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

Introducing a Central African Primate Vocalisation Dataset for Automated Species Classification

Joeri A. Zwerts[1], *Jelle Treep*[2], *Casper S. Kaandorp*[2], *Floor Meewis*[1], *Amparo C. Koot*[1], *Heysem Kaya*[3]

[1] Department of Biology, Utrecht University, Utrecht, The Netherlands
[2] Information and Technology Services, Utrecht University, Utrecht, The Netherlands
[3] Department of Information and Computing Sciences, Utrecht University, Utrecht, The Netherlands

j.a.zwerts@uu.nl, h.kaya@uu.nl

Abstract

Automated classification of animal vocalisations is a potentially powerful wildlife monitoring tool. Training robust classifiers requires sizable annotated datasets, which are not easily recorded in the wild. To circumvent this problem, we recorded four primate species under semi-natural conditions in a wildlife sanctuary in Cameroon with the objective to train a classifier capable of detecting species in the wild. Here, we introduce the collected dataset, describe our approach and initial results of classifier development. To increase the efficiency of the annotation process, we condensed the recordings with an energy/change based automatic vocalisation detection. Segmenting the annotated chunks into training, validation and test sets, initial results reveal up to 82% unweighted average recall test set performance in four-class primate species classification.
Index Terms: acoustic primate classification, wildlife monitoring, computational paralinguistics

1. Introduction

Wildlife is declining at unprecedented rates, and monitoring trends in biodiversity is key to engage in effective conservation actions [1]. Using acoustic recordings to identify and count species is a promising non-invasive and cost-effective monitoring tool [2]. This can be particularly useful in environments with limited visibility such as tropical forests, or for arboreal, shy or nocturnal species that are more easily heard than seen. Acoustic monitoring, especially in conjunction with other monitoring methods, has the potential to profoundly change ecological research by opening up new ways of studying community composition, species interactions and behavioral processes [3]. For efficient analysis of audio recordings however, automated detection is pivotal. In addition to relieving a data processing bottleneck, machine learning methods allow for consistency in terms of quality, increasing the comparability and reproducibility of the output.

Training robust classifiers requires sizable amounts of annotated data, which can require substantial efforts to compile from natural forest recordings. To circumvent this problem, we recorded several primate species in a sanctuary in Cameroon, including chimpanzees (*Pan troglodytes*, n=20), mandrills (*Mandrillus sphinx*, n=17), red-capped mangabeys (*Cercocebus torquatus*, n=6) and a mixed group of guenon species (*Cercopithecus spp.*, n=20). The primates in the sanctuary live in semi-natural conditions with background noise that is somewhat, although not wholly, comparable to natural background noise. The ultimate objective of these efforts is to train a classifier capable of detecting species in the wild. This may also provide insights into whether this approach, of using sanctuary recordings, can be used to train classifiers for other species as well, to aid in the development of cost-effective monitoring to meet modern conservation challenges.

In this paper, we present the dataset, the semi-automatic annotation process that we used to speed up the manual annotation process, and a benchmark species classification system.

1.1. Related Work

Multiple studies have applied automatic acoustic monitoring for a variety of taxa including cetaceans [4], birds [5], bats [6], insects [7], amphibians [8], and forest elephants [9]. However, they have so far only been sporadically been used for primates [10, 11, 12, 13, 14, 15]. A brief summary of recent works on classification of primate vocalisations is given in Table 1. We observe that Mel-Frequency Cepstral Coefficients (MFCC) are commonly used in classifying primate vocalisations, in most cases without other acoustic descriptors. In our study, we also use MFCCs (together with temporal delta coefficients) and combine them with RASTA-style Perceptual Linear Prediction Cepstral Coefficients (RASTA-PLPC). There are also off-the-shelf applications like Kaleidoscope Pro (Wildlife Acoustics, MA, USA) based on Hidden Markov Models that were used in recent works for call type classification of Japanese macaques (*Macaca fuscata*) [15].

2. Central African Primate Dataset

2.1. Acoustic Data Collection

The acoustic data is collected in the Mefou Primate Sanctuary (Ape Action Africa) in Cameroon in December 2019 and January 2020. The sanctuary, which houses the primates in a semi-natural forest setting, cares for rescued primates and engages in conservation and education initiatives. Recordings were made using Audiomoth (v1.1.0) recorders [16]. Devices recorded 1-min segments continuously at 48 kHz and 30.6 dB gain, storing the data in one minute WAVE-files, with interruptions from two to five seconds between recordings for the recorder to save the files. For all species, the recorders were installed either directly on the fence of their respective enclosures, or maximally up to 3 meters away from it. Per species, the enclosures differed in size and were approximately 40 $\times$ 40 meters in size for the guenons and red-capped mangabeys, 50 $\times$ 50 meters for the mandrills and 70 $\times$ 70 meters for the chimpanzees. Distance between the recorder and the animals naturally varied depending on the location of the animal within the enclosure. The smallest distance between two enclosures having different species was 30 meters.

http://dx.doi.org/10.21437/Interspeech.2021-154

Table 1: *Summary of recent works on automatic primate vocalisation classification. k-NN: k-Nearest Neighbors, LPF: Linear Prediction Filter, MLP: Multi-Layer Perceptron, SVM: Support Vector Machines, OPF: Optimum Path Forest, ZCR: Zero Crossing Rate.*

Work	**Task(s)**	**Species**	**Features**	**Classifiers**
Mielke et al. [10]	Three recognition tasks (individual, call type and species)	Blue monkey (*Cercopithecus mitis stuhlmanni*), Olive baboon (*Papio anubis*), Redtail monkey (*Cercopithecus ascanius schmidti*), Guereza colobus (*Colobus guereza occidentalis*)	MFCC [1-32] and Deltas	MLP
Heinicke et al. [11]	5-class primate classification	Chimpanzee (*Pan troglodytes*), Diana monkey (*Cercopithecus diana*), King colobus (*Colobus polykomos*) and Western red colobus (*Procolobus badius*)	MFCCs, loudness, spectral crest factor, spectral flatness measure, and ZCR	SVM and GMM
Fedurek et al. [12]	Age, Context, Identity, Social Status	Chimpanzee (*Pan troglodytes*)	MFCCs	SVM
Turesson et al. [13]	8-class classification of Marmoset vocalisations	Common marmoset (*Callithrix jacchus*)	LPC with LPF orders of 10, 15, 20 and 25	AdaBoost, Bayesian Classifier, k-NN, Logistic regression, MLP, SVM, OPF
Clink et al. [14]	Distinguishing individuals	Bornean gibbon (*Hylobatidae muelleri*)	MFCC [1-12]	SVM

Due to the limited distance between some of the enclosures and the loudness of the vocalisations, some level of interference (i.e. the existence of a distant call of an unintended species) between the species' vocalisations is present, particularly in the mandrill recordings. Recordings can also contain noise from dogs, humans talking, or other human activities. The chimpanzees were recorded in two separate enclosures with two recorders per enclosure recording simultaneously. Hence, there may be overlap in vocalisations for recordings 1 and 2 as well as for recordings 3 and 4. This issue is considered in the chronological ordering based segmentation of the data into the training, validation and test sets. The total dataset amounts to a duration of 1112 hours, 358 GBs of original audio collected over a time span of 32 days.

2.2. Annotation

The recordings were annotated by two experts who manually reviewed the sound recordings and corresponding spectrograms in Raven Pro® software. Inter-annotator agreement was reached by doubly annotating recordings until discrepancies were negligible. To speed up the annotation process, we 'condensed' the data with an energy/change based automatic vocalisation detection using the first batch of manual annotations to estimate the detection performance. An overview of the semi-automatic annotation process is illustrated in Figure 1. The detection comprises obtaining the power distribution from the power spectrum. From a species-specific frequency sub-band, we collect chunks (time-intervals) in which the registered signal loudness exceeds a species-specific threshold, or in which the local cumulative power distribution deviates from a global counterpart. The species specific thresholds are optimized to include close to all (>95%) initial annotations and to remove as much background sections as possible. The 'condensed' collection represents a set of timestamps, where we expect to hear disruptions in the ambient noise. The time-intervals are used to extract the corresponding signal fragments from our raw data. These fragments are bundled into a new audio file containing a high density of vocalisations that can be annotated more efficiently.

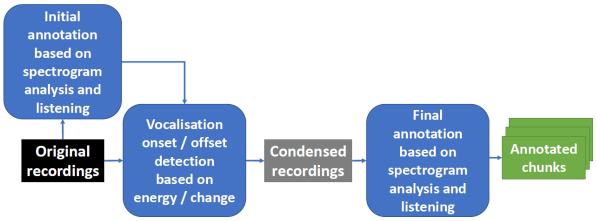

Figure 1: *The semi-automatic annotation pipeline used in the study.*

Each species produces several vocalisation types, each varying in relative frequency, loudness and spectral properties. The experts consider these cues while observing the spectrogram (see Figure 2 for exemplar spectrograms of various call types), spotting a candidate chunk and then listening to the selected chunk. This process yields over 10K annotated primate vocalisations with a class distribution of 6652 chimpanzee, 2623 mandrill, 627 red-capped mangabey and 476 of the mixed guenon group.

3. Benchmark Vocalisation Classification System

To assess how well the species vocalisations can be automatically classified in the presented dataset, we present an acoustic primate classification system. The first stage is acoustic feature extraction, where we extract a standard set of acoustic descriptors from the signal and then summarize them using the statistical functionals (such as mean and standard deviation) over each chunk. This stage produces suprasegmental features of equal length. The next stage is machine learning, where the acoustic features and corresponding primate classes are input to a supervised learner. The details of these stages are given in the subsequent subsections.

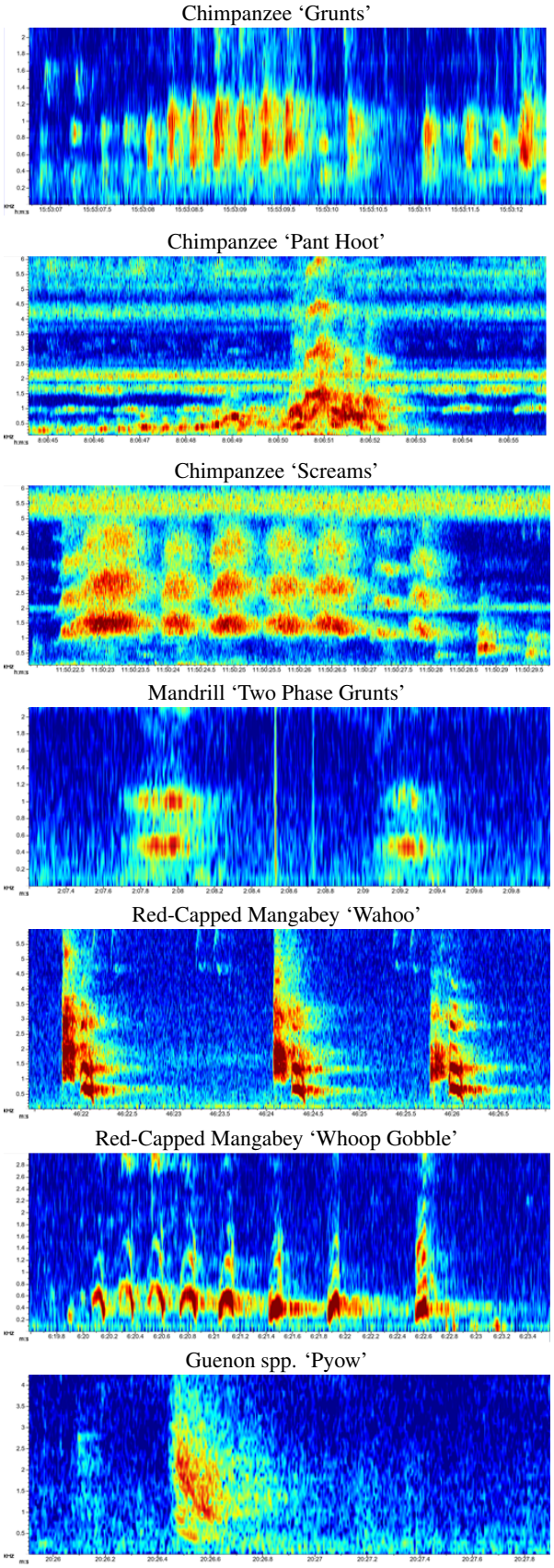

Figure 2: *Exemplar spectrograms for different vocalisations of the annotated primate species.*

3.1. Acoustic Feature Extraction

As acoustic Low-Level Descriptors (LLDs), we extract Mel-Frequency Cepstral Coefficients (MFCCs) 0-24 and Relative Spectral Transform (RASTA) [17] - Perceptual Linear Prediction (PLP) [18] cepstrum for 12^{th} order linear prediction, together with their first and second order temporal coefficients (Δ and $\Delta\Delta$), making an LLD vector of 114 dimensions. The descriptors are then summarized using 10 functionals, based on the success observed in former paralinguistic studies [19, 20]. The functionals used are: mean, standard deviation, slope and offset from the first order polynomial, the curvature (the leading coefficient) from the second order polynomial fit to the LLD contour, minimum value and its relative position, maximum value and its relative position, zero crossing rate of the LLD contour normalized into [-1,1] range. This process yields $114 \times 10 = 1140$ supra-segmental acoustic features for each chunk, regardless of the number of frames.

3.2. Model Learning

In our work, we employ Kernel Extreme Learning Machine (ELM) [21] method, since this is a fast and accurate algorithm that previously produced state-of-the-art results on several paralinguistic problems [22, 23].

Here, we opt to provide a brief explanation of ELM. Initially, ELM is proposed as a fast learning method for Single Hidden Layer Feedforward Networks (SLFN): an alternative to back-propagation [24]. To increase the robustness and the generalisation capability of ELM, a regularisation coefficient $\mathbf{C}$ is included in the optimisation procedure. Therefore, given a kernel $\mathbf{K}$ and the label vector $\mathbf{T} \in \mathbb{R}^{N \times 1}$ where N denotes the number of instances, the projection vector β is learned as follows [21]:

$$\beta = (\frac{\mathbf{I}}{C} + \mathbf{K})^{-1}\mathbf{T}. \quad (1)$$

In order to prevent parameter over-fitting, we use the linear kernel $\mathbf{K}(x, y) = x^T y$, where x and y are the (normalised) feature vectors. With this approach, the only parameter of our model is the regularisation coefficient C, which we optimize on the validation set.

4. Preliminary Experiments on the Primate vocalisation Dataset

In this section we present our spectral analysis and the results of the preliminary classification experiments using the proposed benchmark system.

4.1. Spectral Analysis of vocalisations

During the semi-automatic annotation process, we have analysed the spectral characteristics of vocalisations and the background noise per species. Based on domain knowledge and initial experimentation, we focused on spectral bands up to 2KHz. For this analysis, we have combined all annotated chunks for each primate class, obtained the power spectrum and then summarized the power in decibels (dB) using mean over time. We applied the same procedure for corresponding background portions for each species. The difference between the two means (see Figure 3) in dB provides an idea about the signal-to-noise ratio (SNR) of each species, and as such the relative difficulty of distinguishing each species' vocalisations in the given acoustic background conditions. In the figure, we observe multiple modes for mandrills and red-capped mangabeys, which cor-

respond to different call types (c.f. Figure 2). In line with the acoustic observations during the annotations, vocalisations from mandrills and red-capped mangabeys have lower SNR values, making both the annotation and automated detection a harder problem.

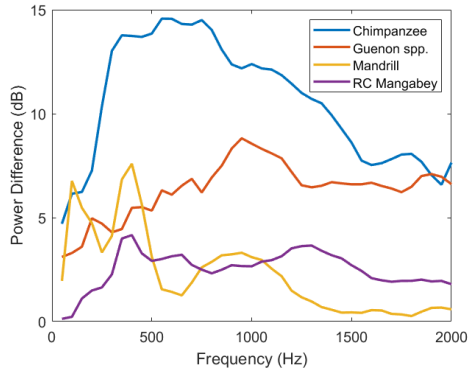

Figure 3: *Average power (dB) difference between the mean vocalisation (signal) and background (noise) spectrum.*

4.2. Classification Results

For the classification experiments, we partitioned the dataset into training, validation and test sets using a temporal ordering (i.e. training correspond to the oldest, test to the newest recordings) with a ratio of 3:1:1, respectively. We set up two classification tasks 1) four-class classification of the species, 2) the four species classes plus the background chunks from the recordings of all species as the fifth class. To generate the background chunks, we sampled from the recordings not annotated as vocalisation, to exactly match the duration distribution of the annotated chunks of each species. This makes the five class problem highly imbalanced, as half of the chunks are of background class. However, such an imbalance is not extra-ordinary, if the final aim is to train classifiers for wildlife monitoring.

The models are trained on the training set, optimizing the Kernel ELM complexity hyper-parameter on the validation set. Then using the optimal hyper-parameter, the combination of the training and the validation sets are re-trained, and the corresponding model's predictions are checked against the ground truth test set labels. We use both accuracy and unweighted average recall (UAR), to report the predictive performance.

Using the acoustic features described in Section 3.1, we then trained the Kernel ELM models with z-normalisation (ZN - standardising each feature such that they have zero mean and unit variance) and a combination of ZN with feature-vector level L2 normalisation, as suggested in [25]. When used with a linear kernel, L2 normalisation effectively converts the linear kernel into a cosine similarity kernel. The hyper-parameters of Kernel ELM method is optimized in the set $10^{\{-6,-5,-4,-3,-2,-1,0,1\}}$ with ZN and in in the set $10^{\{-1,0,1,2,3,4,5,6\}}$ with ZN+L2 normalisation combination. The respective validation and test performance of the trained models are summarized in Table 2. Note that we optimize for UAR due to class imbalance, while reporting both accuracy and UAR measures.

From the table, we observe that the test set (single probe for each normalisation option and task combination) performances are always better than the corresponding validation set performance. Moreover, all results are dramatically higher than chance-level UAR, which is 0.25 for the four-class and 0.2 for the five-class classification task. The results show that 1) the collected acoustic recordings have clear distinction for automatic discrimination of primate vocalisations, and 2) the proposed system has a good generalisation, reaching test set UAR scores of 0.82 and 0.70 in four-class and five-class classification tasks, respectively.

Table 2: *Validation and test set performances of KELM models for four and five-class classification tasks.*

		Validation		Test	
Task	Norm	Accuracy	UAR	Accuracy	UAR
Four-cls	ZN	0.554	0.697	0.735	0.821
	ZN+L2	0.595	0.705	0.767	0.823
Five-cls	ZN	0.603	0.610	0.682	0.707
	ZN+L2	0.617	0.627	0.697	0.698

5. Discussion and Conclusions

Initial results showed that we attain relatively high classification performance using our proposed system combining functionals of MFCC and RASTA-PLPC descriptors and modeling them using Kernel ELM. Data condensation also proved to be a valuable addition to the workflow for reducing the annotation workload. Our future aim is to apply the model on acoustic recordings of natural forests.

Natural forest sounds pose the additional challenge of containing far fewer vocalisations compared to the sanctuary, and significantly higher levels of background noise, in particular in less relevant frequency bands. Moreover, similar to humans, primates can have varying vocal behavior across sex and age, including sex-specific call types, differences in frequency of specific vocalisation types, and differences in acoustic structures of shared call types [26, 10]. There is also some extent of inter-individual variation, especially for chimpanzees [12]. Considering the limited group sizes from which we derive our data, such variation may inevitably result in low generalisation when applied to the natural variation of individuals and group composition. Finally, not all species and species call types will be equally suitable for automated detection. Louder species such as chimpanzees will be more easily distinguished from background noise than for instance mandrills, and will consequently also have wider detection areas. Chimpanzees, however, often scream simultaneously, making it difficult to distinguish separate calls.

Future work lies in overcoming these challenges, which are partly caused because of the mismatch of acoustic conditions between sanctuary and natural data. Nonetheless, using sanctuary data has the advantage to provide relatively low-cost and accessible training data for classifiers, which may in turn boost the development and increased adoption of semi-automatic acoustic wildlife monitoring methods. To aid this development, the presented dataset is made publicly available in the context of the INTERSPEECH 2021 Computational Paralinguistics Challenge (ComParE 2021) [27].

6. Acknowledgements

This research was funded by the focus area Applied Data Science at Utrecht University, The Netherlands.

7. References

[1] R. Almond, M. Grooten, and T. Peterson, *Living Planet Report 2020-Bending the curve of biodiversity loss*. World Wildlife Fund, 2020.

[2] L. S. M. Sugai, T. S. F. Silva, J. W. Ribeiro Jr, and D. Llusia, "Terrestrial passive acoustic monitoring: review and perspectives," *BioScience*, vol. 69, no. 1, pp. 15–25, 2019.

[3] R. T. Buxton, P. E. Lendrum, K. R. Crooks, and G. Wittemyer, "Pairing camera traps and acoustic recorders to monitor the ecological impact of human disturbance," *Global Ecology and Conservation*, vol. 16, p. e00493, 2018.

[4] M. Bittle and A. Duncan, "A review of current marine mammal detection and classification algorithms for use in automated passive acoustic monitoring," in *Proceedings of Acoustics*, vol. 2013, 2013.

[5] 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.

[6] D. Russo and C. C. Voigt, "The use of automated identification of bat echolocation calls in acoustic monitoring: A cautionary note for a sound analysis," *Ecological Indicators*, vol. 66, pp. 598–602, 2016.

[7] T. Ganchev and I. Potamitis, "Automatic acoustic identification of singing insects," *Bioacoustics*, vol. 16, no. 3, pp. 281–328, 2007.

[8] C. L. Brauer, T. M. Donovan, R. M. Mickey, J. Katz, and B. R. Mitchell, "A comparison of acoustic monitoring methods for common anurans of the northeastern united states," *Wildlife Society Bulletin*, vol. 40, no. 1, pp. 140–149, 2016.

[9] 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.

[10] A. Mielke and K. Zuberbühler, "A method for automated individual, species and call type recognition in free-ranging animals," *Animal Behaviour*, vol. 86, no. 2, pp. 475–482, 2013.

[11] 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.

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

[13] H. K. Turesson, S. Ribeiro, D. R. Pereira, J. P. Papa, and V. H. C. de Albuquerque, "Machine learning algorithms for automatic classification of marmoset vocalizations," *PloS one*, vol. 11, no. 9, p. e0163041, 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] H. Enari, H. S. Enari, K. Okuda, T. Maruyama, and K. N. Okuda, "An evaluation of the efficiency of passive acoustic monitoring in detecting deer and primates in comparison with camera traps," *Ecological Indicators*, vol. 98, pp. 753–762, 2019.

[16] 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.

[17] H. Hermansky and N. Morgan, "RASTA Processing of Speech," *Speech and Audio Processing, IEEE Transactions on*, vol. 2, no. 4, pp. 578–589, 1994.

[18] H. Hermansky, "Perceptual Linear Predictive (PLP) Analysis of Speech," *the Journal of the Acoustical Society of America*, vol. 87, no. 4, pp. 1738–1752, 1990.

[19] E. Çiftçi, H. Kaya, H. Güleç, and A. A. Salah, "The Turkish audio-visual bipolar disorder corpus," in *2018 First ACII Asia*. IEEE, 2018, pp. 1–6.

[20] H. Kaya, D. Fedotov, D. Dresvyanskiy, M. Doyran, D. Mamontov, M. Markitantov, A. A. Akdag Salah, E. Kavcar, A. Karpov, and A. A. Salah, "Predicting depression and emotions in the cross-roads of cultures, para-linguistics, and non-linguistics," in *Audio/Visual Emotion Challenge and Workshop*, ser. AVEC '19, 2019, p. 27–35.

[21] G.-B. Huang, H. Zhou, X. Ding, and R. Zhang, "Extreme Learning Machine for Regression and Multiclass Classification," *IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics*, vol. 42, no. 2, pp. 513–529, 2012.

[22] H. Kaya and A. Karpov, "Fusing acoustic feature representations for computational paralinguistics tasks," in *INTERSPEECH*, San Francisco, USA, 2016, pp. 2046–2050.

[23] H. Kaya and A. A. Karpov, "Introducing weighted kernel classifiers for handling imbalanced paralinguistic corpora: Snoring, addressee and cold," in *INTERSPEECH*, Stockholm, Sweden, 2017, pp. 3527–3531.

[24] G.-B. Huang, Q.-Y. Zhu, and C.-K. Siew, "Extreme learning machine: a new learning scheme of feedforward neural networks," in *Proc. IEEE Intl. Joint Conf. on Neural Networks*, vol. 2. IEEE, 2004, pp. 985–990.

[25] H. Kaya, A. A. Karpov, and A. A. Salah, "Robust acoustic emotion recognition based on cascaded normalization and extreme learning machines," in *13th International Symposium on Neural Networks - ISNN'16, LNCS 9719*. St. Petersburg, Russia: Springer, 2016, pp. 115–123. [Online]. Available: http://dx.doi.org/10.1007/978-3-319-40663-3_14

[26] J. Soltis, K. Leong, and A. Savage, "African elephant vocal communication ii: rumble variation reflects the individual identity and emotional state of callers," *Animal Behaviour*, vol. 70, no. 3, pp. 589–599, 2005.

[27] B. W. Schuller, A. Batliner, C. Bergler, C. Mascolo, J. Han, I. Lefter, H. Kaya, L. J. Rothkrantz, J. Zwerts, J. Treep, C. Kaandorp, and et al., "The INTERSPEECH 2021 Computational Paralinguistics Challenge: COVID-19 Cough, COVID-19 Speech, Escalation & Primates," in *INTERSPEECH*, Brno, Czechia, September 2021, p. 5 pages, to appear.