Typicality and Emotion in the Voice of Children with Autism Spectrum Condition: Evidence Across Three Languages

Erik Marchi[1], *Björn Schuller*[2,3], *Simon Baron-Cohen*[4], *Ofer Golan*[5], *Sven Bölte*[6], *Prerna Arora*[7], *Reinhold Häb-Umbach*[7]

[1]Machine Intelligence & Signal Processing Group, MMK, Technische Universität München, Germany
[2]Department of Computing, Imperial College London, UK
[3]Chair of Complex & Intelligent Systems, University of Passau, Germany
[4]Autism Research Centre, University of Cambridge, UK
[5]Department of Psicology, Bar-Ilan University, Israel
[6]Center of Neurodelopmental Disorders, Karolinska Intitute, Sweden
[7]Department of Communication Engineering, University of Paderborn, Germany
erik.marchi@tum.de

Abstract

Only a few studies exist on automatic emotion analysis of speech from children with Autism Spectrum Conditions (ASC). Out of these, some preliminary studies have recently focused on comparing the relevance of selected acoustic features against large sets of prosodic, spectral, and cepstral features; however, no study so far provided a comparison of performances across different languages. The present contribution aims to fill this white spot in the literature and provide insight by extensive evaluations carried out on three databases of prompted phrases collected in English, Swedish, and Hebrew, inducing nine emotion categories embedded in short-stories. The datasets contain speech of children with ASC and typically developing children under the same conditions. We evaluate automatic diagnosis and recognition of emotions in atypical children's voice over the nine categories including binary valence/arousal discrimination.

Index Terms: Emotion Recognition, Feature Analysis, Autism Spectrum Conditions, Knowledge Based Systems, Speech Classification

1. Introduction

The last three decades of research, showed that children and adults with Autism Spectrum Conditions (ASC) may experience significant difficulties in recognising and expressing emotions from facial expressions, speech, gestures, and body language. Attempts to teach emotion and mental state recognition, either on an individual basis or as a part of social skills group training, have shown mixed results. A solution for the shortage of trained therapists for individuals with ASC may be found in Information and Communication Technology (ICT), which enables users everywhere to enjoy state-of-the-art professional support on-line. The computerised environment is especially appealing for individuals with ASC, due to its predictable, controllable and structured nature, which facilitates them to use their strong systemizing skills. Existing systems, such as the Rachel Embodied Conversational Agent (ECA) [1] and the Mind-Reading software [2], aim to elicit the targeted emotion through an interactive agent in order to study the interaction patterns of children with ASC and to teach people in the spectrum to recognise complex emotions using interactive multimedia. The ASC-Inclusion platform [3–5] created an internet-based platform that assists children with ASC to improve their socio-emotional communication skills. Unlike the past ICT solutions, the platform addresses the recognition and the expression of socio-emotional cues, by providing an interactive-game that gives scores on the prototypicality and on the naturalness of child's expressions. It combines several state-of-the-art technologies in one comprehensive virtual environment (VE), combining voice, face and body gesture analysis, giving corrective feedback as for the appropriateness of the child's expressions. The automatic processing of children's speech is however highly challenging. Indeed, children have very specific voices compared to adults (acoustic, lexical and linguistic characteristics of solicited and spontaneous children's speech have been correlated with age and gender [6]), and even more when they are affected by ASC. When one adds in the background noise of children's homes and doctor's offices, it makes it even harder for automatic recognition systems to perform accurately. Yet, in spite of these challenges, we do, however, consider the integration of voice-based technologies in ICT-based virtual platform as an important component to enable multimodal interaction between children with ASC and VE. The ability to convey socio-affective behaviours through speech is probably the most natural way to engage social interactions [7]. Ringeval et al. proposed a system to quantify differences in the imitation of prosodic contours between different groups of ASC children, using the CPSD database [8,9]. Bone et al. investigated how prosodic speech cues of children with ASC can be quantified during spontaneous interaction [10], using the USC CARE Corpus [11]. These two databases – except for the database dealt herein – are the only other existing corpora containing speech material from children with ASC used for computational research purposes. In [12, 13] an analysis of selected prosodic features with respect to their relevance in the classification of the emotion in the voice of children with ASC was performed.

1.1. Contribution of this work

The present study focuses on the recognition of emotional expressions in the voice of children with ASC, in order to investigate the classification performances with expert-based reduced feature sets against large sets of features that include a vast number of acoustic, spectral and cepstral features. This study further focuses on an other aspect, such as discrimination of typicality between typically developing children and children with ASC.

Table 1: Number of utterances per emotion category (# Emotion), binary arousal/valence, and overall number of utterances (# All) per group for the three languages. Emotion classes: afraid (Af), angry (An), happy (Ha), sad (Sa), surprised (Su), ashamed (As), proud (Pr), calm (Ca) and neutral (Ne). Diagnosis categories: Typically developing children (TD), and children with Autism Spectrum Condition (ASC). Gender: number of male (m) and female (f) subjects.

# utterances	**Diagnosis**	**# Subjects**		**# Emotion**									**Arousal**		**Valence**		**All**
		m	**f**	**Af**	**An**	**Ha**	**Sa**	**Su**	**As**	**Pr**	**Ca**	**Ne**	**+**	**-**	**+**	**-**	
English	TD	5	4	40	40	50	40	40	40	50	28	40	220	148	208	160	368
-	ASC	5	3	31	30	39	31	31	31	39	24	32	170	118	165	123	288
Swedish	TD	6	5	40	40	49	48	39	28	48	30	38	216	144	204	156	360
-	ASC	9	-	36	36	45	44	36	27	45	27	36	198	134	189	143	332
Hebrew	TD	5	5	38	38	49	38	38	37	46	27	40	209	142	200	151	351
-	ASC	6	1	18	20	30	21	21	17	22	13	16	111	67	102	76	178

Given our interest in the classification of children's emotional vocal expression, a database of prompted phrases was collected in English, Swedish, and Hebrew, inducing up to nineteen emotions embedded in short-stories. The utterances were produced by children with ASC as well as by typically developing children.

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

2. ASC-Inclusion children's emotional speech database

As an evaluation database for the recognition of emotions and for the analysis of speech features that are modulated by emotion, a set of prototypical emotional utterances containing sentences spoken in English, Swedish, and Hebrew by children with ASC and typically developing children has been created [12–14]. All children with an autism spectrum condition were diagnosed by trained clinicians, based on established criteria (DSM IV/ICD 10). In order to limit the effort of the children, the experimental task was designed to focus on the six "basic" emotions except *disgust*: *happy, sad, angry, surprised, afraid* plus other three mental states: *ashamed, calm, proud,* and *neutral*. During a 2 hour meeting with the child and his/her 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 children's home according to the following set-up: the child and the examiner sat at a table in front of a laptop. The examiner read to the child a sequence of short stories and the child 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:

Hebrew dataset – The Hebrew dataset consists of seven children (6 male, 1 female) at the age of 6 to 10 (M=8.1, SD=1.6), all diagnosed with an autism spectrum condition by trained clinicians. 10 typically developing children (5 female, 5 male) at the age of 5 to 9 (M=7.2, SD=1.8) were selected to form the control group. As recording device, a Zoom H1 Handy Recorder was used. Recordings were taken at a sampling rate of 96 kHz and a quantization of 16 bits. Details on the number of utterances per emotion and group are given in Table 1.

Swedish dataset – A total number of 20 children took part in the recordings held in Sweden. The language throughout recordings is Swedish and all children are native speakers. The focus group consists of 9 children (9 male) at the age of 7 to 11 (M=9.1, SD=1.2). The control group consists of 11 children (5 male, 6 female) at the age of 5 to 9 (M=6.8, SD=1.7). The recording protocol adopted for the dataset is identical to the one used in the Hebrew dataset. As recording device, a Zoom H4 with RØDE NTG-2 microphone was used. Recordings were taken at a sampling rate of 96 kHz and a quantization of 16 bits. Details on the number of utterances per emotion and group are given in Table 1.

English dataset –A total number of 18 children (cf. Table 1) took part in the recordings held in England. The recordings are in English and all children are native speakers. The focus group consists of 9 children (5 male, 4 female) at the age of 7 to 11 (M=8.8, SD=1.5). The control group consists of 10 children (5 male, 5 female) at the age of 5 to 10 (M=7.9, SD=1.6). As recording device, a Zoom H1 Handy Recorder was used. Recordings were taken in wav format at a sampling rate of 96 kHz and a quantization of 16 bits.

Compared to the standards of present day databases used for automatic speech processing, this is a small database; however, taking into account the difficulties to recruit children from the envisaged population, to successfully conduct all the experimental tasks, and in comparison to other studies within the fields of ASC and emotion modelling for specific and less-studied populations, it can be taken as fairly representative, especially considering it contains recordings in three languages of both typically developing children and children with ASC under the same conditions.

3. Experiments

Four tasks were evaluated: typicality, emotion, valence, and arousal. The **typicality** task concerns the classification of typically developing children and children with ASC. The **emotion** task covers the recognition of the nine target classes (eight emotions plus "neutral"). We further evaluated the discrimination between high and low **arousal** as well as between positive and negative **valence**. The typicality task was performed on the full language-dependent database. All the emotion related tasks (emotion, valence, arousal) were performed on the focus and control group subsets separately.

3.1. Acoustic feature sets

Acoustic low-level descriptors (LLD) were automatically extracted from the speech waveform on a per-chunk level by using our open-source openSMILE feature extractor in its recent 2.1 release [15]. Three different feature sets were applied: a

large brute-forced feature set (ComParE), a smaller, expert-knowledge based feature set (GeMAPS), and its extended version (eGeMAPS). A detailed description and implementation of these feature set is given in [16].

The INTERSPEECH 2013 **ComParE** Challenge [9] feature set includes energy, spectral, cepstral (MFCC) and voicing related low-level descriptors (LLD) as well as logarithmic harmonic-to-noise ratio (HNR), spectral harmonicity, and psychoacoustic spectral sharpness. Altogether, the 2013 ComParE feature set contains 65 LLD and their first order derivatives – 130 LLD in total. Functionals are then applied to the LLDs obtaining a total of 6 373 features. The ComParE feature set is the result of a continuous refinement of acoustic descriptors used for the analysis of paralinguistics in speech and language. It has been successfully employed for the automatic recognition of various paralinguistic traits and states, such as those investigated in the INTERSPEECH Computational Paralinguistic Challenge (ComParE), e. g. personality [17], pathology [9], cognitive and physical load [18] and eating condition [19].

In contrast to large scale brute-force feature sets, two smaller, expert-knowledge based feature sets have also been applied. In fact, a minimalistic standard parameter set aims at reducing the risk of over-fitting in the training phase as compared to brute-forced large features sets, which in our task is of great interest. Recently, a recommended minimalistic standard parameter set for the acoustic analysis of speaker states and traits has been proposed in [20]. The proposed feature set is the so-called Geneva Minimalistic Acoustic Parameter Set (**GeMAPS**). Features were mainly selected based on their potential to index affective physiological changes in voice production, for their proven value in former studies, and for their theoretical definition. A detailed list of the LLD is provided in Table 2. We also applied the extended Geneva Minimalistic Acoustic Parameter Set (**eGeMAPS**) which basically contains 7 cepstral and dynamic-related LLD in addition to the 18 LLD in the minimalistic set.

3.2. Setup and evaluation

Since all data sets are unbalanced (i.e. one class is underrepresented in the data), the unweighted average recall (UAR) of the classes is used as scoring metric. Adopting the Weka toolkit [21], Support Vector Machines (SVMs) with linear kernel were trained with the Sequential Minimal Optimization (SMO) algorithm. SVMs have been chosen as classifier since they are a well known standard method for emotion recognition due to their capability to handle high and low dimensional data. The SVM training has been made at different complexity constant values $C \in \{0.001, 0.005, 0.01, 0.05\}$. To ensure speaker independent evaluations, Leave-One-Speaker-Out (LOSO) cross-validation has been performed. In order to balance the class distribution, we applied upsampling in all the evaluation experiments. Furthermore, we adopt the speaker z-normalisation (SN) method since it is known to improve the performance of speech-related recognition tasks, as described in [22]. With such a method, the feature values are normalised to a mean of zero and a standard deviation of one for each speaker. For the typicality task, we do not apply speaker z-normalisation since centring and scaling the feature space in such tasks is not effective because the phenomena considerably vary in the range across subjects. By applying this technique the relevant features able to characterise the subject are flattened, making the classification performances not acceptable and below the chance level. For the emotion-related task we first perform LOSO cross-validation within the control group. Then, we performed a mismatched evaluation by training only with material from TD children and testing on the focus group to quantify how emotion recognition from TD children's voice can be generalised on ASC children.

Table 2: GeMAPS (eGeMAPS) acoustic feature sets: 18(25) low-level descriptors (LLDs).

6(8) frequency related LLD	**Group**
F_0 (linear & semi-tone)	Prosodic
Jitter (local), Formant 1 (bandwidth)	Voice qual.
Formants 1, 2, 3 (frequency)	Vowel qual.
Formant 2, 3 (bandwidth) (eGeMAPS)	Voice qual.
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 (eGeMAPS)	Spectral
MFCC 1–4 (eGeMAPS)	Cepstral

4. Results

This section shows evaluation for the targeted tasks: typicality (Section 4.1), emotion, arousal, and valence (Section 4.2).

4.1. Typicality

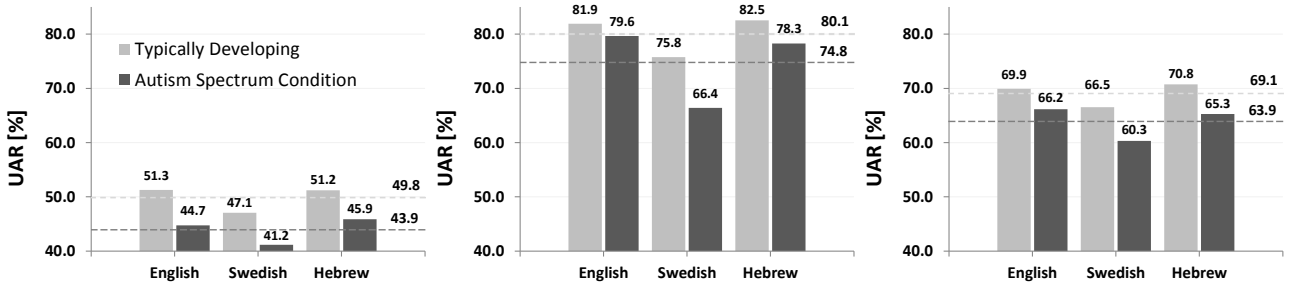
For the classification of typicality, we perform the task on each language dependent full dataset. Table 3 shows the best results obtained over the different complexities among the three feature sets. Applying the large set of features (ComParE), we obtain up to 78.3%, 86.4%, and 82.7% UAR for English, Swedish and Hebrew, respectively. However, reducing the feature space led to an expected – albeit moderate – decrease of performance on each language. In particular, we observe similar performances between GeMAPS and eGeMAPS, indicating that cepstral features (only included in eGeMAPS) are interestingly not prominently relevant in this task. One may note that all obtained performance are far above the chance level (50%), which was also found in [9] for French language.

4.2. Emotion related tasks

For emotion classification, we perform three different tasks: a 9-class emotion task, and a 2-class arousal and valence task. All the tasks were performed first within the control group with LOSO cross-validation; then we performed cross-group evaluation by training on the control group and testing on the focus group. We show the best results achieved over the different complexities among the three feature sets with speaker z-normalisation (cf. Table 4). In addition we analyse the differences between the three languages across the three tasks (cf. Figure 1).

Table 3: *Unweighted Average Recall for typicality task, respectively, on each language-dependent dataset. Typicality classes: typically developing children (TD), children with ASC (ASC).*

Language	**ComParE**	**eGeMAPS**	**GeMAPS**
English	**78.3**	75.5	75.2
Swedish	**86.4**	83.8	82.7
Hebrew	**82.7**	78.7	77.3



(a) **Emotion** (b) **Arousal** (c) **Valence**

Figure 1: Comparison over the three languages of the classification performances for (a) Emotion, (b) Arousal, and (c) Valence tasks. Results are given for LOSO cross-validation on typically developing children and for cross-group evaluation testing on children with ASC, using the ComParE features set. Dotted lines indicate average UAR over all the tasks per group.

Emotion {**9-class task**} – On the control group (cf. columns marked as TD in Table 4), we observe higher performances with the large set of features (ComParE) across the three languages. We obtain up to 51.3%, 47.1%, and 51.2% UAR for English, Swedish, and Hebrew; however, reducing the features space led to lower performance, which seems to be more prominent for the Swedish-dependent dataset down to 27.6%. The mismatched evaluation on the focus group revealed the difference between typically developing and ASC children in terms of performances; in fact, we obtained lower performances down to 44.7%, 41.2%, and 45.9% respectively for English, Swedish, and Hebrew. This difference is however quite small, and let supposes that the way TD children are encoding emotions in speech, e. g. using parametric variations of prosodic features, can be somehow generalised to ASC children, in terms of automatic recognition performance. The eGeMAPS and GeMAPS sets perform quite close, however a significant decrease in performance is observed across all languages.

Arousal {**2-class task**} – On the control group, we observe reasonably high performances with the large set of features (ComParE) across the three languages. We obtain up to 82.5% UAR for Hebrew; despite what we observed in the emotion task, reducing the features space led to similar performances. The mismatched evaluation on the focus group revealed less differences in terms of performances; in fact, we obtained performances up to 79.6%, 66.4%, and 78.3% respectively for English, Swedish, and Hebrew. The eGeMAPS and GeMAPS sets perform quite close to the large feature set, confirming that prosodic and voicing related features are relevant for the arousal.

Valence {**2-class task**} – Looking at the results obtained on the control group (cf. Table 4), the large set of features (ComParE) seems to perform better across the three languages. However, the eGeMAPS performed best up to 72.9% in the English-dependent dataset. Reducing the features space led to slightly lower performances, showing that GeMAPS and eGeMAPS features set can be robustly applied also for the valence task as well. As observed in the arousal task, the mismatched evaluation on the focus group revealed less differences in terms of performances; in fact, we obtained performances up to 66.2%, 60.3%, and 65.3% (using ComParE) respectively for English, Swedish, and Hebrew.

Evidence across languages – The cross-group evaluation across the three languages evidently show the mismatch between the two groups (cf. Figure 1). A delta of 5.9%, 5.3%, and 5.2% for emotion, arousal, and valence tasks is observed across the three languages. The Swedish dataset seems to show lower performance on average against the other two languages, which might stimulate future studies and cross-language analysis.

Table 4: *Unweighted Average Recall for emotion (9-class), and binary arousal and valence tasks, respectively, on each language-dependent dataset. Results are given for LOSO cross-validation on typically developing children (TD) and for cross-group evaluation training on the control group and testing on children with ASC. Dotted line indicate average UAR over all the tasks per group.*

Task	**ComParE**		**eGeMAPS**		**GeMAPS**	
	TD	ASC	TD	ASC	TD	ASC
English						
Emotion	**51.3**	**44.7**	40.0	39.0	38.5	36.7
Arousal	**81.9**	**79.6**	80.8	76.3	79.6	76.0
Valence	69.9	**66.2**	**72.9**	63.7	69.8	65.2
Swedish						
Emotion	**47.1**	**41.2**	28.1	22.7	27.6	21.8
Arousal	**75.6**	**66.4**	69.7	65.7	70.0	65.6
Valence	**66.5**	**60.3**	55.8	50.4	54.5	49.3
Hebrew						
Emotion	**51.2**	**45.9**	34.9	30.6	32.9	27.0
Arousal	**82.5**	**78.3**	76.1	75.2	76.3	74.2
Valence	**70.8**	**65.3**	63.8	60.6	62.2	61.3

5. Conclusions

Summing up, we evaluated a speech emotion database which is unique in collecting speech data of children with ASC and a typically developing control group in three different languages under the same conditions. Then, we discussed results concerning the classification of typicality, and of ASC children's emotional expressions, evaluating the 9-emotions task and the binary arousal/valence discrimination task across English, Swedish, and Hebrew languages. Since no study so far provided a comparison of performances across different languages, we aimed to fill this white spot in the literature and provide insight by extensive evaluations. The caveat has to be made that this is still a study, with a rather small number of cases per class; the results will be reviewed, verified or falsified, with larger databases collected in the future, including other languages. Future work will focus on cross-cultural/language analysis.

6. Acknowledgements

The research leading to these results has received funding from the EC Seventh Framework Programme under grant agreement no. 338164 (ERC iHEARu), and no. 645378 (RIA ARIA-VALUSPA).

7. References

[1] 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. of ICMCS/ICME*, 2011, pp. 1–6.

[2] 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. 02, pp. 591–617, 2006.

[3] B. Schuller, E. Marchi, S. Baron-Cohen, H. O'Reilly, P. Robinson, I. Davies, O. Golan, S. Friedenson, S. Tal, S. Newman, N. Meir, R. Shillo, A. Camurri, S. Piana, S. Bölte, D. Lundqvist, S. Berggren, A. Baranger, and N. Sullings, "ASC-Inclusion: Interactive Emotion Games for Social Inclusion of Children with Autism Spectrum Conditions," in *Proceedings 1st International Workshop on Intelligent Digital Games for Empowerment and Inclusion (IDGEI 2013) held in conjunction with the 8th Foundations of Digital Games 2013 (FDG)*, B. Schuller, L. Paletta, and N. Sabouret, Eds., ACM. Chania, Greece: SASDG, May 2013.

[4] B. Schuller, E. Marchi, S. Baron-Cohen, H. O'Reilly, D. Pigat, P. Robinson, I. Davies, 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, and N. Sullings, "The state of play of ASC-Inclusion: An Integrated Internet-Based Environment for Social Inclusion of Children with Autism Spectrum Conditions," in *Proceedings 2nd International Workshop on Digital Games for Empowerment and Inclusion (IDGEI 2014)*, L. Paletta, B. Schuller, P. Robinson, and N. Sabouret, Eds., ACM. Haifa, Israel: ACM, February 2014, held in conjunction with the 19th International Conference on Intelligent User Interfaces (IUI 2014).

[5] 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. 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. of the 3rd International Workshop on Intelligent Digital Games for Empowerment and Inclusion (IDGEI 2015) as part of the 20th ACM International Conference on Intelligent User Interfaces, IUI 2015*, L. Paletta, B. Schuller, P. Robinson, and N. Sabouret, Eds., ACM. Atlanta, GA: ACM, March 2015.

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

[7] 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)*, A. Neustein, Ed. Berlin: De Gruyter, 2014, invited contribution.

[8] 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, no. 5, pp. 1328–1342, July 2011.

[9] 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 *Proceedings INTERSPEECH 2013, 14th Annual Conference of the International Speech Communication Association*, ISCA. Lyon, France: ISCA, August 2013, 5 pages.

[10] 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.

[11] 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. of Interspeech*, Aug. 2011.

[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 *Proceedings 3rd Workshop on Child, Computer and Interaction (WOCCI 2012), Satellite Event of INTERSPEECH 2012*, ISCA. Portland, OR: ISCA, September 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 *Proceedings First International Workshop on Wide Spectrum Social Signal Processing (WS3P 2012), ASE/IEEE SocialCom 2012*, ASE/IEEE. Amsterdam, The Netherlands: IEEE, September 2012.

[14] 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. S. Elfström, "Voice Emotion Games: Language and Emotion in the Voice of Children with Autism Spectrum Condition," in *Proc. of the 3rd International Workshop on Intelligent Digital Games for Empowerment and Inclusion (IDGEI 2015) as part of the 20th ACM International Conference on Intelligent User Interfaces, IUI 2015*, L. Paletta, B. Schuller, P. Robinson, and N. Sabouret, Eds., ACM. Atlanta, GA: ACM, March 2015.

[15] 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*, ACM. Barcelona, Spain: ACM, October 2013, pp. 835–838.

[16] F. Eyben, "Real-time speech and music classification by large audio feature space extraction," Ph.D. dissertation, Technische Universität München, 2014, submitted, to appear.

[17] B. Schuller, S. Steidl, A. Batliner, E. Nöth, A. Vinciarelli, F. Burkhardt, R. van Son, F. Weninger, F. Eyben, T. Bocklet, G. Mohammadi, and B. Weiss, "The INTERSPEECH 2012 Speaker Trait Challenge," in *Proceedings INTERSPEECH 2012, 13th Annual Conference of the International Speech Communication Association*, ISCA. Portland, OR: ISCA, September 2012, 4 pages.

[18] 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. of INTERSPEECH 2014, 15th Annual Conference of the International Speech Communication Association (ISCA)*, Singapore, Republic of Singapore, September 2014, pp. 427–431.

[19] 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. of INTERSPEECH 2015, 16th Annual Conference of the International Speech Communication Association (ISCA)*, Dresden, Germany, September 2015.

[20] F. Eyben, K. Scherer, B. Schuller, J. Sundberg, E. André, C. Busso, L. Devillers, J. Epps, P. Laukka, S. Narayanan, and K. Truong, "The Geneva Minimalistic Acoustic Parameter Set (GeMAPS) for Voice Research and Affective Computing," *IEEE Transactions on Affective Computing*, 2015, to appear.

[21] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten, "The weka data mining software: an update," *SIGKDD Explor. Newsl.*, vol. 11, no. 1, pp. 10–18, Nov. 2009. [Online]. Available: http://doi.acm.org/10.1145/1656274.1656278

[22] B. Schuller, B. Vlasenko, F. Eyben, M. Wöllmer, A. Stuhlsatz, A. Wendemuth, and G. Rigoll, "Cross-Corpus Acoustic Emotion Recognition: Variances and Strategies," *IEEE Transactions on Affective Computing*, vol. 1, no. 2, pp. 119–131, July-December 2010.