INTERSPEECH 2021
30 August – 3 September, 2021, Brno, Czechia

The INTERSPEECH 2021 Computational Paralinguistics Challenge: COVID-19 Cough, COVID-19 Speech, Escalation & Primates

Björn W. Schuller[1,2], Anton Batliner[2,3], Christian Bergler[3], Cecilia Mascolo[4], Jing Han[4], Iulia Lefter[5], Heysem Kaya[6], Shahin Amiriparian[2], Alice Baird[2], Lukas Stappen[2], Sandra Ottl[2], Maurice Gerczuk[2], Panagiotis Tzirakis[1], Chloë Brown[4], Jagmohan Chauhan[4], Andreas Grammenos[4], Apinan Hasthanasombat[4], Dimitris Spathis[4], Tong Xia[4], Pietro Cicuta[4], Leon J. M. Rothkrantz[5], Joeri A. Zwerts[6], Jelle Treep[6], Casper Kaandorp[6]

[1]GLAM – Group on Language, Audio & Music, Imperial College London, UK
[2]EIHW – Chair of Embedded Intelligence for Health Care and Wellbeing, University of Augsburg, Germany
[3]Pattern Recognition Lab, FAU Erlangen-Nuremberg, Germany
[4]University of Cambridge, UK
[5]Delft University of Technology, The Netherlands
[6]Faculty of Science, Utrecht University, The Netherlands

schuller@IEEE.org

Abstract

The INTERSPEECH 2021 Computational Paralinguistics Challenge addresses four different problems for the first time in a research competition under well-defined conditions: In the *COVID-19 Cough* and *COVID-19 Speech* Sub-Challenges, a binary classification on COVID-19 infection has to be made based on coughing sounds and speech; in the *Escalation* Sub-Challenge, a three-way assessment of the level of escalation in a dialogue is featured; and in the *Primates* Sub-Challenge, four species vs background need to be classified. We describe the Sub-Challenges, baseline feature extraction, and classifiers based on the 'usual' ComParE and BoAW features as well as deep unsupervised representation learning using the auDeep toolkit, and deep feature extraction from pre-trained CNNs using the Deep Spectrum toolkit; in addition, we add deep end-to-end sequential modelling, and partially linguistic analysis.

Index Terms: Computational Paralinguistics, Challenge, COVID-19, Escalation, Primates

1. Introduction

In this INTERSPEECH 2021 COMputational PARAlinguistics challengE (ComParE) – the thirteenth since 2009 [1], we address four new problems within the field of Computational Paralinguistics [2] in a challenge setting:

In the **COVID-19 Cough** Sub-Challenge (**CCS**) and **COVID-19 Speech** Sub-Challenge (**CSS**), coughing sounds or speech are used to binary classify COVID-19 (or not) infection. In the present pandemic situation, great potential lies in low-cost, anywhere and anytime accessible real-time pre-diagnosis of COVID-19 infection. To date, the possibility has been shown [3], yet a controlled challenge test-bed is lacking. In the **Escalation** Sub-Challenge (**ESS**), participants are faced with three-way classification of the level of escalation in human dialogues. A range of applications exists including human-to-computer interaction, computer mediated human-to-human conversation, or public security. Finally, in the **Primate** Sub-Challenge (**PRS**), we classify four species of primates versus background noise. Real-life applications include wild-life monitoring in habitats, e. g., to save species from extinction.

For all tasks, a target 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. Participants have to use the pre-defined partitions for each Sub-Challenge. They may report results obtained from the Train(ing)/Dev(elopment) set – preferably with the supplied evaluation setups, 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 and has to be accepted for the conference in order to participate. The organisers preserve the right to re-evaluate the findings, but will not participate in the Challenge. As evaluation measure, we employ in all Sub-Challenges **Unweighted Average Recall (UAR)** as used since the first Challenge from 2009 [1], especially because it is more adequate for (unbalanced) multi-class classifications than Weighted Average Recall (i. e., accuracy) [2, 4]. Ethical approval for the studies has been obtained from the pertinent committees. In section 2, we describe the challenge corpora. Section 3 details baseline experiments, metrics, and baseline results; concluding remarks are given in section 4.

2. The Four Sub-Challenges

2.1. The COVID-19 Cough Sub-Challenge (CCS) and the COVID-19 Speech Sub-Challenge (CSS)

For the **CCS** and **CSS**, we employ two subsets from the Cambridge COVID-19 Sound database [5, 6]. The database was collected via the COVID-19 Sounds App since its launch in April 2020, aiming at collecting data to inform the diagnosis of COVID-19 based primarily on voice, breathing, and coughing. Participants were able to provide audio samples together with their COVID-19 test results via multiple platforms (a webpage, an Android app, and an iOS app). The participants also provided basic demographic, medical information, and reported symptoms. For the **CCS** and the **CSS**, only cough sounds and voice recordings with COVID-19 positive/ negative test results were included separately, and only audio data and the corresponding COVID-19 test labels are provided. The quality of these data was manually checked. As they were crowd-sourced, the original audio data had varying sampling rates and formats; all of them were resampled (in a few cases, upsampled) and converted to 16 kHz and mono/16 bit, and further normalised recording-wise

http://dx.doi.org/10.21437/Interspeech.2021-19

to eliminate varying loudness. For the **CCS**, 725 recordings from 343 participants were provided, in total 1.63 hrs. In each cough recording, the participant provided one to three forced coughs. For the **CSS**, we use 893 recordings from 366 participants. in total 3.24 hrs. In each speech recording, the participant recorded speech content ("I hope my data can help to manage the virus pandemic.") in one language (English, Italian, or German, etc), one to three times. For each recording, a COVID-19 test result was available which was self-reported by the participant. To create the two-class classification task, the original COVID-19 test results were mapped onto either positive (denoted as 'P') or negative ('N'). Note that Train, Dev, and Test sets contain mutually different speakers; within each set, however, speakers can occur more than once; thus, it is essential to stick to the partitioning provided.

2.2. The Escalation Sub-Challenge (ESS)

For the **ESS**, the INTERSPEECH COMPARE Escalation Corpus is provided, consisting of the Dataset of Aggression in Trains (TR) [7] and the Stress at Service Desk Dataset (SD) [8]. Both present unscripted interactions between actors, where friction appears as they spontaneously react to each other based on short scenario descriptions. While the datasets share the same procedure for eliciting interactions, the topics, the number of participants in the scene, and amount of overlapping speech, as well as the recording quality differ. The TR dataset consists of 21 scenarios of unwanted behaviours in trains and train stations (e. g., harassment, theft, travelling without a ticket) played by 13 subjects. It was annotated based on aggression levels on a 5 point scale by 7 raters (Krippendorff's alpha $= 0.77$). Here, the annotation based on audio footage is used. The SD dataset contains scenarios of problematic interactions situated at a service desk (e. g., a slow and incompetent employee while the customer has an urgent request). It contains 8 subjects and the recordings were annotated for stress levels on a 5 point scale by 4 raters (Krippendorff's alpha $= 0.74$), based on audio-visual footage. All original labels were mapped onto a 3 point scale: SD classes 1 and 2 and TR class 1 onto **L**ow, SD class 3 and TR class 2 onto **M**edium, and the rest of the data onto **H**igh escalation. The language spoken in the Escalation Corpus is Dutch (two scenarios from SD where English was spoken were excluded). Manual transcriptions are provided. The corpus has been re-segmented based on linguistic information, resulting in 413 and 501 (test) segments, of an average length of 5 seconds. The challenge task is to use the SD dataset for training, and to recognise escalation levels in the TR dataset.

2.3. The Primates Sub-Challenge (PRS)

For the **PRS**, the Primate Vocalisations Corpus described in Zwerts et al. [9] is used. The global biodiversity crisis calls for effective monitoring methods to measure, manage and conserve wildlife. Using acoustic recordings is a non-invasive and potentially cost-effective way to identify and count species for environments like tropical forests, where opportunities for visual monitoring are limited. Several studies have applied automatic acoustic monitoring for a variety of taxa, ranging from birds [10] to forest elephants [11], and sporadically also for primates [12, 13, 14]. Zwerts et al. [9] recently collected acoustic data from a primate sanctuary in Cameroon. The recorded species were **C**himpanzees (*Pan troglodytes*), **M**andrills (*Mandrillus sphinx*), **R**ed-capped mangabeys (*Cercocebus torquatus*) and a mixed group of **G**uenons (*Cercopithecus spp.*). The sanctuary houses primates under semi-natural conditions making background noise relatively comparable to natural forests, albeit less rich in biodiversity and also containing human related noise. Recordings were made between December 2019 and January 2020 with a timespan of 32 days, using Audiomoth (v1.1.0) recorders [15], mounted either on the fence or nearby the respective species' enclosure, with 48 kHz sampling rate and 30.6 dB gain, yielding 358 GBs of acoustic data, with a total duration of 1 112 hours [9]. A semi-automatic annotation process speeded up the manual annotation efforts, with 1) initial annotation based on spectrogram analysis and listening, 2) vocalisation detection based on energy/variation in certain frequency sub-bands (150 Hz - 2 KHz), and 3) final annotation based on spectrogram analysis and listening, yielding over 10 k annotated vocalisations. For the background class, the recordings not annotated as vocalisation were sampled so as to exactly match the duration distribution of the annotated chunks of each species [9].

Table 1: *Databases: Number of instances per class in the Train/Dev/Test splits: Test split distributions are blinded during the ongoing challenge and will be given in the final version.*

#	Train	Dev	Test	Σ
CCS: COVID-19 COUGH (C19C) corpus				
no COVID-19	215	183	169	567
COVID-19	71	48	39	158
Σ	286	231	208	725
CSS: COVID-19 SPEECH (C19S) corpus				
no COVID-19	243	153	189	585
COVID-19	72	142	94	308
Σ	315	295	283	893
ESS: Escalation at Service-desks and in Trains (CEST)				
L	156	69	260	485
M	75	34	191	300
H	64	16	50	130
Σ	295	119	501	915
PRS: Primate Vocalisations Corpus (PVC)				
C	2 217	2 217	2 218	6 652
M	874	874	875	2 623
R	208	209	210	627
G	158	159	159	476
Background	3 458	3 459	3 461	10 378
Σ	6 915	6 918	6 923	20 756

3. Experiments and Results

For all corpora, the segmented audio was converted to single-channel 16 kHz, 16 bits PCM format. Table 1 shows the number of cases for Train, Dev, and Test for the databases; partitions for CCS, CSS, and ESS were gender-balanced.

3.1. Approaches

COMPARE Acoustic Feature Set: The official baseline feature set is the same as has been used in the eight previous editions of the COMPARE challenges, starting from 2013 [16]. It contains 6 373 static features resulting from the computation of functionals (statistics) over low-level descriptor (LLD) contours [17, 16]. A full description of the feature set can be found in [18].

Bag-of-Audio-Words (BoAWs): These have been applied successfully for, e. g., acoustic event detection [19] and speech-

Table 2: *Results for the four Sub-Challenges. The* ***official baselines*** *for Test are highlighted (bold and greyscale); there are* ***no*** *official baselines for Dev. C: Complexity parameter of the SVM, for all from 10^{-5} to 1, only best result. N: Codebook size for Bag-of-Audio-Words (BoAW) splitting the input into two codebooks (*ComParE*-LLDs/*ComParE*-LLD-deltas) of the same given size, with 50 assignments per frame. $DenseNet121$: pre-trained CNN used for extraction of* Deep Spectrum *features. X: Threshold power levels for $S2SAE$ under which was clipped.* DiFE*: Linguistic feature extraction pipeline and SVM.* End2You*: End-to-end learning with convolutional recurrent neural network hidden units N_h.* ***UAR****: Unweighted Average Recall.* **CCS***: COVID-19 Coughing.* **CSS***: COVID-19 Speech.* **ESS***: Escalation Sub-Challenge.* ***PRS****: Primates Sub-Challenge. CI on Test: confidence intervals for Test, see explanation in text.*

	CCS UAR [%]			**CSS** UAR [%]			**ESS** UAR [%]			**PRS** UAR [%]		
	Dev	Test	CI on Test	Dev	Test	CI on Test	Dev	Test	CI on Test	Dev	Test	CI on Test
C	**openSMILE: ComParE functionals+SVM**											
	61.4	65.5	56.1-74.3 / 66.1-67.2	57.9	**72.1**	66.0-77.8 / 70.2-71.1	70.5	58.6	53.5-63.3 / 55.2-58.3	82.4	82.2	80.5-83.9 / 78.8-79.6
N	**openXBOW: ComParE BoAW+SVM**											
125	60.7	66.7	59.5-75.3 / 64.5-65.5	66.0	63.6	57.6-69.6 / 62.0-63.2	72.2	55.8	50.2-61.0 / 52.6-56.4	–	–	
250	60.7	63.3	54.1-72.3 / 60.8-62.0	60.6	60.4	54.5-66.3 / 60.9-61.9	69.0	53.0	47.8-57.8 / 50.9-53.3	80.0	80.9	79.2-82.5 / 78.8-79.5
500	66.4	67.6	59.3-76.7 / 65.7-66.8	64.2	64.7	58.7-70.4 / 62.6-63.7	70.1	49.4	44.4-54.0 / 47.3-49.3	83.1	82.4	80.6-84.0 / 80.1-80.8
1000	66.2	69.1	60.6-77.5 / 69.3-70.2	62.6	68.7	62.9-74.2 / 66.0-67.0	69.7	56.8	52.0-61.8 / 55.7-56.9	83.3	83.9	82.2-85.5 / 81.4-81.9
2000	64.7	72.9	64.4-80.5 / 71.5-72.2	66.3	68.7	62.9-74.2 / 64.4-66.4	70.6	**59.8**	54.8-64.7 / 56.3-58.2	–	–	
Network	**DeepSpectrum+SVM**											
DenseNet121	63.3	64.1	55.7-72.8 / 65.9-67.1	56.0	60.4	55.9-64.9 / 57.8-58.7	64.2	56.4	51.5-61.3 / 53.6-55.2	81.3	78.8	76.9-80.6 / 76.1-76.8
X [dB]	**auDeep: S2SAE+SVM**											
-30	60.7	55.2	47.6-61.9 / 51.9-53.5	65.8	59.9	53.6-65.4 / 58.2-59.3	39.1	35.3	30.0-40.4 / 34.8-37.3	70.6	69.7	67.7-71.8 / 69.1-69.5
-45	64.1	60.5	51.8-69.5 / 61.0-62.0	66.3	55.2	49.1-61.0 / 54.1-55.2	41.3	43.1	37.8-48.6 / 38.5-42.0	80.3	82.3	80.6-83.8 / 80.5-81.3
-60	67.6	67.6	60.3-75.4 / 64.9-65.8	59.4	53.3	47.4-59.4 / 52.2-53.5	42.0	44.3	39.2-49.6 / 41.7-44.1	81.6	84.1	82.5-85.6 / 82.4-83.2
-75	64.0	64.6	56.1-72.6 / 61.0-62.3	58.4	52.2	45.9-57.7 / 52.0-52.9	49.0	52.2	47.2-56.9 / 50.1-52.0	80.7	83.0	81.5-88.0 / 81.1-82.0
Fused	65.4	64.2	57.0-72.2 / 62.1-63.1	62.2	64.2	63.1-74.3 / 62.3-64.2	46.8	45.0	39.8-50.4 / 45.1-47.5	84.6	86.6	85.1-88.0 / 84.6-85.2
Features	**DiFE: Transformer+SVM**											
plain	–	–	–	–	–	–	51.2	36.8	32.2-41.7 / 38.8-41.2	–	–	
plain-BlAtt	–	–	–	–	–	–	50.3	45.2	39.4-50.8 / 44.0-45.3		–	
sent	–	–	–	–	–	–	56.5	44.1	38.4-49.7/ 40.9-44.2	–	–	
sent-BlAtt	–	–	–	–	–	–	47.3	47.2	41.8-52.9 / 46.9-47.8	–	–	
tuned-BlAtt	–	–	–	–	–	–	43.5	44.9	40.0-50.3 / 43.7-45.3	–	–	
N_h RNN	**End2You: CNN+LSTM RNN**											
64	61.8	64.7	56.2-73.5 / –	70.5	68.7	63.1-74.3 / –	64.1	54.0	48.8-59.5 / –	72.70	70.8	68.8-72.9 / –
	Fusion of Best											
	–	**73.9**	66.0-82.6 / –	–	71.1	65.4-76.3 / –	–	59.7	55.0-64.4 / –	–	**87.5**	86.0-88.9 / –

based emotion recognition [20]. 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 another 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/deltas in the training data. Each LLD/delta is assigned to the 10 audio words from the codebooks with the lowest Euclidean distance. Both BoAW representations, one from the LLDs and one from their deltas, are concatenated. Finally, a logarithmic term frequency weighting is applied to compress the numeric range of the histograms. LLDs are extracted with the openSMILE toolkit, BoAWs are computed using openXBOW [21].

Deep Spectrum: The feature extraction Deep Spectrum toolkit[1] is applied to obtain first deep representations from the input audio data utilising pre-trained convolutional neural networks (CNNs) [22]. Deep Spectrum features have been shown to be effective, e. g., for speech processing [23]. First, audio signals are transformed into mel-spectrogram plots using a Hanning window of width 32 ms and an overlap of 16 ms. From these, 128 Mel frequency bands are computed. The spectrograms are then forwarded through DenseNet121 [24], a pre-trained CNN, and the activations of the 'avg_pool' layer of the network are extracted, resulting in a 2 048 dimensional feature vector.

auDeep: Another feature set is obtained through unsupervised representation learning with recurrent sequence to sequence autoencoders, using auDeep[2] [25, 26]. These explicitly model the inherently sequential nature of audio with Recurrent Neural Networks (RNNs) within the encoder and decoder networks [25, 26]. Here, 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 different 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.

DiFE: Escalation is marked by an increase in arousal coming from acoustic rather than linguistic features; yet, semantic connotations might additionally play a role [27, 28]. To this aim, we developed a lightweight Dutch Linguistic Feature Extractor (DiFE) pipeline similar to [29] and last year's challenge [30] to utilise linguistic features for ESS[3]. Transformer language embeddings recently showed tremendous success over a wide range of Natural Language Processing tasks. For the vectorisation, DiFE either utilises a) a standard pre-trained Dutch BERT model ($plain$), b) a fine-tuned version on an external sentiment ($sent$) task [31], or c) a fine-tuned version on the escalation $train$ and $validation$ partitions ($tuned$). Next, a 768-dimensional context embedding vector for each word of a segment of the last 4 layers is extracted and summed up over the last four layers [32]. The sequence of encoded words is then either summed up again across the time dimension, or fed into a feature compression block to obtain a single feature vector for the entire segment. For

[1] https://github.com/DeepSpectrum/DeepSpectrum

[2] https://github.com/auDeep/auDeep

[3] https://github.com/lstappen/DiFE

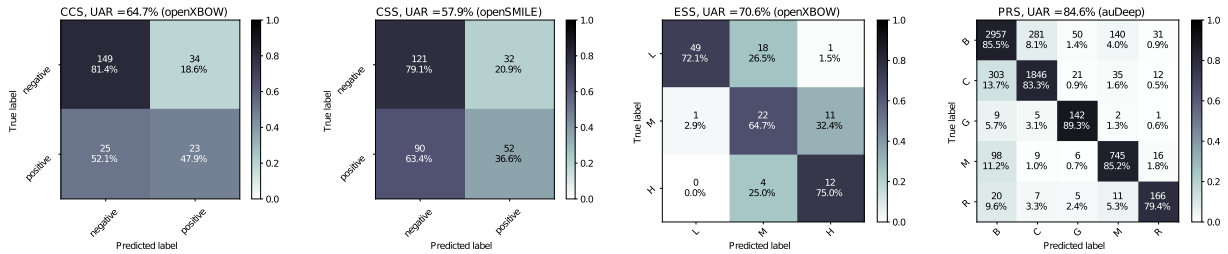

Figure 1: *Confusion matrices for **CCS**, **CSS**, **ESS**, and **PRS**. The individual approach/hyperparameters performing on Dev for the best Test result (without fusion) were chosen – given on top of each figure. In the cells, absolute number and percent of 'classified as' of the class displayed in the respective row; percentage also indicated by colour-scale: the darker, the higher.*

compression, the pipeline uses a bidirectional Long Short-Term Memory (LSTM) RNN with an attention module (BLAtt), followed by two feedforward layers. The output of this last layer is used as feature input for the SVM evaluation.

End2You: We utilise the multimodal profiling toolkit End2You [33][4] to perform end-to-end learning. For our purposes, we utilise the Emo-18 [34] deep neural network that uses 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 100 ms each (except for the PRS Sub-Challenge with chunks of 70 ms). These are fed into a three layer convolutional network comprised of a series of convolution and pooling operations which try to find a robust representation of the original signal. The extracted features are passed to a two layer GRU to capture the temporal dynamics in the raw waveform.

3.2. Challenge Baselines and Interpretation

For the sake of transparency and reproducibility of the baseline computation, in line with previous years, we use an open-source implementation of SVMs with linear kernels. The provided scripts employ the SCIKIT-LEARN toolkit with its class LINEARSVC for the classification based on functionals, BoAW, AUDEEP, DIFE, and DEEP SPECTRUM features. All feature representations were scaled to zero mean and unit standard deviation (STANDARDSCALER of SCIKIT-LEARN), using the parameters from the respective training set (when Train and Dev were fused for the final classifier, the parameters were calculated on this fusion). The complexity parameter C was always 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 on Dev, and scoring by the competition and further measures). This year, we provide the six approaches outlined above. The same way as in the last three years, we chose the highest results on Test for defining the baselines, irrespective of the corresponding results on Dev, 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 best configurations (each different approach with its best parameters) with *Majority Voting* is given in the last row. As can be seen in Table 2, for **CCS**, the baseline is fusion of best with **UAR** $= 73.9\,\%$; for **CSS**, the baseline is based on COMPARE with **UAR** $= 72.1\,\%$; for **ESS**, BoAWs define the baseline with **UAR** $= 59.8\,\%$; and for **PRS**, the baseline is fusion of best with **UAR** $= 87.5\,\%$.

We provide two types of 95 % confidence intervals, see the column 'CI on Test' in Table 2: First, we did 1000x bootstrapping for Test (random selection with replacement) and computed UARs, based on the same model that was trained with Train and Dev; the CI for these UARs is given before the slash. Then, we did 100x bootstrapping[5] for the corresponding combination of Train and Dev, and employed the different models obtained from these combinations to get UARs for Test[6] and subsequently, CIs, as displayed after the slash. Note that for this type of CI, the Test results are often above the CI, sometimes within and in a few cases below. Obviously, reducing the variability of the samples in the training phase with bootstrapping results on average in somehow lower performance.

Figure 1 displays the confusion matrices for the four sub-challenges for Dev corresponding to the best result on Test; e. g., for CCS, best Test result (without fusion) is 72.9 % UAR for N $= 2000$; displayed is the confusion matrix corresponding to the UAR of 64.7 %. Especially for CCS but for CSS as well, positive is frequently confused with negative, which may be tuned in a use case. For ESS, confusion between the extreme classes L and H are almost non-existent. The high UAR for PRS is mirrored by the high values in the diagonal – all five classes are predicted in a range of 10 % absolute, from 79 % to 89 %.

4. Concluding Remarks

This year's challenge is new by four new tasks (COVID-19 Cough and Speech, Escalation, and Primates), all of them highly relevant for applications. Besides the by now 'classic' approaches **COMPARE** and **Bag-of-Audio-Words (BoAWs)**, we further featured sequence-to-sequence autoencoder-based audio features by the **AUDEEP** toolkit, **DEEP SPECTRUM**, a Dutch LinguistIc Feature Extractor (**DIFE**) as well as **End2End Deep Sequence Modelling**. 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. Acknowledgements

We acknowledge funding from the DFG's Reinhart Koselleck project No. 442218748 (AUDI0NOMOUS), the EU's HORIZON 2020 Grant No. 115902 (RADAR CNS), and the ERC project No. 833296 (EAR).

[4] https://github.com/end2you/end2you

[5] For PRS, only 10x was executed, because of the large number of data points.

[6] This holds apart from End2You that would have required too time-consuming computation cycles.

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] M. A. Ismail, S. Deshmukh, and R. Singh, "Detection of covid-19 through the analysis of vocal fold oscillations," arXiv preprint arXiv:2010.10707, 2020.

[4] A. Rosenberg, "Classifying skewed data: Importance weighting to optimize average recall," in Proc. Interspeech, Portland, OR, 2012, pp. 2242–2245.

[5] C. Brown, J. Chauhan, A. Grammenos, J. Han, A. Hasthanasombat, D. Spathis, T. Xia, P. Cicuta, and C. Mascolo, "Exploring Automatic Diagnosis of COVID-19 from Crowdsourced Respiratory Sound Data," in Proc. KDD, San Diego, CA, 2020, pp. 3474–3484.

[6] J. Han, C. Brown, J. Chauhan, A. Grammenos, A. Hasthanasombat, D. Spathis, T. Xia, P. Cicuta, and C. Mascolo, "Exploring Automatic COVID-19 Diagnosis via Voice and Symptoms from Crowdsourced Data," in Proc. ICASSP, Toronto, Canada, 2021, 5 pages, to appear.

[7] I. Lefter, L. J. Rothkrantz, and G. J. Burghouts, "A comparative study on automatic audio–visual fusion for aggression detection using meta-information," Pattern Recognition Letters, vol. 34, no. 15, pp. 1953–1963, 2013.

[8] I. Lefter, G. J. Burghouts, and L. J. Rothkrantz, "An audio-visual dataset of human–human interactions in stressful situations," Journal on Multimodal User Interfaces, vol. 8, no. 1, pp. 29–41, 2014.

[9] J. A. Zwerts, J. Treep, C. S. Kaandorp, F. Meewis, A. C. Koot, and H. Kaya, "Introducing a central african primate vocalisation dataset for automated species classification," arXiv preprint arXiv:2101.10390, 2021.

[10] N. Priyadarshani, S. Marsland, and I. Castro, "Automated birdsong recognition in complex acoustic environments: a review," Journal of Avian Biology, vol. 49, no. 5, pp. jav–01 447, 2018.

[11] P. H. Wrege, E. D. Rowland, S. Keen, and Y. Shiu, "Acoustic monitoring for conservation in tropical forests: examples from forest elephants," Methods in Ecology and Evolution, vol. 8, no. 10, pp. 1292–1301, 2017.

[12] S. Heinicke, A. K. Kalan, O. J. Wagner, R. Mundry, H. Lukashevich, and H. S. Kühl, "Assessing the performance of a semi-automated acoustic monitoring system for primates," Methods in Ecology and Evolution, vol. 6, no. 7, pp. 753–763, 2015.

[13] P. Fedurek, K. Zuberbühler, and C. D. Dahl, "Sequential information in a great ape utterance," Scientific reports, vol. 6, p. 38226, 2016.

[14] D. J. Clink, M. C. Crofoot, and A. J. Marshall, "Application of a semi-automated vocal fingerprinting approach to monitor Bornean gibbon females in an experimentally fragmented landscape in Sabah, Malaysia," Bioacoustics, vol. 28, no. 3, pp. 193–209, 2019.

[15] A. P. Hill, P. Prince, J. L. Snaddon, C. P. Doncaster, and A. Rogers, "Audiomoth: A low-cost acoustic device for monitoring biodiversity and the environment," HardwareX, vol. 6, p. e00073, 2019.

[16] 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. Interspeech, Lyon, France, 2013, pp. 148–152.

[17] F. Eyben, F. Weninger, F. Groß, and B. Schuller, "Recent Developments in openSMILE, the Munich Open-Source Multimedia Feature Extractor," in Proc. ACM Multimedia, Barcelona, Spain, 2013, pp. 835–838.

[18] 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, pp. 1–12, 2013.

[19] H. Lim, M. J. Kim, and H. Kim, "Robust Sound Event Classification Using LBP-HOG Based Bag-of-Audio-Words Feature Representation," in Proc. Interspeech, Dresden, Germany, 2015, pp. 3325–3329.

[20] 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. Interspeech, San Francisco, CA, 2016, pp. 495–499.

[21] M. Schmitt and B. W. Schuller, "openXBOW – Introducing the Passau Open-Source Crossmodal Bag-of-Words Toolkit," Journal of Machine Learning Research, vol. 18, pp. 1–5, 2017.

[22] S. Amiriparian, M. Gerczuk, S. Ottl, N. Cummins, M. Freitag, S. Pugachevskiy, and B. Schuller, "Snore sound classification using image-based deep spectrum features," in Proc. Interspeech 2017, Stockholm, Sweden, 2017, pp. 3512–3516.

[23] S. Amiriparian, M. Gerczuk, S. Ottl, N. Cummins, S. Pugachevskiy, and B. Schuller, "Bag-of-deep-features: Noise-robust deep feature representations for audio analysis," in Proc. IJCNN, Rio de Janeiro, Brazil, 2018, pp. 2419–2425.

[24] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in Proc. CVPR, 2017, pp. 4700–4708.

[25] S. Amiriparian, M. Freitag, N. Cummins, and B. Schuller, "Sequence to Sequence Autoencoders for Unsupervised Representation Learning from Audio," in Proc. DCASE 2017, Munich, Germany, 2017, pp. 17–21.

[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. 18, pp. 1–5, 2018.

[27] L. Stappen, A. Baird, E. Cambria, and B. W. Schuller, "Sentiment analysis and topic recognition in video transcriptions," IEEE Intelligent Systems, vol. 36, no. 2, 2021.

[28] L. Stappen, B. Schuller, I. Lefter, E. Cambria, and I. Kompatsiaris, "Summary of muse 2020: Multimodal sentiment analysis, emotion-target engagement and trustworthiness detection in real-life media," in Proceedings of the 28th ACM International Conference on Multimedia, 2020, pp. 4769–4770.

[29] L. Stappen, G. Rizos, M. Hasan, T. Hain, and B. W. Schuller, "Uncertainty-aware machine support for paper reviewing on the interspeech 2019 submission corpus," Proc. Interspeech 2020, pp. 1808–1812, 2020.

[30] B. W. Schuller, A. Batliner, C. Bergler, E.-M. Messner, A. Hamilton, S. Amiriparian, A. Baird, G. Rizos, M. Schmitt, L. Stappen et al., "The interspeech 2020 computational paralinguistics challenge: Elderly emotion, breathing & masks," Proc. INTERSPEECH. Shanghai, China: ISCA, 2020.

[31] W. de Vries, A. van Cranenburgh, A. Bisazza, T. Caselli, G. van Noord, and M. Nissim, "Bertje: A dutch bert model," arXiv preprint arXiv:1912.09582, 2019.

[32] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," in Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 2019, pp. 4171–4186.

[33] 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.

[34] P. Tzirakis, J. Zhang, and B. W. Schuller, "End-to-end speech emotion recognition using deep neural networks," in Proc. ICASSP, 2018, pp. 5089–5093.