Interspeech 2018
2-6 September 2018, Hyderabad

The INTERSPEECH 2018 Computational Paralinguistics Challenge: Atypical & Self-Assessed Affect, Crying & Heart Beats

Björn W. Schuller[1,2,3], *Stefan Steidl*[4], *Anton Batliner*[2,4], *Peter B. Marschik*[5,6,7], *Harald Baumeister*[8], *Fengquan Dong*[9], *Simone Hantke*[2,10], *Florian B. Pokorny*[5,10], *Eva-Maria Rathner*[8], *Katrin D. Bartl-Pokorny*[5], *Christa Einspieler*[5], *Dajie Zhang*[5,6], *Alice Baird*[2], *Shahin Amiriparian*[2,10], *Kun Qian*[2,10], *Zhao Ren*[2], *Maximilian Schmitt*[2], *Panagiotis Tzirakis*[1], *Stefanos Zafeiriou*[1,11]

[1]GLAM – Group on Language, Audio & Music, Imperial College London, UK
[2]ZD.B Chair of Embedded Intelligence for Health Care and Wellbeing, University of Augsburg, Germany
[3]audEERING GmbH, Gilching, Germany
[4]Pattern Recognition Lab, FAU Erlangen-Nuremberg, Germany
[5]iDN – interdisciplinary Developmental Neuroscience, Medical University of Graz, Austria
[6]University Medical Center Göttingen, Germany
[7]Department of Women's and Children's Health, Karolinska Institutet, Stockholm, Sweden
[8]Department of Clinical Psychology and Psychotherapy, University of Ulm, Germany
[9]Shenzhen University General Hospital, Shenzhen, P.R. China
[10]Machine Intelligence & Signal Processing Group, Technische Universität München, Germany
[11]University of Oulu, Finland

schuller@IEEE.org

Abstract

The INTERSPEECH 2018 Computational Paralinguistics Challenge addresses four different problems for the first time in a research competition under well-defined conditions: In the *Atypical Affect* Sub-Challenge, four basic emotions annotated in the speech of handicapped subjects have to be classified; in the *Self-Assessed Affect* Sub-Challenge, valence scores given by the speakers themselves are used for a three-class classification problem; in the *Crying* Sub-Challenge, three types of infant vocalisations have to be told apart; and in the *Heart Beats* Sub-Challenge, three different types of heart beats have to be determined. We describe the Sub-Challenges, their conditions, and baseline feature extraction and classifiers, which include data-learnt (supervised) feature representations by end-to-end learning, the 'usual' ComParE and BoAW features, and deep unsupervised representation learning using the AUDEEP toolkit for the first time in the challenge series.

Index Terms: Computational Paralinguistics, Challenge, Atypical Affect, Self-Assessed Affect, Crying, Heart Beats

1. Introduction

In this INTERSPEECH 2018 COMPUTATIONAL PARALINGUISTICS CHALLENGE (COMPARE) – the tenth since 2009 [1], we address four new problems within the field of Computational Paralinguistics [2] in a challenge setting: In the **Atypical Affect (A) Sub-Challenge**, four basic emotions annotated in the speech of people with disabilities have to be classified. A possible application is a speech-driven, emotionally sensitive assistance system to support disabled individuals. In the **Self-Assessed Affect (S) Sub-Challenge**, valence scores given by speakers themselves are used for a three-class classification task. These experiments want to lay the ground for applications that support individuals with affective disorders and monitor synchronisation between therapists and clients. In the **Crying (C) Sub-Challenge**, three mood-related types of infant vocalisation have to be told apart: neutral/positive, fussing, and crying. This allows automatic (mood) monitoring of babies not only for research purposes, but also for clinical or home applications ('intelligent baby-phone'). Finally, in the **Heart Beats (H) Sub-Challenge**, normal, mild, and moderate/severe types of heart beats have to be classified. By including these acoustic – albeit non-vocal – signals, we contribute to heart sound analysis [3]; applications are self-evident, e. g., the monitoring of patients with unclear symptoms.

For all tasks, a target value/class has to be predicted for each case. Contributors can employ their own features and machine learning algorithms; standard feature sets and procedures are provided that may be used. Participants have to use predefined training/development/test splits for each Sub-Challenge. They may report development results obtained from the training set (preferably with the supplied evaluation setups), but have only a limited number of five trials to upload their results on the test sets per Sub-Challenge, whose labels are unknown to them. Each participation must be accompanied by a paper presenting the results, which undergoes peer-review and has to be accepted for the conference in order to participate in the Challenge. The organisers preserve the right to re-evaluate the findings, but will not participate in the Challenge. As evaluation measure, we employ Unweighted Average Recall (UAR) as used since the first Challenge held in 2009 [1], especially because it is more adequate for (more or less unbalanced) multi-class classifications than Weighted Average Recall (i. e., accuracy). Ethical approval for the studies has been obtained from the pertinent committees. In the next section 2, we describe the challenge corpora. Section 3 details the baseline experiments and metrics as well as the baseline results; concluding remarks are given in section 4.

2. The Four Sub-Challenges

2.1. The Atypical Affect (A) Sub-Challenge

For the **EmotAsS (EMOTional Sensitivity ASsistance System for people with disabilities)** database [4], 15 mentally, neuro-

10.21437/Interspeech.2018-51

logically, and/or physically disabled individuals were recorded in a familiar room at their workplace while speaking about their personal and health issues, including their form of disability: 8 f, 7 m; age 20-58 years; mean age 33 years; std. dev. 11.8 years; 12 mentally and 2 neurologically disabled, 1 with multiple disabilities. Due to strict ethical restrictions, no further details on the disabilities can be given. Audio was recorded with a Zoom H6 and a Jabra Speak 510 microphone, both at 44.1 kHz, 24 bit in mono. A recording setup was developed in order to avoid undue stress. An experimental supervisor, an occupational therapist, and a psychologist were sitting next to the participants all the time to communicate and help them through the five tasks: (i) describing images, e. g., from persons, beaches, or catastrophe scenes; (ii) talking on specific topics (e. g., their favourite travel destination, or sports activity); (iii) telling a story of a picture book; (iv) answering questions about professional life; and (v) playing together games like "Ludo (Do not get angry)". A typical session did not last more than one hour and the participants could make a break whenever needed. Overall, 10 627 segmented chunks representing 9.2 hours of speech were collected. The data were annotated on average by 12 volunteering annotators using the gamified crowdsourcing platform iHEARu-PLAY[1] [5]. The labellers had to choose between the six basic emotions (anger, disgust, fear, happiness, sadness, and surprise) and neutral. Since only very few chunks were annotated as disgust, fear, and surprise, these chunks were discarded.

2.2. The Self-Assessed Affect (S) Sub-Challenge

According to Russell's theory of a core affect [6], every emotional state is a combination of valence and arousal values. These aspects of the core affect influence attention [7], perception [8], cognition, judgement [9], memory retrieval and behaviour [10], i. e., the state of mind. The general principle is mood congruency; positive core affect shifts attention to positive material, negative core affect to negative material, and vice versa [11]. The aim of the **Ulm State-of-Mind in Speech (USoMS)** corpus and study is to distinguish state of mind, more specifically core affect via valence in free speech. From the 127 students recorded, due to technical reasons or subjects' withdrawal, 100 remained (85 f, 15 m, age 18-36 years, mean 22.3 years, std. dev. 3.6 years). Audio was captured in Stereo, converted to mono, at 44.1 kHz, 32 bit, and manually cleaned. The students told two negative and two positive stories ('narratives'), each with a duration of some 5 min. Before recording, and after each narrative, they self-assessed valence and arousal on a 10-point Likert scale. This yields global scores, given for a certain period of time where emotions surely fluctuate [12]; more fine-grained attentional shifts towards emotions would, however, change the subject's perception [13]. Valence has been chosen as the most relevant domain of interest for this task. Segments of 8 seconds each were selected from the cleaned recordings in a semi-automatic way. This resulted in 2 313 chunks. To create the three-class classification task, the raw values for the five self-assessed valence scores have been mapped onto (i) (L)ow: 0-4, (ii) (M)edium: 5-7, (iii) (H)igh: 8-10.

2.3. The Crying (C) Sub-Challenge

The **Cry Recognition In Early Development (CRIED)** database comprises 5 587 vocalisations of 20 healthy infants (10 f, 10 m, no pre-terms) recorded within a study on postnatal neuro-functional and neuro-behavioural changes and adaptations [14]. In a multi-device setup, the infants were recorded 7 times in bi-weekly intervals, with the first assessment at 4 weeks and the last one at 16 weeks of age (post-term). All vocalisations were extracted from sequences of up to 5 minutes in duration in which the infants were awake, lying in supine position in a cot. During these sequences, the infants were not exposed to external stimuli or manipulation. The CRIED corpus is based on audio-video recordings with a standard HD camcorder (Panasonic HC-V707) stored in MPEG-4 format (audio: 2 channels (LR), 44.1 kHz, AAC (LC)). The camcorder was mounted 0.8 m above the foot of the cot. Vocalisation segmentation was based on the concept to assign a vocalisation to a distinct vocal breathing group [15]. Vegetative sounds, such as breathing sounds, smacking sounds, hiccups, etc., were not segmented and thus not included in the dataset. The data has been partitioned into two splits: one partition (train-LOSO) for training and optimisation of the models' hyperparameters through a leave-one-subject-out cross-validation (LOSO) and a test partition. The vocalisations of both partitions were categorised into the following three classes: (i) neutral/positive mood vocalisations (2 292 cases in train-LOSO), (ii) fussing vocalisations (368 cases in train-LOSO), and (iii) crying vocalisations (178 cases in train-LOSO). The categorisation process was done on the basis of audio-video clips by two experts in the field of early speech-language development.

2.4. The Heart Beats (H) Sub-Challenge

The **Heart Sounds Shenzhen (HSS)** corpus, provided by the Shenzhen University General Hospital, comprises heart sounds gathered from 170 subjects (55 f, 115 m; ages from 21 to 88 years (mean age 65.4 years, std. dev. 13.2 years) with varied health conditions (including coronary heart disease, heart failure, arrhythmia, hypertension, hyperthyroid, valvular heart disease, congenital heart disease, etc.). Audio was recorded in .wav format, using an Electronic Stethoscope (Eko CORE, USA) set up via Bluetooth 4.0, with a 4 kHz sampling rate and a 20 Hz–2 kHz frequency response, in four locations on the body, i e., auscultatory mitral, aortic valve auscultation, pulmonary valve auscultation, and auscultatory areas of the tricuspid valve. In each area, a duration of 30 seconds on average within [29.808 s; 30.152 s] in a sitting or supine position of the subjects was recorded, resulting in 845 recordings representing 422.82 min from the 170 participants.

Three classes have to be recognised: (i) normal, (ii) mild, and (iii) moderate/severe (heart disease), as diagnosed by physicians specialised in heart disease. Annotations were confirmed by echocardiography (i. e., cardiac ultrasound). The recordings were split into subject-independent train/development/test sets. We aimed at an approximately equal distribution of gender, class, and age according to three age groups: (i) 21-36 years, (ii) 37-62 years, and (iii) 63-88 years, ending up with 502/180/163 instances for the train/development/test sets collected from 100/35/35 subjects.

3. Experiments and Results

For all Sub-Challenges, except for HSS (unchanged with original sampling rate of 4 kHz), the segmented and categorised audio was converted to single-channel 16 kHz, 16 bits PCM format.

3.1. End-to-end Learning

For the second time in the COMPARE challenge, we provide results using end-to-end learning (e2e) models. An attractive characteristic of these models is that the optimal features for a

[1] https://www.ihearu-play.eu

Table 1: *Databases: Number of instances per class in the train/devel/test splits (or LOSO for CRIED); Test split distributions are blinded during the ongoing challenge.*

#	Train	Devel	Test	Σ
Emotion-sensitive Assistance Systems (EmotAsS)				
Angry	125	50	272	447
Happy	743	965	650	2 358
Neutral	2 287	2 842	2 024	7 153
Sad	187	329	153	669
Σ	3 342	4 186	3 099	10 627
Ulm State-of-Mind in Speech (USoMS)				
Low (l)	95	79	75	249
Medium (m)	388	310	353	1 051
High (h)	363	353	297	1 013
Σ	846	742	725	2 313
Cry Recognition In Early Development (CRIED)				
Neutral/positive (0)	2 292		2 172	4 464
Fussing (1)	368		441	809
Crying (2)	178		136	314
Σ	2 838		2 749	5 587
Heart Sounds Shenzhen (HSS)				
Normal (0)	84	32	28	144
Mild (1)	276	98	91	465
Moderate/Severe (2)	142	50	44	236
Σ	502	180	163	845

given task can be learnt purely from the data at hand, i. e., we aim to learn simultaneously the optimal features and the classifier in a single optimisation problem. Similar to [16], we use a convolutional network to extract features from the raw time representation and then a subsequent recurrent network with Gated Recurrent Units (GRUs) which performs the final classification. For training the network, we split the raw waveform into chunks of 40 ms each. These are fed into a convolutional network comprised by a series of alternating convolution and pooling operations which try to find a robust representation of the original signal (cf. participant scripts). The extracted features are then fed to M GRU modules (cf. Table 2) which compress the temporal signal to a single final hidden state of the recurrent network which is then used to perform the final classification. For our purposes the END2YOU toolkit was utilised [17][2].

3.2. COMPARE Acoustic Feature Set

The official baseline feature set is the same as has been used in the five previous editions of the INTERSPEECH COMPARE challenges. This feature set contains 6 373 static features resulting from the computation of various functionals over low-level descriptor (LLD) contours [18]. The configuration file is the ComParE_2016.conf, which is included in the 2.3 public release of OPENSMILE [19, 20]. A full description of the feature set can be found in [21].

3.3. Bag-of-Audio-Words

In addition to the default ComParE feature set, where functionals (statistics) are applied to the acoustic LLDs, we provide Bag-of-Audio-Words (BoAW) features. BoAW has already been applied successfully for, e. g., acoustic event detection [22], speech-based emotion recognition [23], and classification of snore sounds [24]. Audio chunks are represented as histograms of acoustic LLDs, after quantisation based on a codebook. One codebook is learnt for the 65 LLDs from the COMPARE feature set and one for the 65 deltas of these LLDs. In Table 2, results are given for different codebook sizes. Codebook generation is done by *random sampling* from the LLDs in the training data. When fusing training and development data for the final model, the codebook is learnt again from the fused data. The LLDs have been extracted with the OPENSMILE toolkit, BoAW have been computed using OPENXBOW [25].

3.4. AUDEEP

Another feature set is obtained through unsupervised representation learning with recurrent sequence to sequence autoencoders, using the AUDEEP toolkit[3] [26]. Representation learning commonly requires less human intervention than manually engineering a feature set such as the COMPARE acoustic feature set. The recurrent sequence to sequence autoencoders which are employed by AUDEEP, in particular, explicitly model the inherently sequential nature of audio with RNNs within the encoder and decoder networks [27, 26]. In the AUDEEP approach, Mel-scale spectrograms are first extracted from the raw waveforms in a data set. In order to eliminate some background noise, power levels are clipped below four given thresholds in these spectrograms, which results in four separate sets of spectrograms per data set. Subsequently, a distinct recurrent sequence to sequence autoencoder is trained on each of these sets of spectrograms in an unsupervised way, i. e., without any label information. The learnt representations of a spectrogram are then extracted as feature vectors for the corresponding instance. Finally, these feature vectors are concatenated to obtain the final feature vector. For the results shown in Table 2, the autoencoders' hyperparameters were not optimised.

3.5. Challenge Baselines

The primary evaluation measure for the Sub-Challenges (all being classification tasks) is Unweighted Average Recall (UAR). The motivation 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 for all four Sub-Challenges of this year).

For the sake of transparency and reproducibility of the baseline computation and in line with the previous years, we use an open-source implementation of Support Vector Machines (SVM) with linear kernels and Sequential Minimal Optimisation (SMO) [28] as training algorithm from WEKA 3 (revision 3.8.2) [29] for the classification based on functionals, BoAW, and AUDEEP features. Features were scaled to zero mean and unit standard deviation (option `-N 1` for Weka's SMO), using the parameters from the respective training set (when multiple folds where used for development, the parameters were calculated on the training set of each fold). For all tasks, the complexity parameter C was optimised during the development phase.

Each Sub-Challenge package includes scripts that allow participants to reproduce the baselines and perform the testing in a reproducible and automatic way (including pre-processing, model training, model evaluation, and scoring by the competition and further measures).

This year, we provide the four above outlined approaches to computational paralinguistics: besides the usual COMPARE features plus SVM, we employ for the second time e2e and BoAW

[2] A detailed implementation of these models can be found at `https://github.com/end2you/end2you`

[3] `https://github.com/auDeep/auDeep`

Table 2: *Results for the four Sub-Challenges. The **official baselines** for test are highlighted (bold and greyscale). Dev: Development. LOSO: Leave-one-subject-out; for e2e, 3 subjects were held out as a validation partition instead of using the LOSO split used with the other approaches. M: Number of LSTM layers in end-to-end (e2e) learning. C: Complexity parameter of the SVM. N: Codebook size of Bag-of-Audio-Words (BoAW) splitting the input into two codebooks (ComParE-LLDs/ComParE-LLD-Deltas), with 10 assignments per frame, optimised complexity parameter of SVM. X: Power levels which are clipped below four given thresholds. **UAR**: Unweighted Average Recall.*

UAR [%]	**A**typical	**S**elf-Ass.	**C**rying	**H**eart
	Dev/Test	Dev/Test	Loso/Test	Dev/Test
M	**END2YOU: CNN + LSTM**			
2	41.8/28.0	49.7/45.8	–/63.5	41.2/37.7
3	40.1/27.9	48.2/46.6	–/61.3	37.5/37.7
C	**OPENSMILE: COMPARE functionals + SVM**			
10^{-6}	34.2/41.3	54.2/60.0	72.8/67.3	41.1/44.8
10^{-5}	37.8/43.1	54.2/61.2	75.6/71.9	44.5/45.6
10^{-4}	28.2/38.4	56.5/65.2	67.5/57.5	50.3/46.4
10^{-3}	29.9/35.4	53.9/64.7	67.9/66.1	44.5/40.4
10^{-2}	29.9/33.1	53.9/64.9	74.5/67.6	43.2/41.7
N	**OPENXBOW: COMPARE BoAW + SVM**			
250	38.7/36.4	56.5/61.7	75.6/73.2	43.1/43.4
500	40.5/36.5	55.8/59.1	75.9/71.8	42.3/47.2
1000	39.8/38.1	52.5/63.2	76.9/67.7	43.7/41.0
2000	38.1/41.3	51.1/62.2	75.5/69.8	42.6/52.3
4000	37.9/39.2	56.7/60.9	75.4/68.8	39.9/43.3
X	**AUDEEP: RNN + SVM**			
$-30\,dB$	32.5/35.4	39.1/46.6	64.4/58.5	34.1/40.3
$-45\,dB$	33.3/33.3	43.5/53.7	74.4/62.1	40.3/40.0
$-60\,dB$	38.1/33.6	48.8/57.0	69.0/68.0	38.3/46.3
$-75\,dB$	39.1/34.9	44.8/54.4	73.2/71.1	34.1/42.1
fused	40.4/35.6	49.9/57.3	70.5/68.6	38.6/47.9
n-best	**Fusion (Majority Vote)**			
2-best	–/42.9	–/65.4	–/70.4	**–/56.2**
3-best	–/42.0	**–/66.0**	**–/74.6**	–/51.1
4-best	–/41.0	–/62.2	–/71.3	–/53.0
n-best	**Fusion (Confidence-based)**			
2-best	**–/43.4**	–/64.6	–/73.1	–/49.3
3-best	–/42.0	–/64.7	–/73.9	–/53.6
4-best	–/42.0	–/64.7	–/73.9	–/53.6

plus SVM; additionally, we present novel sequence-to-sequence autoencoder (AUDEEP) learnt acoustic features, classified with an SVM. The same way as last year, we choose the highest results on test for defining the baselines, irrespective of the corresponding results on Dev/LOSO, in order to prevent participants from surpassing the official baseline by simply repeating or slightly modifying other constellations that can be found in Table 2. A fusion of the n-best models has been made in the following two ways: 1) *Majority Vote:* The label predicted most often by the n approaches performing best in the respective Sub-Challenge and split is considered. In case of a same number of votes for more than one class, the majority prediction for the whole split is taken into account. 2) *Confidence-based:* From the n approaches performing best in the respective Sub-Challenge and split, the confidence scores are added together for each class and the class with maximum confidence sum is considered for each instance. As the confidence scores obtained by the e2e approach are not comparable to those of the SVM, the confidence score of e2e predictions is assumed to be 1.0.

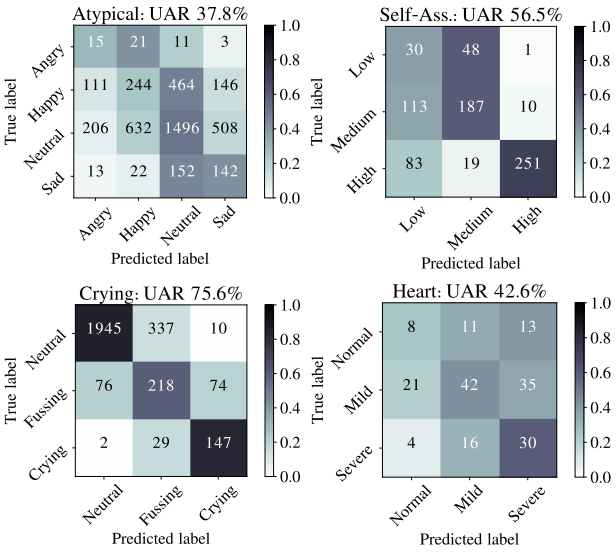

Figure 1: *Confusion matrices of the development set. For each Sub-Challenge, the individual approach/hyperparameters performing best on the test set has been chosen.*

As can be seen in Table 2, for the Atypical Affect Sub-Challenge, the baseline is UAR $= 43.4\,\%$; for the Self-Assessed Affect Sub-Challenge, it is UAR $= 66.0\,\%$, for the Crying Sub-Challenge, it is UAR $= 74.6\,\%$, and for the Heart Beats Sub-Challenge, it is UAR $= 56.2\,\%$. All baselines have been reached with a fusion of the predictions of 2 or 3 models, respectively.

Figure 1 displays a 'good' confusion for Crying (high frequencies in diagonal cells) and the difficulty of the other tasks (low frequencies in some of the diagonal cells, high frequencies in some of the off-diagonal cells).

4. Concluding Remarks

This year's challenge is new in several respects – four new tasks (Atypical and Self-Assessed Affect, Crying, and Heart Beats, all of them highly relevant for applications) and a new procedure: sequence-to-sequence autoencoder-based audio features by the AUDEEP toolkit using deep learning for audio classification. We further featured the END2YOU toolkit providing for the second time end-to-end learning baselines and the popular OPENXBOW toolkit. For all computation steps, scripts are provided that can, but need not be used by the participants. We expect participants to obtain considerably better performance measures by employing novel (combinations of) procedures and features including such tailored to the particular tasks.

5. Acknowledgements

We acknowledge funding from the EU's HORIZON 2020 Grants No. 115902 (RADAR CNS) and No. 645378 (ARIA-VALUSPA), the EU's 7$^{\text{th}}$ ERC Starting Grant No. 338164 (iHEARu), the German national BMBF IKT2020-Grant No. 16SV7213 (EmotAsS), the EPSRC Centre for Doctoral Training in High Performance Embedded and Distributed Systems (HiPEDS, Grant Reference EP/L016796/1), the Bavarian State Ministry of Education, Science and the Arts in the framework of the Centre Digitisation.Bavaria (ZD.B), and BioTechMed-Graz, Austria. We thank the sponsor of the Challenge: audEERING GmbH.

6. References

[1] B. Schuller, A. Batliner, S. Steidl, and D. Seppi, "Recognising realistic emotions and affect in speech: State of the art and lessons learnt from the first Challenge," *Speech Communication*, vol. 53, pp. 1062–1087, 2011.

[2] B. Schuller and A. Batliner, *Computational Paralinguistics – Emotion, Affect, and Personality in Speech and Language Processing*. Chichester, UK: Wiley, 2014.

[3] G. D. Clifford, C. Liu, B. Moody, J. Millet, S. Schmidt, Q. Li, I. Silva, and R. G. Mark, "Recent advances in heart sound analysis," *Physiological Measurement*, vol. 38, pp. E10–E25, 2017.

[4] S. Hantke, H. Sagha, N. Cummins, and B. Schuller, "Emotional Speech of Mentally and Physically Disabled Individuals: Introducing The EmotAsS Database and First Findings," in *Proc. of INTERSPEECH 2017*. Stockholm, Sweden: ISCA, 2017, pp. 3137–3141.

[5] S. Hantke, F. Eyben, T. Appel, and B. Schuller, "iHEARu-PLAY: Introducing a game for crowdsourced data collection for affective computing," in *Proc. 1st Workshop on Automatic Sentiment Analysis in the Wild (WASA) held in conjunction with ACII 2015*. Xi'an, P. R. China: IEEE, 2015, pp. 891–897.

[6] J. Russel, "Core affect and the psychological construction of emotions," *Phychological Review*, vol. 110, no. 1, pp. 145–172, 2003.

[7] S. Shapiro and D. Maclnnis, "Understanding program-induced mood effects: Decoupling arousal from valenc," *Journal of Advertising*, vol. 31, no. 4, pp. 15–26, 2013.

[8] M. Mather and M. R. Sutherland, "Arousal-biased competition in perception and memory," *Perspectives on Psychological Science*, vol. 6, no. 2, pp. 114–133, 2011.

[9] E. Peters, D. VastfJall, and T. Garling, "Affect and decision making: A "hot" topic," *Journal of Behavioral Decision Making*, vol. 19, no. 1, pp. 79–85, 2006.

[10] B. Hommel, A. Moors, D. Sander, and J. Deonna, "Emotion meets action: Towards an integration of research and theory," *Emotion Review*, vol. 9, no. 4, pp. 295–298, 2017.

[11] N. Shwarz and G. L. Clore, "Mood, misattribution, and judgments of well-being: Informative and directive functions of aggective states," *Journal of Personality and Social Psychology*, vol. 45, no. 3, pp. 513–523, 1983.

[12] P. Koval and P. Kuppens, "Changing emotion dynamics: Individual differences in the effect of anticipatory social stress on emotional inertia," *Emotion*, vol. 22, no. 2, pp. 256–267, 2012.

[13] P. Koval, E. A. Butler, T. Hollenstein, D. Lanteigne, and P. Kuppens, "Emotion regulation and the temporal dynamics of emotions: Effects of cognitive reappraisal and expressive suppression on emotional inertia," *Cognition and Emotion*, vol. 29, no. 5, pp. 831–851, 2014.

[14] P. B. Marschik, F. B. Pokorny, R. Peharz, D. Zhang, J. O'Muircheartaigh, H. Roeyers, S. Bölte, A. J. Spittle, B. Urlesberger, B. Schuller, L. Poustka, S. Ozonoff, F. Pernkopf, T. Pock, K. Tammimies, C. Enzinger, M. Krieber, I. Tomantschger, K. D. Bartl-Pokorny, J. Sigafoos, L. Roche, G. Esposito, M. Gugatschka, K. Nielsen-Saines, C. Einspieler, W. E. Kaufmann, and The BEE-PRI study group, "A Novel Way to Measure and Predict Development: A Heuristic Approach to Facilitate the Early Detection of Neurodevelopmental Disorders," *Current Neurology and Neuroscience Reports*, vol. 17, no. 43, 2017, 15 pages.

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

[16] G. Trigeorgis, F. Ringeval, R. Brueckner, E. Marchi, M. A. Nicolaou, B. Schuller, and S. Zafeiriou, "Adieu features? End-to-end speech emotion recognition using a deep convolutional recurrent network," in *Proc. of ICASSP*. Shanghai, China: IEEE, 2016, pp. 5200–5204.

[17] P. Tzirakis, S. Zafeiriou, and B. W. Schuller, "End2you–the imperial toolkit for multimodal profiling by end-to-end learning," *arXiv preprint arXiv:1802.01115*, 2018.

[18] B. Schuller, S. Steidl, A. Batliner, A. Vinciarelli, K. Scherer, F. Ringeval, M. Chetouani, F. Weninger, F. Eyben, E. Marchi, M. Mortillaro, H. Salamin, A. Polychroniou, F. Valente, and S. Kim, "The INTERSPEECH 2013 Computational Paralinguistics Challenge: Social Signals, Conflict, Emotion, Autism," in *Proc. of INTERSPEECH*. Lyon, France: ISCA, 2013, pp. 148–152.

[19] F. Eyben, M. Wöllmer, and B. Schuller, "openSMILE – The Munich Versatile and Fast Open-Source Audio Feature Extractor," in *Proc. of ACM Multimedia*. Florence, Italy: ACM, 2010, pp. 1459–1462.

[20] F. Eyben, F. Weninger, F. Groß, and B. Schuller, "Recent developments in openSMILE, the Munich open-source multimedia feature extractor," in *Proc. of ACM Multimedia*. Barcelona, Spain: ACM, 2013, pp. 835–838.

[21] F. Weninger, F. Eyben, B. Schuller, M. Mortillaro, and K. R. Scherer, "On the Acoustics of Emotion in Audio: What Speech, Music and Sound have in Common," *Frontiers in Emotion Science*, vol. 4, no. 292, pp. 1–12, May 2013.

[22] H. Lim, M. J. Kim, and H. Kim, "Robust sound event classification using LBP-HOG based bag-of-audio-words feature representation," in *Proc. of INTERSPEECH*. Dresden, Germany: ISCA, 2015, pp. 3325–3329.

[23] M. Schmitt, F. Ringeval, and B. Schuller, "At the border of acoustics and linguistics: Bag-of-audio-words for the recognition of emotions in speech," in *Proc. of INTERSPEECH*. San Francisco, USA: ISCA, 2016, pp. 495–499.

[24] M. Schmitt, C. Janott, V. Pandit, K. Qian, C. Heiser, W. Hemmert, and B. Schuller, "A bag-of-audio-words approach for snore sounds' excitation localisation," in *Proc. of ITG Speech Communication*. Paderborn, Germany: IEEE, 2016, pp. 230–234.

[25] M. Schmitt and B. W. Schuller, "openXBOW – introducing the passau open-source crossmodal bag-of-words toolkit," *The Journal of Machine Learning Research*, vol. 18, pp. 1–5, 2017.

[26] M. Freitag, S. Amiriparian, S. Pugachevskiy, N. Cummins, and B. Schuller, "auDeep: Unsupervised learning of representations from audio with deep recurrent neural networks," *Journal of Machine Learning Research*, vol. 19, 2018, 5 pages, to appear.

[27] S. Amiriparian, M. Freitag, N. Cummins, and B. Schuller, "Sequence to sequence autoencoders for unsupervised representation learning from audio," in *Proc. of the Detection and Classification of Acoustic Scenes and Events 2017 Workshop (DCASE)*. IEEE, 2017, pp. 17–21.

[28] J. Platt, "Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods," in *Advances in large margin classifiers*, A. Smola, P. Bartlett, B. Schölkopf, and D. Schuurmans, Eds. Cambridge, MA: MIT Press, 1999, pp. 61–74.

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