Affective and behavioural computing: Lessons learnt from the First Computational Paralinguistics Challenge

Schuller, Bjorn; Weninger, Felix; Zhang, Yue; Ringeval, Fabien; Batliner, Anton; Steidl, Stefan; Eyben, Florian; Marchi, Erik; Vinciarelli, Alessandro; Scherer, Klaus R.; Chetouani, Mohamed; Mortillaro, Marcello

Available online at www.sciencedirect.com
ScienceDirect
Computer Speech & Language xxx (2018) xxx-xxx
www.elsevier.com/locate/csl

Affective and behavioural computing: Lessons learnt from the First Computational Paralinguistics Challenge

Q1 Björn Schuller[a,b,c,*], Felix Weninger[d], Yue Zhang[a,d], Fabien Ringeval[e,f], Anton Batliner[b,g], Stefan Steidl[g], Florian Eyben[e], Erik Marchi[d,e], Alessandro Vinciarelli[h], Klaus Scherer[c], Mohamed Chetouani[i], Marcello Mortillaro[c]

Q2 [a] *Imperial College London, GLAM — Group on Language, Audio, & Music, UK*
[b] *Chair of Embedded Intelligence for Health Care and Wellbeing, University of Augsburg, Germany*
[c] *Université de Genève, Swiss Center for Affective Sciences, Switzerland*
[d] *Machine Intelligence & Signal Processing group, Technische Universität München, Germany*
[e] *audEERING GmbH, Germany*
[f] *Université Grenoble Alpes, CNRS, Grenoble INP, LIG, Grenoble, France*
[g] *Pattern Recognition Lab, Friedrich-Alexander-Universität Erlangen-Nürnberg, Germany*
[h] *School of Computing Science, University of Glasgow, UK*
[i] *Sorbonne Universités, Université Pierre-et-Marie-Curie, Paris, France*

Received 10 December 2016; received in revised form 1 October 2017; accepted 19 February 2018
Available online xxx

Abstract

In this article, we review the INTERSPEECH 2013 Computational Paralinguistics ChallengE (ComParE) — the first of its kind — in light of the recent developments in affective and behavioural computing. The impact of the first ComParE instalment is manifold: first, it featured various new recognition tasks including social signals such as laughter and fillers, conflict in dyadic group discussions, and atypical communication due to pervasive developmental disorders, as well as enacted emotion; second, it marked the onset of the ComParE, subsuming all tasks investigated hitherto within the realm of computational paralinguistics; finally, besides providing a unified test-bed under well-defined and strictly comparable conditions, we present the definite feature vector used for computation of the baselines, thus laying the foundation for a successful series of follow-up Challenges. Starting with a review of the preceding INTERSPEECH Challenges, we present the four Sub-Challenges of ComParE 2013. In particular, we provide details of the Challenge databases and a meta-analysis by conducting experiments of logistic regression on single features and evaluating the performances achieved by the participants.

Keywords: Computational Paralinguistics; Social Signals; Conflict; Emotion; Autism; Survey; Challenge

Q3 * Corresponding author at: Chair of Embedded Intelligence for Health Care and Wellbeing, University of Augsburg, Germany.
E-mail address: bjoern.schuller@imperial.ac.uk (B. Schuller).

http://dx.doi.org/10.1016/j.csl.2018.02.004
0885-2308/ 2018 Published by Elsevier Ltd.

1. Introduction

Affective Computing, focusing on the emotional mechanisms in natural human-machine interaction, has been an active topic for two decades now since its early emergence in the second quinquennium of the 1990s (Picard, 1997). Affective computers are aimed to recognise, express, model, communicate, and respond to emotional information, thus providing better performance in collaboration and communication with human beings (Picard, 1997). Propelled by the advances in speech processing technology, many of the suggested applications of affective computing to computer-assisted learning, perceptual information retrieval, arts and entertainment, and human health and interaction as envisioned in Picard's pioneering work have already become reality, e.g., wearable computer devices, interactive emotion games for social inclusion of people with autism spectrum condition (ASC), and big data analytic systems.

From a psychological point of view, the realm of affect extends beyond the domain of emotions and moods (Russell, 2003; Beedie et al., 2005); in current studies, the terms affect, mood, and emotion are often used interchangeably, without much effort at conceptual differentiation (Ekkekakis, 2013). In an attempt to draw some lines of demarcation, Russell (2009) advocated the concept of *core affect* as a neurophysiological state, accessible to consciousness as a simple non-reflective feeling: feeling good or bad, feeling lethargic or energised, with the two underlying dimensions of pleasure−displeasure and activation−deactivation.

Most importantly, in spite of the paramount importance of affect, it only presents one facet of human beings, thus the paradigm of affective computing has been shifting towards a more holistic understanding of human social intelligence (Albrecht, 2006). In this context, Pentland (2007) and Vinciarelli et al. (2012a) pioneered the domain of social signal processing, with the aim to endow machines with human-like emotional, social perceptual and behavioural abilities.

For speech processing, the paradigm shift has led to an increasing attention to the automatic recognition of speaker characteristics beyond affective states, which has enabled a new broad spectrum of applications such as virtual assistants with personalised aspects, safety and security monitoring services, and speaker identification systems. There is currently a wealth of loosely connected studies, mostly on affect recognition (including emotion, depression, and stress level), but also recognition of other speaker states and traits such as sleepiness, alcohol intoxication (Schiel and Heinrich, 2009), health condition (Maier et al., 2009), personality (Mohammadi et al., 2010), and biological primitives in terms of age, gender, height, weight (Krauss et al., 2002; Schuller et al., 2013). From the plethora of well studied and currently under-researched speech phenomena, a new major field of speech technology research has been emerging, termed 'computational paralinguistics' by Schuller (2012) and Schuller and Batliner (2014).

2. The INTERSPEECH challenges

Along with the growing maturity of this field, different research challenges have been established, allowing researchers to compare their affect recognition systems with benchmark performances, and at the same time, addressing the different channels of affect manifestations such as facial expression, body gesture, speech, and physiological signals (e.g., heart rate, skin conductivity) (Tao and Tan, 2005). For instance, the Audio/Visual Emotion Challenge and Workshop (AVEC) is aimed at bridging between different modalities by featuring audio, visual, and audiovisual analysis for spontaneous emotion recognition (Ringeval et al., 2015). Likewise, the Emotion Recognition In The Wild Challenge and Workshop (EmotiW) scopes multimodal emotion recognition, while focusing on snippets of movies (Dhall et al., 2013). The MediaEval Benchmarking Initiative for Multimedia Evaluation[1] sets a special focus on the social and human aspects of multimedia access and retrieval, while emphasising the 'multi' in multimedia involving speech recognition, content analysis, music and audio analysis, user-contributed information (tags, tweets), viewer affective response, social networks, temporal and geo-coordinates.

The INTERSPEECH Challenges 2009 to 2012 were held in conjunction with the annual INTERSPEECH conference, one of the prime venues in speech signal processing. In the following, we detail the task specifications, data, features, Challenge conditions and evaluations of this Challenge series. The first INTERSPEECH 2009 Emotion Challenge (IS09EC) (Schuller et al., 2009; 2011a) featured a binary (idle vs negative) and a five-way (anger, emphatic, neutral, positive, and rest) classification task on the FAU Aibo Emotion Corpus of naturalistic children's

[1] http://www.multimediaeval.org/.

Please cite this article as: B. Schuller et al., Affective and behavioural computing: Lessons learnt from the First Computational Paralinguistics Challenge, Computer Speech & Language (2018), http://dx.doi.org/10.1016/j.csl.2018.02.004

speech (Steidl, 2009). In light of the Challenge, the first widely used open-source affect analysis toolkit openEAR (Eyben et al., 2009) was introduced. A follow-up effort, the INTERSPEECH 2010 Paralinguistic Challenge (IS10PC) (Schuller et al., 2010; 2013), evaluated the continuous-valued level of interest ([-1,+1]) and the biological primitives age (child, youth, adult, and senior) and gender/age (female, male, and children). In the ensuing INTERSPEECH 2011 Speaker State Challenge (IS11SSC) (Schuller et al., 2011b; 2014), intoxication (above or below 0.5 per mill blood alcohol concentration) and sleepiness (above or below 7.5 on the Karolinska sleepiness scale) had to be detected. Finally, in the INTERSPEECH 2012 Speaker Trait Challenge (IS12STC) (Schuller et al., 2012; 2015), personality (openness, conscientiousness, extraversion, agreeableness, and neuroticism), likability, and intelligibility of pathological speakers were investigated, where all tasks were binarised to above or below average.

Specifically, high realism was fostered in the choice of all Challenge data, e.g., genuine intoxication and sleep deprivation was given, and spontaneous speech was considered for tasks based on subjective perception. Furthermore, partitioning is strictly subject-independent, whenever possible. Only the first Challenge did not feature a development partition. The subsequent Challenges defined roughly a 40:30:30 partitioning for the training, the development, and the test set, where training and development were united for the baseline computation. Test data − without target labels − were provided to the participants, who had limited trials of result submissions per competing site. To uphold the quality and validity of research, the individual paper submissions undergo the regular INTERSPEECH peer-review process and have to be accepted for the conference in order to participate in the Challenge. In each Challenge, an acoustic feature set was specified, comprising 384, 1582, 4368, and 6125 attributes, respectively (2009−2012), which were obtained by applying statistical functionals to low-level descriptors. For transparency, the openSMILE feature extraction toolkit has been consistently used over the years (Eyben et al., 2010; 2013); openEAR is a release of openSMILE including models for emotion recognition as targeted in the IS09EC Challenge. Another distinguishing mark of this Challenge series is the reproducibility for the learning algorithms by consistently using the data mining toolkit WEKA 3 (Witten and Frank, 2005). Last but not least, the popularity of these events has steadily increased from 33 to 52 registered participants. An overview of the Challenge results is given in Table 1. It can be seen from the table that the baselines always were competitive but could be surpassed by the winners, and that in all but one cases, the majority vote could surpass the single best vote by a small margin.

3. The First Computational Paralinguistics Challenge (ComParE)

Fig. 1 depicts an exemplary space of speaker characteristics spanned by the axes of subjectivity and time, ranging from temporary speaker states to long-term speaker traits, and from objective measures (ground truth) to subjective gold standards determined through inter-rater procedures.

As can be seen from the taxonomic representation in Fig. 1, the tasks investigated in the INTERSPEECH Challenges represent specific sub-domains and much scope is left for exploration in the broad field of paralinguistic speech phenomena. Based on this motivation, the first Challenge of the ComParE series was aimed at illuminating a cross-section of closely connected tasks of high relevance for affective and behavioural research, and subsuming different kinds of investigated and potential new tasks under the umbrella of computational paralinguistics (Schuller

Table 1
Results of the INTERSPEECH 2009−2012 Challenges. Evaluation measures: unweighted average recall (UAR [%]), Pearson's correlation coefficient (CC). Base: baseline results. Best: best participant results. Vote: majority vote over the optimal number (shown in parentheses) of the participants' results.

Challenge	Tasks	Classes	Base	Best	Vote
IS12STC	Personality	2	68.3	71.6 (Ivanov and Chen)	70.4 (5)
	Likability	2	59.0	65.8 (Montacié and Caraty)	68.7 (3)
	Intelligibility	2	68.9	76.8 (Kim et al.)	76.8 (1)
IS11STC	Intoxication	2	65.9	70.5 (Bone et al.)	72.2 (3)
	Sleepiness	2	70.3	71.7 (Huang et al.)	72.5 (3)
IS10PC	Age	4	48.9	52.4 (Kockmann et al.)	53.6 (4)
	Gender	3	81.2	84.3 (Meinedo and Trancoso)	85.7 (5)
	Interest	[-1,1]	0.421	0.428 (Jeon et al.)	–
IS09EC	Emotion	5	38.2	41.6 (Dumouchel et al.)	44.0 (5)
	Negativity	2	67.7	70.3 (Lee et al.)	71.2 (7)

Please cite this article as: B. Schuller et al., Affective and behavioural computing: Lessons learnt from the First Computational Paralinguistics Challenge, Computer Speech & Language (2018), http://dx.doi.org/10.1016/j.csl.2018.02.004

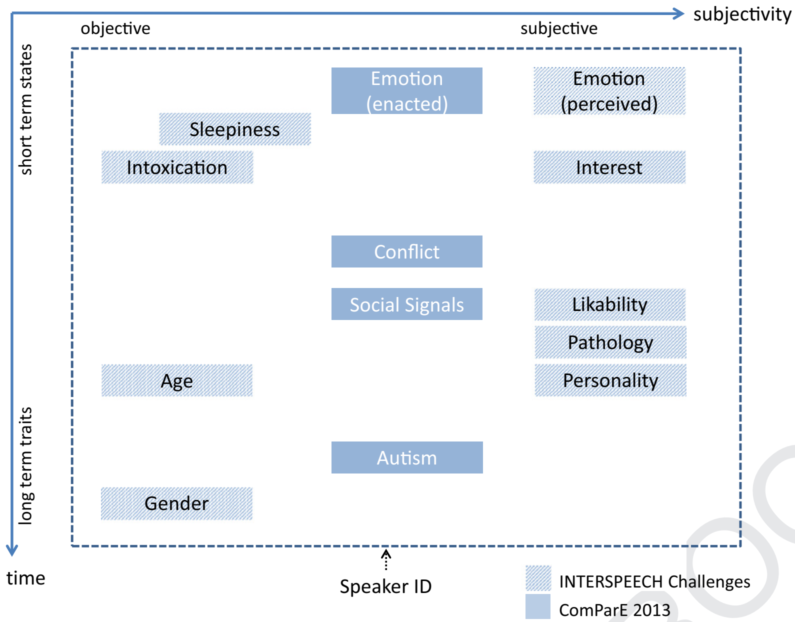


Fig. 1. Speaker characteristics investigated in the INTERSPEECH Challenges 2009−2012 and the First Computational Paralinguistics Challenge (ComParE) 2013.

and Batliner, 2014). Thus, in response to the growing popularity of the Challenge series (Schuller, 2012), ComParE 2013 broadened the scope with a larger variety of tasks compared to previous years. In line with INTERSPEECH 2013's theme *Speech in Life Sciences and Human Societies*, social signals (Vinciarelli et al., 2009) and conflicts in communication (Roth and Tobin, 2010) as occurring in real-life were detected and localised. In addition, we re-addressed the topics emotion and intelligibility from IS09EC and IS12STC by introducing new databases and task definitions.

3.1. Challenge corpora

3.1.1. SSPNet Vocalisation Corpus (SVC)

The Social Signals Sub-Challenge was carried out on the "SSPNet Vocalisation Corpus" (SVC), which contains 2 763 audio clips of 11 seconds (total duration: 8.4 h) annotated in regard to laughter and fillers. Laughter (Bachorowski et al., 2001; Vettin and Todt, 2004; Tanaka and Campbell, 2011) in terms of vocal outbursts can be regarded as an indicator for amusement, joy, scorn, or embarrassment. Fillers such as *um, er, uh* in English are frequently used delays in speaking when the speaker needs to bridge the time when searching for a word or deciding what to say next (Clark and Fox Tree, 2002). The corpus was extracted from a collection of 60 phone calls involving 120 subjects (63 female, 57 male) (Vinciarelli et al., 2012b).

The fillers were identified manually by an individual annotator and then validated (accepted or discarded) by a second, independent listener. Thus, the corpus includes only fillers for which there is agreement between annotator and listener. The identification of the fillers was performed with a tool allowing one to manually set beginning and end of a given filler. In case of ambiguity, start or end point were set in correspondence of the earliest or latest point, respectively, where the signal actually corresponded to a filler for both annotator and listener. The tool allows one to set a point with an error as small as the sampling period of the signal (the time interval between two consecutive samples). However, the tool was used with a precision of 30 ms, a value sufficiently good for automatic processing like the one described in this work.

The participants of each call were fully unacquainted and never met face-to-face before or during the experiment. The calls revolved around the Winter Survival Task: The two participants had to identify objects (out of a predefined list) that increase the chances of survival in a polar environment. The subjects were not given instructions on how to conduct the conversation, the only constraint was to discuss only one object at a time. The conversations were recorded on both phones (model Nokia N900) used during the call. The clips were extracted from the microphone recordings of the phones. Thus, clips from the same speaker never overlap, whereas clips from two subjects participating in the same call may overlap (for example in the case of simultaneous laughter). However, they do not contain the same audio data because they are recorded with separate microphones. Each clip was selected in such a way that it contains at least one laughter or filler event between $t = 1.5$ s and $t = 9.5$ s. In total, the database contains 2988 filler events and 1158 laughter events. Both types of vocalisation in this database can be considered fully spontaneous. Given this layout, the Social Signals Sub-Challenge introduced for the first time a frame-wise detection and localisation task instead of supra-segmental classification as in the other Sub-Challenges and all previous Challenges. The data were divided into speaker disjoint subsets for training, development, and testing. For transparency, this was simply done by using calls 1−35 (70 speakers) for training, calls 36−45 (20 speakers) for development, and calls 46−60 for testing. The Challenge data were delivered with a manual segmentation of the training and development data into 'garbage', 'laughter', and 'filler' segments, in the 'master label file' (MLF) format used by the Hidden Markov Model Toolkit (HTK) (Young et al., 2006). Further meta data were not provided. The resulting partitioning by numbers of utterances, number of vocalisation segments (filler, laughter) as well as vocalisation and garbage frames (100 per second) is shown in Table 2.

3.1.2. SSPNet Conflict Corpus (SC^2)

In the Conflict Sub-Challenge, the "SSPNet Conflict Corpus" (SC^2) was used (Kim et al., 2012b). It contains 1 430 clips of 30 seconds (total duration: 11.9 h) extracted from the Canal9 Corpus − a collection of 45 Swiss political debates (in French). For the Challenge, 110 subjects in total: 18 females (1 moderator and 17 participants) and 92 males (1 moderators and 91 participants) were considered. The clips were annotated in terms of conflict level by 551 assessors recruited via Amazon Mechanical Turk. The annotation was performed using a questionnaire fully described by Kim et al. (2012b). As the goal of the corpus was the study of nonverbal communication, only non-French speakers were involved. In this way it was possible to avoid, or at least to limit, the effect of the content (Kim et al., 2014). Every clip was rated by 10 randomly assigned annotators and the agreement was measured in terms of *effective reliability* R (Rosenthal, 2005):

$$R = \frac{Nr}{1+(N-1)r} \quad ; \quad r = 2\frac{\sum_{i=1}^{N}\sum_{j=i+1}^{N} r_{ij}}{N(N-1)} \tag{1}$$

where N is the number of assessors and r is the average of the correlations between all possible pairs of assessors (r_{ij} is the correlation between assessors i and j). The observed value of R for the corpus was 0.91, above the threshold of 0.90 that the literature considers to be sufficient in experimental practice (Rosenthal, 2005).

Each clip is associated with a continuous conflict score in the range [–10, +10], giving rise to a straightforward regression task ('Score' task). A classification task was specified based on these labels, which were binarised into 'high' (≥ 0) or 'low' (< 0) level of conflict ('Class' task). As several subjects were involved in debates with

Table 2
Partitioning of the SSPNet Vocalisation Corpus into train, dev(elopment), and test set: numbers of utterances, vocalisation segments (laughter, filler), and vocalisation/'garbage' frames.

#	Train	Dev	Test	Σ
Utterances				
Σ	1583	500	680	2763
Segments				
Laughter	649	225	284	1158
Filler	1710	556	722	2988
Frames				
Laughter	59,294	25,750	23,994	109,038
Filler	85034	29,432	35,459	149,925
Garbage	1,591,442[a]	492,607	684,937	2,768,986
Σ	1,735,770	547,789	744,390	3,027,949

[a] 79,572 frames after training set balancing by re-sampling.

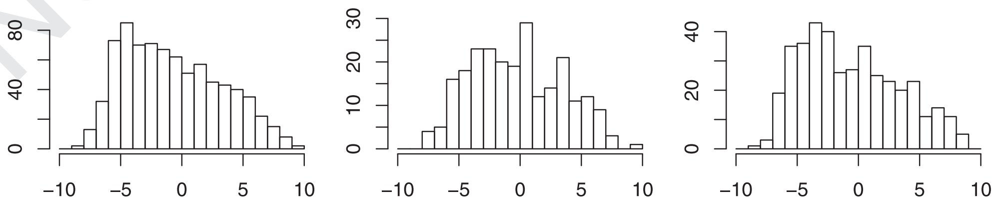
different moderators, a truly speaker independent partitioning was not possible for these data. Considering the fact that all participants except the moderators are not present more than a few times (mostly only once), the following strategy was followed to reduce speaker dependency to a minimum. All broadcasts with the female moderator (speaker # 50) were assigned to the training set. The development set consists of all broadcasts moderated by the (male) speaker # 153, and the test set comprises the remaining male moderators. This also ensures that the development and test sets are similar in case that the gender of the moderator had an influence. The resulting partitioning is shown in Table 3, along with the distribution of binary class labels and continuous ratings (Fig. 2) among the partitions. The training set comprises 55% of the data, the development 17% and the test set 28%. A drawback of this partitioning is the rather small development set, but participants were encouraged to use both training and development set for data analysis. As meta data, manual speaker segmentation, as well as role (participant/moderator) and gender of the subjects were provided for the training and development sets. Participants were encouraged to use the manual speaker segmentation for the development of features extraction, but an automatic speaker diarisation system had to be used for the test set.

3.1.3. Geneva Multimodal Emotion Portrayals (GEMEP)

For the Emotion Sub-Challenge, the “Geneva Multimodal Emotion Portrayals” (GEMEP) corpus (Bänziger et al., 2012) was selected. It comprises 1 260 instances of emotional speech (total duration: 8.9 h) from ten professional actors (five female) in 18 categories. Specifically, prompted speech, which contains sustained vowel phonations and two ‘nonsensical’ phrases (phrase #1: ‘ne kal ibam soud molen!’, phrase #2: ‘koun se mina lod belam?’) with two different intended sentence modalities were pronounced by each actor in various degrees of regulation (emotional intensity) ranging from ‘high’ to ‘masked’ (hiding the true emotion). As a partitioning that is both text and speaker disjoint is not feasible, we used vowels and phrase #2 subdivided by speaker ID for training and development, and phrase #1 for testing, to ensure text independence. Masked regulation utterances are only included in the test set in order to alleviate potential model distortions. This is similar to typical automatic speech recognition tasks where the lowest signal-to-noise ratios are only encountered in the test set. As six of the 18 emotional categories are extremely sparse ($\leq$ 30 instances in total), we restricted the evaluation to the 12 most frequent ones in the multi-class classification task. The classification labels for each utterance correspond to the emotions intended to be acted; no manual annotation is done. For the binary tasks, mappings of the original labels were only applied on those emotion categories such as to obtain a balanced distribution of positive/negative instances for the dimensions arousal and valence. Nevertheless, the remaining data were given to the participants (with labels in 18 categories for the training and development sets), which could be used, e.g., to train ‘background’ or ‘garbage’ models. The resulting partitioning is shown in Table 4. As meta data, actor IDs, prompts, and intended regulation were released for the training and the development set.

Table 3
Partitioning of the SSPNet Conflict Corpus into train, dev(elopment), and test set for binary classification (‘low’ $\equiv[-10,0[$, ‘high’ $\equiv[0,+10]$).

#	Train	Dev	Test	Σ
Low	471	127	226	824
High	322	113	171	606
Σ	793	240	397	1430



(a) train (b) devel (c) test

Fig. 2. Level of conflict ($\in[-10,+10]$) histograms for the Challenge partitions of the SSPNet Conflict Corpus.

Please cite this article as: B. Schuller et al., Affective and behavioural computing: Lessons learnt from the First Computational Paralinguistics Challenge, Computer Speech & Language (2018), http://dx.doi.org/10.1016/j.csl.2018.02.004

3.1.4. Child Pathological Speech Database (CPSD)

The AUTISM SUB-CHALLENGE used the “Child Pathological Speech Database” (CPSD) (Ringeval et al., 2011), created at two university departments of child and adolescent psychiatry (Université Pierre et Marie Curie/Pitié-Salpêtière Hospital and Université René Descartes/Necker Hospital), located in Paris, France. The recordings are prompted sentence imitation of 26 sentences representing different modalities (declarative, exclamatory, interrogative, and imperative) and four types of intonations (descending, falling, floating, and rising); another version of this database including emotional speech (CPESD) has been recently studied and released (Ringeval et al., 2016; Schmitt et al., 2016). The CPSD dataset used in the Sub-Challenge comprises 2 542 instances of speech recordings (total duration: 1 h) from 99 children aged 6 to 18 years; 35 of these children show either pervasive development disorders of autism spectrum condition (PDD, 10 male, 2 female), specific language impairment such as dysphasia (DYS, 10 male, 3 female), or PDD non-otherwise specified (NOS, 9 male, 1 female), according to the DSM-IV criteria[2] (First, 1994), which distinguish ASC subtypes: e.g., Autism Disorders (AD), with symptoms in all areas that characterise PDD; or PDD-NOS, which is characterised by social, communicative and/or stereotypical impairments that are less severe than in AD. Further, a monolingual control group of 64 typically developing children (TYP, 52 male, 12 female) is included. None of the TYP subjects had a history of speech, language, hearing or general learning problems (Demouy et al., 2011).

Typically developing children were recorded in two different places according to their age (middle/high school), whereas children with developmental conditions were either recorded at home or at the clinic (DYS: Necker Hospital, PDD and PDD-NOS: Pitié-Salpêtière Hospital), depending on their availability. Various acoustic conditions are thus present in the data due to the use of different places for the recordings of the children; two different places for TYP, and at least four different places for the three groups of children suffering developmental conditions.

Two evaluation tasks were specified: a binary ‘Typicality’ task (typically vs atypically developing children), and a four-way ‘Diagnosis’ task (classifying into the above named categories). Note that by ‘Diagnosis’, we refer to the classification of the children’s developmental condition in the four classes reported by the clinicians using DSM-IV criteria. Performance reported by the automatic classification of those conditions thus reflect the agreement of the system with the diagnosis provided by the clinicians on the children from the CPSD database, which can evolve

Table 4
Partitioning of the GEMEP database into train, dev(elopment), and test set for 12-way classification by emotion category, and binary classification by pos(itive)/neg(ative) arousal (A) and valence (V).

#	Train	Dev	Test	A	V	Σ
Admiration[+]	20	2	8	pos	pos	30
Amusement	40	20	30	pos	pos	90
Anxiety	40	20	30	neg	neg	90
Cold anger	42	12	36	neg	neg	90
Contempt[+]	20	6	4	neg	neg	30
Despair	40	20	30	pos	neg	90
Disgust[+]	20	2	8	−*	−*	30
Elation	40	12	38	pos	pos	90
Hot anger	40	20	30	pos	neg	90
Interest	40	20	30	neg	pos	90
Panic fear	40	12	38	pos	neg	90
Pleasure	40	20	30	neg	pos	90
Pride	40	12	38	pos	pos	90
Relief	40	12	38	neg	pos	90
Sadness	40	12	38	neg	neg	90
Shame[+]	20	2	8	pos	neg	30
Surprise[+]	20	6	4	−*	−*	30
Tenderness[+]	20	6	4	neg	pos	30
Σ	602	216	442			1260

[+] Mapped to ‘other’ and excluded from evaluation in 12-class task.
* Mapped to ‘undefined’ and excluded from evaluation in binary tasks.

[2] Even though the recent DSM-V adopted a single diagnosis of ASC based on dimensional features, we kept the definition of DSM-IV for this study, since ASC children from the CPSD database were originally diagnosed with the criteria of the DSM-IV.

Please cite this article as: B. Schuller et al., Affective and behavioural computing: Lessons learnt from the First Computational Paralinguistics Challenge, Computer Speech & Language (2018), http://dx.doi.org/10.1016/j.csl.2018.02.004

over time. Speaker independent partitioning into training, development, and test data was performed on stratified data according to the children's age and gender. The respective class distribution is shown in Table 5. As additional meta data, age and gender of the children were enclosed.

Because evaluations performed in this study are speaker-independent, it is probable that some tested subjects present acoustic conditions that have not been seen either during the training or the optimisation of the hyper-parameters of the classifier (e.g., a child recorded at home). In a practical perspective, such conditions for system training would be ideal for the development of health care systems that would work well at home on unseen children, while taking additional benefits from recordings collected at the hospital.

3.2. The overall scope of the ComParE 2013

Ideally, we could choose for each year's sub-challenges amongst many database candidates the ones that fit together under a clearly defined umbrella. However, suitable candidates are rather scarce because they have to meet several conditions, i. e., they have to be new (especially the test set), large enough for experimental purposes, and of considerable interest for the community. Nevertheless, the four sub-challenges in this first ComParE Challenge reflect pivotal aspects of human communication – to be more precise, of specific 'non-communications' and problems of a-typicality, according to the type of speaker and speech phenomenon:

- In the SSPNet Vocalisation corpus SVC (social signal sub-challenge), laughter and fillers represent 'non-semantic' phenomena, which are very helpful for characterising speakers and gaining a deeper understanding of dialogues beyond the sole exchange of semantic messages. They can be modelled and detected together with words, but have been disregarded in 'classic' Automatic Speech Recognition (ASR).
- In the SSPNet Conflict Corpus (SC^2), conflict occurs as a disruptive event that frequently results in speech overlaps, thus creating problems for ASR and speech modelling.
- In the Geneva Multimodal Emotion Portrayals (GEMEP), pronounced but unrealistic portrayals of frequent and less frequent emotions, serves as a upper baseline for modelling a many-class problem and demonstrates the difficulty of this task even in 'ideal' conditions; it has been shown that, when going over to realistic, spontaneous data, performance considerably deteriorates (Batliner et al., 2000; Vogt and André, 2005).
- In the Child Pathological Speech Database (CPSD), a-typical speech, which often forms a obstacle for standard ASR, can used for modelling these specific types of speech pathologies (Bone et al., 2012; Marchi et al., 2015; McCann and Peppe, 2003; Van Santen et al., 2010; Demouy et al., 2011).

Following the preceding INTERSPEECH Challenges' example, strict comparability, transparency and reproducibility, as well as research validation through peer-review were maintained. From this ComParE Challenge onwards, a 'recipe' for re-producing the baseline classification and regression results on the development set in an automated fashion has been supplied, embedding the entire workflow from pre-processing, over model training and evaluation, to scoring by the according measures.

Table 5
Partitioning of the Child Pathological Speech Database into train, dev(elopment), and test set for four-way classification by diagnosis, and binary classification by typical / atypical development. Diagnosis classes: typically developing (TYP), pervasive developmental disorders (PDD), pervasive developmental disorders non-otherwise specified (NOS), and specific language impairment such as dysphasia (DYS).

#	Train	Dev	Test	Σ
Typically developing				
TYP	566	543	542	1651
Atypically developing				
PDD	104	104	99	307
NOS	104	68	75	247
DYS	129	104	104	337
Σ	903	819	820	2542

3.3. Challenge features

As standard acoustic feature set to be used as the new reference in the ComParE series, we modified the feature set adapted from the INTERSPEECH 2012 Speaker Trait Challenge (Schuller et al., 2012) − the most effective one up to that point (cf. Section 2). In detail, voice quality features (jitter and shimmer) were slightly improved, slight modifications of the F_0 extraction algorithms were made (i. e., the non-greedy peak detection was replaced by a greedy one), and the rules for applying functionals to low-level descriptors (LLD) were simplified. Altogether, the ComParE feature set contains 6 373 attributes, including energy, spectral, cepstral (MFCC), and voicing related LLDs as well as a few other LLDs (e.g., logarithmic harmonic-to-noise ratio (HNR), spectral harmonicity, and psychoacoustic spectral sharpness). Different sets of functionals are applied to two groups of LLDs. Group A of LLDs consists of four energy related LLDs and 55 spectral LLDs; group B consists of the remaining 6 voicing related LLDs. A set of 54 functionals is applied to the LLDs of group A, and 46 functionals are applied to the ΔLLDs of this group, resulting in $59 \cdot (54 + 46) = 5\,900$ acoustic features. A smaller set of only 39 functionals is applied to the LLDs of group B and their ΔLLDs, resulting in $6 \cdot (39 + 39) = 468$ acoustic features. In addition, five temporal statistic descriptors are computed for voiced segments: the mean length, the standard deviation of the segment length, the minimum length, and the maximum length of the voiced segments, and the ratio of non-zero F_0 values. In total, the final feature set consists of $5900 + 468 + 5 = 6373$ features. The sets of LLDs and applied functionals are given in Tables 6 and 7, respectively. For a more detailed description of the functionals and LLDs as well as the underlying algorithms, please refer to Eyben (2015).

For the Social Signals Sub-Challenge that requires localisation, a frame-wise feature set was derived. Taking into account space and memory requirements, only a small set of descriptors was calculated per frame, following a sliding window scheme to combine frame-wise LLDs and functionals. In particular, frame-wise MFCCs 1−12 and logarithmic energy were computed along with their first and second order delta (Δ) regression coefficients as typically processed in speech recognition. They were augmented by voicing probability, HNR, F_0, and zero-crossing rate, as well as their first order Δs. Subsequently, each frame-wise LLD is augmented by the arithmetic mean and standard deviation across the frame itself and eight of its neighbouring frames (four before and four after), resulting in $47 \cdot 3 = 141$ descriptors per frame.

3.4. Challenge baselines

As primary evaluation measure, we retained the choice of unweighted average recall (UAR) as used since IS09EC (Schuller et al., 2011a). The reason to consider *unweighted* rather than weighted average recall (‘conventional’ accuracy) is that it is also meaningful for highly unbalanced distributions of instances among classes, as is the case in, e. g., the Autism Sub-Challenge. Given the nature of the Social Signals Sub-Challenge as a detection-oriented task, we

Table 6
ComParE acoustic feature set: 65 provided **low-level descriptors** (LLD).

4 Energy Related LLD	Group
Sum of Auditory Spectrum (Loudness)	Prosodic
Sum of RASTA-Style Filtered Auditory Spectrum	Prosodic
RMS Energy, Zero-Crossing Rate	Prosodic
55 Spectral LLD	**Group**
RASTA-Style Auditory Spectrum, Bands 1−26 (0−8 kHz)	Spectral
MFCC 1−14	Cepstral
Spectral Energy 250−650 Hz, 1 k−4 kHz	Spectral
Spectral Roll Off Point 0.25, 0.50, 0.75, 0.90	Spectral
Spectral Flux, Centroid, Entropy, Slope, Harmonicity	Spectral
Spectral Psychoacoustic Sharpness	Spectral
Spectral Variance, Skewness, Kurtosis	Spectral
6 Voicing Related LLD	**Group**
F_0 (SHS & Viterbi Smoothing)	Prosodic
Probability of Voicing	Sound Quality
Log. HNR, Jitter (Local, Delta), Shimmer (Local)	Sound Quality

Please cite this article as: B. Schuller et al., Affective and behavioural computing: Lessons learnt from the First Computational Paralinguistics Challenge, Computer Speech & Language (2018), http://dx.doi.org/10.1016/j.csl.2018.02.004

Table 7
ComParE acoustic feature set: **functionals** applied to LLDs as defined in Table 6.

Mean Values
Arithmetic Mean$^{A\triangleq,B}$, Arithmetic Mean of Positive Values$^{A\Delta,B}$,
Root-Quadratic Mean, Flatness
Moments: Standard Deviation, Skewness, Kurtosis
Temporal Centroid$^{A\triangleq,B}$
Percentiles
Quartiles 1−3, Inter-Quartile Ranges 1−2, 2−3, 1−3,
1%-tile, 99%-tile, Range 1−99%
Extrema
Relative Position of Maximum and Minimum, Full Range (Maximum−Minimum)
Peaks and Valleys^{A}
Mean of Peak Amplitudes,
Difference of Mean of Peak Amplitudes to Arithmetic Mean,
Mean of Peak Amplitudes Relative to Arithmetic Mean,
Peak to Peak Distances: Mean and Standard Deviation,
Peak Range Relative to Arithmetic Mean
Range of Peak Amplitude Values,
Range of Valley Amplitude Values Relative to Arithmetic Mean,
Valley-Peak (Rising) Slopes: Mean and Standard Deviation,
Peak-Valley (Falling) Slopes: Mean and Standard Deviation
Up-Level Times: 25%, 50%, 75%, 90%
Rise and Curvature Time
Relative Time in which Signal is Rising,
Relative Time in which Signal has Left Curvative
Segment Lengths^{A}
Mean, Standard Deviation, Minimum, Maximum
Regression$^{A\triangleq,B}$
Linear Regression: Slope, Offset, Quadratic Error,
Quadratic Regression: Coefficients *a* and *b*, Offset *c*, Quadratic Error
Linear Prediction
LP Analysis Gain (Amplitude Error), LP Coefficients 1−5

^{A}Functionals applied only to energy related and spectral LLDs (group A)
^{B}Functionals applied only to voicing related LLDs (group B)
$^{\Delta}$Functionals applied only to ΔLLDs
$\triangleq$ Functionals **not** applied to ΔLLDs

also considered the Area Under the Curve measure (Witten and Frank, 2005) for laughter and filler detection on frame level (100 frames per second), with the unweighted average (UAAUC) as the official competition measure of this Sub-Challenge. In this respect, participants were required to also submit posterior class probabilities ('confidences') per frame in this Sub-Challenge. Besides, in the Conflict Sub-Challenge, we additionally chose the Pearson correlation coefficient (CC) as evaluation criterion for regression on the 'continuous-valued' original labels, following the IS10PC, which also featured a regression task (Schuller et al., 2013).

3.4.1. SVM baselines

In order to provide a standard evaluation measure, linear SVMs were used, where logistic functions map hyperplane distances to class pseudo-posteriors (Platt, 1999),

$$d_{\mathrm{SVM}}(\boldsymbol{x}) = \frac{1}{1+\exp\left(-\left(a(\boldsymbol{w}^T\boldsymbol{x}+b_1)+b_2\right)\right)}, \tag{2}$$

where $\boldsymbol{w}$ is the normal vector of the SVM hyperplane, $\boldsymbol{x}$ is an acoustic feature vector, b_1 is the SVM bias and a and b_2 are parameters of the logistic function, which are fitted to the SVM outputs on the training set in analogy to the method described in Section 3.4.2 on univariate logistic regression. A convenient property of linear support vector machines (SVMs) is that they are robust against overfitting in high dimensional feature spaces. The complexity parameter C weighs the trade-off between classification error and the L2-norm of $\boldsymbol{w}$. For each task, we chose the SVM complexity parameter $C \in \{10^{-3}, 10^{-2}, 10^{-1}, 1\}$ that achieved best UAR on the development set. The weight vector $\boldsymbol{w}$ was determined with sequential minimal optimisation (SMO). Multi-way classification was reduced to

Please cite this article as: B. Schuller et al., Affective and behavioural computing: Lessons learnt from the First Computational Paralinguistics Challenge, Computer Speech & Language (2018), http://dx.doi.org/10.1016/j.csl.2018.02.004

pair-wise binary classification in the same way as for logistic regression (see Section 3.4.2). In case of regression (only in the Conflict Sub-Challenge), SMO-trained support vector regression (SVR) was used.

To cope with imbalanced class distribution in the Autism Sub-Challenge, instance upsampling was applied. The instances of the under-represented categories (PDD, PDD-NOS, SLI) in the four-way 'Diagnosis' task were replicated five times in order to increase their effective weight in the loss function; in the binary 'Typicality' task a factor of two was applied. Note that we found this simple method to achieve similar performance for our tasks as more elaborate techniques such as SMOTE (Chawla et al., 2002). Conversely, for the Social Signals Sub-Challenge, downsampling was used, where only 5% of the 'garbage' frames were kept. No resampling of the training instances was done for the other Sub-Challenges. The baseline recipe provided to the participants performs training set resampling in a reproducible way. For evaluation on the test set, we retrained the models using the training and development set, applying resampling as above.

Let us now briefly summarise the baseline results as displayed in Table 8. In the Social Signals Sub-Challenge, detection of fillers seemed slightly 'easier' than detection of laughter, and for both a somewhat acceptable performance in terms of AUC (83.3% baseline UAAUC on test) was achieved – yet, showing the challenge of vocalisation localisation in naturalistic recordings of spontaneous speech. Note that the chance level baseline for AUC – obtained as mean and standard deviation over 25 random trials using random class posteriors – is at 50% with small standard deviation, as would be expected.

In the Conflict Sub-Challenge, it turned out that the SVM baseline did not significantly outperform univariate logistic regression on the classification task (cf. the results in Section 3.4.2). This might be due to the fact that the features and classification do not respect the multi-party conversation scenario (e.g., mean F_0 is calculated on average across all participants). However, in the regression task, a CC of above 81% was achieved, which is significantly ($p < 0.05$ according to a one-tailed z-test) higher than the CC of any single feature (cf. Table 10).

In the Emotion Sub-Challenge, the SVM baseline again showed arousal to be easier to be classified than valence – this is a well known phenomenon when using acoustic features only. On the test set, a performance drop was observed for the binary tasks. In the 12-way Category task there is a large room for improvement (40.9% baseline UAR on test), indicating the challenge of classifying subtle emotional differences even in enacted emotional speech. While the SVM baseline was tied by the logistic regression baseline on the development set (cf. Table 10), it clearly outperformed it on the test set, where some utterances are 'masked'. This can motivate the investigation of feature robustness in masked emotion in future work.

Table 8
Official Challenge baselines using support vector methods. *C*: Complexity parameter in SVM/SVR training (tuned on development set). Dev: Result on the development set, by training on the training set. Test: Result on the test set, by training on the training and development sets. Chance: Expected measure by chance (cf. text). UAAUC: Unweighted average of AUC for detection of the laughter and filler events. Official Challenge competition measures are highlighted.

[%]	*C*	Dev	Test	Chance
Social Signals Sub-Challenge				
AUC [Laughter]	0.1	86.2	82.9	50.0 ± 0.18
AUC [Filler]	0.1	89.0	83.6	50.0 ± 0.21
UAAUC		87.6	**83.3**	50.0 ± 0.13
Conflict Sub-Challenge				
CC [Score]	0.001	81.6	82.6	−0.8 ± 2.3
UAR [Class]	0.1	79.1	**80.8**	50.0
Emotion Sub-Challenge				
UAR [Arousal]	0.01	82.4	75.0	50.0
UAR [Valence]	0.1	77.9	61.6	50.0
UAR [Category]	1.0	40.1	**40.9**	8.33
Autism Sub-Challenge				
UAR [Typicality]	0.01	92.8	90.7	50.0
UAR [Diagnosis]	0.001	52.4	**67.1**	25.0

Please cite this article as: B. Schuller et al., Affective and behavioural computing: Lessons learnt from the First Computational Paralinguistics Challenge, Computer Speech & Language (2018), http://dx.doi.org/10.1016/j.csl.2018.02.004

Finally, in the Autism Sub-Challenge, the binary Typicality task can again alternatively be solved by mapping from the 4-way task leading to 92.6% UAR on test (not shown in Table 8). However, this high classification performance must be taken with caution, since channel recording conditions were different between typically and atypically developing children (Bone et al., 2013), and results are reported for relatively small groups of children (35 ASC vs 64 TYP). Reported results are therefore indicative pointers rather than strong markers of ASC deficiencies in speech production (Marchi et al., 2014). Better algorithms are clearly sought after for the Diagnosis task (67.1% baseline UAR on test).

In order to bring insights into the impact of recording conditions on performance, we performed additional experiments. In the first experiment, we removed all spectrum-related features from the feature set, as they convey most of the acoustic changes due to the use of different rooms. In the second experiment, we removed all static features from the feature set, and only kept derivates, which is likely to reduce the impact of stationary noises from the recordings. Results show that performance is increased for the binary Typicality task when spectrum-related features are removed from the feature set, whereas the removal of static features slightly degrades the performance, cf. Table 9. Therefore, features related to voice quality, pitch and loudness appear more robust for the Typicality task than spectrum-related features, which are indeed directly computed from the spectrum, and thus reflect the acoustic of the rooms used for the recordings, e.g., reverberation, environmental noise. Regarding the 4-way classification task, i.e., the Diagnosis task, a small degradation is again observed when only the first-order derivate of the acoustic features is kept in the feature set, whereas the removal of all spectrum-related features degrades more severely the performance. This supposes that a fine classification task like the diagnosis requires the use of a larger feature space, including spectral-related features, in order to achieve a better performance. As this might be related to specific room conditions, the use of dynamic features instead could be a suitable compromise for robustness.

3.4.2. Univariate logistic regression

We now introduce − for the first time in such a challenge − a univariate evaluation measure, i. e., we look for a single best feature. This serves two purposes: we can see whether at all and to which extent such a univariate reference value is beaten by our standard baseline procedure, and the other way round, how far we can get with one single feature as reference. To this aim, we used logistic functions of the form

$$d_i(x_i) = \frac{1}{1 + \exp\Big(-(a_i x_i + b_i)\Big)}\,, \tag{3}$$

where x_i is the value of feature i. For each feature and binary recognition task, the parameters a_i and b_i are fitted to the training set by the least squares method, modelling one of the classes as the positive, and the other as the negative outcome of a Bernoulli trial. A decision for the positive class is taken whenever $d_i > 0.5$. This baseline serves both for verification of the acoustic feature extraction procedure and as a reference for the results obtained with more sophisticated machine learning algorithms. In contrast to test statistics such as the t- or the Wilcoxon W-statistic, the

Table 9
Impact of recording conditions on performance for the Autism-Sub-Challenge tasks (typicality and diagnosis). Baseline: full acoustic features set. Only deltas: static features removed; No spectrum: all spectrum-related features removed. Dev: Result on the development set, by training on the training set. Test: Result on the test set, by training on the training and development sets.

[% UAR]	Dev	Test
Typicality		
Baseline	92.8	90.7
Only deltas	86.1	89.2
No spectrum	87.4	**91.8**
Diagnosis		
Baseline	52.4	**67.1**
Only deltas	42.8	66.6
No spectrum	45.2	58.9

Please cite this article as: B. Schuller et al., Affective and behavioural computing: Lessons learnt from the First Computational Paralinguistics Challenge, Computer Speech & Language (2018), http://dx.doi.org/10.1016/j.csl.2018.02.004

Table 10
Challenge results by logistic regression on single features. Multi-way classification (Category, Diagnosis) by pairwise coupling of 1-vs-1 classifiers. Dev: Result on the development set, by training on the training set. Test: Result on the test set, by training on the training and development set. Chance: Expected measure by chance (cf. text). Official Challenge competition measures are highlighted.

[%]	Feature	Dev	Test	Chance
Conflict Sub-Challenge				
CC [Score]	Mean of Positive Log. HNR	57.2	64.6	–0.8
UAR [Class]	Mean of Positive Log. HNR	74.5	**76.2**	50.0
Emotion Sub-Challenge				
UAR [Arousal]	Q3 of 25% Spectral Roll-Off	69.9	71.0	50.0
UAR [Valence]	Skewness of MFCC 1	68.3	57.2	50.0
UAR [Category]	(*Pairwise coupling*)	42.5	**29.9**	8.33
Autism Sub-Challenge				
UAR [Typicality]	Flatness of RMS Energy	84.7	82.2	50.0
UAR [DYS vs NOS]	IQR $1-3$ of ZCR	78.4	70.4	50.0
UAR [DYS vs PDD]	Flatness of F_0	49.5	51.1	50.0
UAR [NOS vs PDD]	Mean Dist. of Peak Mean from Mean in ΔLoudness	73.3	66.3	50.0
UAR [DYS vs TYP]	Flatness of RMS Energy	88.2	89.8	50.0
UAR [NOS vs TYP]	Flatness of RMS Energy	77.3	76.6	50.0
UAR [PDD vs TYP]	Flatness of RMS Energy	81.6	88.5	50.0
UAR [Diagnosis]	(*Pairwise coupling*)	52.2	**49.0**	25.0

UAR achieved by logistic regression is a realistic performance measure of a discriminatively trained classifier, yet it does not tell us whether feature values are positively or negatively correlated with the class label (0 or 1). However, this can be easily seen from the sign of a_i: $a_i > 0$ indicates that higher feature values are related to the class with label 1. For multi-way classification tasks (emotion category and developmental disorder diagnosis), logistic regression functions are trained for each pair of classes, and posterior probabilities are estimated by pairwise coupling (Hastie and Tibshirani, 1998), which is an iterative method that estimates multi-class posteriors from the ones provided by binary classifiers for each pair of classes.

For selecting the best suited logistic model among those obtained on the individual features, we chose different strategies for the Sub-Challenges. For the Conflict and Emotion Sub-Challenges, we used the one that achieved the highest UAR on the union of training and development set (i. e., reclassification of the training set, and classification of the development set). For the Autism Sub-Challenge, we manually selected prosodic features (cf. 6) that achieved a high UAR. As there are sometimes differences in the recording conditions across the classes (cf. Section 3.1.4), one could argue that spectral features from the ComParE feature set also reflect acoustic conditions apart from paralinguistic content, a hypothesis put forth by Bone et al. (2013). On the contrary, prosodic features are known to be robust against effects of reverberation Schuller (2011). Thus, the manual feature selection serves to show that the baseline feature set does indeed capture the task of interest.

Results of the single feature evaluation are shown in Table 10. There, we also compared against chance level. For UAR, they are defined as an equal class distribution (50% for 2, 25% for 4, and 8.33% for 12 classes). For CC (Conflict Sub-Challenge only), these are obtained as mean and standard deviation over 25 random trials prediction of Gaussian random numbers with mean and standard deviation of the training set labels.

For the Conflict Sub-Challenge, we found the mean of HNR to be indicative: if the HNR is low, there is a high degree of conflict. Logistic regression delivers 76.2% UAR on the test set. This might indicate a higher tension of the speakers in situations of conflict, resulting, for example, in more pressed/harsh voice. In the regression task, if we suppose that the (negated) mean HNR feature, which delivers the best CC (64.5%) on the training + development set, is our regressor, we obtain a similar CC of 64.6% on the test set.

In the Emotion Sub-Challenge, arousal can be classified relatively robustly on both the development and the test set, with around 70% UAR when considering the third quartile of the 25% spectral roll-off point $-$ portending that the speech contains a large portion of higher frequencies. Note that this feature is related to F_0, but much easier to compute and robust (being a percentile based feature); it mirrors the expected higher effort when arousal is high (positive). For valence, single features are less effective, as can be generally expected. The skewness of the first

Please cite this article as: B. Schuller et al., Affective and behavioural computing: Lessons learnt from the First Computational Paralinguistics Challenge, Computer Speech & Language (2018), http://dx.doi.org/10.1016/j.csl.2018.02.004

MFCC delivers above chance accuracy on the development set and the test set, and is hard to interpret as well. In pairwise coupling, the performance is relatively high on the development set (42.5%), but lower on the test set. This can be explained by the fact that in the test set, some of the utterances are spoken with ‘masked’ emotion.

In the Autism Sub-Challenge, we found that typicality can be classified with 82.2% UAR on the test set if using the flatness of RMS energy. A low flatness (‘spiky’ energy curve) is indicative of language impairments due to difficulties in regulating the speech, while a high flatness implies smooth speech output. For DYS against NOS, we observed 70.4% UAR on the test set by considering the inter-quartile range (IQR) 1−3 of the zero-crossing rate, which is particularly low for NOS. The DYS vs PDD task seems to be very hard with just a single feature, and only chance level UAR is obtained on held out data (development/test set). For the NOS vs PDD task, we observed that the mean distance of the loudness change peaks from the average loudness change is higher for autism (PDD), and this feature delivers 66.3% UAR on the test set. This result is particularly interesting for the purpose of eliminating possible acoustic confounders, as (most of) the NOS and PDD group were recorded in the same acoustic conditions. For the classification of any language-impaired group against typical children, we used the flatness of RMS energy as for the typicality task, delivering UAR way above chance in all three cases. Pairwise coupling of the above-named logistic regression functions delivers 52.2% and 49.0% UAR on the development and the test set, respectively, which is highly and significantly above chance ($p \ll .001$ according to a one-sided z-test). This suggests that it is feasible to classify language impairments using only low-level acoustic features which are robust against channel effects.

Summing up, we have demonstrated the general feasibility of the univariate approach, and at the same time, the superiority of the multi-feature approach as employed in the computation of the baselines. Certainly, we can imagine further promising avenues of research: the curve shape from the best to the n-best features (n being a number like 10, 50, or 100, or meeting some stop criterion) will most likely be rather flat, and interpreting these features (or feature types) will be interesting. Feature selection can be extended from single-best to a combination of n-best features. Yet, our experience from the Challenges tells us that most likely, we will not get a real boost of performance when using a well-suited classifier such as SVM with a rather complete (yet highly redundant) feature vector due to its robustness to the curse of dimensionality.

3.5. Participants and results

One of the requirements for participation in the Challenge was the acceptance of a paper submitted to ComParE and undergoing peer-review. Following the increasing trend of participant numbers and due to the fact that more tasks were featured, 65 research groups registered for the Challenge, and finally, 19 papers were accepted for the INTERSPEECH conference proceedings. All participants were encouraged to compete in all Sub-Challenges. Table 11 shows the individual participants for each Sub-Challenge. In summary, eleven teams took part in only one Sub-Challenge, one team in two, and two teams in three Sub-Challenges. Furthermore, the majority vote of the n best systems shows that the performances of the winning team can still be improved. Fig. 3 depicts the results of this fusion for values of n between six and fifteen. Note that not all the systems that were used for majority vote could be considered in the official Challenge in course of the peer-review process. As the number n of fused systems is optimised on the test set by selecting the combination with maximum performance on test, this fusion result is an upper limit of what can be reached by combining different systems, but is not meant to compete with the participants' results.

3.5.1. Contributions to the social signals sub-challenge

The studies on social signals detection are mainly based on two approaches, focusing on either features or classifiers. An et al. (2013) and Oh et al. (2013) both used syllabic-level features. Wagner et al. (2013) included phonetic features extracted from raw speech transcriptions obtained with the CMU Sphinx toolkit for speech recognition. All these groups retain the choice of using SVM as classifier. In contrast, Gosztolya et al. (2013) and Gupta et al. (2013) applied their own algorithms to the task, while using the official ComParE features. Specifically, Gosztolya et al. (2013) successfully applied the meta-algorithm AdaBoost to the Social Signals, but also Emotion and Autism Sub-Challenges. In particular, the probabilistic time-series smoothing and masking approach by Gupta et al. has proven to be highly efficient, achieving 6.1% absolute improvement over the baseline. Janicki (2013) adjusted both the features and the algorithm by advocating a hybrid Gaussian Mixture Models (GMM) - SVM approach, combining three

Table 11
Features, algorithms, and ‘gimmicks’ used by the participants; performances (UAAUC/UAR) on the test set

Participant	Features	Algorithms	Gimmick	[%]
Social Signals Sub-Challenge				UAAUC
An et al.	Frame- + syllabic-level	SVM	Rescoring of segment-internal frames	84.8
Oh et al.	Syllabic-level features	SVM	Syllabic-level segmentation	85.3
Wagner et al.	ComParE (141)+phonetic features	SVM	Phonetic transcription by ASR	88.4
Janicki	MFCCs + log-likelihoods	GMMs + SVM	Hybrid GMM-SVM approach	89.8
Gosztolya et al.	ComParE (141)	AdaBoost	Feature analysis	89.9
Gupta et al.	ComParE (141)	DNN	Probabilistic time-series smoothing and masking	**91.5**
Conflict Sub-Challenge				UAR
Grèzes et al.	ComParE + overlap ratio	SVR + SVM	Reducing conflict classification to overlap regression	83.1
Räsänen and Pohjalainen	ComParE subsets	*k*NN	Random subset feature selection	**83.9**
Emotion Sub-Challenge				UAR
Räsänen and Pohjalainen	ComParE subsets	*k*NN	Random subset feature selection	31.7
Sethu et al.	MFCC + ΔMFCC	GMMs	Sub-system fusion	35.7
Lee et al.	ComParE	SVM, DNN, *k*NN, acoustic segment model	Ensemble of classifiers	41.0
Gosztolya et al.	ComParE	AdaBoost	Feature analysis	**42.3**
Autism Sub-Challenge				UAR
Bone et al.	Spectral energy and smoothness features	SVM, *k*NN	Prosodic template and pronunciation quality modelling	60.2
Räsänen and Pohjalainen	ComParE subsets	*k*NN	Random subset feature selection	61.9
Gosztolya et al.	ComParE	AdaBoost	Feature analysis	62.1
Kirchhoff et al.	ComParE subsets	MLPs	Submodular feature selection and ranking	64.4
Lee et al.	ComParE	SVM, DNN, *k*NN, acoustic segment model	Ensemble of classifiers	64.8
Mart$\imath$nez et al.	ComParE + prosody, formants, shifted-delta cepstrum, amplitude modulation index	SVM	iVectors	66.1
Asgari et al.	Voice quality, energy, spectrum, cepstrum	SVM + SVR	Harmonic model of voiced speech	**69.4**

GMMs working in the 36-dimensional MFCC space and the discriminative SVM working in the 4-dimensional log-likelihood space. The majority vote of the best two systems leads to 92.7%.

3.5.2. Contributions to the conflict sub-challenge

Grèzes et al. (2013) suggested that the ratio of overlapping speech to non-overlapping speech is a useful feature for the detection of conflict levels, thus efficiently reducing the classification task to an overlap detection problem. Using this feature, they obtained 83.1% on the test set. Räsänen and Pohjalainen (2013) performed feature selection by using a new variant of random subset sampling methods with *k*-nearest neighbours (*k*NN) as a classifier, despite some effects of overfitting the feature set to finite data. It is noted that their approach has also proven to be effective in the Emotion and Autism Challenge. The best result obtained by majority voting of the best three participants is 85.9%.

3.5.3. Contributions to the emotion sub-challenge

In this Sub-Challenge, the teams Lee et al. (2013) and Gosztolya et al. (2013) both used the ComParE feature set, while applying different algorithms. In particular, the fusion of sub-systems and classifiers leads to superior results over the baseline, as shown by Sethu et al. (2013) and Lee et al. (2013). The best fusion result of twelve systems is 46.1%, considering all systems uploaded for evaluation. Although the number n of fused systems is optimized on test, the fusion results are always better than the winner for $n \geq 5$ (s. Fig. 3c).

3.5.4. Contributions to the autism sub-challenge

Most of the participants (Räsänen and Pohjalainen, 2013; Gosztolya et al., 2013; Kirchhoff et al., 2013; Lee et al., 2013) in the Autism Sub-Challenge applied different algorithms on the ComParE acoustic feature set, achieving mediocre results. Bone et al. (2012); Mart$\imath$nez et al. (2013); Asgari et al. (2013) applied SVM on individual feature sets, where the sets comprising prosodic and ceptral features used by the latter two groups led to the best results.

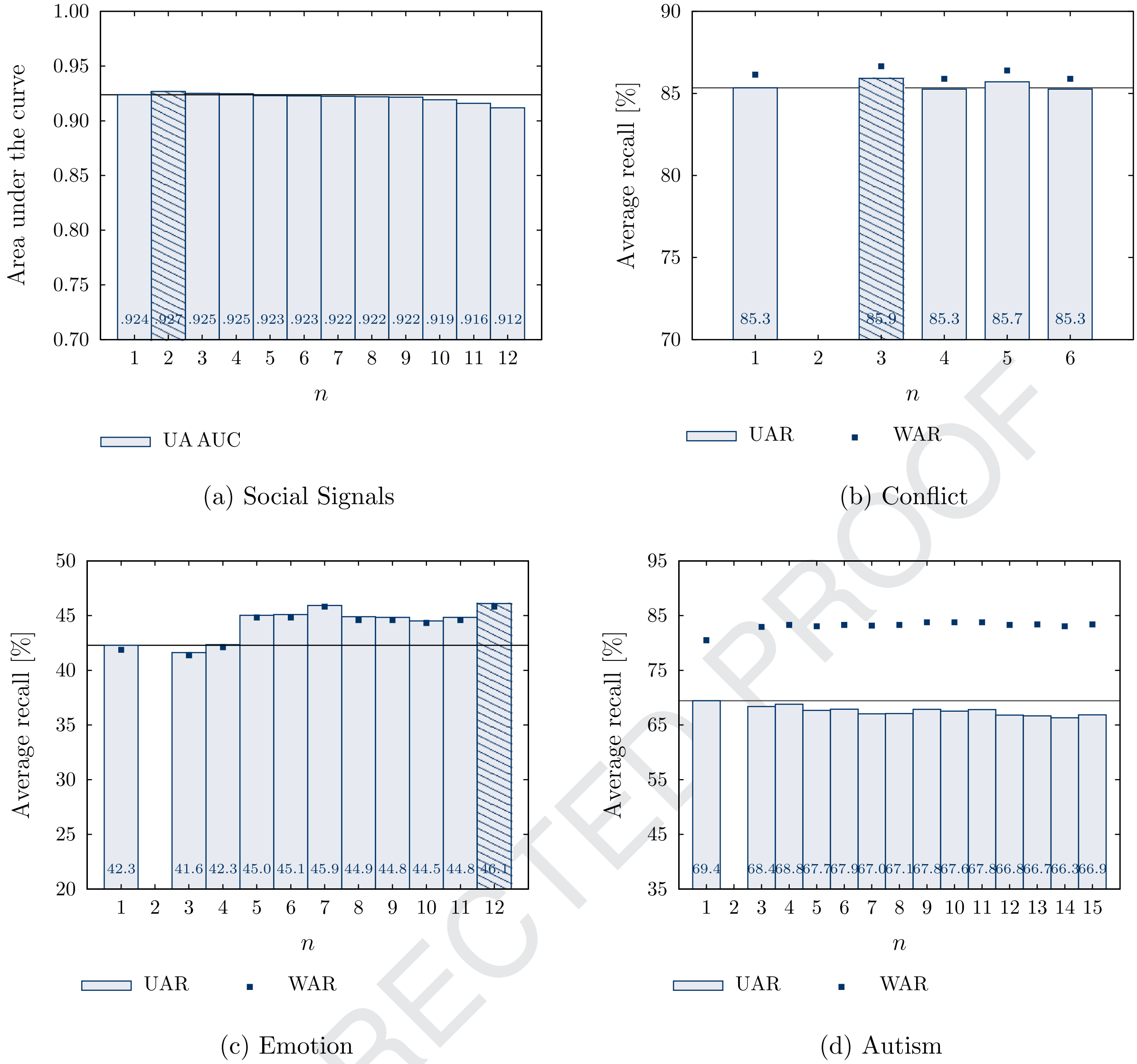


Fig. 3. Fusion of the results of the *n* best participants by majority vote.

Asgari et al. (2013) achieved the best UAR at 69.4%, which could not be outperformed by fusion of the best *n* participants' systems.

3.5.5. *Regions of significance*

Fig. 4 shows which absolute improvements over the result obtained in a given experiment could be considered as being significantly better for the four levels of significance $\alpha = .050$, .010, .005, and .001 in a one-sided test (Dietterich, 1998). For instance, to outperform the baseline at a significance level of $\alpha = .05$, the participants had to achieve a minimum absolute improvement of 4.4 % over the baseline of the Conflict Sub-Challenge 80.8%, 5.5 % compared to the baseline of the Emotion Sub-Challenge 40.9%, and 3.8 % compared to the baseline of the Autism Sub-Challenge 67.1%. A one-sided test can be applied if there is a substantial alternative hypothesis H1 over the null hypothesis H0; without such an H1, we had to use a two-sided test which means for Fig. 4 that the α level displayed has to be divided in half.

Please cite this article as: B. Schuller et al., Affective and behavioural computing: Lessons learnt from the First Computational Paralinguistics Challenge, Computer Speech & Language (2018), http://dx.doi.org/10.1016/j.csl.2018.02.004

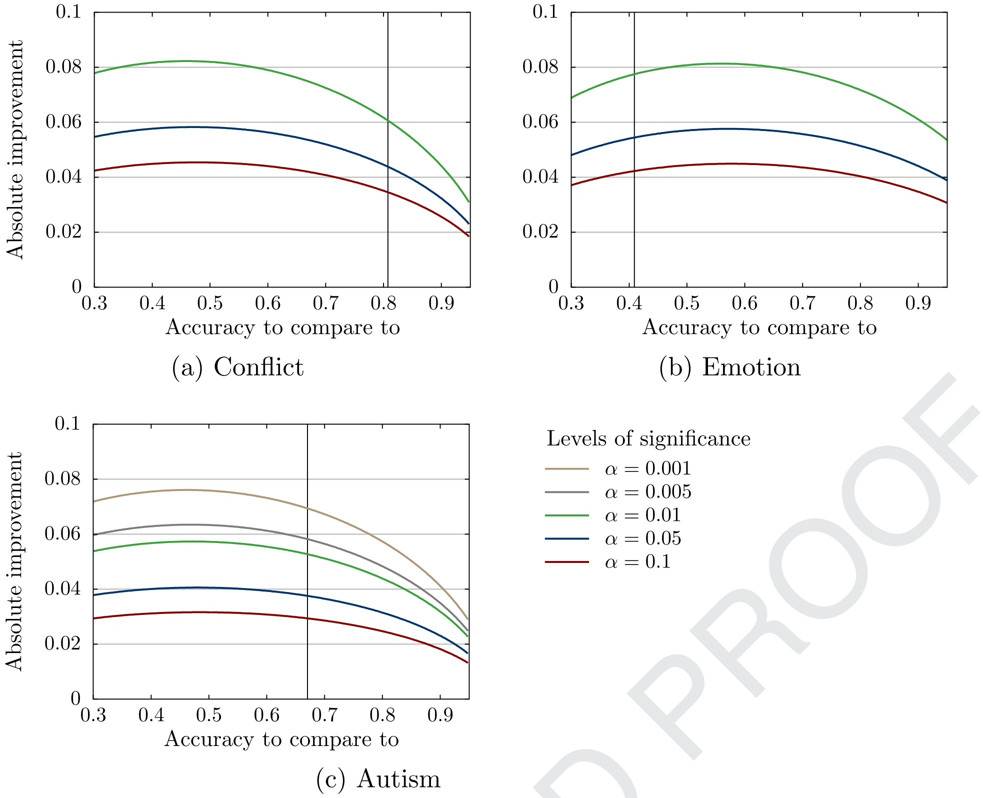


Fig. 4. Significance of results.

3.5.6. Meta-analysis

Let us now provide some meta-analysis of the participants' results beyond simple accuracy measures. For instance, in the Emotion Sub-Challenge, it is interesting to see the performances depending on emotion regulation. The figures displayed in Fig. 5 show that systems have most difficulties in understanding highly regulated arousal ('low' and 'masked' intensities), as would be expected. However, it is interesting that high intensity is not easier to recognise than normal intensity (Fig. 5a). We might speculate that high intensity stimuli produced by actors are definitely pronounced (clear) but might vary due to speaker idiosyncrasies whereas normal intensity might be less pronounced but more 'standard'. Thus, a higher stronger manifestation is counterbalanced by a more regular manifestation. In contrast, valence seems to be hard to recognise from acoustics in general – although 'masked'

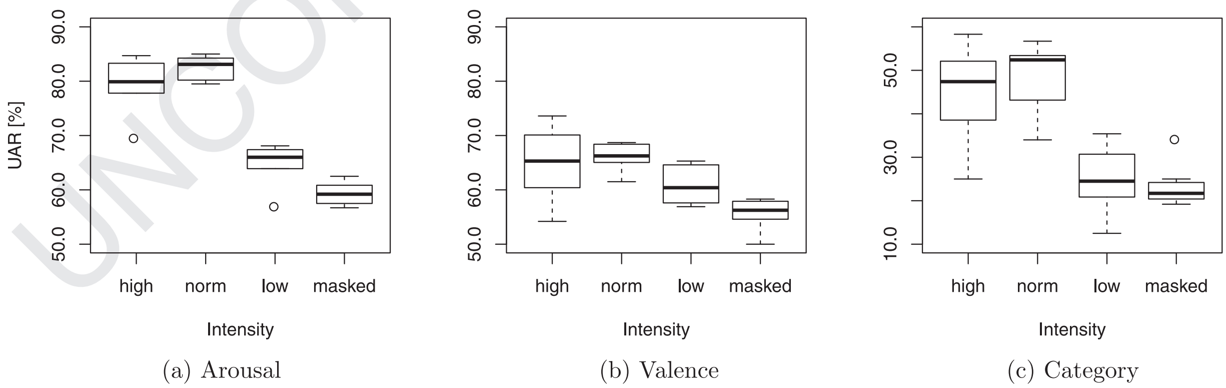


Fig. 5. Boxplots of participants' performances (UAR) by regulated intensity of emotion.

intensity leads to worst results again, the differences are less pronounced than for arousal (Fig. 5b). This we know from practically all studies on valence recognised from speech. The trend in the 12-class category discrimination (Fig. 5c) is very similar to the one observed for arousal recognition.

Furthermore, let us now investigate the results of the two multi-way classification Sub-Challenges more closely. Here, we are interested in the most frequently occurring confusions per class.

To shed light on this question, we computed the average confusion matrix of the participants' predictions and the SVM baseline predictions for the Category task (Emotion Sub-Challenge) as well as for the Diagnosis task (Autism Sub-Challenge). Table 12 shows the results for the emotion category task. The most easily recognised categories are sadness, amusement, relief, hot anger, and interest. Most difficult to recognise are pride (14.7%) and elation (15.3%). Confusions of one category with another specific category are rather low, the highest being 25%, namely pleasure confounded with sadness; due to the rather small number of cases per category, we should not over-interpret single confusions, though. The confusions are distributed across many categories and not especially across categories sharing the same dimension values (either plus OR minus for arousal and/or valence).

As cases with masked regulation (hiding the true emotion) are only represented in the test set, they could not be learned in the training. Of course, this fact contributes to a higher overall confusion between categories. To illustrate the different degrees of confusions between one category and all others, we give in the last line of Table 12 the sum of all percentages by columns to show the tendency of hits and false alarms in each category. High values above 100% imply that the category has been recognised well (hits) and/or there exists a bias towards this category (false alarms). To put it the other way round, lower values than 100% indicate that this category is rather imprecisely recognised and/or there is a negative bias 'away from' this category. We can see a positive bias towards sadness, interest, and anxiety, and a negative bias towards elation, despair, and pride. All these categories are obviously less distinct than, for instance, amusement that is recognised relatively well. All in all, the high percentage of confusions − only sadness is classified with a recall clearly above 50% (amusement at 50.7%) − demonstrates the difficulty of such a multi-class task and the challenge when facing realistic − even more noisy − data.

Table 13 shows the corresponding result for the autism diagnosis task. It is notable that there is a strong bias towards predicting the majority class (typically developing children), which might be remedied by threshold

Table 12
Average confusion matrix of participants' systems for the Category task (Emotion Sub-Challenge).

[%]	am	an	co	de	el	ho	in	pa	pl	pr	re	sa
am(usement)	50.7	14.0	5.7	6.0	7.0	0.3	6.0	3.7	2.0	2.7	1.0	1.3
an(xiety)	5.3	30.3	5.7	4.0	2.3	5.7	16.0	8.7	3.7	6.0	1.0	12.0
co(ld anger)	4.2	7.5	27.5	0.3	0.3	3.9	20.3	2.2	10.3	8.6	10.0	5.3
de(spair)	6.3	8.0	5.3	27.3	11.7	6.7	7.0	8.7	1.3	5.3	1.7	10.7
el(ation)	12.9	5.0	7.1	8.4	15.3	11.8	10.0	7.4	5.8	7.4	4.7	4.7
ho(t anger)	3.3	7.7	11.0	2.7	1.7	42.7	6.7	5.7	1.3	11.3	2.3	4.0
in(terest)	0.7	11.0	11.7	2.7	0.0	1.0	41.0	1.0	9.3	2.0	4.7	14.7
pa(nic fear)	10.3	7.6	3.4	4.2	6.6	17.6	2.6	38.7	0.3	2.6	6.1	0.0
pl(easure)	2.0	8.0	5.7	3.0	0.7	0.7	9.0	0.7	39.3	1.7	5.0	25.0
pr(ide)	0.5	13.7	16.6	3.2	4.2	8.2	10.8	2.1	8.4	14.7	6.3	11.1
re(lief)	0.5	7.4	9.7	0.3	0.3	2.6	4.5	0.0	15.0	4.5	47.9	7.6
sa(dness)	0.3	8.9	3.9	0.5	0.0	0.3	8.4	0.3	9.2	1.3	3.7	62.9
$\sum$	97.0	129.1	113.3	62.6	50.1	101.5	142.3	79.2	105.9	68.1	94.4	159.3

Table 13
Average confusion matrix of participants' systems for the Diagnosis task (Autism Sub-Challenge).

	DYS	NOS	PDD	TYP
DYS	64.0	5.9	18.8	11.3
NOS	1.2	58.4	14.3	26.1
PDD	32.3	25.4	31.0	11.4
TYP	1.0	2.0	1.4	95.6
$\sum$	98.5	91.7	65.5	144.4

Please cite this article as: B. Schuller et al., Affective and behavioural computing: Lessons learnt from the First Computational Paralinguistics Challenge, Computer Speech & Language (2018), http://dx.doi.org/10.1016/j.csl.2018.02.004

optimisation (it is not possible to give results because the participants did not have to submit posteriors for this task). Among the language impairment conditions, dysphasia seems easiest to recognise from acoustics, while the manifestation of autism (PDD) or unspecific impairments is harder, which is expected.

4. Conclusions and future challenges

In this work, we reviewed the first of its kind Computational Paralinguistics Challenge, which has been initialised to overcome comparability issues regarding data sets, partitioning, evaluation measures, baseline systems, and test-beds. The introduction of the common ComParE feature set, designed to tackle various paralinguistic recognition tasks, has proven very successful, as can be seen from the fact that most of successful participants' submissions employed the feature set or parts of it, and at the same time it has contributed to utmost comparability of results.

Along with SVM, the ComParE features introduced here yielded competitive performance in the participants' field of the Conflict, the Emotion, and the Autism Sub-Challenge; yet, no single feature from the ComParE set was competitive on its own. In line with the other challenges, combining classifier results (late fusion, cf. Fig. 3) normally gives some boost to performance.

The Conflict Sub-Challenge was the first Challenge task in the INTERSPEECH series to feature speech from multiple speakers in a single instance, and hence speech overlap − a mid-level feature whose extraction is usually studied in the neighbouring field of speaker diarisation − performed very respectably. In a similar vein, the Social Signals Sub-Challenge was the first INTERSPEECH Challenge task requiring segmentation, and hence methods known from the field of ASR, where this is a well understood issue, prevailed over the ComParE baseline approach. All in all, these results show a promising avenue for further Challenges: exploring a greater variety of paralinguistic recognition tasks that differ in nature from previously tackled ones is likely to lead to more diverse methodologies being successful.

In this Challenge, we introduced four paralinguistic tasks which are important for the realm of affective human-computer interaction, yet some of them go beyond the traditional tasks of emotion recognition. Thus, as a milestone, ComParE 2013 laid the foundation for a successful series of follow-up ComParEs to date, exploring more and more the paralinguistic facets of human speech in tomorrow's real-life information, communication and entertainment systems.

5. Acknowledgement

The research work has received funding from the European Community's Seventh Framework Programme [ASC-Inclusion, grant No. 289021], the European Union's Framework Programme for Research and Innovation HORIZON 2020 [ARIA-VALUSPA, grant No. 645378], and the European Union Seventh Framework Programme ERC Starting Grant [iHEARu, grant No. 338164]. This research has been also supported by the Laboratory of Excellence SMART (ANR-11-LABX-65) supported by French State funds managed by the ANR within the Investissements d'Avenir programme (ANR-11-IDEX-0004-02). The authors would further like to thank the sponsors of the Challenge, the Association for the Advancement of Affective Computing (former HUMAINE Association) and the Social Signal Processing Network (SSPNet). The responsibility lies with the authors.

References

Albrecht, K., 2006. Social Intelligence: The New Science of Success 2005. John Wiley & Sons.
An, G., Brizan, D.-G., Rosenberg, A., 2013. Detecting laughter and filled pauses using syllable-based features. In: Proceedings of Interspeech. ISCA, Lyon, France, pp. 178–181.
Asgari, M., Bayestehtashk, A., Shafran, I., 2013. Robust and accurate features for detecting and diagnosing autism spectrum disorders. In: Proceedings of Interspeech. ISCA, Lyon, France, pp. 191–194.
Bachorowski, J.-A., Smoski, M.J., Owren, M.J., 2001. The acoustic features of human laughter. J. Acoust. Soc. Am. 110, 1581–1597.
Bänziger, T., Mortillaro, M., Scherer, K.R., 2012. Introducing the Geneva multimodal expression corpus for experimental research on emotion perception. Emotion 12, 1161–1179.
Batliner, A., Fischer, K., Huber, R., Spilker, J., Nöth, E., 2000. Desperately seeking emotions: Actors, wizards, and human beings. In: Proceedings of ISCA Workshop on Speech and Emotion. Newcastle, Northern Ireland, pp. 195–200.
Beedie, C., Terry, P., Lane, A., 2005. Distinctions between emotion and mood. Cognit. Emot. 19 (6), 847–878.

Please cite this article as: B. Schuller et al., Affective and behavioural computing: Lessons learnt from the First Computational Paralinguistics Challenge, Computer Speech & Language (2018), http://dx.doi.org/10.1016/j.csl.2018.02.004

Bone, D., Black, M.P., Lee, C.-C., Williams, M.E., Levitt, P., Lee, S., Narayanan, S.S., 2012. Spontaneous-speech acoustic-prosodic features of children with autism and the interacting psychologist. In: Proceedings of Interspeech, pp. 1043–1046.
Bone, D., Black, M. P., Li, M., Metallinou, A., Lee, S., Narayanan, S., 2011. Intoxicated speech detection by fusion of speaker normalized hierarchical features and GMM supervectors, 3217−3220.
Bone, D., Chaspari, T., Audhkhasi, K., Gibson, J., Tsiartas, A., Van Segbroeck, M., Li, M., Lee, S., Narayanan, S., 2013. Classifying language-related developmental disorders from speech cues: the promise and the potential confounds. In: Proceedings of Interspeech. ISCA, Lyon, France, pp. 182–186.
Chawla, N.V., Bowyer, K.W., Hall, L.O., Kegelmeyer, W.P., 2002. SMOTE: Synthetic Minority Over-sampling Technique. J. Artif. Intel. Res. 16, 321–357.
Clark, H., Fox Tree, J., 2002. Using “uh” and “um” in spontaneous speaking. Cognition 84 (1), 73–111.
Demouy, J., Plaza, M., Xavier, J., Ringeval, F., Chetouani, M., Périsse, D., Chauvin, D., Viaux, S., Golse, B., Cohen, D., Robel, L., 2011. Differential language markers of pathology in autism, pervasive developmental disorders not otherwise specified and specific language impairment. Res. Autism Spectrum Disorders 5 (4), 1402–1412.
Dhall, A., Göcke, R., Joshi, J., Wagner, M., Gedeon, T., 2013. Emotion recognition in the wild challenge (EmotiW) challenge and workshop summary. In: Proceedings of ICMI. ACM, Sydney, Australia, pp. 371–372.
Dietterich, T.G., 1998. Approximate statistical tests for comparing supervised classification learning algorithms. Neural Comput. 10, 1895–1923.
Q4 Dumouchel, P., Dehak, N., Attabi, Y., Dehak, R., Boufaden, N.
Ekkekakis, P., 2013. The Measurement of Affect, Mood, and Emotion: A Guide for Health-Behavioral Research. Cambridge University Press.
Eyben, F., 2015. Real-time Speech and Music Classification by Large Audio Feature Space Extraction. Springer Theses, Springer International Publishing, Switzerland.
Eyben, F., Weninger, F., Gro, F., Schuller, B., 2013. Recent developments in openSMILE, the munich open-source multimedia feature extractor. In: Proceedings of ACM Multimedia. ACM, Barcelona, Spain, pp. 835–838.
Eyben, F., Wöllmer, M., Schuller, B., 2009. openEAR − introducing the munich open-source emotion and affect recognition toolkit. In: Proceedings of ACII. IEEE, Amsterdam, Netherlands, pp. 576–581.
Eyben, F., Wöllmer, M., Schuller, B., 2010. openSMILE − the munich versatile and fast open-source audio feature extractor. In: Proceedings of ACM Multimedia. ACM, Florence, Italy, pp. 1459–1462.
First, M.B., 1994. Diagnostic and Statistical Manual of Mental Disorders. 4 American Psychiatric Association Publishing, Arlington, VA.
Gosztolya, G., Busa-Fekete, R., Tóth, L., 2013. Detecting autism, emotions and social signals using adaboost. In: Proceedings of Interspeech. ISCA, Lyon, France, pp. 220–224.
Grèzes, F., Richards, J., Rosenberg, A., 2013. Let me finish: automatic conflict detection using speaker overlap. In: Proceedings of Interspeech. ISCA, Lyon, France, pp. 200–204.
Gupta, R., Audhkhasi, K., Lee, S., Narayanan, S., 2013. Paralinguistic event detection from speech using probabilistic time-series smoothing and masking. In: Proceedings of Interspeech. ISCA, Lyon, France, pp. 173–177.
Hastie, T., Tibshirani, R., 1998. Classification by pairwise coupling. Annals Stat. 26 (2), 451–471.
Huang, D.-Y., Zhang, Z., Ge, S.S., 2014. Speaker state classification based on fusion of asymmetric simple partial least squares (simpls) and support vector machines. Comput. Speech Lang. 28 (2), 392–419.
Ivanov, A., Chen, X., 2012. Modulation spectrum analysis for speaker personality trait recognition. In: Proceedings of Interspeech. ISCA, Portland, OR, pp. 278–281.
Janicki, A., 2013. Non-linguistic vocalisation recognition based on hybrid GMM-SVM approach. In: Proceedings of Interspeech. ISCA, Lyon, France, pp. 153–157.
Jeon, J.H., Xia, R., Liu, Y., 2014. Level of interest sensing in spoken dialog using decision-level fusion of acoustic and lexical evidence. Comput. Speech Lang. 28 (2), 420–433.
Kim, J., Kumar, N., Tsiartas, A., Li, M., Narayanan, S., 2012a. Intelligibility classification of pathological speech using fusion of multiple subsystems. In: Proceedings of Interspeech. ISCA, Portland, OR, pp. 534–537.
Kim, S., Filippone, M., Valente, F., Vinciarelli, A., 2012b. Predicting the conflict level in television political debates: an approach based on crowdsourcing, nonverbal communication and gaussian processes. In: Proceedings of ACM Multimedia. ACM, Nara, Japan, pp. 793–796.
Kim, S., Valente, F., Filippone, M., Vinciarelli, A., 2014. Predicting continuous conflict perceptionwith bayesian gaussian processes. IEEE Trans. Affect. Comput. 5 (2), 187–200.
Kirchhoff, K., Liu, Y., Bilmes, J.A., 2013. Classification of developmental disorders from speech signals using submodular feature selection.. In: Proceedings of Interspeech. ISCA, Lyon, France, pp. 187–190.
Kockmann, M., Burget, L., Černockỳ, J., 2010. Brno university of technology system for interspeech 2010 paralinguistic challenge. In: Proceedings of Interspeech. ISCA, Makuhari, Japan, pp. 2822–2825.
Krauss, R.M., Freyberg, R., Morsella, E., 2002. Inferring speakers’ physical attributes from their voices. J. Exper. Social Psychol. 38 (6), 618–625.
Lee, C.-C., Mower, E., Busso, C., Lee, S., Narayanan, S., 2011. Emotion recognition using a hierarchical binary decision tree approach. Speech Commun. 53 (9), 1162–1171.
Lee, H.-y., Hu, T.-y., Jing, H., Chang, Y.-F., Tsao, Y., Kao, Y.-C., Pao, T.-L., 2013. Ensemble of machine learning and acoustic segment model techniques for speech emotion and autism spectrum disorders recognition. In: Proceedings of Interspeech. ISCA, Lyon, France, pp. 215–219.
Maier, A., Haderlein, T., Eysholdt, U., Rosanowski, F., Batliner, A., Schuster, M., Nöth, E., 2009. Peaks−a system for the automatic evaluation of voice and speech disorders. Speech Commun. 51 (5), 425–437.
Marchi, E., Ringeval, F., Schuller, B., 2014. Voice-enabled Assistive Robots for Handling Autism Spectrum Conditions: An Examination of the Role of Prosody. Speech and Automata in Health Care (Speech Technology and Text Mining in Medicine and Healthcare). De Gruyter, Boston/Berlin/Munich, pp. 207–236.

Please cite this article as: B. Schuller et al., Affective and behavioural computing: Lessons learnt from the First Computational Paralinguistics Challenge, Computer Speech & Language (2018), http://dx.doi.org/10.1016/j.csl.2018.02.004

Marchi, E., Schuller, B., Baron-Cohen, S., Golan, O., Bölte, S., Arora, P., Häb-Umbach, R., 2015. Typicality and emotion in the voice of children with autism spectrum condition: Evidence across three languages. In: Proceedings of Interspeech. ISCA, Dresden, Germany, pp. 115–119.

Martınez, D., Ribas, D., Lleida, E., Ortega, A., Miguel, A., 2013. Suprasegmental information modelling for autism disorder spectrum and specific language impairment classification. In: Proceedings of Interspeech. ISCA, Lyon, France, pp. 195–199.

McCann, J., Peppe, S., 2003. Prosody in autism spectrum disorders: a critical review. Int. J. Lang. Commun. Disorder 38, 325–350.

Meinedo, H., Trancoso, I., 2010. Age and gender classification using fusion of acoustic and prosodic features. In: Proceedings of Interspeech. ISCA, Makuhari, Japan, pp. 2818–2821.

Mohammadi, G., Vinciarelli, A., Mortillaro, M., 2010. The voice of personality: mapping nonverbal vocal behavior into trait attributions. In: Proceedings of ACM Multimedia Workshop on Social Signal Processing. ACM, pp. 17–20.

Montacié, C., Caraty, M.-J., 2012. Pitch and intonation contribution to speakers' traits classification. In: Proceedings of Interspeech. ISCA, Portland, OR, pp. 526–529.

Oh, J., Cho, E., Slaney, M., 2013. Characteristic contours of syllabic-level units in laughter. In: Proceedings of Interspeech. ISCA, Lyon, France, pp. 158–162.

Pentland, A., 2007. Social signal processing [exploratory dsp]. IEEE Signal Process. Mag. 24 (4), 108–111.

Picard, R.W., 1997. Affective Computing. MIT Press, Cambridge, MA, USA.

Platt, J.C., 1999. Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. Advances in Large Margin Classifiers. MIT Press, pp. 61–74.

Räsänen, O., Pohjalainen, J., 2013. Random subset feature selection in automatic recognition of developmental disorders, affective states, and level of conflict from speech. In: Proceedings of Interspeech. ISCA, Lyon, France, pp. 210–214.

Ringeval, F., Demouy, J., Szaszák, G., Chetouani, M., Robel, L., Xavier, J., Cohen, D., Plaza, M., 2011. Automatic intonation recognition for the prosodic assessment of language impaired children. IEEE Trans. Audio Speech Lang. Process. 19, 1328–1342.

Ringeval, F., Marchi, E., Grossard, C., Xavier, J., Chetouani, M., Cohen, D., Schuller, B., 2016. Automatic analysis of typical and atypical encoding of spontaneous emotion in the voice of children. In: Proceedings of Interspeech. ISCA, San Fransisco, CA, pp. 1210–1214.

Ringeval, F., Schuller, B., Valstar, M., Jaiswal, S., Marchi, E., Lalanne, D., Cowie, R., Pantic, M., 2015. AV+EC 2015 − the first affect recognition challenge bridging across audio, video, and physiological data. In: Proceedings of International Workshop on Audio/Visual Emotion Challenge (AVEC), co-Located with ACM Multimedia. ACM, Brisbane, Australia, pp. 3–8.

Rosenthal, R., 2005. Conducting judgment studies: some methodological issues. In: Harrigan, J., Rosenthal, R., Scherer, K. (Eds.), The New Handbook of Methods in Nonverbal Behavior Research. Oxford University Press, pp. 199–234.

Roth, W.-M., Tobin, K., 2010. Solidarity and conflict: aligned and misaligned prosody as a transactional resource in intra- and intercultural communication involving power differences. Cult. Stud. Sci. Educ. 5, 807.

Russell, J.A., 2003. Core affect and the psychological construction of emotion. Psychol. Rev. 110 (1), 145.

Russell, J.A., 2009. Emotion, core affect, and psychological construction. Cognit. Emotion 23 (7), 1259–1283.

Schiel, F., Heinrich, C., 2009. Laying the foundation for in-car alcohol detection by speech. In: Proceedings of Interspeech. ISCA, Brighton, UK, pp. 983–986.

Schmitt, M., Marchi, E., Ringeval, F., Schuller, B., 2016. Towards cross-lingual automatic diagnosis of autism spectrum condition in children's voices. In: Proceedings of ITG Conference on Speech Communication. IEEE, Paderborn, Germany, p. 264—268.

Schuller, B., 2011. Affective speaker state analysis in the presence of reverberation. Int. J. Speech Technol. 14 (2), 77–87.

Schuller, B., 2012. The computational paralinguistics challenge. IEEE Signal Process. Mag. 29 (4), 97–101.

Schuller, B., Batliner, A., 2014. Computational Paralinguistics: Emotion, Affect and Personality in Speech and Language Processing. John Wiley & Sons, New York, NY.

Schuller, B., Batliner, A., Steidl, S., Seppi, D., 2011a. Recognising realistic emotions and affect in speech: State of the art and lessons learnt from the first challenge. Speech Commun. 53 (9/10), 1062–1087.

Schuller, B., Steidl, S., Batliner, A., 2009. The INTERSPEECH 2009 Emotion Challenge. In: Proceedings of Interspeech. ISCA, Brighton, UK, pp. 312–315.

Schuller, B., Steidl, S., Batliner, A., Burkhardt, F., Devillers, L., MüLler, C., Narayanan, S., 2013. Paralinguistics in speech and language − state-of-the-art and the challenge. Special Issue on Paralinguistics in Naturalistic Speech and Language, Comput. Speech Lang. 27 (1), 4–39.

Schuller, B., Steidl, S., Batliner, A., Burkhardt, F., Devillers, L., Müller, C.A., Narayanan, S., 2010. The INTERSPEECH 2010 Paralinguistic Challenge − Age, Gender, and Affect. In: Proceedings of Interspeech. ISCA, Makuhari, Japan, pp. 2794–2797.

Schuller, B., Steidl, S., Batliner, A., Nöth, E., Vinciarelli, A., Burkhardt, F., van Son, R., Weninger, F., Eyben, F., Bocklet, T., et al., 2015. A survey on perceived speaker traits: Personality, likability, pathology, and the first challenge. Comput. Speech Lang. 29 (1), 100–131.

Schuller, B., Steidl, S., Batliner, A., Nöth, E., Vinciarelli, A., Burkhardt, F., Van Son, R., Weninger, F., Eyben, F., Bocklet, T., et al., 2012. The INTERSPEECH 2012 speaker trait challenge. In: Proceedings of Interspeech. ISCA, Portland, OR, pp. 254–257.

Schuller, B., Steidl, S., Batliner, A., Schiel, F., Krajewski, J., 2011b. The INTERSPEECH 2011 speaker state challenge. In: Proceedings of Interspeech. ISCA, Florence, Italy, pp. 3201–3204.

Schuller, B., Steidl, S., Batliner, A., Schiel, F., Krajewski, J., Weninger, F., Eyben, F., 2014. Medium-term speaker states − a review on intoxication, sleepiness and the first challenge. Special Issue on Broadening the View on Speaker Analysis, Comput. Speech Lang. 28 (2), 346–374.

Sethu, V., Epps, J., Ambikairajah, E., Li, H., 2013. Gmm based speaker variability compensated system for interspeech 2013 compare emotion challenge. In: Proceedings of Interspeech. ISCA, Lyon, France, pp. 205–209.

Steidl, S., 2009. Automatic Classification of Emotion-Related User States in Spontaneous Children's Speech. Logos Verlag, Berlin.

Tanaka, H., Campbell, N., 2011. Acoustic features of four types of laughter in natural conversational speech. In: Proceedings of International Congress of Phonetic Sciences (ICPhS). Hong Kong, China, pp. 1958–1961.

Please cite this article as: B. Schuller et al., Affective and behavioural computing: Lessons learnt from the First Computational Paralinguistics Challenge, Computer Speech & Language (2018), http://dx.doi.org/10.1016/j.csl.2018.02.004

Tao, J., Tan, T., 2005. Affective computing: A review. In: Proceedings of the International Conference on Affective Computing and Intelligent Interaction, pp. 981–995.

Van Santen, J.P., Prud'hommeaux, E.T., Black, L.M., Mitchell, M., 2010. Computational prosodic markers for autism. Autism 14, 215–236.

Vettin, J., Todt, D., 2004. Laughter in conversation: Features of occurrence and acoustic structure. J. Nonverbal Behav. 28 (2), 93–115.

Vinciarelli, A., Pantic, M., Bourlard, H., 2009. Social signal processing: survey of an emerging domain. Image Vis. Comput. 27, 1743–1759.

Vinciarelli, A., Pantic, M., Heylen, D., Pelachaud, C., Poggi, I., D'Errico, F., Schröder, M., 2012a. Bridging the gap between social animal and unsocial machine: a survey of social signal processing. IEEE Trans. Affect. Comput. 3 (1), 69–87.

Vinciarelli, A., Salamin, H., Polychroniou, A., Mohammadi, G., Origlia, A., 2012b. From nonverbal cues to perception: personality and social attractiveness. Cognitive Behav. Syst. 60–72.

Vogt, T., André, E., 2005. Comparing feature sets for acted and spontaneous speech in view of automatic emotion recognition. In: Proceedings of the International Conference on Multimedia and Expo. IEEE, pp. 474–477.

Wagner, J., Lingenfelser, F., André, E., 2013. Using phonetic patterns for detecting social cues in natural conversations. In: Proceedings of the Interspeech. ISCA, Lyon, France, pp. 168–172.

Witten, I.H., Frank, E., 2005. Data Mining: Practical Machine Learning Tools and Techniques. 2nd Edition Morgan Kaufmann, San Francisco.

Young, S., Evermann, G., Gales, M., Hain, T., Kershaw, D., Liu, X., Moore, G., Odell, J., Ollason, D., Povey, D., Valtchev, V., Woodland, P., 2006. The HTK book (v3.4). Cambridge University Press.

Please cite this article as: B. Schuller et al., Affective and behavioural computing: Lessons learnt from the First Computational Paralinguistics Challenge, Computer Speech & Language (2018), http://dx.doi.org/10.1016/j.csl.2018.02.004

Björn Schuller is a Reader (Associate Professor) at Imperial College London/UK since 2015, being a Senior Lecturer since 2013, Full Professor and Chair of Embedded Intelligence for Health Care and Wellbeing at the University of Augsburg/Germany, and the co-founding CEO of audEERING GmbH. Previously, he headed Chairs at the University of Passau/Germany (2013 − 2017), and a Group at TUM/Germany (2006 − 2014). Dr. Schuller is elected member of the IEEE Speech and Language Processing Technical Committee, Senior Member of the IEEE, member of the ACM and ISCA, and President-emeritus of the AAAC. He (co-)authored 600+ publications (16 000+ citations, h-index = 62), and is Editor in Chief of the IEEE Transactions on Affective Computing, Associate Editor of Computer, Speech and Language amongst many other journals, a Program Chair of Interspeech 2019, a General Chair of ACII 2019, organizer of the INTERSPEECH 2009 − 2017 annual Computational Paralinguistics Challenges and the 2011−2017 annual Audio/Visual Emotion Challenges amongst many other commitments.

Felix Weninger received his diploma and his doctoral degree, both in computer science, from TUM in 2009 and 2015. He is currently a senior research scientist at Nuance Communications, Ulm, Germany. In 2013/14, he was an intern at Mitsubishi Electric Research Laboratories (MERL), Cambridge, MA. His research interests include automatic speech recognition, speech analysis, and deep learning. Dr. Weninger has published more than 80 peer-reviewed papers in books, journals and conference proceedings.

Yue Zhang received her master's degree in Electrical Engineering and Information Technology (M.Sc.) from Technische Universität München (TUM) in 2013. From 2014 to 2015, she was a research assistant in the Machine Intelligence and Signal Processing Group at TUM's Institute for Human-Machine Communication. Currently, she is working towards her PhD degree at Imperial College London, Department of Computing. Her research focus is on novel machine learning approaches for holistic speech analysis in computational paralinguistics.

Fabien Ringeval received the M.S. degree in speech and image signal processing in 2006, and the PhD degree for his researches on the automatic recognition of acted and spontaneous emotions from speech in 2011, both from the Université Pierre et Marie Curie (UPMC), Paris, France. He is an Associate Professor at the Université Grenoble Alpes, CNRS, LIG, France, since 2016. Dr. Ringeval is also a senior researcher at audEERING GmbH. His research interests concern digital signal processing and machine learning, with applications on the automatic recognition of paralinguistic information from multimodal data. Dr. Ringeval (co-)authored more than 50 publications in peer-reviewed books, journals and conference proceedings in the field. He co-organised workshops and international challenges, including the INTERSPEECH 2013 ComParE challenge, the Alpine Rendez-vous (ARV) 2013 Workshop on Tools and Technologies for Emotion Awareness in Computer-Mediated Collaboration and Learning, the International Audio/Visual Emotion Challenge and Workshop (AVEC 15−17), and also serves as Publication Chair for the 7th AAAC International Conference on Affective Computing and Intelligent Interaction (ACII 2017), and as Grand Challenge Chair for the 20th ACM International Conference on Multimodal Interaction (ICMI 2018).

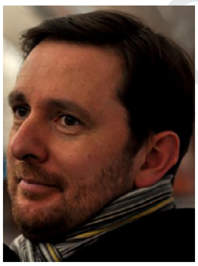
Please cite this article as: B. Schuller et al., Affective and behavioural computing: Lessons learnt from the First Computational Paralinguistics Challenge, Computer Speech & Language (2018), http://dx.doi.org/10.1016/j.csl.2018.02.004

Anton Batliner received his doctoral degree in Phonetics in 1978 at LMU Munich. His main research interests are all (cross-linguistic) aspects of prosody and (computational) paralinguistics. He is co-editor/author of two books and author/co-author of more than 300 technical articles, with an h-index of > 40 and > 8000 citations.

Stefan Steidl received his diploma degree in computer science in 2002 from Friedrich-Alexander-Universität Erlangen-Nürnberg (FAU). In 2008, he received his doctoral degree from FAU for his work in the area of vocal emotion recognition. In 2010/2011, he spent one year as a research scholar at the International Computer Science Institute (ICSI) at Berkeley in the group of Nelson Morgan. Currently, Stefan Steidl is a lecturer at the FAU Pattern Recognition Lab and head of the Medical Image Segmentation Group. He is (co-)author of 1 book, 2 book chapters, 17 journal articles, and 72 peer-reviewed conference papers; his current h-index is 30 with 4 500+ citations. Since 2009, he is co-organizer of the annual computational paralinguistics challenges at the INTERSPEECH conferences. He is a reviewer for many international journals and conferences and associate editor for IEEE Transactions on Affective Computing. In 2012, he was a co-organizer of the Workshop on Child Computer Interaction (WOCCI 2012) in Portland. In 2015, he was the general chair of the Workshop on Speech and Language Technology for Education (SLaTE 2015) and served as publication chair for INTERSPEECH 2015.

Florian Eyben is co-founder and Chief Technical Officer (CTO) at audEERING GmbH. Holding a doctorate in electrical engineering from Technische Universität München (TUM), he is an expert in the field of digital signal processing, speech and music analysis and machine learning. His 100+ publications have 6000+ citations (h-index 37). Florian Eyben is responsible for the analytics tools openSMILE and openEAR which are capable of something not all humans are able to accomplish: listening between the lines.

Erik Marchi received his M.Sc. degree in electronic engineering in 2011 from Università Politecnica delle Marche in Ancona/Italy. He is currently working towards his PhD degree in the Machine Intelligence and Signal Processing group of Technische Universität München in Munich/Germany. His research focusses on affective computing, speech recognition, acoustic novelty detection, and music information retrieval. His further area of involvement is centered around the EU-FP7 project ASC-Inclusion to teach children with autism how to recognise and express emotion. There, he led the development of a vocal expression evaluation system providing corrective feedback. He is also involved in the EU-H2020 project DE-ENIGMA to realize robust, context-sensitive, multi-modal and naturalistic human-robot interaction aimed at enhancing the social imagination skills of children with autism. He is a member of the IEEE/ACM and (co-)authored more than 50 publications (1k citations) in peer-reviewed journals and conference proceedings.



Alessandro Vinciarelli is a Professor at the School of Computing Science and Associate Academic of the Institute of Neuroscience and Psychology. His research interest is the analysis of nonverbal-social behavior in real world situations like debates and meetings. In particular his work focuses on four major social phenomena: emergence and dynamics of conflicts, display of status and power relationships, role recognition, automatic personality perception and communication effectiveness. His 200+ publications found 5k citations (h-Index = 34).

Please cite this article as: B. Schuller et al., Affective and behavioural computing: Lessons learnt from the First Computational Paralinguistics Challenge, Computer Speech & Language (2018), http://dx.doi.org/10.1016/j.csl.2018.02.004

Klaus Scherer (PhD Harvard University) has held professorships at the University of Pennsylvania and the Universities of Kiel, Giessen, and Geneva. He is currently an emeritus professor at the University of Geneva and an honorary professor at the University of Munich. His extensive work on different aspects of emotion, in particular vocal and facial expression and emotion induction by music, has been widely published in international peer-reviewed journals. Klaus Scherer is a fellow of several international scientific societies and a member of several learned academies. He founded and directed the Swiss Center for Affective Sciences, held an Advanced Grant of the European Research Council and has been awarded honorary doctorates by the University of Bologna and the University of Bonn.

Mohamed Chetouani is the head of the IMI2S (Interaction, Multimodal Integration and Social Signal) research group at the Institute for Intelligent Systems and Robotics (CNRS UMR 7222), University Pierre and Marie Curie-Paris 6. He received the M.S. degree in Robotics and Intelligent Systems from the UPMC, Paris, 2001. He received the PhD degree in Speech Signal Processing from the same university in 2004. In 2005, he was an invited Visiting Research Fellow at the Department of Computer Science and Mathematics of the University of Stirling (UK). Prof. Chetouani was also an invited researcher at the Signal Processing Group of Escola Universitaria Politecnica de Mataro, Barcelona (Spain). He is currently a Visiting Researcher at the Human Media Interaction Lab of the University of Twente. He is now a full professor in Signal Processing, Pattern Recognition and Machine Learning at the UPMC. His research activities, carried out at the Institute for Intelligent Systems and Robotics, cover the areas of social signal processing and personal robotics through non-linear signal processing, feature extraction, pattern classification and machine learning. He is also the co-chairman of the French Working Group on Human- Robots/Systems Interaction (GDR Robotique CNRS) and a Deputy Coordinator of the Topic Group on Natural Interaction with Social Robots (euRobotics). He is the Deputy Director of the Laboratory of Excellence SMART Human/Machine/Human Interactions In The Digital Society.

Marcello Mortillaro is senior scientist and Head of Applied Research at University of Geneva – Swiss Center for Affective Sciences – one the largest centers in the world entirely devoted to the study of emotion and other affective phenomena. He authored several publications in scientific international journals and has been awarded several grants from public (Swiss National Science Foundation) and private institutions (among others Wrigley Inc.) to investigate the role of emotions in product development and service optimisation, as well as for development of new instruments to assess emotions and emotion-related skills.

Please cite this article as: B. Schuller et al., Affective and behavioural computing: Lessons learnt from the First Computational Paralinguistics Challenge, Computer Speech & Language (2018), http://dx.doi.org/10.1016/j.csl.2018.02.004