Predicting Biological Signals from Speech: Introducing a Novel Multimodal Dataset and Results

Alice Baird
Augsburg University, Germany
alice.baird@informatik.uni-augsburg.de

Shahin Amiriparian
Augsburg University, Germany
shahin.amiriparian@informatik.uni-augsburg.de

Miriam Berschneider
Augsburg University, Germany
miriam.berschneider@student.uni-augsburg.de

Maximilian Schmitt
Augsburg University, Germany
maximilian.schmitt@informatik.uni-augsburg.de

Björn Schuller
Augsburg University, Germany
bjoern.schuller@informatik.uni-augsburg.de

***Abstract*— In recent years, diagnosis and awareness of mental health conditions, e. g., chronic stress, have been increasing globally. Biological signals can be an effective way to monitor such conditions, yet acquisition can be cumbersome and invasive. Alternatively, acoustic features offer non-invasive and efficient monitoring of an array of health and wellbeing characteristics. This study presents the BioSpeech Database (BIOS-DB), a novel database of audio and biological signals – blood volume pulse (BVP) and skin conductance (SC) – from 55 individuals speaking aloud in front of others, whilst having their emotional state annotated in real time. Through a variation of conventional and state-of-the-art approaches, initial experiments have shown for the first time that acoustic features can be applied for the task of BVP prediction. Notably, using deep representations of audio and a sequence-to-sequence auto-encoders with a GRU-RNN as a time-dependent regressor achieved at best 0.075 and 0.123 RMSE for [0 ; 1] normalised BVP and SC, respectively.**

I. INTRODUCTION

Individual wellbeing is an ever present aspect of modern life and in recent years a variety of research fields have begun focusing their efforts upon this. An array of diagnosed conditions can fall into the definition of sub-optimal wellbeing, including stress disorders that are becoming particularly prevalent due to the increasing modern pressures such as the working environment [1]. The ability to continually monitor such characteristics has shown economic benefits, as employers are more easily able to facilitate support [2].

With this in mind, speaking aloud is an aspect of daily-life, i. e., in the workplace, that can result in what is known as performance anxiety [3]. Performance anxiety is a condition particularly know to performance, e. g., musicians, resulting in symptoms of increased heart rate, sweating, dry mouth and shortness of breath [4].

When an individual is in an overwhelmed state, speech can be one characteristic which is altered, with attributes such as articulation rate and filled pauses varying [5]. As well as speech, a multitude of biological signals can be monitored during such situations, in particular blood volume pulse (BVP) and skin conductance (SC) have been shown to be of value in the affective computing community, to continuously monitor emotional states [6].

All authors are affiliated to the ZD.B Chair of Embedded Intelligence for Health Care and Wellbeing. Björn Schuller is also affiliated to GLAM – Group on Language, Audio and Music, Imperial College London, UK.

In this regard, combining multiple modalities is now the state-of-the-art for human-like interpretations of states, e. g., emotion [7], or traits, e. g., gender [8]. Within the health domain, subtle manifestations of conditions including depression can be recognised more accurately with combined (audio and video) approaches and the annual Audio-Visual Emotion Challenge (AVEC) recently focused on this [9]. Acoustic features alone offer non-intrusive monitoring, which has shown great success for many health related tasks from, characteristics of autism [10] to snore sound classification [11], and previously using conventional acoustic features has shown success for automatic recognition of biological signals including heart rate and SC [12]. In a similar way, acoustic features fused with biological signals have also been successfully applied for emotion tasks [13].

With this in mind, this study presents a novel database – the BioSpeech database (BIOS-DB) – collected from 55 individuals speaking aloud in both German and English, with real-time continuous emotional annotations. To the best of our knowledge, it is the first time that such emotion-based annotations have been made in this continuous manner, at the exact time of the participants' speech. For initial experiments on the BIOS-DB, we utilise multiple modalities and present a prediction task in which acoustic and emotional features are utilised to predict biological signals. For the first time using blood volume pulse (BVP) as the target, conventional expert designed acoustic features are extracted, as well as using state-of-the-art sequence to sequence autoencoders (S2SAE) for deep representations. Both of aforementioned methods have shown promise for tasks in the health domain [14].

The paper is structured as follows: the dataset is described in Section II, followed by a description of the experimental setup Section III and then a discussion of the results Section IV. Concluding remarks are then given in Section V.

II. THE BIOSPEECH DATABASE

The BioSpeech database (BIOS-DB) is a novel database of individuals in a public speaking scenario. The BIOS-DB obtained full ethical approval[1] and was collected over

[1]The BIOS-DB data acquisition obtained full ethical approval from the University of Augsburg Ethics Commission under the project title 'Multimodal Signal Recording Techniques and Emotional Analysis'. BIOS-DB is available at `http://doi.org/10.5281/zenodo.3346632` (restricted access). Cite this paper and [15] if using the database.

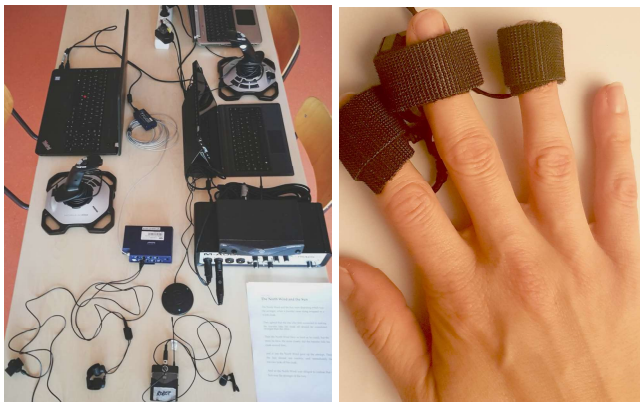

Fig. 1: Left: The BIOS-DB study equipment from speaker perspective - speakers text, room microphone placement, utilised biosignal sensors and annotator joysticks. Right: The placement of the sensors on the hand of the participant.

a period of one month, with subjects voluntarily participating. Based on aspects of the recognised Tier Social Stress Test [16] (a study protocol developed in the field of psychology to evoke states of controlled stress), in the BIOS-DB individuals are reciting a text (the North Wind and the Sun) in both German and English language and in real time have their emotional states annotated by three individuals, whilst audio and biosignals (BVP and SC) are captured.

A. Participants

The BIOS-DB contains 55 individuals (33 male and 22 female), with a mean age of 28.9 years ($\pm$ 10.5 years). Individuals were predominately German Natives (33)[2] and either students (30) or staff from the computer science department at the University of Augsburg, Germany. Participants were speaking in German and English. The average speech length was 45 s for German and 42 s for English.

Subjective self assessment measures were gathered from the participants, before and after speaking. These measures were taking into account the current state of wellbeing and took the form of a questionnaire, which was partially based on the on the short-form of the the Spielberger State-Trait Anxiety Inventory (STAI) by Marteau and Bekker [17]. The evaluation was gathered on a 5-point Likert Scale (weak to strong) and the mean and standard deviation across individuals for each parameter of the STAI evaluation is given in Table I. Of most prominence emotions which could be linked to poor wellbeing, i. e., stress and sadness are reduced after the activity and factors such as relaxation (positive wellbeing) increase. Although no significance was found, there seems to be a tendency of task apprehension.

B. Audio

To capture a realistic acoustic environment and comprehensible speech, two channels of audio were recorded during the study. One close-talk lapel microphone was placed on the individuals garments below the chin and a second room microphone was placed ca. 50 cm from the speaker on the table directly in front of them Figure 1). Using a 4-channel audio interface and the open-source software Audacity, audio was captured at 44.1k Hz and 16 bit in *WAV* format.

[2]Other nationalities in the BIOS-DB include, American (1), British (1), Chinese (3), Indian (1), Israeli (1), Russian (1) and Spanish (1).

TABLE I: (m)ean and standard deviation ($\pm$) values of the (B)efore and (A)fter values from participant State-Trait Anxiety Inventory (STAI) evaluation. The evaluation was gathered on a 5-point Likert Scale (weak to strong).

STAI	m	$\pm$	STAI	m	$\pm$
B-Stress	2.14	1.16	B-Relaxed	3.10	1.10
A-Stress	1.98	1.02	A-Relaxed	3.30	1.10
B-Excited	2.45	1.10	B-Sadness	1.31	0.63
A-Excited	2.30	1.20	A-Sadness	1.24	0.54

C. Biological signals

The biological signals (biosignals) included in the BIOS-DB were gathered using the Thought Technology ProComp Infiniti System and include blood volume pulse (BVP) and skin conductance (SC), signals discussed in the affective computing community to be fruitful signals measure human states [6]. In Figure 2, the normalised distribution of the signals used within the studies experimental setup are shown.

Blood volume pulse (BVP) is defined as the flow of blood through the vascular system and results from changes in the amount of blood in the capillaries, caused by the contraction of the heart [18]. Within the BIOS-DB , the raw BVP values were captured at 2 048 samples per second and values were consistent across all speakers ranging from $22.4 - 41.8$ BVP amplitude (amp) (mean: $35.7\,amp$, $\pm$: $0.77\,amp$).

Skin conductance (SC) is captured via electrodermal activity (EDA). SC depends on the activity of the sweat glands and the pore size of the skin, which are both controlled by the sympathetic nervous system. The sympathetic nervous system is activated especially during stress or anxiety inducing activities and often leads to increased sweat production and thus decrease skin conductivity [19]. As compared to BVP the raw SC values have more variance, captured at 256 samples per second, values range from $0.0 - 42.0$ microSiemens (μS) (mean: $8.2\,\mu S$, $\pm$: $5.7\,\mu S$).

D. Dimensional emotional annotations

Annotations of the dimensional emotions of arousal and valence were gathered in real-time. Using the conventional two dimensional circumplex model of emotion [20], three trained individuals (1 male and 2 female, students) were using a joystick and continually observing both dimensions. Arousal indicates an emotions active/ passive value and valence is how positive/ negative it is [20].

When creating a gold standard from the three individual time-continuous annotations, the bias that may be present depending on the weighting of a single annotator was considered [21]. Conventional approaches include the maximum likelihood estimation (MLE), however, a weighted metric such as the evaluator weighted estimation (EWE) could be

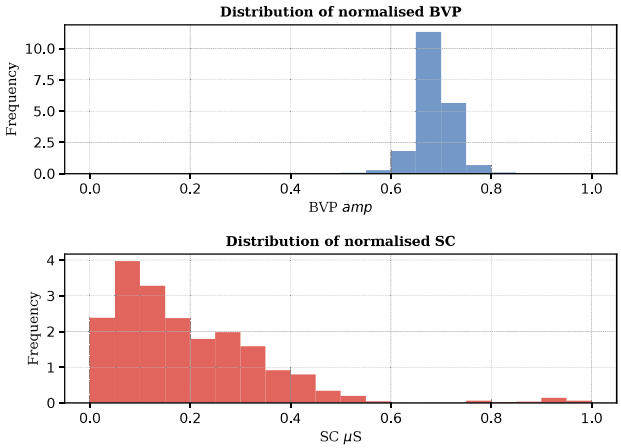

Fig. 2: Distribution of blood volume pulse (BVP) amplitude (amp) and skin conductance (SC) in microSiemens (μS) for all instances. Post normalisation and feature space reduction.

TABLE II: Speaker independent partitions, Train, and Test applied for the baseline results of the BIOS-DB. Including raw audio of combined speech scenarios (German, English,) and recording channels (close-talk and overhead), as well as instances extracted from the audio files, for each of the 9 feature sets described in Section III.

	Train	Test	$\sum$
Speakers	33	22	55
Gender M:F	11:22	10:12	21:34
Audio	132	88	220
Instances	11 364	7 906	19 270

more fairly produced for time-continuous ratings [22]. EWE considers the pairwise pearson's correlation (PC) weighting between each annotator and from the BIOS-DB, the mean EWE across the three annotators was 0.36 and 0.47 for valence and arousal, respectively.

III. EXPERIMENTAL SETTINGS

To evaluate the BIOS-DB, experiments to predict both BVP and SC have been performed, including conventional and state-of-the-art approaches. Speakers have been divided across Train and Test partitions, 33 in Train (11 female, 22 male), 22 in Test (10 female, 12 male). For the experiments to reduce the feature space and for concatenation, both biosignals were resampled to 2 samples per second and normalised to a range of $[0; 1]$.

In total 9 features sets have been extracted (2 COMPARE, 2 DEEP SPECTRUM and 5 S2SAE) from both audio channels. All extracted with a window size of 500 ms and an overlap of 250 ms, after which additional features were fused. The gold standard of the two-dimensional emotional features are then fused with a selection of feature sets to explore the emotionality which may be present in the biological signals.

A. Conventional acoustic features

Conventional acoustic features were extracted from speech through the use of our open-source OPENSMILE feature extractor [23]. The COMPARE feature set from the *INTERSPEECH 2016 Computational Paralinguistics Challenge* has been used [24]. This configuration includes features which have previously been used to classify speech corpora including social signals and emotion [24].

B. Deep spectrum features

The feature extraction DEEP SPECTRUM toolkit[3] is applied to obtain deep representations from the input audio data using pre-trained convolutional neural networks (CNNs) [25]. First, audio signals are transformed into mel-spectrogram plots using a Hanning window of width 500 ms and an overlap 250 ms. From these, 128 mel frequency bands are then computed. The generated spectrograms are then forwarded through AlexNet [26], a pre-trained CNN and the activations of the third-last fully connected layer (*fc6*) of the network are extracted, resulting in a 4 096 dimensional DEEP SPECTRUM feature set. These features can be considered as being a high-level representation of the mel-spectrograms [25].

C. Recurrent sequence to sequence autoencoders

For the second deep learning approach, AUDEEP[4] is utilised, which applies sequence-to-sequence autoencoders (S2SAEs) for unsupervised representation learning from audio signals [27], [28]. In the first step, input audio signals were chunked into 500 ms segments, from which power spectra are generated. An S2SAE is then trained on these spectra [28]. Afterwards, the learnt representations of the input spectra are extracted for use as feature vectors. The power spectra are created using periodic Hanning windows of width 80 ms and overlap 40 ms. Afterwards 128 log-scaled mel frequency bands are computed. Finally, the mel-spectrograms are normalised to have values in $[-1; 1]$, as the outputs of the S2SAE are constrained to this [27], [28].

The applied S2SAE architecture has two recurrent layers ($N_{layer}^{S2SAE} = 2$) each with $N_{unit}^{S2SAE} = 256$ Gate Recurrent Units (GRUs), a unidirectional encoder RNN and a bidirectional decoder RNN. Autoencoders are then trained on the created mel-spectrograms using Adam optimiser with a fixed learning rate of 0.001 [29] for 32 epochs in batches of 512 samples. A dropout of 20 % is applied to the output of each recurrent layer in order to reduce overfitting [30]. Furthermore, as noise may be present amplitude clipping as applied in [14] is explored. For the experiments, the amplitude of the spectrograms have been clipped using four thresholds $T_{amp} \in \{-30, -45, -60, -75\}\, dB$, resulting in four S2SAE feature sets plus one fusion set of all four representations.

[3] https://github.com/DeepSpectrum/DeepSpectrum
[4] https://github.com/auDeep/auDeep

TABLE III: Baseline results for the task of blood volume pulse (BVP) and skin conductance (SC) prediction from speech-based features utilising the BIOS-DB . The COMPARE , COMPARE + emotion (valence and arousal) (A), DEEP SPECTRUM , DEEP SPECTRUM + emotion (B) and AUDEEP (C) feature set (Dim.)ensions results are shown. We report the 2 best optimisation (C)omplexities for the SMOReg model. The targets have been normalised to $[0, 1]$, and Root Mean Squared Error (RMSE) is the evaluation metric. Emphasised results discussed in Section IV.

C	Dim.	BVP	SC
COMPARE (SMOReg)			
$C = 10^{-5}$	4 096	**0.146**	0.259
$C = 10^{-4}$	4 096	0.147	0.263
COMPARE + Emo (SMOREg)			
$C = 10^{-5}$	4 096 + 2	0.146	0.260
$C = 10^{-4}$	4 096 + 2	0.147	**0.265**

(A)

C	Dim.	BVP	SC
DEEP SPECTRUM (SMOReg)			
$C = 10^{-5}$	4 096	0.146	0.262
$C = 10^{-5}$	4 096	0.152	**0.279**
DEEP SPECTRUM + Emo (SMOReg)			
$C = 10^{-5}$	4 096 + 2	0.146	0.262
$C = 10^{-4}$	4 096 + 2	0.153	0.281

(B)

T_{amp}	Dim	BVP	SC
AUDEEP (GRU-RNN)			
-30	1 024	0.116	0.163
-45	1 024	0.121	0.124
-60	1 024	**0.075**	**0.123**
-75	1 024	0.084	0.137
Fused	4 096	0.207	0.133

(C)

D. Regression algorithms

As a conventional approach and to set a baseline, the Weka open-source machine learning environment was used with the Sequential Minimal Optimisation (SMO) algorithm for support vector regression (SMOreg) [31], [32]. Complexity was applied at 2 values 10^{-5} and 10^{-4}, feature normalisation was applied and all other parameters were the algorithms default.

For a state-of-the-art approach, a recurrent neural network (RNN) as a time dependent regressor was applied, which included long short-term memory (LSTM) cells and gated recurrent units (GRUs). The best configuration included one hidden layer with 128 GRUs, a dropout of 0.2, learning rate of 0.001 and 200 Epochs. This configuration used the conventional Adam optimiser, the activation functions for the hidden layers was rectified linear unit (ReLU) [33] and Sigmoid was used for the output layer.

IV. RESULTS AND DISCUSSION

A summary of results from all approaches for the task of biosignal prediction is given in Table III, with results under discussion in this section emphasised. Results from the conventional COMPARE features sets are given as the BIOS-DB baseline system reporting Root Mean Squared Error (RMSE) as the evaluation metric.

From the conventional approach with COMPARE features, we see results are consistently better for BVP, speculating that this could be due to the limited distribution of values for BVP: mean 0.55 ± 0.05, as compared to SC: mean 0.20 ± 0.16, ((cf. Figure 2 for an overview). In this same way, DEEP SPECTRUM features are achieving almost identical results, yet we see slight improvement when focusing on the second complexity for DEEP SPECTRUM of 0.01 RMSE as compared to COMPARE .

Fusing the emotional features to COMPARE and DEEP SPECTRUM has had little impact on the result. For COMPARE , we do see a slight improvement for SC, compared to the next best COMPARE only result. However, this is too slight to conclude any specific advantages to the emotional fusion. Given this small improvement, we suggest that emotional features are more closely correlated to SC values as there is a more of a physical response from this signal, however, this would require further research.

The GRU-RNN architecture in which various AUDEEP feature sets were used does show our best results – with $T_{amp} = -60$ dB. Amplitude clippings appear to improve predictions until threshold $(T_{amp}) = -75$ dB and it can be assumed that at this point more than noise is removed and the speech is obscured. We see from $T_{amp} = -30$ dB to $T_{amp} = -60$ dB for BVP quite a difference in the result, thus allowing us to assume that T_{amp} is a valuable hyper parameter.

V. CONCLUSIONS AND FUTURE WORK

In this study, the BIOS-DB has been presented and initial experiments for the task of biological signal prediction have been made. From these first step results, we see promise from the BIOS-DB and an abundance of future work can be made, such as deeper focus on the vocalisation types.

Results have shown that deep representation learning is the most effective approach for biological signal prediction from speech, particularly in the case of BVP, in which results were improved upon substantially. Despite this, the low resource features and algorithms are still showing to be competitive, against state-of-the-art deep learning, however, it could be stated that the data set itself may not be large enough for the deep architectures to learn as much meaningful information as it is capable of. Expanding the BIOS-DB through data augmentation would be a next step.

Additionally, emotional features did not appear to give any substantial benefit for the task of biosignal prediction, however there is a slight tendency from the SC results which indicate promise. The limited improvement could be due to annotation quality and a next step for the BIOS-DB would be to evaluate these annotation values more closely, making secondary offline annotations, in which arousal and valence dimension are annotated separately.

VI. ACKNOWLEDGEMENTS

This work is funded by the Bavarian State Ministry of Education, Science and the Arts in the framework of the Centre Digitisation.Bavaria (ZD.B).

REFERENCES

[1] Terry A Beehr, *Psychological stress in the workplace (psychology revivals)*, Routledge, 2014.

[2] Yasuhiko Deguchi, Shinichi Iwasaki, Akihito Konishi, Hideyuki Ishimoto, Koichiro Ogawa, Yuichi Fukuda, Tomoko Nitta, and Koki Inoue, "The usefulness of assessing and identifying workers temperaments and their effects on occupational stress in the workplace," *PloS one*, vol. 11, no. 5, pp. 1–12, 2016.

[3] Christopher R Jones, Russell H Fazio, and Michael W Vasey, "Attentional control buffers the effect of public-speaking anxiety on performance," *Social psychological science*, vol. 3, no. 5, pp. 556–561, 2012.

[4] Glenn D Wilson and David Roland, "Performance anxiety," *The science and psychology of music performance: Creative strategies for teaching and learning*, pp. 47–61, 2002.

[5] Christian Müller, Barbara Großmann-Hutter, Anthony Jameson, Ralf Rummer, and Frank Wittig, "Recognizing time pressure and cognitive load on the basis of speech: An experimental study," in *User Modeling 2001*, Mathias Bauer, Piotr J. Gmytrasiewicz, and Julita Vassileva, Eds., Berlin, Heidelberg, 2001, pp. 24–33, Springer Berlin Heidelberg.

[6] Rosalind W. Picard, E. Vyzas, and Jennifer. Healey, "Toward machine emotional intelligence: analysis of affective physiological state," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 23, no. 10, pp. 1175–1191, 2001.

[7] Panagiotis Tzirakis, George Trigeorgis, Mihalis A Nicolaou, Björn W Schuller, and Stefanos Zafeiriou, "End-to-end multimodal emotion recognition using deep neural networks," *IEEE Journal of Selected Topics in Signal Processing*, vol. 11, no. 8, pp. 1301–1309, 2017.

[8] Xiaoguang Lu, Hong Chen, and Anil K Jain, "Multimodal facial gender and ethnicity identification," in *Proc. International Conference on Biometrics*. Springer, 2006, pp. 554–561.

[9] Michel Valstar, Jonathan Gratch, Björn Schuller, Fabien Ringeval, Denis Lalanne, Mercedes Torres Torres, Stefan Scherer, Giota Stratou, Roddy Cowie, and Maja Pantic, "Avec 2016: Depression, mood, and emotion recognition workshop and challenge," in *Proc of International Workshop on Audio/Visual Emotion Challenge*. ACM, 2016, pp. 3–10.

[10] Shahin Amiriparian, Alice Baird, Sahib Julka, Alyssa Alcorn, Sandra Ottl, Suncica Petrović, Eloise Ainger, Nicholas Cummins, and Björn Schuller, "Recognition of Echolalic Autistic Child Vocalisations Utilising Convolutional Recurrent Neural Networks," in *Proc. INTERSPEECH*, Hyderabad, India, 2018, pp. 2334–2338, ISCA.

[11] Shahin Amiriparian, Maurice Gerczuk, Sandra Ottl, Nicholas Cummins, Michael Freitag, Sergey Pugachevskiy, Alice Baird, and Björn Schuller, "Snore sound classification using image-based deep spectrum features," in *Proc. INTERSPEECH*, 2017, pp. 3512–3516.

[12] Björn Schuller, Felix Friedmann, and Florian Eyben, "Automatic recognition of physiological parameters in the human voice: Heart rate and skin conductance," in *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*. IEEE, 2013, pp. 7219–7223.

[13] Gil Keren, Tobias Kirschstein, Erik Marchi, Fabien Ringeval, and Bjorn Schuller, "End-to-end learning for dimensional emotion recognition from physiological signals," in *Proc. ICME*, Hong Kong, China, 2017, IEEE, pp. 985–990.

[14] Shahin Amiriparian, Maximilian Schmitt, Nicholas Cummins, Kun Qian, Fengquan Dong, and Björn Schuller, "Deep unsupervised representation learning for abnormal heart sound classification," in *2018 40th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*. IEEE, 2018, pp. 4776–4779.

[15] Miriam Berschnider Maximilian Schmitt Alice Baird, Shahin Amiriparian and Björn Schuller, "The biospeech database (bios-db) [data set]," *Zenodo*, 2019.

[16] Clemens Kirschbaum, Karl-Martin Pirke, and Dirk H Hellhammer, "The trier social stress test–a tool for investigating psychobiological stress responses in a laboratory setting," *Neuropsychobiology*, vol. 28, no. 1-2, pp. 76–81, 1993.

[17] Theresa M Marteau and Hilary Bekker, "The Development of a Six-Item Short-Form of the State Scale of the Spielberger State-Trait Anxiety Inventory (STAI)," *British Journal of Clinical Psychology*, vol. 31, no. 3, pp. 301–306, 1992.

[18] Azadeh Kushki, Jillian Fairley, Satyam Merja, Gillian King, and Tom Chau, "Comparison of Blood Volume Pulse and Skin Conductance Responses to Mental and Affective Stimuli at different Anatomical Sites," *Physiological Measurement*, vol. 32, no. 10, pp. 1529–1541, 2011.

[19] Shalom Greene, Himanshu Thapliyal, and Allison Caban-Holt, "A Survey of Affective Computing for Stress Detection: Evaluating Technologies in Stress Detection for Better Health," *IEEE Consumer Electronics Magazine*, vol. 5, no. 4, pp. 44–56, 2016.

[20] James A Russell, "A Circumplex Model of Affect," *Journal of Personality and Social Psychology*, vol. 39, no. 6, pp. 1161–1178, 1980.

[21] Felix Weninger, Florian Eyben, and Bjorn Schuller, "On-line continuous-time music mood regression with deep recurrent neural networks," in *Proc. ICASSP*, Florence, Italy, 2014, IEEE, pp. 5412–5416.

[22] Simone Hantke, Erik Marchi, and Björn W Schuller, "Introducing the weighted trustability evaluator for crowdsourcing exemplified by speaker likability classification.," in *LREC*, Portoro, Slovenia, 2016.

[23] Florian Eyben, Felix Weninger, Florian Gross, and Björn Schuller, "Recent Developments in openSMILE, the Munich Open-Source Multimedia Feature Extractor," in *Proc. ACM*, Barcelona, Spain, 2013, pp. 835–838.

[24] Björn Schuller, Stefan Steidl, Anton Batliner, Alessandro Vinciarelli, Klaus Scherer, Fabien Ringeval, Mohamed Chetouani, Felix Weninger, Florian Eyben, Erik Marchi, Marcello Mortillaro, Hugues Salamin, Anna Polychroniou, Fabio Valente, and Samuel Kim, "The INTERSPEECH 2013 Computational Paralinguistics Challenge: Social Signals, Conflict, Emotion, Autism," in *Proc. INTERSPEECH*, Lyon, France, 2013, pp. 148–152.

[25] Shahin Amiriparian, Maurice Gerczuk, Sandra Ottl, Nicholas Cummins, Michael Freitag, Sergey Pugachevskiy, and Björn Schuller, "Snore Sound Classification Using Image-based Deep Spectrum Features," in *Proc. of INTERSPEECH*, Stockholm, Sweden, 2017, ISCA, 5 pages.

[26] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in Neural Information Processing Systems 25*, F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger, Eds., pp. 1097–1105. Curran Associates, Inc., 2012.

[27] Michael Freitag, Shahin Amiriparian, Sergey Pugachevskiy, Nicholas Cummins, and Björn Schuller, "audeep: Unsupervised learning of representations from audio with deep recurrent neural networks," *Journal of Machine Learning Research*, vol. 18, no. 173, pp. 1–5, 2018.

[28] Shahin Amiriparian, Michael Freitag, Nicholas Cummins, and Björn Schuller, "Sequence to sequence autoencoders for unsupervised representation learning from audio," in *Proc. DCASE*, 2017.

[29] Diederik Kingma and Jimmy Ba, "Adam: A method for stochastic optimization," in *Proceedings of the 3rd International Conference on Learning Representations*, Banff, CA, 2014.

[30] Nitish Srivastava, Geoffrey E Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov, "Dropout: A simple way to prevent neural networks from overfitting.," *Journal of Machine Learning Research*, vol. 15, no. 1, pp. 1929–1958, 2014.

[31] Mark Hall, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, and Ian H Witten, "The weka data mining software: an update," *ACM SIGKDD explorations newsletter*, vol. 11, no. 1, pp. 10–18, 2009.

[32] John C Platt, "12 fast training of support vector machines using sequential minimal optimization," *Advances in kernel methods*, pp. 185–208, 1999.

[33] Vinod Nair and Geoffrey E Hinton, "Rectified linear units improve restricted boltzmann machines," in *Proceedings of the 27th international conference on machine learning (ICML-10)*, 2010, pp. 807–814.