (5th Ed.)," Washington, DC: American Psychiatric Association, 2013.

[2] "Diagnostic and statistical manual of mental disorders (4th Ed.)," Washington, DC: American Psychiatric Association, 2000.

[3] M. Helt, E. Kelley, M. Kinsbourne, J. Pandey, H. Boorstein, M. Herbert, and D. Fein, "Can children with autism recover? if so, how?," *Neuropsychology Review*, vol. 18, pp. 339–366, December 2008.

[4] E. Marchi, F. Ringeval, and B. Schuller, "Perspectives and Limitations of Voice-controlled Assistive Robots for Young Individuals with Autism Spectrum Condition," in *Speech and Automata in Health Care (Speech Technology and Text Mining in Medicine and Healthcare)*, Berlin: De Gruyter, 2014.

[5] J. McCann and S. Peppe, "Prosody in autism spectrum disorders: a critical review," *International Journal of Lan-*

guage & Communication Disorders, vol. 38, no. 4, pp. 325–350, 2003.

[6] E. Mower, M. P. Black, E. Flores, M. Williams, and S. Narayanan, "Rachel: Design of an emotionally targeted interactive agent for children with autism," in *Proc. ICMCS/ICME*, pp. 1–6, 2011.

[7] O. Golan and S. Baron-Cohen, "Systemizing empathy: Teaching adults with asperger syndrome or high-functioning autism to recognize complex emotions using interactive multimedia," *Development and Psychopathology*, vol. 18, no. 2, pp. 591–617, 2006.

[8] B. Schuller *et al.*, "ASC-Inclusion: Interactive Emotion Games for Social Inclusion of Children with Autism Spectrum Conditions," in *Proc. IDGEI*, (Chania, Greece), ACM, SASDG, 2013.

[9] B. Schuller, E. Marchi, S. Baron-Cohen, H. O'Reilly, D. Pigat, P. Robinson, and I. Daves, "The state of play of ASC-Inclusion: An Integrated Internet-Based Environment for Social Inclusion of Children with Autism Spectrum Conditions," in *Proc. IDGEI*, (Haifa, Israel), ACM, 2014.

[10] 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. Fridenson, S. Tal, S. Newman, N. Meir, R. Shillo, A. Camurri, S. Piana, A. Staglianò, 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 *Proc. IDGEI*, (Atlanta, USA), ACM, 2015.

[11] A. Potamianos and S. Narayanan, "A review of the acoustic and linguistic properties of children's speech," in *Proc. IEEE 9th Workshop on Multimedia Signal Processing*, pp. 22–25, 2007.

[12] E. Marchi, B. Schuller, A. Batliner, S. Fridenzon, S. Tal, and O. Golan, "Emotion in the Speech of Children with Autism Spectrum Conditions: Prosody and Everything Else," in *Proc. WOCCI*, (Portland, USA), ISCA, 2012.

[13] E. Marchi, A. Batliner, B. Schuller, S. Fridenzon, S. Tal, and O. Golan, "Speech, Emotion, Age, Language, Task, and Typicality: Trying to Disentangle Performance and Feature Relevance," in *Proc. WS3P*, (Amsterdam, The Netherlands), ASE/IEEE, 2012.

[14] F. Ringeval, J. Demouy, G. Szaszák, M. Chetouani, L. Robel, J. Xavier, D. Cohen, and M. Plaza, "Automatic intonation recognition for prosodic assessment of language impaired children," *IEEE Transactions on Audio, Speech & Language Processing*, vol. 19, pp. 1328–1342, July 2011.

[15] B. Schuller *et al.*, "The INTERSPEECH 2013 Computational Paralinguistics Challenge: Social Signals, Conflict, Emotion, Autism," in *Proc. INTERSPEECH*, (Lyon, France), ISCA, 2013.

[16] 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," in *Proc. INTERSPEECH*, (San Fransisco (CA), USA), ISCA, 2016. 5 pages, to appear.

[17] M. P. Black, D. Bone, M. E. Williams, P. Gorrindo, P. Levitt, and S. S. Narayanan, "The usc care corpus: Child-psychologist interactions of children with autism spectrum disorders," in *Proc. INTERSPEECH*, (Florence, Italy), pp. 1497–1500, ISCA, 2011.

[18] D. Bone, C.-C. Lee, M. P. Black, M. E. 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, no. 4, pp. 1162–1177, 2014.

[19] 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), pp. 115–119, ISCA, 2015.

[20] E. Marchi, B. Schuller, S. Baron-Cohen, A. Lassalle, H. O'Reilly, D. Pigat, O. Golan, S. Friedenson, S. Tal, S. Bölte, S. Berggren, D. Lundqvist, and M. Elfström, "Voice Emotion Games: Language and Emotion in the Voice of Children with Autism Spectrum Condition," in *Proc. IDGEI*, (Atlanta, USA), ACM, 2015.

[21] M. Mayer, *Frog where are you?* New York: Dial Books for young readers, 1969.

[22] F. Eyben, F. Weninger, F. Groß, and B. Schuller, "Recent Developments in openSMILE, the Munich Open-Source Multimedia Feature Extractor," in *Proc. ACM MM*, (Barcelona, Spain), pp. 835–838, ACM, 2013.

[23] F. Eyben, *Real-time speech and music classification by large audio feature space extraction*. PhD thesis, Technische Universität München, 2015.

[24] B. Schuller *et al.*, "The INTERSPEECH 2012 Speaker Trait Challenge," in *Proc. INTERSPEECH*, (Portland, USA), ISCA, 2012.

[25] B. Schuller, S. Steidl, A. Batliner, J. Epps, F. Eyben, F. Ringeval, E. Marchi, and Y. Zhang, "The INTERSPEECH 2014 Computational Paralinguistics Challenge: Cognitive & Physical Load," in *Proc. INTERSPEECH*, (Singapore), pp. 427–431, ISCA, 2014.

[26] B. Schuller, S. Steidl, A. Batliner, S. Hantke, F. Hönig, J. R. Orozco-Arroyave, E. Nöth, Y. Zhang, and F. Weninger, "The INTERSPEECH 2015 Computational Paralinguistics Challenge: Nativeness, Parkinson's & Eating Condition," in *Proc. INTERSPEECH*, (Dresden, Germany), pp. 478–482, ISCA, 2015.

[27] F. Eyben *et al.*, "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.

[28] D. Bone, C.-C. Lee, and S. S. Narayanan, "Robust Unsupervised Arousal Rating: A Rule-Based Framework with Knowledge-Inspired Vocal Features," *IEEE Transactions on Affective Computing*, vol. 5, no. 2, pp. 201–213, 2014.

[29] F. Ringeval, S. Amiriparian, F. Eyben, K. Scherer, and B. Schuller, "Emotion recognition in the wild: Incorporating voice and lip activity in multimodal decision-level fusion," in *Proc. EmotiW*, (Istanbul, Turkey), pp. 473–480, 2014.

[30] F. Ringeval, E. Marchi, M. Méhu, K. Scherer, and B. Schuller, "Face reading from speech – predicting facial action units from audio cues," in *Proc. INTERSPEECH*, (Dresden, Germany), pp. 1977–1981, ISCA, 2015.

[31] R.-E. Fan, K.-W. Chang, C.-J. Hsieh, X.-R. Wang, and C.-J. Lin, "Liblinear: A library for large linear classification," *The Journal of Machine Learning Research*, vol. 9, pp. 1871–1874, 2008.

[32] E. Coutinho, J. Deng, and B. Schuller, "Transfer Learning Emotion Manifestation Across Music and Speech," in *Proc. IJCNN*, (Beijing, China), pp. 3592–3598, IEEE, 2014.

[33] J. Deng, R. Xia, Z. Zhang, Y. Liu, and B. Schuller, "Introducing Shared-Hidden-Layer Autoencoders for Transfer Learning and their Application in Acoustic Emotion Recognition," in *Proc. ICASSP*, (Florence, Italy), pp. 4851–4855, IEEE, 2014.

[34] J. Pohjalainen, F. Ringeval, Z. Zhang, and B. Schuller, "Spectral and Cepstral Audio Noise Reduction Techniques in Speech Emotion Recognition," in *Proc. ACM MM*, (Amsterdam, The Netherlands), ACM, 2016. 5 pages.

[35] Z. Zhang, F. Ringeval, J. Han, J. Deng, E. Marchi, and B. Schuller, "Facing realism in spontaneous emotion recognition from speech: Feature enhancement by autoencoder with LSTM neural networks," in *Proc. INTERSPEECH*, (San Fransisco (CA), USA), ISCA, 2016. 5 pages.