The ACM Multimedia 2022 Computational Paralinguistics Challenge: vocalisations, stuttering, activity, & mosquitoes

Björn Schuller, Anton Batliner, Shahin Amiriparian, Christian Bergler, Maurice Gerczuk, Natalie Holz, Pauline Larrouy-Maestri, Sebastien Bayerl, Korbinian Riedhammer, Adria Mallol-Ragolta, Maria Pateraki, Harry Coppock, Ivan Kiskin, Marianne Sinka, Stephen Roberts

Angaben zur Veröffentlichung / Publication details:

Schuller, Björn, Anton Batliner, Shahin Amiriparian, Christian Bergler, Maurice Gerczuk, Natalie Holz, Pauline Larrouy-Maestri, et al. 2022. "The ACM Multimedia 2022 Computational Paralinguistics Challenge: vocalisations, stuttering, activity, & mosquitoes." In *Proceedings of the 30th ACM International Conference on Multimedia, Lisboa, Portugal, October 10-14, 2022*, edited by João Magalhães, Alberto del Bimbo, Shin'ichi Satoh, Nicu Sebe, Xavier Alameda-Pineda, Qin Jin, Vincent Oria, and Laura Toni, 7120–24. New York, NY: ACM. https://doi.org/10.1145/3503161.3551591.

The ACM Multimedia 2022 Computational Paralinguistics Challenge: Vocalisations, Stuttering, Activity, & Mosquitoes

Björn W. Schuller
Imperial College London
London, United Kingdom

Anton Batliner
University of Augsburg
Augsburg, Germany

Shahin Amiriparian
University of Augsburg
Augsburg, Germany

Christian Bergler
FAU
Erlangen-Nuremberg, Germany

Maurice Gerczuk
University of Augsburg
Augsburg, Germany

Natalie Holz
MPI
Frankfurt, Germany

Pauline Larrouy-Maestri
MPI
Frankfurt, Germany

Sebastian P. Bayerl
TH Nürnberg
Nürnberg, Germany

Korbinian Riedhammer
TH Nürnberg
Nürnberg, Germany

Adria Mallol-Ragolta
University of Augsburg
Augsburg, Germany

Maria Pateraki
FORTH
Heraklion, Greece

Harry Coppock
Imperial College London
London, United Kingdom

Ivan Kiskin
University of Surrey
Guildford, United Kingdom

Marianne Sinka
University of Oxford
Oxford, United Kingdom

Stephen Roberts
University of Oxford
Oxford, United Kingdom

ABSTRACT

The ACM Multimedia 2022 Computational Paralinguistics Challenge addresses four different problems for the first time in a research competition under well-defined conditions: In the *Vocalisations* and *Stuttering* Sub-Challenges, a classification on human non-verbal vocalisations and speech has to be made; the *Activity* Sub-Challenge aims at beyond-audio human activity recognition from smartwatch sensor data; and in the *Mosquitoes* Sub-Challenge, mosquitoes need to be detected. We describe the Sub-Challenges, baseline feature extraction, and classifiers based on the 'usual' ComParE and BoAW features, the auDeep toolkit, and deep feature extraction from pre-trained CNNs using the DeepSpectrum toolkit; in addition, we add end-to-end sequential modelling, and a log-mel-128-BNN.

CCS CONCEPTS

• **Information systems** → **Multimedia and multimodal retrieval**; • **Computing methodologies** → **Artificial intelligence**.

KEYWORDS

Computational Paralinguistics; Vocalisations; Stuttering; Human Activity Recognition; Mosquito Detection; Challenge; Benchmark

ComParE '22, October 22, Lisbon, Portugal

ACM ISBN 978-1-4503-8678-4
https://doi.org/10.1145/3503161.3551591

ACM Reference Format:
Björn W. Schuller, Anton Batliner, Shahin Amiriparian, Christian Bergler, Maurice Gerczuk, Natalie Holz, Pauline Larrouy-Maestri, Sebastian P. Bayerl, Korbinian Riedhammer, Adria Mallol-Ragolta, Maria Pateraki, Harry Coppock, Ivan Kiskin, Marianne Sinka, and Stephen Roberts. 2022. The ACM Multimedia 2022 Computational Paralinguistics Challenge: Vocalisations, Stuttering, Activity, & Mosquitoes. In *Proceedings of the ACM Multimedia (MM '22), October, 2022, Lisboa, Portugal.* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3503161.3551591

1 INTRODUCTION

In this ACM Multimedia 2022 COMputational PARalinguistics challengE (ComParE) – the 14th since 2009 [28], we address four new problems within the field of Computational Paralinguistics [27] in a challenge setting:

In the **Vocalisations** Sub-Challenge, non-verbal vocal expressions from the Variably Intense Vocalizations of Affect and Emotion Corpus [12, 13] are used (**VOC-C**) for classifying the expression of six different emotions. Such human non-verbals are still understudied but are ubiquitous in human communication [24].

In the **Stuttering** Sub-Challenge, parts (**KSF-C**) of the Kassel State of Fluency corpus [6, 7] are used. Stuttering is a complex speech disorder with a crude prevalence of about 1 % of the population [34]. Monitoring of stuttering would allow objective feedback to persons who stutter (PWS) and speech therapists, thus facilitating tailored speech therapy, with the automatic detection of different stuttering phenomena as a necessary prerequisite.

The human activity recognition corpus harAGE as used in the **Activity** Sub-Challenge (**HAR-C**), provided by the EU Horizon 2020 project sustAGE [19], is a multimodal dataset collected with the smartwatch Garmin Vivoactive 3 [17, 18]. The monitoring of

different types of physical activity vs inactivity is of vital importance to promote healthier and active life styles in the population, improving their overall physical health and wellbeing [10, 23].

The Mosquito corpus as used in the **Mosquitoes** Sub-Challenge (**MOS-C**), provided by the HumBug Project, is a large-scale audio database consisting of over 20 hours of mosquito flight recordings (HumBugDB [15]). Mosquitoes are responsible for more human deaths than any other creature; e. g., in 2020 malaria caused around 241 million cases of disease across more than 100 countries resulting in an estimated 627 000 deaths [21]. It is imperative therefore to accurately locate and identify dangerous mosquitoes to achieve efficient mosquito control.

For all tasks, a target class has to be predicted for each case. Contributors can employ their own features and machine learning (ML) algorithms; standard feature sets and procedures are provided. Participants have to use the pre-defined partitions for each Sub-Challenge. They may report results that they obtain from the **Train**(ing)/**Dev**(elopment) set but have only five trials to upload their results on the **Test** set per Sub-Challenge, whose labels are unknown to them. Each participation must be accompanied by a paper presenting the results, which undergoes peer-review. The organisers preserve the right to re-evaluate the findings, but will not participate in the Challenge. As evaluation measure, we employ for all Sub-Challenges but Mosquitoes the **Unweighted Average Recall (UAR)** as used since the first Challenge from 2009 [28, 29]; it is more adequate for (unbalanced) multi-class classifications than Weighted Average Recall (i. e., accuracy) [27]. The Mosquitoes Sub-Challenge is an audio event detection task; hence, we utilise the **Polyphonic Sound Event Detection Score (PSDS)** [8] – an extension for a classifier threshold-independent event-based F-score. Ethical approval for the studies has been obtained.

2 THE FOUR SUB-CHALLENGES

Vocalisations – The Vocalisation Corpus VOC-C:
It is provided by the MPI for Empirical Aesthetics, Frankfurt am Main, featuring vocalisations – such as laughter, cries, moans, or screams – with different affective intensities, expressing different emotional states. The data from the female speakers have been made available to the public, see [12, 13]; the male speakers are so far unseen. We partition the female vocalisations into Train (6 speakers, 625 samples) and Dev(elopment) (5 speakers, 460 samples), and the male vocalisations (2 speakers, 276 samples) into Test, modelling a 6-class problem with the emotional classes *achievement, anger, fear, pain, pleasure,* and *surprise*.

Stuttering – The Kassel State of Fluency Corpus KSF-C:
The corpus provided by the TH Nürnberg and the Kasseler Stottertherapie is derived from the Kassel State of Fluency (KSoF) corpus [6, 7]. The original corpus features 5 597 typical and nontypical (stuttering) 3 s segments from 37 German speakers with an overall duration of 4.6 h. The segments were annotated by three labellers as one of 7 classes (*block, prolongation, sound repetition, word/phrase repetition, modified speech technique, interjection, no disfluency*) and with some additional information, e. g., about the recording quality. Annotators were able to assign more than one label per segment. For this challenge, we removed all the ambiguously labelled segments, thus only featuring 4 601 segments. The task proposed in this challenge is the classification of speech segments as one of 8 classes – the seven stuttering-related classes and an eighth *garbage* class, denoting segments that are unintelligible, contain no speech, or are negatively affected by background noise. The dataset is split into three speaker-independent partitions (Train: 23 speakers, Dev: 6 speakers, Test: 8 speakers).

Activity – The Human Activity Recognition Corpus HAR-C:
The harAGE corpus[1] [17, 18] contains 17 h 37 m 20 s of triaxial accelerometer, heart rate, and pedometer sensor measurements from 30 (14 f, 16 m) participants with a mean age of 40.0 years and a standard deviation of 8.3 years. Sensor measurements from eight activities are included: *lying*, *sitting*, *standing*, *washing hands*, *walking*, *running*, *stairs climbing*, and *cycling*. The dataset is split into three participant-independent and gender-balanced partitions. The Train, Dev, and Test partitions contain a total of 10 h 41 m 20 s, 2 h 16 m 0 s, and 4 h 40 m 0 s of data from 17 (8 f, 9 m), 6 (3 f, 3 m), and 7 (3 f, 4 m) participants, respectively. Each sample in the harAGE corpus contains 20 s of continuous sensor measurements from one participant performing one of the different activities considered in the dataset. The task in this Sub-challenge consists in the development of unimodal and/or multimodal systems able to analyse these 20 s of sensor measurements and infer the corresponding activity.

Mosquitoes – The Mosquito Corpus MOS-C:
It is provided by the HumBug Project[2] and is strongly based on HumBugDB [15]. In a revision for this Sub-challenge[3], the former test set A is expanded with more challenging negatives and now forms Dev A. The former test set B forms Dev B, and the training set is identical. The task is to detect *timestamps for acoustic mosquito events* – Mosquito Event Detection (MED). The challenge is therefore scored in the time domain with the PSDS package [8]. Details of Train and Dev are given in [15, Sec. 4]. To summarise, Dev A represents semi-field conditions, where mosquitoes were manually released near recording setups that feature traditional housing constructions, equipped with mosquito bednets [33, Sec. 2.1.2]. Dev B is a low-SNR recording set of free-flying mosquitoes within culture cages. The data vary in sample rate, recording devices, ambient conditions, and experimental assumptions. As these factors can introduce confounding, they are given as metadata and documented in [15, Appx. C]. The test set consists of recordings conducted in South East Tanzania by volunteers in people's homes. It is therefore not included in the hosted Zenodo dataset due to the sensitive nature of the data. Please note that participants will not receive the test data but will submit dockerised versions of their code using the help of the provided templates for either Tensorflow 2.0 [1], or PyTorch [22], that participants are free to choose as they wish.

3 EXPERIMENTS AND RESULTS

For the VOC-C and the KSF-C, the segmented audio was converted to single-channel 16 kHz, 16 bits PCM format. Table 1 shows the number of data for Train, Dev, and Test for the different corpora. MOS-C Dev was split into two sets with differing conditions.

[1] https://zenodo.org/record/6517688
[2] The full list of authors contributing to HumBugDB is in [15] and associated Zenodo repository
[3] `v0.0.2 HumBugDB:` https://zenodo.org/record/6478589

Table 1: Summary of the databases presented per Sub-Challenge. Number of instances per class in the Train/Dev/Test splits. The test split distributions are blinded during the ongoing challenge and will be given in the final version.

VOC-C: classification task (#)					**KSF-C**: classification task (#)					**HAR-C**: classification task (#)					**MOS-C**: detection (in hours)				
Class	Train	Dev	Test	Σ	Class	Train	Dev	Test	Σ	Class	Train	Dev	Test	Σ	Class	Train	Dev A/B	Test	Σ
achiev.	89	72	–	–	Block	310	102	–	–	lying	257	61	–	–	mosquito	17.0	1.1/0.25	–	–
anger	101	73	–	–	Fillers	205	104	–	–	sitting	238	57	–	–	non-mosquito	13.4	2.7/0.56	–	–
fear	103	73	–	–	Garbage	52	33	–	–	standing	244	57	–	–					
pain	114	71	–	–	Modified	687	185	–	–	wash. hands	133	35	–	–					
pleasure	109	93	–	–	Prolong.	183	53	–	–	walking	302	57	–	–					
surprise	109	78	–	–	SoundRep.	169	38	–	–	running	301	40	–	–					
					WordRep.	53	23	–	–	stairs climb.	263	43	–	–					
					no_disfl.	830	444	–	–	cycling	186	58	–	–					
Σ	625	460	276	1 361	Σ	2 489	982	1 130	4 601	Σ	1 924	408	840	3 172	Σ	30.4	3.8/0.81	18	53.01

Table 2: Results for the Sub-Challenges. The official baselines for Test are highlighted (bold and greyscale); there are no official baselines for Dev. UAR: Unweighted Average Recall. CI on Test: Confidence Intervals on Test, see explanation in the text.

%	Vocalisations: **VOC-C**: UAR			Stuttering: **KSF-C**: UAR			Activity: **HAR-C**: UAR				Mosquitoes: **MOS-C**: PSDS			
Approach	Dev	Test	CI on Test	Dev	Test	CI on Test	Approach	Dev	Test	CI on Test	Approach	Dev A/B	Test	CI on Test
ComParE	39.8	32.7	27.7 – 38.0	30.2	37.6	33.5 – 41.4	HR	34.4	30.2	28.0 – 32.3	mel-BNN	61.4/3.4	**6.4**	6.0 – 6.9
DeepSpectrum	35.0	34.1	29.5 – 39.2	28.1	**40.4**	36.4 – 44.2	Steps	36.4	32.1	30.6 – 33.7				
auDeep	31.0	31.2	26.1 – 36.6	17.7	25.9	21.9 – 30.3	XYZ	65.8	69.3	66.6 – 72.2				
BoAWs	39.6	**37.4**	32.6 – 42.8	26.7	32.1	28.2 – 36.0	HR⊕Steps	52.2	42.1	39.6 – 44.5				
Fusion	39.8	36.1	31.3 – 31.3	28.7	38.3	34.3 – 41.9	HR⊕XYZ	74.5	63.9	61.0 – 66.9				
							Steps⊕XYZ	66.6	65.5	62.5 – 68.4				
							HR⊕Steps⊕XYZ	77.7	**72.2**	69.2 – 75.0				

3.1 Approaches

ComParE Acoustic Feature Set: The official baseline feature set from openSMILE is the same as has been used in previous editions of the ComParE challenges, starting from 2013 [30]. It is described in [9, 30].

DeepSpectrum:[4] It is applied to obtain deep representations from the input audio data utilising image pre-trained Convolutional Neural Networks (CNNs) [3]. It has been used in previous challenges [31, 32] and is described in [3]. A lightweight version of DeepSpectrum for audio signal processing on-device can be found in [5][5].

auDeep:[6] This feature set is obtained through unsupervised representation learning with recurrent sequence-to-sequence autoencoders [2, 11]; it has as well been employed in previous challenges [31, 32]. Learnt representations of a spectrogram are extracted and then concatenated to obtain a final feature vector.

Bag-of-Audio-Words (BoAWs): Audio chunks are represented as histograms of ComParE LLDs, after quantisation based on a codebook. They have been used in previous challenges [31, 32] and other studies [4, 16, 25]; the toolkit openXBOW is described in [26].

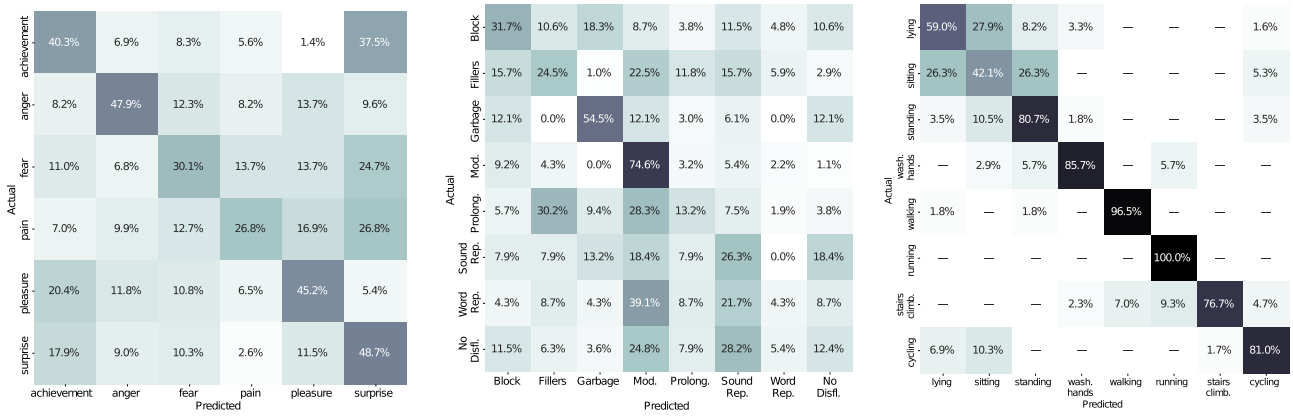
End-to-end sequential modelling: The HAR-C implements an end-to-end approach exploiting the sensor data as input. As described in [18], 3-dimensional, 2-dimensional, and 9-dimensional traces are generated from the raw heart rate, pedometer, and triaxial accelerometer measurements, respectively. As opposed to [18], herein, we do not debias the accelerometer measurements to ease the deployment of the presented approach in real-life applications. The network implemented is composed of a dedicated feature extraction block for each modality – responsible for extracting deep learnt representations from the input traces – followed by a classification block – in charge of performing the actual inference. The feature extraction block implements a 1-dimensional convolutional layer, and the classification layer two fully connected layers. The dimensionality of the resulting features at the output of the feature extraction block depends on the number of modalities to be fused, concatenating the embedded representations learnt separately from each modality.

log-mel-128-BNN: MOS-C utilises a Bayesian Convolutional Neural Network with four convolutional, two max-pooling, and one fully connected layer augmented with dropout layers [15, Appx. B.4]. Its structure is based on prior models that have been successful in assisting domain experts for mosquito tagging [14]. As features, the baseline uses 128 log-mel spectrogram coefficients with a time window of 30 feature frames and a stride of 5 frames for training. Each frame spans 64 ms, forming a single training example $\mathbf{X}_i \in \mathbb{R}^{128\times30}$ with a temporal window of 1.92 s. Dev and Test events may, however, be scored on shorter time windows.

3.2 Challenge Baselines and Interpretation

We provide a branch on the official challenge repository[7] for each Sub-Challenge, which includes scripts allowing participants to fully reproduce the baselines (including pre-processing, model training, and model evaluation on Dev). For VOC-C, KSF-C, and HAR-C, the 95 % Confidence Intervals (CI) were computed by 1 000x bootstrapping (random sampling with replacement) and UARs for Test, based on the same model that was trained with Train and Dev. For

[4] https://github.com/DeepSpectrum/DeepSpectrum
[5] https://github.com/DeepSpectrum/DeepSpectrumLite
[6] https://github.com/auDeep/auDeep
[7] https://github.com/EIHW/ComParE2022



(a) VOC-C Confusion Matrix **(b) KSF-C Confusion Matrix** **(c) HAR-C Confusion Matrix**

Figure 1: Confusion matrices for VOC-C, KSF-C, and HAR-C on Dev. The individual approach/hyperparameters performing best on Dev (without fusion) are chosen; see Table 2. In the cells, the percent of 'classified as' of the class displayed in the respective row are given; percentage also indicated by colour-scale: the darker, the higher. Cases per class are given in Table 1.

MOS-C, as appropriate for sound event detection tasks, the 95 % CI intervals are constructed using the jackknife method (leave-one-out sampling) [20] with the number of samples equal to the number of test audio recordings. Due to space restrictions, for VOC-C and KSF-C, we leave out the results for every hyperparameter configuration evaluated and only provide the best results obtained. The baselines for both VOC-C and KSF-C consist of using Support Vector Machines with linear kernels on four different audio feature representations – ComParE, DeepSpectrum, auDeep, and BoAWs. All feature representations are scaled to zero mean and unit standard deviation, using the parameters from the respective training set (when Train and Dev are fused for the final classifier, the parameters are calculated on this fusion). The SVM complexity parameter C is always optimised during the development phase.

Vocalisations – VOC-C: We obtain the best **UAR=37.4 %** on Test with BoAWs, see Table 2. Figure 1(a) shows, for the best Dev result given in Table 2, that the two classes *pain* and *fear* fall behind the other four classes, and that they are mostly confused with surprise.

Stuttering – KSF-C: We achieve **UAR=40.4 %** on Test with DeepSpectrum. Looking at the confusion matrix of our best result on Dev in Figure 1(b), word repetitions seem to be the hardest to detect and differentiate, especially from instances of modified speech and sound repetitions.

Activity – HAR-C: The best approach on Test fuses the heart rate, the pedometer, and the accelerometer modalities, scoring a **UAR=72.2 %**. The results highlight the importance of the accelerometer information for this task, as the models exploiting this modality outperform those using the heart rate and the pedometer information, either unimodally or multimodally. Analysing the confusion matrix given in Figure 1(c) for the best result on Dev in Table 2, we observe that the main confusions take place among the 'non-moving' activities lying, sitting, and standing.

Mosquitoes – MOS-C: Table 2 shows the PSDS of 61.4 % and 3.4 % achieved on Dev A and B, respectively. We note that the provided model is unable to achieve a good score on Dev B, which features a lower SNR, more challenging dataset. The baseline scores **6.4 % PSDS** on the Test partition, which can be thought of as an approximate combination in recording conditions of Dev A and B. The results highlight the need to train a model that is able to perform well and generalise across different deployment scenarios. Each of the Dev A, Dev B, and Test sets features considerably different audio backgrounds, as they are recorded in different environments. Additional *feature window-based* metrics are supplied in the repository, which give a breakdown of performance by precision-recall, ROC, and confusion matrices. These may be helpful for developing with the ultimate aim of maximising the PSDS on Test.

4 CONCLUDING REMARKS

This year's challenge is new by four new tasks, all of them highly relevant for applications. We feature our 'classic' approaches **ComParE** and **Bag-of-Audio-Words (BoAWs)**, **auDeep**, and **DeepSpectrum** for VOC-C and KSF-C, and two new ones, tailored for HAR-C and MOS-C. For all computation steps, scripts are provided that can, but need not be used by the participants. We expect participants to obtain better performance measures by employing novel (combinations of) procedures and features, including such tailored to the particular tasks.

5 ACKNOWLEDGMENTS

We acknowledge funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No. 826506 (sustAGE), from the Deutsche Forschungsgemeinschaft (DFG) under grant agreement No. 421613952 (ParaStiChaD), from the DFG's Reinhart Koselleck project No. 442218748 (AUDI0NOMOUS), and from the Gates Foundation No. opp1209888, as well as the contributions of all authors in [15, HumBugDB] and MOS-C Zenodo repository.

REFERENCES

[1] Martín Abadi, Ashish Agarwal, Paul Barham, Eugene Brevdo, Zhifeng Chen, Craig Citro, Greg S. Corrado, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Ian Goodfellow, Andrew Harp, Geoffrey Irving, Michael Isard, Yangqing Jia, Rafal Jozefowicz, Lukasz Kaiser, Manjunath Kudlur, Josh Levenberg, Dandelion Mané, Rajat Monga, Sherry Moore, Derek Murray, Chris Olah, Mike Schuster, Jonathon Shlens, Benoit Steiner, Ilya Sutskever, Kunal Talwar, Paul Tucker, Vincent Vanhoucke, Vijay Vasudevan, Fernanda Viégas, Oriol Vinyals, Pete Warden, Martin Wattenberg, Martin Wicke, Yuan Yu, and Xiaoqiang Zheng. 2016. TensorFlow: Large-Scale Machine Learning on Heterogeneous Systems. *arXiv* (2016).

[2] Shahin Amiriparian, Michael Freitag, Nicholas Cummins, and Björn Schuller. 2017. Sequence to Sequence Autoencoders for Unsupervised Representation Learning from Audio. In *Proc. DCASE 2017*. Munich, Germany, 17–21.

[3] Shahin Amiriparian, Maurice Gerczuk, Sandra Ottl, Nicholas Cummins, Michael Freitag, Sergey Pugachevskiy, and Björn Schuller. 2017. Snore Sound Classification Using Image-based Deep Spectrum Features. In *Proc. Interspeech 2017*. ISCA, Stockholm, Sweden, 3512–3516.

[4] Shahin Amiriparian, Maurice Gerczuk, Sandra Ottl, Nicholas Cummins, Sergey Pugachevskiy, and Björn Schuller. 2018. Bag-of-Deep-Features: Noise-Robust Deep Feature Representations for Audio Analysis. In *Proc. IJCNN*. IEEE, Rio de Janeiro, Brazil, 2419–2425.

[5] Shahin Amiriparian, Tobias Hübner, Vincent Karas, Maurice Gerczuk, Sandra Ottl, and Björn W. Schuller. 2022. DeepSpectrumLite: A Power-Efficient Transfer Learning Framework for Embedded Speech and Audio Processing From Decentralized Data. *Frontiers in Artificial Intelligence* 5 (2022).

[6] Sebastian P. Bayerl, Florian Hönig, Joëlle Reister, and Korbinian Riedhammer. 2020. Towards Automated Assessment of Stuttering and Stuttering Therapy. In *Proc. TSD*. Brno, Czech Republic, 386–396.

[7] Sebastian P. Bayerl, Alexander Wolff von Gudenberg, Florian Hönig, Elmar Nöth, and Korbinian Riedhammer. 2022. KSoF: The Kassel State of Fluency Dataset - A Therapy Centered Dataset of Stuttering. In *Proc. LREC*. Marseille, France.

[8] Çağdaş Bilen, Giacomo Ferroni, Francesco Tuveri, Juan Azcarreta, and Sacha Krstulović. 2020. A framework for the robust evaluation of sound event detection. In *Proc. ICASSP*. IEEE, Barcelona, Spain, 61–65.

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

[10] Kenneth R Fox. 1999. The influence of physical activity on mental well-being. *Public Health Nutrition* 2, 3a (1999), 411–418.

[11] Michael Freitag, Shahin Amiriparian, Sergey Pugachevskiy, Nicholas Cummins, and Björn Schuller. 2018. auDeep: Unsupervised Learning of Representations from Audio with Deep Recurrent Neural Networks. *Journal of Machine Learning Research* 18 (2018), 1–5.

[12] Natalie Holz, Pauline Larrouy-Maestri, and David Poeppel. 2021. The paradoxical role of emotional intensity in the perception of vocal affect. *Scientific reports* 11, 1 (2021), 1–10.

[13] Natalie Holz, Pauline Larrouy-Maestri, and David Poeppel. 2022. The Variably Intense Vocalizations of Affect and Emotion (VIVAE) Corpus prompts new perspective on nonspeech perception. *Emotion* 22, 1 (2022), 213–225.

[14] Ivan Kiskin, Adam D Cobb, Marianne Sinka, Kathy Willis, and Stephen J Roberts. 2021. Automatic Acoustic Mosquito Tagging with Bayesian Neural Networks. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*. Springer, 351–366.

[15] I. Kiskin, M. Sinka, A.D. Cobb, W. Rafique, L. Wang, D. Zilli, B. Gutteridge, R. Dam, T. Marinos, Y. Li, and D. Msaky. 2021. HumBugDB: A Large-scale Acoustic Mosquito Dataset. In *Proc. NeurIPS Track on Datasets and Benchmarks*. New Orleans, USA, 1–13.

[16] Hyungjun Lim, Myung Jong Kim, and Hoirin Kim. 2015. Robust Sound Event Classification Using LBP-HOG Based Bag-of-Audio-Words Feature Representation. In *Proc. Interspeech*. ISCA, Dresden, Germany, 3325–3329.

[17] A. Mallol-Ragolta, A. Semertzidou, M. Pateraki, and B. Schuller. 2021. harAGE: A Novel Multimodal Smartwatch-based Dataset for Human Activity Recognition. In *Proc. FG*. IEEE, Jodhpur, India – Virtual Event, 1–7.

[18] A. Mallol-Ragolta, A. Semertzidou, M. Pateraki, and B. Schuller. 2022. Outer Product-Based Fusion of Smartwatch Sensor Data for Human Activity Recognition. *Frontiers in Computer Science, section Mobile and Ubiquitous Computing* 4 (2022), 1–10. Article ID 796866.

[19] Adria Mallol-Ragolta, Iraklis Varlamis, Maria Pateraki, Manolis Lourakis, Georgios Athanassiou, Michail Maniadakis, Konstantinos Papoutsakis, Thodoris Papadopoulos, Anastasia Semertzidou, Nicholas Cummins, Björn Schuller, Ion-Anastasios Karolos, Christos Pikridas, Petros Patias, Spyros Vantolas, Leonidas Kallipolitis, Frank Werner, Antonio Ascolese, and Vito Nitti. 2022. sustAGE 1.0 – First Prototype, Use Cases, and Usability Evaluation. In *Proc. 7th International Conference on Human Interaction & Emerging Technologies: Artificial Intelligence & Future Applications*. Springer, Lausanne, Switzerland – Virtual Event. 10 pages, to appear.

[20] Annamaria Mesaros, Aleksandr Diment, Benjamin Elizalde, Toni Heittola, Emmanuel Vincent, Bhiksha Raj, and Tuomas Virtanen. 2019. Sound Event Detection in the DCASE 2017 Challenge. *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 27, 6 (2019), 992–1006.

[21] World Health Organization et al. 2021. World malaria report 2021. (2021).

[22] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. 2019. PyTorch: An Imperative Style, High-Performance Deep Learning Library. In *Advances in Neural Information Processing Systems 32*, H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett (Eds.). Curran Associates, Inc., 8024–8035.

[23] Frank J Penedo and Jason R Dahn. 2005. Exercise and well-being: a review of mental and physical health benefits associated with physical activity. *Current Opinion in Psychiatry* 18, 2 (2005), 189–193.

[24] Katarzyna Pisanski, Gregory A Bryant, Clément Cornec, Andrey Anikin, and David Reby. 2022. Form follows function in human nonverbal vocalisations. *Ethology Ecology & Evolution* (2022), 1–19.

[25] Maximilian Schmitt, Fabien Ringeval, and Björn Schuller. 2016. At the Border of Acoustics and Linguistics: Bag-of-Audio-Words for the Recognition of Emotions in Speech. In *Proc. Interspeech*. ISCA, San Francisco, USA, 495–499.

[26] M. Schmitt and B. W. Schuller. 2017. openXBOW – Introducing the Passau Open-Source Crossmodal Bag-of-Words Toolkit. *Journal of Machine Learning Research* 18 (2017), 1–5.

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

[28] Björn Schuller, Anton Batliner, Stefan Steidl, and Dino Seppi. 2011. Recognising Realistic Emotions and Affect in Speech: State of the Art and Lessons Learnt from the First Challenge. *Speech Communication* 53 (2011), 1062–1087.

[29] B. Schuller, S. Steidl, and A. Batliner. 2009. The INTERSPEECH 2009 Emotion Challenge. In *Proc. Interspeech*. ISCA, Brighton, UK, 312–315.

[30] 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. 2013. The Interspeech 2013 Computational Paralinguistics Challenge: Social Signals, Conflict, Emotion, Autism. In *Proc. Interspeech*. ISCA, Lyon, France, 148–152.

[31] Björn W. Schuller, Anton Batliner, Christian Bergler, Cecilia Mascolo, Jing Han, Iulia Lefter, Heysem Kaya, Shahin Amiriparian, Alice Baird, Lukas Stappen, Sandra Ottl, Maurice Gerczuk, Panagiotis Tzirakis, Chloë Brown, Jagmohan Chauhan, Andreas Grammenos, Apinan Hasthanasombat, Dimitris Spathis, Tong Xia, Pietro Cicuta, Leon J. M. Rothkrantz, Joeri Zwerts, Jelle Treep, and Casper Kaandorp. 2021. The INTERSPEECH 2021 Computational Paralinguistics Challenge: COVID-19 Cough, COVID-19 Speech, Escalation & Primates. In *Proc. Interspeech*. ISCA, Brno, Czechia, 431–435.

[32] Björn W Schuller, Anton Batliner, Christian Bergler, Eva-Maria Messner, Antonia Hamilton, Shahin Amiriparian, Alice Baird, Georgios Rizos, Maximilian Schmitt, Lukas Stappen, et al. 2020. The INTERSPEECH 2020 Computational Paralinguistics Challenge: Elderly Emotion, Breathing & Masks. In *Proc. Interspeech*. ISCA, Shanghai, China, 2042–2046.

[33] Marianne E Sinka, Davide Zilli, Yunpeng Li, Ivan Kiskin, Daniel Kirkham, Waqas Rafique, Lawrence Wang, Henry Chan, Benjamin Gutteridge, Eva Herreros-Moya, et al. 2021. HumBug–An Acoustic Mosquito Monitoring Tool for use on budget smartphones. *Methods in Ecology and Evolution* 12, 10 (2021), 1848–1859.

[34] Martin Sommer, Andrea Waltersbacher, Andreas Schlotmann, Helmut Schröder, and Adam Strzelczyk. 2021. Prevalence and Therapy Rates for Stuttering, Cluttering, and Developmental Disorders of Speech and Language: Evaluation of German Health Insurance Data. *Frontiers in Human Neuroscience* 15 (2021).