Emotional speech of mentally and physically disabled individuals: introducing the EmotAsS database and first findings

Simone Hantke, Hesam Sagha, Nicholas Cummins, Björn Schuller

Angaben zur Veröffentlichung / Publication details:

Hantke, Simone, Hesam Sagha, Nicholas Cummins, and Björn Schuller. 2017. "Emotional speech of mentally and physically disabled individuals: introducing the EmotAsS database and first findings." In *Situated interaction: 18th Annual Conference of the International Speech Communication Association (INTERSPEECH 2017), Stockholm, Sweden, 20-24 August 2017; Volume 5*, 3137–41. Red Hook, NY: Curran Associates, Inc. https://doi.org/10.21437/interspeech.2017-409.

INTERSPEECH 2017
August 20–24, 2017, Stockholm, Sweden

Emotional Speech of Mentally and Physically Disabled Individuals: Introducing the EmotAsS Database and First Findings

Simone Hantke[1,2], *Hesam Sagha*[1,3], *Nicholas Cummins*[1] *and Björn Schuller*[1,3,4]

[1]Chair of Complex & Intelligent Systems, University of Passau, Germany
[2]Machine Intelligence & Signal Processing Group, Technische Universität München, Germany
[3]audEERING GmbH, Gilching, Germany
[4]Department of Computing, Imperial College London, UK

simone.hantke@uni-passau.de

Abstract

The automatic recognition of emotion from speech is a mature research field with a large number of publicly available corpora. However, to the best of the authors knowledge, none of these datasets consist solely of emotional speech samples from individuals with mental, neurological and/or physical disabilities. Yet, such individuals could benefit from speech-based assistive technologies to enhance their communication with their environment and to manage their daily work process. With the aim of advancing these technologies, we fill this void in emotional speech resources by introducing the EmotAsS (Emotional Sensitivity Assistance System for People with Disabilities) corpus consisting of spontaneous emotional German speech data recorded from 17 mentally, neurologically and/or physically disabled participants in their daily work environment, resulting in just under 11 hours of total speech time and featuring approximately 12.7 k utterances after segmentation. Transcription was performed and labelling was carried out in seven emotional categories, as well as for the intelligibility of the speaker. We present a set of baseline results, based on using standard acoustic and linguistic features, for arousal and valence emotion recognition.

Index Terms: Emotional Corpus, Disabilities, Speech-driven Assistive Technology, Neural Networks

1. Introduction

Individuals with mental and/or physical disabilities and disorders can react with particularly sensitively to stress factors [1, 2]. This can cause emotional fluctuations, which in turn, affects their abilities and cognitive processes. Therefore, they can often have difficulties with remembering what they have learnt, retrieving their knowledge, or being able to concentrate. These problems can introduce difficulties associated with completing daily tasks, and therefore, there is a need for constant supervision and help from others. To reduce this burden, they can benefit from advanced assistance techniques and systems to enhance their communication with their environment and improve their daily life situation [3, 4].

The rapid growth of technology – especially the release of smartphones and tablets – offers a myriad of advanced communication and computing capabilities. This growth opens a new world of opportunities in different areas of healthcare by proposing different applications for many health conditions such as dementia [5], autism [6], depression [7], or Parkinson's disease [8]. In this context, the word "mHealth" has been introduced [9, 10], which can be described as the application of mobile communications and network technologies for healthcare [11]. Specifically this entails the usage of portable devices with the capability to create, analyse, store, retrieve, and transmit data in real-time between the users, for the purpose of improving individuals' safety and quality of life [9].

As well as being used to improve life quality, mHealth systems facilitate communication between clinicians and patients [8]. Remote monitoring systems, based on smart technologies, have been proposed for asthma patients [12], the tracking of patients with dementia [13], and to support treatment of sleep apnoea [14]. An overview of smartphone use in behavioural healthcare for integrating mobile technologies into clinical practices is provided in [10]. Such technologies can also be used in other kinds of assistance systems, such as for unobtrusively recognising stress from human voice [15], or to perform suicide prevention [16]. Further, they can be used to enable individuals to both join the workforce and aid them at their job [17].

Within the *Emotional Sensitivity Assistance System for People with Disabilities* (EmotAsS) project[1], a language-driven, workplace integrated, assistance system is being developed, that supports individuals with mental, neurological, and/or physical disabilities in the handling of certain activities while taking into account their emotional-cognitive constitution and state. The system being developed is designed to recognise the emotional state – derived from factors such as the choice of words, loudness or pitch – of users with disabilities to help divide their work process into individually manageable small work steps that can vary in the degree of difficulty, depending on the user's condition. In addition, the system may recommend users to take a break if there is an increased risk of injury. The overall aim of this assistance system is to help avoid stressful situations, to facilitate independent work, and thus to strengthen the self-confidence of disabled individuals in the workplace.

1.1. Contributions of this Work

This paper introduces the *EmotAsS database*, a new, first-of-its kind, dataset which includes the spontaneous recorded emotional speech of 17 mentally, neurologically, and/or physically disabled individuals[2]. The corpus contains approximately 12.7 k utterances and just under 11 hours of emotional German speech recorded in the wild. Baseline emotion recognition results are also presented, giving the first insights into the development of the novel assistance system discussed above. To the best of our knowledge, this is the first study that performs the

[1] http://www.emotass.de

[2] The EmotAsS database will be available upon request for scientific research purposes.

http://dx.doi.org/10.21437/Interspeech.2017-409

automatic analysis of spontaneous emotion in speech from mentally, neurologically, and physically disabled individuals.

2. Emotional Speech Data Collection

To develop the EmotAsS assistance system, a database of emotional speech from individuals with a mental, neurological, or physical disability is required. Until now, such a database has not been collected, even though the kind of disability often influences the speech of the individuals. Therefore, we recorded speech data from disabled employees at a shelter for disabled individuals.

2.1. Participants

Within the EmotAsS project, an ethics approval was granted by an external ethics committee (datenschutz nord GmbH), who gave permission to collect the proposed data. Seventeen participants agreed to take part in the experiment and to record, store, and distribute data for the scientific purposes, by the signing of a written consent form themselves or by their legal guardian. The recording procedure and the consent forms have also been checked and approved by the external ethic committee. All given data are stored in an anonymous form, with no identifying information collected from the participants.

The participants provided data relating to their personal and health issues including their form of disability. Out of the 17 participants, 10 are female and 7 male, ages range from 19 to 58 with a mean age of 31.6 years, and standard deviation of 11.7 years. As there are strict ethic restrictions on the data, no further details on the disability of the subjects are given, but we can cluster them into mental, neurological, and physical disabilities. Thirteen participants are mentally disabled, 3 neurological, and one has multiple disabilities (cf. Table 1).

2.2. Setup

Taking into account the daily or even hourly mood changes of the participants, a recording setup was developed, as not to add undue stress. For our recordings, the participants were invited into a familiar room of the shelter. To achieve a high but also realistic audio quality, the recordings took place in a working room with equal set-up and conditions for each recording session. The participants had to sit down in front of the recording equipments; an experimental supervisor, an internal occupational therapist, and an internal psychologist of the shelter were sitting next to the participants all the time to communicate and help participants through the given tasks. Before the recordings, the participants were instructed on the procedure by the experimental supervisor and were able to ask questions also during the tasks.

For technical realisation of our recordings, we used a Zoom H6 and a Jabra Speak 510 microphone, both with a sampling frequency of 44.1 kHz in mono with 24 bits per sample.

2.3. Tasks

The recorded data consists of spontaneous speech and was recorded by giving the participants different tasks related to different contents. In this way, questions were raised about professional life and certain tasks to be accomplished. The tasks were designed in such a way that the mood of the participants can be shown and their emotions are being provoked. To ensure a professional management of possibly expressed emotions, the tasks were performed in supervision of a psychologist. Five different tasks were performed by:

Table 1: *Demographics on the 17 participants (Gen(der): f(emale), m(ale) and Eth(nicity): G(erman), M(acedonian)), with different Dis(abilities): Me(ntal), Neu(rolocical) and/or Mu(ltiple), showing the allover Dur(ation) and number of instances (#).*

ID	Gen.	Age	Eth	Dis.	Dur.	#
01	m	28	G	Me.	15’15”	441
02	f	58	G	Me.	30’22”	563
03	m	31	G	Me.	50’35”	435
04	m	22	G	Me.	31’54”	1063
05	f	54	G	Me.	5’44”	279
06	m	43	G	Me.	57’57”	746
07	f	37	G	Me.	30’22”	530
08	f	32	G	Me.	37’33”	1043
09	f	46	G	Me.	19’34”	1003
10	m	30	G	Neu.	43’23”	354
11	f	24	G	Neu.	34’43”	1116
12	f	22	G	Me.	53’51”	1649
13	f	20	G	Neu.	29’23”	521
14	f	19	G	Me.	56’34”	843
15	f	23	G	Mu.	16’42”	186
16	m	20	M	Me.	1°06’31”	796
17	m	29	G	Me.	1°14’13”	1184

1. Showing images, e. g., from persons, beaches, catastrophe scenes, or an injured dog, which typically trigger certain emotions (cf. Figure 1). The participants were asked to assess and describe what they see on a given picture and how they feel when looking at it.

2. Asking the participants to talk on specific topics (e. g., their favourite travel destination, genre of music, or sports activity).

3. Telling a story of the pictured book “Frog, where are you?”[18], wherein a little boy tries to find his escaped frog and has different adventures (cf. Figure 2), which was earlier successfully used for recordings with children with Autism Spectrum Disorders [19]. In total, the book included 15 negative, 6 neutral, and 5 emotionally positive pictures.

4. Asking the participants about their everyday business, e. g., employees reported cleaning the sanitary facilities, ironing, or putting together the laundry.

5. Playing together games like “Ludo (Do not get angry)” [20], depending on their mental and physical possibilities.

Due to the different tasks and recording situations (1-5), it was possible to record the participants in different emotional moments. In this regard, we found that very different emotional responses could be observed for the same task; for example, a woman laughed with joy when she had to describe a picture with a hurt dog because she simply liked the dog. Another woman wept over the picture because she had just lost her dog.

We ensured, that a typical session of one participant did not last more than one hour – always taking into account the individuals condition – and the participants were able to make a break whenever needed. Overall, 12 752 instances (turns) representing 10.54 hours of speech were recorded coming from the 17 participants.

Figure 1: *Exemplary pictures shown to mentally, neurologically, and physically disabled individuals to record spontaneous speech correlated with emotion production [21, 22, 23, 24].*

Figure 2: *An extract from the book "Frog where are you?' [18] that was used for recording spontaneous emotional speech from mentally, neurologically and physically disabled individuals. We hypothesised that images with neutral valence (left picture, the beginning of the story: the boy is sleeping in his bed while his frog escaped), negative valence (middle picture, the middle of the story: the boy is searching for his frog), and positive valence (right picture, the end of the story: the boy finally finds his frog being conjunct with his family) are correlated with emotion production.*

2.4. Annotations

The data was annotated using our gamified crowdsourcing platform iHEARu-PLAY [3] [25]. For our performed emotion recognition tasks, we gathered labels giving the annotators the choice to select from the six basic emotions (anger, disgust, fear, happiness, sadness, and surprise), as well as 'neutral' [26]. Moreover, each utterance was annotated on a 5-point likert scale, to represent the intelligibility of the speech (cf. Table 2), so the data can be used for tasks such as automatic speech recognition. The transcription was performed by two fluent German speakers.

The annotators had the possibility to repeat listening to each utterance as often as they needed, before submitting their final answer. Overall, 29 annotators labelled the 12 752 instances and each instance was labelled on average by 12 annotators. The annotators could give their metadata voluntarily within iHEARu-PLAY. Out of the 29 overall annotators, 21 gave us their metadata. Out of them we had 8 female and 13 male annotators – their age ranged from 19 to 42 years, with a mean of 24.1 years and a standard deviation of 5.47 years.

Table 2: *Data distribution according to the participants intelligibility rated on a 5-point likert scale (1 = not intelligible at all, 5 = totally intelligible). Note, not all files had been rated at the time of publication.*

Intelligibility	**1**	**2**	**3**	**4**	**5**
# Instances	769	900	1376	3886	3104

Table 3: *Distribution of the acoustic feature sets, covering Spec(tral)/energy-related, Sou(rce)/excitation-related and Dur(ation)-related features with different levels of detail for different configurations of openSMILE [27], introduced in Interspeech (IS) Challenges.*

Feature-Set (IS)	**Spec.**	**Sou.**	**Dur.**	**Total**
IS09 Emotion [28]	336	48	0	384
IS10 Paralinguistic [29]	1216	212	154	1582
IS11 Speaker State [30]	2808	272	1288	4368
IS16 ComParE [31]	4366	397	1610	6373
eGeMAPS [32]	48	48	6	102

3. Automatic Emotion Recognition Experiments

In this section, we provide baseline results on automatic emotion recognition on the recorded data. We perform acoustic emotion recognition using five popular feature sets on converted emotions (from seven categories to arousal/valence dimensions).

3.1. Acoustic Features

The acoustic features of the audio files were automatically extracted by using the openSMILE toolkit [27]. We investigated four standard large brute-forced feature sets (IS09 [28], IS10 [29], IS11 [30], and ComParE2016 [31]), which have all been used for paralinguistic information retrieval and showed clear tendencies in enlarging the feature space over the years. Furthermore, we investigate the smaller, expert knowledge-based feature set eGeMAPS [32]. Detailed information on these five feature sets, which cover spectral-, source-, and duration-related feature space with different levels of detail, can be found in Table 3, and a detailed description is given in [33].

3.2. Setup

To provide a baseline, we performed a leave-one-subject-out cross-validation. Moreover, since labels were not perfectly balanced within the emotion classes, without loss of generality, we converted the categorical emotions into dimensional arousal and valence representations, to yield more uniform distribution. Therefore, we converted each emotional category to its corresponding continuous arousal and valence values, and took the average of all the ratings of each utterance. We used the conversion values from [34], which are summarised in Table 4. Finally, we normalised emotions on positive and negative values separately, so that maximum and minimum values are -1 and 1, respectively. The histogram of arousal and valence is provided in Figure 3.

[3] https://www.ihearu-play.fim.uni-passau.de/

Table 4: *Ar(ousal) and Val(ence) conversion values for each categorical emotion, following [34].*

Emotion	**Ar.**	**Val.**	**Emotion**	**Ar.**	**Val.**
Sadness	-0.40	-0.80	Surprise	0.90	0.10
Fear	0.80	-0.20	Happiness	0.15	0.95
Anger	0.80	-0.40	Disgust	0.50	-0.65
Neutral	0.00	0.00			

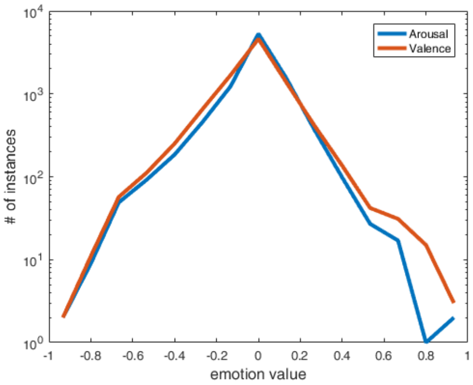

Figure 3: *Logarithmic histogram of the converted and normalised emotions into arousal and valence dimensions.*

3.3. Evaluation

As the baseline regressor, we used a feed-forward neural network with two hidden layers[4]. At each step of leave-one-subject-out cross-validation, training data is divided into development (20%) and training (80%) sets to find the best number of hidden nodes in each layer. For the first layer, 10 to 100 nodes with the step of 10 are evaluated, while for the second layer 5, 10, and 20 nodes were considered. All the activation functions are set to 'hyperbolic tangent sigmoid' transfer function, and the training function is set to 'scaled conjugate gradient backpropagation'. The performance is measured with the correlation coefficient (CC) as well as the concordance correlation coefficient (CCC) which considers shift and variance in the labels [35].

4. Results

The results of acoustic emotion recognition using the five different feature sets is provided in Table 5. As the results indicate, all the feature sets except IS09 are appropriate for the recognition of arousal and the IS10 feature set yields slightly higher performance of valence recognition. Least performances are obtained using the IS09 feature set with 384 features. On the classification side, the average best selected number of nodes for the neural networks for the feature sets are provided in Table 6. It can be seen that arousal recognition needs slightly more training weights than valence recognition.

On the analysis of the transcriptions, we did not find satisfiable correlation with arousal and valence, using standard techniques such as Bag-of-Words, auto-encoders, and word affect nets. The highest correlation (app. $CC = 0.12$) obtained by using Berlin Affective Word List-Reloaded (BAWL-R) which maps about 3 K German words into their arousal and valence values [36]. The low performance could be due to the limited number of spoken words (∼9 K) in the corpus, and/or due to speech impairments associated with the mental, neurological, and/or physical disabled participants [37]. However, it can be improved using external large datasets for building sentiment recognition by considering contextual information (e. g., techniques such as word embedding).

Table 5: *Results of regression on the EmotAsS database using different acoustic feature sets.*

Feature set	**Arousal**		**Valence**	
	CC	CCC	CC	CCC
IS09	0.417	0.331	0.262	0.191
IS10	0.462	0.372	**0.315**	**0.239**
IS11	**0.468**	**0.386**	0.293	0.228
IS16	0.466	0.378	0.278	0.211
eGeMAPS	0.460	0.382	0.298	0.220

Table 6: *Average selected number of nodes for the 1^{st} and 2^{nd} layers of our feed-forward neural network regressor.*

Feature set	**Arousal**		**Valence**	
	1^{st}	2^{nd}	1^{st}	2^{nd}
IS09	62	9	58	13
IS10	49	12	43	8
IS11	50	8	47	9
IS16	56	11	51	9
eGeMAPS	66	13	65	9

5. Conclusion and Outlook

We introduced the novel emotional German speech database EmotAsS, containing speech samples from 17 individuals with mental, neurological, and/or physical disabilities. All in all, the corpus consists of just under 11 hours of total speech time and contains approximately 12.7 k speech samples recorded during performing five different tasks. We presented a set of baseline results and showed that using acoustic features, we could yield concordance correlation coefficients of 0.386 and 0.239 for arousal and valence recognition, respectively, using feed-forward neural networks. For the baseline results, we converted the emotion representations from categorical six basic emotions to the arousal/valence dimensions. Future work will focus on gathering continuous labels directly from annotators for arousal and valence for better comparison with the class labels and running further emotion recognition experiments. Finally, we plan to perform automatic speech recognition experiments, to analyse the content of speech for emotion recognition (with word embedding techniques) in a fully automated manner.

6. Acknowledgements

The research leading to these results has received funding from the European Union's Seventh Framework Programme under grant agreement No 338164 (ERC Starting Grant iHEARu), from the German national BMBF IKT2020-Grant under grant agreement No 16SV7213 (EmotAsS), and from the Innovative Medicines Initiative 2 Joint Undertaking under grant agreement No 115902. This Joint Undertaking receives support from the European Union's Horizon 2020 research and innovation programme and EFPIA. We thank all iHEARu-PLAY users for donating their annotations.

[4] The choice of the feed-forward neural network, over other standard classification methods such as Support Vector Machines, is due to its scalability property.

7. References

[1] H. McConachie, "Implications of a model of stress and coping for services to families of young disabled children," *Child: care, health and development*, vol. 20, pp. 37–46, 1994.

[2] W. C. Gamble and S. M. McHale, "Coping with stress in sibling relationships: A comparison of children with disabled and nondisabled siblings," *Journal of Applied Developmental Psychology*, vol. 10, pp. 353–373, 1989.

[3] H. Hoenig, D. H. Taylor Jr, and F. A. Sloan, "Does assistive technology substitute for personal assistance among the disabled elderly?" *American Journal of Public Health*, vol. 93, pp. 330–337, 2003.

[4] L. M. Verbrugge, C. Rennert, and J. H. Madans, "The great efficacy of personal and equipment assistance in reducing disability." *American journal of public health*, vol. 87, pp. 384–392, 1997.

[5] N. K. Vuong, S. Chan, and C. T. Lau, "mHealth sensors, techniques, and applications for managing wandering behavior of people with dementia: A review," in *Mobile Health*. Springer, 2015, pp. 11–42.

[6] C. Tryfona, G. Oatley, A. Calderon, and S. Thorne, "M-Health solutions to support the national health service in the diagnosis and monitoring of autism spectrum disorders in young children," in *Proc. of Int. Conference on Universal Access in Human-Computer Interaction*. Toronto, Canada: Springer, July 2016, pp. 249–256.

[7] N. Cummins, B. Vlasenko, H. Sagha, and B. Schuller, "Enhancing speech-based depression detection through gender dependent vowel-level formant," in *Proc. of Conference on Artificial Intelligence in Medicine*. Springer, 2017, p. 5.

[8] B. C. Zapata, J. L. Fernández-Alemán, A. Idri, and A. Toval, "Empirical studies on usability of mHealth apps: a systematic literature review," *Journal of medical systems*, vol. 39, p. 1, 2015.

[9] R. Istepanian, S. Laxminarayan, and C. S. Pattichis, *M-Health*. Springer, 2006.

[10] D. D. Luxton, R. A. McCann, N. E. Bush, M. C. Mishkind, and G. M. Reger, "mhealth for mental health: Integrating smartphone technology in behavioral healthcare." *Professional Psychology: Research and Practice*, vol. 42, p. 505, 2011.

[11] E. Jovanov and Y. Zhang, "Introduction to the special section on M-Health: Beyond seamless mobility and global wireless healthcare connectivity," *Transactions on Information Technology in Biomedicine*, vol. 8, pp. 405–414, 2004.

[12] L. S. Namazova-Baranova, A. I. Molodchenkov, E. A. Vishneva, E. V. Antonova, and V. I. Smirnov, "Remote monitoring of children with asthma, being treated in multidisciplinary hospital," in *Proc. of Int. Conference on Biomedical Engineering and Computational Technologies*. Novosibirsk, Russia: IEEE, October 2015, pp. 7–12.

[13] F. Miskelly, "Electronic tracking of patients with dementia and wandering using mobile phone technology," *Age and ageing*, vol. 34, pp. 497–498, 2005.

[14] V. Isetta, M. Torres, K. González, C. Ruiz, M. Dalmases, C. Embid, D. Navajas, R. Farré, and J. M. Montserrat, "A new mHealth application to support treatment of sleep apnoea patients," *Journal of telemedicine and telecare*, vol. 10, 2015.

[15] H. Lu, D. Frauendorfer, M. Rabbi, M. S. Mast, G. T. Chittaranjan, A. T. Campbell, D. Gatica-Perez, and T. Choudhury, "Stresssense: Detecting stress in unconstrained acoustic environments using smartphones," in *Proc. of Conference on Ubiquitous Computing*. Pittsburgh, USA: ACM, September 2012, pp. 351–360.

[16] M. E. Larsen, N. Cummins, T. W. Boonstra, B. O'Dea, J. Tighe, J. Nicholas, F. Shand, J. Epps, and H. Christensen, "The use of technology in suicide prevention," in *Proc. of Int. Conference on Engineering in Medicine and Biology Society*. Milan, Italy: IEEE, August 2015, pp. 7316–7319.

[17] L. M. Verbrugge and P. Sevak, "Use, type, and efficacy of assistance for disability," *The Journals of Gerontology Series B: Psychological Sciences and Social Sciences*, vol. 57, pp. 366–379, 2002.

[18] M. Mayer, *Frog, where are you?* Dial Press New York, 1969.

[19] F. Ringeval, E. Marchi, C. Grossard, J. Xavier, M. Chetouani, D. Cohen, and B. Schuller, "Automatic analysis of typical and atypical encoding of spontaneous emotion in the voice of children," in *Proc. of INTERSPEECH*. San Francisco, USA: ISCA, September 2016, pp. 1210–1214.

[20] E. Glonnegger, C. Voigt, J. Rüttinger, and K. Kappler, *Das Spiele-Buch: Brett-und Legespiele aus aller Welt: Herkunft, Regeln und Geschichte*. Ravensburger Buchverlag, 2009.

[21] "Police operation," May 2016, http://www.bmi.bund.de/SharedDocs/Pressemitteilungen/DE/2015/05/pks-und-pmk-2014.html, [Online].

[22] "Injured dog," May 2016, http://www.bild.de/regional/hannover/prozesse/gassi-attacke-vor-gericht-44762806.bild.html, [Online].

[23] "Beach on the Maledives," May 2016, http://paper4pc.com/relaxing-beach-screensavers.html, [Online].

[24] "War in jenem," May 2016, http://www.faz.net/aktuell/politik/ausland/krieg-im-jemen-luftangriffe-auf-houthi-stellungen-13815693.html, [Online].

[25] S. Hantke, F. Eyben, T. Appel, and B. Schuller, "iHEARu-PLAY: Introducing a game for crowdsourced data collection for affective computing," in *Proc. of Int. Workshop on Automatic Sentiment Analysis in the Wild, satellite of Conference on Affective Computing and Intelligent Interaction*. Xi'an, China: IEEE, September 2015, pp. 891–897.

[26] O. Langner, R. Dotsch, G. Bijlstra, D. H. Wigboldus, S. T. Hawk, and A. van Knippenberg, "Presentation and validation of the radboud faces database," *Cognition and emotion*, vol. 24, pp. 1377–1388, 2010.

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

[28] B. Schuller, S. Steidl, and A. Batliner, "The Interspeech 2009 Emotion Challenge," in *Proc. of INTERSPEECH*. Brighton, UK: ISCA, September 2009, pp. 312–315.

[29] B. Schuller, S. Steidl, A. Batliner, F. Burkhardt, L. Devillers, C. Müller, and S. Narayanan, "The INTERSPEECH 2010 Paralinguistic Challenge," in *Proc. of INTERSPEECH*. Makuhari, Japan: ISCA, September 2010, pp. 2794–2797.

[30] B. Schuller, A. Batliner, S. Steidl, F. Schiel, and J. Krajewski, "The INTERSPEECH 2011 Speaker State Challenge," in *Proc. of INTERSPEECH*. Florence, Italy: ISCA, August 2011, pp. 3201–3204.

[31] B. Schuller, S. Steidl, A. Batliner, J. Hirschberg, J. K. Burgoon, A. Baird, A. Elkins, Y. Zhang, E. Coutinho, and K. Evanini, "The INTERSPEECH 2016 Computational Paralinguistics Challenge: Deception, Sincerity & Native Language," in *Proc. of INTERSPEECH*. San Francisco, USA: ISCA, September 2016, pp. 2001–2005.

[32] F. Eyben, K. Scherer, B. Schuller, J. Sundberg, E. André, C. Busso, L. Devillers, J. Epps, P. Laukka, S. Narayanan, and K. Truong, "The Geneva Minimalistic Acoustic Parameter Set (GeMAPS) for voice research and affective computing," *Transactions on Affective Computing*, vol. 7, pp. 190–202, 2016.

[33] F. Eyben, *Real-time speech and music classification by large audio feature space extraction*. Springer, 2015.

[34] K. R. Scherer, "What are emotions? And how can they be measured?" *Social Science Information*, vol. 44, pp. 695–729, 2005.

[35] L. I.-K. Lin, "A concordance correlation coefficient to evaluate reproducibility," *Biometrics*, vol. 45, pp. 255–268, 1989.

[36] M. L. H. Võ, M. Conrad, L. Kuchinke, K. Urton, M. J. Hofmann, and A. M. Jacobs, "The Berlin Affective Word List Reloaded (BAWL-R)," *Behavior Research Methods*, vol. 41, pp. 534–538, 2009.

[37] W. H. Organization, "ICD-10 guide for mental retardation," 1996.