A Machine Learning Approach for Resource-efficient and Subject-independent Speech Based Stress Detection

1st Roswitha Duwenbeck
Medical Technology Systems
University of Duisburg-Essen
Duisburg, Germany
roswitha.duwenbeck@uni-due.de

2nd Elsa Andrea Kirchner
Medical Technology Systems & Robotics Innovation Center
University of Duisburg-Essen & German Research Center for Artificial Intelligence)
Duisburg & Bremen, Germany
elsa.kirchner@uni-due.de & elsa.kirchner@dfki.de

Abstract—**The aim of this paper is to identify resource-efficient stress recognition methods based on prosodic speech features. Different machine learning classifiers were tested on a two-class problem to distinguish stressed- from non-stressed speakers. This is done using subject-dependent and subject-independent evaluation methods. The resource efficiency was determined by measuring the virtual RAM usage of the prediction, memory consumption of the stored classifiers and classification speed. The best result, a recall of 83.3%, was obtained by Passive Aggressive- and Stochastic Gradient Descent Classifiers when a subject-dependent train-test split with balanced data was used. The results worsened by an average of 28.3 percentage points when the subject-independent leave-one-subject-out cross-validation was used, and by 22.2 points when an also subject-independent balanced GroupKFold evaluation was used. These effects could not be reduced by using Principal Component Analysis on the features, while the inclusion of speaker-specific examples in the previously subject-independent training set brought benefits. Inclusion of speaker-specific examples in the Leave-One-Out training set set led to an average improvement of 6.6 percentage points for three examples, 13.2 percentage points for six examples and even 19.7 percentage points for nine additional examples. The peak recall of train-test split could not be achieved even when nine speaker-specific examples were added. The best classifier in this case was the Passive Aggressive Classifier, with a recall of 79.8% and nine additional samples. In terms of resource efficiency, the RAM consumption per prediction hardly varies between classifiers and lies between 4.6659GB for Aggregated Mondrian Forest Classifier and 4.7853 for One-Vs-Rest SVM, with Dummy Classifier being left out. The space required to store a model is minimal for a Decision Tree, at only 0.0034 MB. The fastest classification is achieved by the Aggregated Mondrian Forest Classifier in 0.0025 seconds. The best preforming classifier in regards of resources, the Passive Aggressive Classifier, needed 0.0495MB for storage, 0.0802s and 4.7677GB virtual Memory for a classification.**

Index Terms—**machine learning, stress detection, stressed speech**

I. Introduction

The term “Stress” has manifold definitions to consider. The first, maybe best known, stems from Hans Seyle and states that “Stress is the non-specific response of the body to any demand” [1]. To alleviate understanding of this definition, it helps to divide types of stress. Therefore Hans Selye introduced the terms “distress”, meaning damaging or unpleasant stress, and “eustress”, meaning pleasurable, satisfying experiences [2]. Stress is studied increasingly and an identified source of health concerns. Barry et al. found that distress brought an increased mortality risk among those reporting high versus low distress [3]. Staufenbiel et al. even state: “The deleterious effects of chronic stress on health and its contribution to the development of mental illness attract broad attention worldwide” [4, p.1].

But there are not only health concerns, stress also impairs the ability to understand each other, as it affects language production. Beginning with increased muscle tension in the vocal cord and tract, the conversion of the linguistic program into neuromuscular commands and even spanning to speech production in the brain. This affects voice quality, and the performance of communication equipment. [5]

Furthermore there is a growing interest in the effects of stress on emotion-recognition, -expression and group dynamics in humans. Nitschke et al. state, that there is evidence of stress-spillover from stressed to unstressed individuals, that acute stress can block affective empathy and emotion contagion, but also that most studies find no effects of acute stress on measures of emotion recognition [6]. It is important to highlight, that acute stress might affect cognitive empathy different for men and women [6]. Paulman et al. showed, how listeners are worse in detecting negative emotions spoken by stressed speakers and also, how stressed listeners are worse at recognizing emotions from (non-stressed) speakers [7]. Van Marle et al. and Li et al. also proved an effect of stress, but on neural responses, to visual emotional stimuli [8] [9]. Overall, stress is not only a health issue, but can also affect effective communication and team dynamics by impairing speech production, empathy and emotion recognition. Therefore a robust and easy-to-use stress detection system could benefit many areas. Detecting stress in speech by using prosodic features could be an easy, non-invasive and inexpensive

This study was founded by the federal ministry for economic affairs and climate action, Grant number: 50RP2260A

TABLE I
State of the Art - Stressed Speech Recognition

Author	Dataset(s)	Best performing Classifier	Classes	Results
Dhole et al. [10]	Own	Custom Neuronal Network	5	97.52% Accuracy
Avila et al. [11]	SUSAS [12]	CNN, DNN	2, 4 and 9	72% Accuracy on Average
Hilmy et al. [13]	Own	Convoluational Neural Network	2	61% Accuracy
Partila et al. [14]	Own (Czech Speech Database)	SVM	2	87.9% Accuracy
Yao et al. [15]	Fujitsu Corporation owned [16]	GMM	2	71.88% Classification rate

solution. The aim of this paper is to examine the ability and resource consumption of different machine learning (ML) methods in recognizing stressed and non-stressed speakers. This should lay the foundation to implement a robust, subject-independent and resource-efficient stress detection method.

II. State of the Art

The detection of stress in speech is not a new topic in ML, so an overview of some results is presented in Table I. Only the best results are shown in this table, others are discussed in the text. Obtained results, used classifiers and the number of classes differ, so it is necessary to analyze their creation.

Dhole et al. [10] achieved a classification accuracy of 97.52% with a custom neural network. They classified five different types of stress: psychological/high workload and sentiments, perceptual/noise, physiological/medical illness, physical/vibration and physical workload, no stress. They used their own dataset, the German database of emotional speech (Emo-DB) [17] and the Toronto Emotional Speech Set (TESS) [18], but achieved the best classification result on their own dataset. Unfortunately, it is unclear how their dataset was created. EMO-DB and TESS are sets of emotional speech, so Dhole et al. had to sort the emotional speech files into stress categories, but there is no description of the methodology, which is also lacking in the case of their own dataset. Furthermore, they achieved the best result with a neural network (NN) and a classical train-test split, of which the division was not explained. Other classifiers such as Support Vector Machine (SVM) or Multilayer Perceptron (MLP) did not perform as well, with accuracies of 61.56% and 85.66% respectively. [10]

Avila et al. [11] used the SUSAS-databse (Speech Under Simulated and Actual Stress), which includes utterances with varying speaking styles (normal, slow, fast, soft, loud, question, clear, angry), different tasks (tracking tasks, motion-fear) and speech from psychiatric analyses [12]. They reached 72% accuracy on average with a Convolutional- (CNN) and a Deep Neuronal Network (DNN). This average was taken from all classification results, including different number of classes, with a 3-fold cross-validation and different feature sets. The CNN achieved the reported accuracy with an individually created feature set, while the DNN used the Interspeech 2010 feature set [19]. Other tests were done with a SVM where the best average accuracy was obtained with an OpenSmile feature set and measured 61%. [11]

Hilmy et al. [13] tested their methods on an own dataset, which was created by interviewing university students and taking their Perceived-Stress-Scales (PSS) [20]. With the PSS audiofiles where sorted as “Stressed” and “Not Stressed”. A train-test split with a division of 75% and 25% was used. They obtained an accuracy of 61% with a CNN. [13]

Partila et al. [14] created their own set of stressed speech, which consists of emergency calls. Callers are automatically sorted as “Stressed”, while the receiving operators are sorted to “No Stress”. They used a train-test split of 75% for training and 25% of testing. A SVM reached the best accuracy with 87.9%, directly followed by CNN with 87.5%. [14]

Yao et al. [15] used an dataset owned by the Fujitsu Corporation [16]. It is comprised of phone calls, in which the callers had to perform different tasks while taking the call, which included concentration, time pressure and risk taking. Tested were Gaussian Mixture Models (GMM) with different numbers of mixture and an own feature set, on a 4-Fold cross-validation. They obtained the best results with four Mixtures and got 71.88% accuracy.

All in all, retained results from the Table I seem promising at first sight, but much has to be done for implementing a real world solution. A successful approach would not only need a ground-truth with a comprehensible method of creation, but also have a good or at least known subject-independent performance. Furthermore, a solution which could run in resource restricted environments would be beneficial, as this could make the usage on low priced devices possible and therefore benefit a broader range of communities.

III. Methodology

The aim of this paper is to take the first steps towards a solution that could be applied in the real world. For this, several ML classifiers have been trained on a dataset of stressed speech. The retained models were evaluated both subject dependent and subject independent. The models' memory consumption, prediction speed and usage of virtual RAM are also taken into account. This chapter presents not only the dataset used, but also the preprocessing steps, the classifiers used and the evaluation methods.

A. Dataset

The used dataset was created by Paulmann et al., to show how induced psychological stress affects the production and recognition of vocal emotions [7]. The set consists of nine female speakers, who were given 15 predefined, neutral sentences. The sentences had to be read in an emotional tone,

TABLE II
Stressed Speech Recognition - Evaluation Results (Recall)

Recall	Dummy	MLP	PA	SGD	SVM-R	GNB	DT	RF	SVM-O	HAT	HT	AMF	ARF
TTS	50.0	81.0	83.3	83.3	78.6	71.4	78.6	69.0	78.6	73.8	71.4	35.7	45.2
GroupK	42.9	53.9	64.1	60.8	48.8	45.5	31.6	46.6	48.8	44.7	45.7	39.5	54.0
LOSOCV	36.5	43.7	59.5	59.1	34.0	43.9	45.9	36.3	34.0	46.2	44.7	24.4	38.1
LOSOCV + 3	40.6	56.7	66.5	66.1	41.3	48.7	47.8	43.9	41.3	47.7	48.9	41.6	38.0
LOSOCV + 6	39.5	61.6	70.2	67.9	51.2	53.4	54.0	47.0	51.2	51.3	53.6	55.3	51.1
LOSOCV + 9	48.5	75.2	79.8	79.0	58.1	56.9	55.4	54.0	58.1	56.8	56.4	53.0	63.3
LOSOCV + PCA 1	41.4	34.6	39.3	32.3	12.6	44.2	43.4	35.7	12.6	0.0	0.0	33.1	40.9
LOSOCV + PCA 2	47.4	31.6	53.9	38.8	9.3	46.1	42.7	38.3	9.3	11.1	0.0	31.6	32.20

TABLE III
Stressed Speech Recognition - Evaluation Results (Balanced Accuracy)

Bal. Acc.	Dummy	MLP	PA	SGD	SVM-R	GNB	DT	RF	SVM-O	HAT	HT	AMF	ARF
TTS	47.6%	84.5%	89.3%	86.9%	76.2%	70.2%	70.2%	83.3%	76.2%	69.0%	69.0%	50%	60.7%

namely angry, disgusted, fearful, happy, pleasantly surprised, sad, or neutral. In order to achieve an emotional tone, the speakers were asked to imagine a situation in which they felt the given emotion. Prior to the reading task, the Trier Social Stress Test [21] was used to induce stress in the speakers. Five speakers were randomly allocated into the “Stressed”-Group, four into “No Stress”. Therefore, all sentences spoken by a speaker are produced either in a stressed or unstressed state. The number of spoken sentences ranges from 13 to 47, as can be seen in Figure 1 and totals 280. The number of unstressed and stressed samples is balanced, with 140 stressed- and 140 unstressed sentences. The success of the stress induction was checked by asking the participants to indicate their stress level before the start, immediately after the Trier Social Stress Test and after reading the emotional sentences. [7]

B. *Preprocessing*

Preprocessing steps consist of scaling, as only the features were provided, and Principal Component Analysis (PCA). The feature extraction was done by Silke Paulmann from the raw audio files, which were recorded for the beforehand mentioned dataset. The Compare_2016-feature set from OpenSmile-Toolbox [22] was chosen. It consists of energy, spectral, cepstral (MFCC) and voicing related low-level descriptors (LLDs), also logarithmic harmonic-to-noise ratio (HNR), spectral harmonicity, and psycho-acoustic spectral sharpness [23]. It combines 6373 features and is optimized regarding pitch, jitter extraction and computation of parameter ratios [24]. An example of features extracted from stressed and unstressed speech is shown in Figure 2. For Figure 2, all features were normalized between -1 and 1. To train and test the models label vectors were extracted from the data, which allowed the models to be classified as “stressed” and “unstressed”. This work therefore shows a two-class problem. Emotions have not yet been analyzed or classified. Scaling was done using the Standard Scaler from scikit-learn [25], after the data was parted into training and testing data. This was also the case for PCA. PCA was not done with every evaluation,

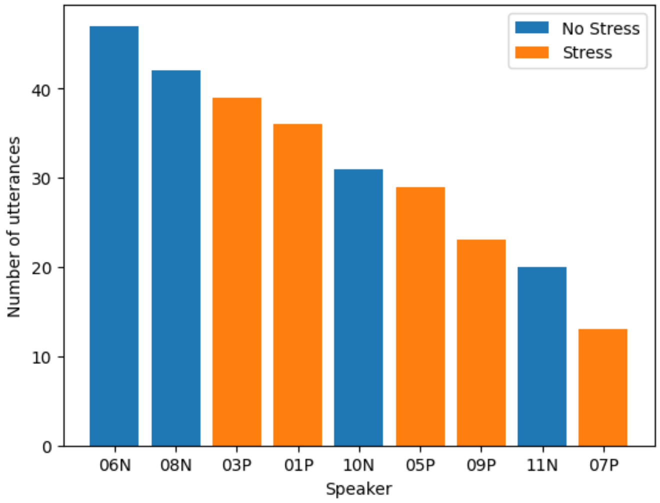

Fig. 1. Number of utterances per speaker

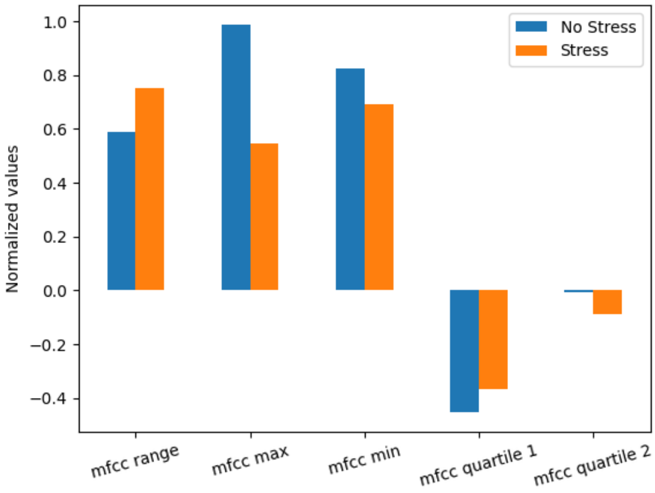

Fig. 2. Features of an utterance with stress and without stress

TABLE IV
STRESSED SPEECH RECOGNITION - STANDARD DEVIATIONS

Standard deviation	Dummy	MLP	PA	SGD	SVM-R	GNB	DT	RF	SVM-O	HAT	HT	AMF	ARF
TTS	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
GroupK	15.2	20.2	8.9	11.7	43.3	29.8	19.0	32.5	43.3	23.1	29.0	37.2	45.0
LOSOCV	8.0	30.9	30.2	31.7	27.2	19.7	22.8	35.2	27.2	20.0	20.6	9.2	24.5
LOSOCV + 3	10.5	28.7	27.9	22.8	28.8	19.5	13.0	41.6	28.8	22.9	19.0	12.8	15.2
LOSOCV + 6	12.1	23.1	21.5	19.7	32.0	24.6	18.5	38.5	32.0	27.3	25.1	23.0	16.3
LOSOCV + 9	17.3	20.4	22.4	26.1	31.5	25.9	15.2	36.6	31.5	27.6	27.9	10.0	17.9
LOSOCV + PCA 1	14.4	22.2	23.3	23.9	16.5	33.6	10.6	9.4	16.5	0.0	0.0	9.5	22.8
LOSOCV + PCA 2	10.0	20.7	20.0	26.3	13.6	34.9	7.7	11.0	13.6	33.3	0.0	15.3	16.5

TABLE V
STRESSED SPEECH RECOGNITION - EVALUATION RESULTS (MEMORY USAGE IN MB)

Dummy	MLP	PA	SGD	SVM-R	SVM-O	GNB	RF	DT	HT	HAT	AMF	ARF
0.0005	14.596	0.0495	0.0495	11.9836	11.9834	0.1951	0.3368	0.0034	47.8473	47.8533	299.2545	24.8498

only in two cases. The decompositional PCA from scikit-learn [25] was chosen. While the first PCA reduced the feature set to 3000 features, the second one reduced it to 1500.

C. Machine Learning Methods

Used ML Methods are from scikit-learn [25] and river [26]. Taken from scikit-learn were Dummy Classifier (DC), Multilayer Perceptron (MLP), Passive Aggressive Classifier (PA), Stochastic Gradient Descent Classifier (SGD), two Support Vector Machines with C-Support and different Ensemble-Methods, which were One-Vs.-One and One-Vs-Rest (SVM-O, SVM-R), a Gaussian Naive-Bayes Classificator (GNB), Random Forest (RF) and Decision Tree (DT). As these Methods use the "-fit"-method for training, their input vectors were either size 280x6373, 280x3000, or 280x1500, depending on whether PCA was used. From River the Hoeffding Tree (HT), Hoeffding Adaptive Tree (HAT), Aggregated Random Forest (ARF) and Aggregated Mondrian Forest Classifiers (AMF) were used. As river classifiers are trained example by example their input vectors were size 1x6373, 1x3000, or 1x1500, depending on whether PCA was used. All classifiers were used with their default settings, except for the DC, which was used with the "stratified" prediction strategy because of the sometimes imbalanced testing data.

D. Evaluation

Different evaluation strategies were used to test robustness and estimate suitability for usage in the real world, with the first one being a subject-dependent train-test split (TTS) and the second one a nested, subject-independent leave-one-subject-out cross validation (LOSOCV) without finetuning. In the TTS 30% of the samples were reserved for testing and 70% for training. Training and testing data was stratified, to avoid bias. After that nested LOSOCV was done, for a subject-independent testing approach.

As a subject-independent approach produces unbalanced data, a GroupKFold evaluation was also performed. The data for GroupKFold was split into three parts, with two speakers in the test set and four in the training set to avoid subject-dependency. The speakers were grouped so that there was always one stressed and one relaxed speaker in the test group, also taking the number of utterances into account. One speaker, namely 07P, was excluded from this approach. Recall was chosen as the evaluation metric for almost all predictions due to the highly unbalanced LOSOCV data. As speakers are either stressed or not, the subject-independent test data contains only one class and there are no true negatives or false positives.
Later, training data from LOSOCV was induced with varying numbers of statements from the test subject to simulate calibration. First, three extra samples were included in the training (LOSOCV + 3), then six (LOSOCV + 6) and finally nine (LOSOCV + 9). These samples were omitted from the test set.
In a final approach, a normal LOSOCV was performed with a feature set reduced py PCA. In the first PCA the number of features was reduced to 3000, in the second to 1500.
The models' memory consumption, virtual RAM usage and the time required to classify a sample were also taken into account. The basis for calculating these values are the models created with the first LOSOCV. To calculate the speed and RAM usage, the models were fed with all samples and the average of all values obtained was taken across all classifiers of a type. All prediction speed and RAM usage calculations were performed on a Raspberry Pi 4 Model B with 8GB of RAM, and all other tasks were performed on a ThinkPad P15 Gen2.

IV. RESULTS AND DISCUSSION

In this sub-chapter results are shown in Tables II to VII and discussed. All values shown in Table II are means of all cross-validations, except for the TTS evaluation. The values shown in IV represent the standard deviation of the classification recalls between all test folds. Therefore, there can be no standard deviation in the first row for TTS. It can be seen that the best results are obtained by PA and SGD on a TTS, with a recall of 83.3%. Since PA has a better balanced accuracy (89.3%), this classifier outperforms the other methods tested,

TABLE VI
STRESSED SPEECH RECOGNITION - EVALUATION RESULTS (PREDICTION SPEED IN SECONDS)

Dummy	MLP	PA	SGD	SVM-R	SVM-O	GNB	RF	DT	HT	HAT	AMF	ARF
0.0067	0.0902	0.0802	0.0670	0.0620	0.0583	0.0058	0.0736	0.0088	0.1026	0.0620	0.0025	0.0127

TABLE VII
STRESSED SPEECH RECOGNITION - EVALUATION RESULTS (RAM USAGE IN GB)

Dummy	MLP	PA	SGD	SVM-R	SVM-O	GNB	RF	DT	HT	HAT	AMF	ARF
4.7150	4.7609	4.7677	4.7781	4.7853	4.7820	4.7125	4.7734	4.7191	4.7521	4.7354	4.6659	4.7273

while SGD comes second with a balanced accuracy of 86.9%. As a TTS does not necessarily resemble real usage, LOSOCV and GroupKFold were used. The results are worse here, with GroupKFold reducing recalls by an average of 22.2 percentage points. The DC was not included in these average calculations. When LOSOCV was used, the average drop was 28.3 percentage points. Also the performance loss with GroupKFold appears to be very sharp. The best classifiers with these evaluation methods were PA with 64.1% recall in GroupKFold and 59.1% recall in LOSOCV.

After observing these results, it was theorized that a reduced feature set or pre-training with subject-specific examples might be useful. Starting with the pre-training, an average improvement of 6.6 percentage points can be seen, compared to the classical LOSOCV, with only 3 subject-specific examples. Performance improved in all cases when more domain-specific training examples were used, on average by 13.2 percentage points with 6 additional domain-specific examples and by 19.7 percentage points with 9 additional examples. Even with nine subject-specific training samples, classification results were worse than with classical TTS. The loss of recall between TTS and training with nine extra samples ranged between 8.7 and 21.8 percentage points. Only AMF and ARF performed better with LOSOCV and extra samples than with TTS. With only six subject-specific extra examples their performance improved by 19.6 and 5.9 percentage points, with nine subject-specific examples by 17.3 and 18.1 percentage points respectively.

PCA did not improve the original LOSOCV results, as with the first LOSOCV the prediction recalls were generally worse, sometimes even dropping to 0% recall. For the first PCA, where the features were reduced to 3000, recalls of the results dropped between 2.6 and 73.8 percentage points compared to TTS. For the second PCA results lost between 2.6 and 71.4 percentage points.

Regarding the memory consumption for storing the classifiers, presented in Table V, the methods from the river library are more expensive, while the classifiers from scikit-learn are leaner. AMF uses an exceptionally large amount of memory with an average of 299.2545 MB, the leanest method is the DT with an average of only 0.0034 MB. This behavior can not be translated to processing speed, presented in Table VI. The river classifiers, especially the AMF, exceed expectations with an average speed of 0.0025s per sample, the slowest model is the MLP with 0.0902s. The average virtual RAM usage of computing one classification is shown in Table VII and does not vary much. Values are within a range of 4.6 and 4.8 GB of RAM.

The balanced accuracies of TTS can be compared with the state of the art. The performance in this paper is similar to that of Partila et al. [14], beating the results of Yao et al. [15] and Hilmy et al. [13]. A comparison with Avila et al. [11] is difficult, because this paper used only two classes, while Avila et al. used 2, 4 and 9, and no deep learning methods were tested in this paper. A result that can be compared is the two class discrimination with SVM and OpenSmile by Avila et al. where they achieved 68% accuracy [11], which is worse than the results obtained in this paper. The classification accuracy of Dhole et al. is still better than the results of this paper. As they achieved their good accuracy using a custom neural network, their performance could be explained by the advantage of deep learning.

Several observations can be made regarding potential real-world use. Firstly, subject-independent testing ensures a sharp reduction in the quality of the performance metrics obtained. This behavior suggests that use with completely unknown data is not yet realistic. Secondly, the training of test data with patient-specific samples improves classification recalls. This suggests that calibration of classifiers may be a way to provide a higher standard. Thirdly, reducing features with PCA does not improve performance and suggests that the most important features for stress prediction from speech may not even be in the set. Finally, while none of the classifiers produced prediction speeds that would be a hindrance in a conversational or diagnostic setting, the relatively high RAM consumption of all the classifiers may preclude their use in smaller devices and the memory consumption for storing the classifiers varies widely and may also be a bottleneck for use in small devices. In general, the PA seems to be the most promising classifier at the moment, with comparably high performance for both subject-dependent and -independent methods, low memory usage for the stored model, RAM usage similar to other models, and only prediction speed at the lower end of performance, but not in a way that would hinder a diagnostic process or conversation.

V. Conclusion and Outlook

This paper presents an approach to resource-efficient stress detection methods based on speech. This could pave the way for making stress detection accessible in the most remote or disadvantaged communities. To achieve this goal, different ML classifiers were trained and tested, starting with a classical TTS and moving to subject-independent scenarios to get more realistic results. Resource-related parameters such as prediction speed, RAM usage and memory consumption when saving the model were also monitored.

PA outperformed most of the researched state-of-the-art methods in the classical train-test split, but could not beat the custom neural network of Dhole et al. [10]. Subject-independent scoring methods reduced the performance of the tested classifiers, and even calibrating with samples from the test subjects improved classification results only slowly. Reducing the number of features with PCA did not increase classification recalls in most cases, suggesting that the most important features for stress detection may not be in the used feature set. Results suggest that stress can be detected in speech without deep networks, but the performance of the methods still needs improvement. A larger dataset might improve classification, as might the use of stress classifiers that are particular to certain emotional subgroups, i.e. using emotion recognition beforehand. Although Paulmann et al. describe that mean pitch and mean amplitude are good predictors for estimating speaker stress levels, they also write that such features do not always seem to tend in the same direction [7]. Rather, features behave differently across sets of emotions, but express more similar patterns in these sets [7]. It can therefore be theorized that the expression of stress is not only highly individual but also dependent on the emotion expressed. Planned future actions in this area of research are therefore to use emotion recognition before or together with stress classification, but also feature importance analysis.

References

[1] H. Selye, *Stress in health and disease*. Butterworth-Heinemann, 1976.

[2] ——, "Stress without distress, mcclelland and steward ltd," *Toronto. Canada*, 1974.

[3] V. Barry, M. E. Stout, M. E. Lynch, S. Mattis, D. Q. Tran, A. Antun, M. J. Ribeiro, S. F. Stein, and C. L. Kempton, "The effect of psychological distress on health outcomes: A systematic review and meta-analysis of prospective studies," *Journal of Health Psychology*, vol. 25, no. 2, pp. 227–239, 2020.

[4] S. M. Staufenbiel, B. W. Penninx, A. T. Spijker, B. M. Elzinga, and E. F. van Rossum, "Hair cortisol, stress exposure, and mental health in humans: a systematic review," *Psychoneuroendocrinology*, vol. 38, no. 8, pp. 1220–1235, 2013.

[5] H. J. Steeneken and J. H. Hansen, "Speech under stress conditions: overview of the effect on speech production and on system performance," in *1999 IEEE International Conference on Acoustics, Speech, and Signal Processing. Proceedings. ICASSP99 (Cat. No. 99CH36258)*, vol. 4. IEEE, 1999, pp. 2079–2082.

[6] J. P. Nitschke and J. A. Bartz, "The association between acute stress & empathy: A systematic literature review," *Neuroscience & Biobehavioral Reviews*, vol. 144, p. 105003, 2023.

[7] S. Paulmann, D. Furnes, A. M. Bøkenes, and P. J. Cozzolino, "How psychological stress affects emotional prosody," *Plos one*, vol. 11, no. 11, p. e0165022, 2016.

[8] H. J. Van Marle, E. J. Hermans, S. Qin, and G. Fernández, "From specificity to sensitivity: how acute stress affects amygdala processing of biologically salient stimuli," *Biological psychiatry*, vol. 66, no. 7, pp. 649–655, 2009.

[9] S. Li, R. Weerda, C. Milde, O. T. Wolf, and C. M. Thiel, "Effects of acute psychosocial stress on neural activity to emotional and neutral faces in a face recognition memory paradigm," *Brain Imaging and Behavior*, vol. 8, pp. 598–610, 2014.

[10] N. P. Dhole and S. N. Kale, "Stress detection in speech signal using machine learning and ai," in *Machine Learning and Information Processing: Proceedings of ICMLIP 2019*. Springer, 2020, pp. 11–26.

[11] A. R. Avila, S. R. Kshirsagar, A. Tiwari, D. Lafond, D. O'Shaughnessy, and T. H. Falk, "Speech-based stress classification based on modulation spectral features and convolutional neural networks," in *2019 27th European Signal Processing Conference (EUSIPCO)*. IEEE, 2019, pp. 1–5.

[12] J. H. Hansen and S. E. Bou-Ghazale, "Getting started with susas: A speech under simulated and actual stress database."

[13] M. S. Hafiy Hilmy, A. Liza Asnawi, A. Z. Jusoh, K. Abdullah, S. N. Ibrahim, H. Adibah Mohd Ramli, and N. F. Mohamed Azmin, "Stress classification based on speech analysis of mfcc feature via machine learning," in *2021 8th International Conference on Computer and Communication Engineering (ICCCE)*, 2021, pp. 339–343.

[14] P. Partila, J. Tovarek, J. Rozhon, and J. Jalowiczor, "Human stress detection from the speech in danger situation," in *Mobile Multimedia/Image Processing, Security, and Applications 2019*, vol. 10993. SPIE, 2019, pp. 179–185.

[15] X. Yao, T. Jitsuhiro, C. Miyajima, N. Kitaoka, and K. Takeda, "Classification of speech under stress based on modeling of the vocal folds and vocal tract," *EURASIP Journal on Audio, Speech, and Music Processing*, vol. 2013, pp. 1–17, 2013.

[16] I. R. Titze and B. H. Story, "Acoustic interactions of the voice source with the lower vocal tract," *The Journal of the Acoustical Society of America*, vol. 101, no. 4, pp. 2234–2243, 1997.

[17] F. Burkhardt, A. Paeschke, M. Rolfes, W. F. Sendlmeier, B. Weiss *et al.*, "A database of german emotional speech." in *Interspeech*, vol. 5, 2005, pp. 1517–1520.

[18] M. K. Pichora-Fuller and K. Dupuis, "Toronto emotional speech set (TESS)," 2020. [Online]. Available: https://doi.org/10.5683/SP2/E8H2MF

[19] B. Schuller, S. Steidl, A. Batliner, F. Burkhardt, L. Devillers, C. Müller, and S. Narayanan, "The interspeech 2010 paralinguistic challenge," in *Proc. INTERSPEECH 2010, Makuhari, Japan*, 2010, pp. 2794–2797.

[20] S. Cohen, T. Kamarck, R. Mermelstein *et al.*, "Perceived stress scale," *Measuring stress: A guide for health and social scientists*, vol. 10, no. 2, pp. 1–2, 1994.

[21] C. Kirschbaum, K.-M. Pirke, and D. 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.

[22] F. Eyben, M. Wöllmer, and B. Schuller, "Opensmile: the munich versatile and fast open-source audio feature extractor," in *Proceedings of the 18th ACM International Conference on Multimedia*, ser. MM '10. New York, NY, USA: Association for Computing Machinery, 2010, p. 1459–1462. [Online]. Available: https://doi.org/10.1145/1873951.1874246

[23] B. Schuller, S. Steidl, A. Batliner, A. Vinciarelli, K. Scherer, F. Ringeval, M. Chetouani, F. Weninger, F. Eyben, E. Marchi *et al.*, "The interspeech 2013 computational paralinguistics challenge: Social signals, conflict, emotion, autism," in *Proceedings INTERSPEECH 2013, 14th Annual Conference of the International Speech Communication Association, Lyon, France*, 2013.

[24] F. Eyben, M. Unfried, G. Hagerer, and B. Schuller, "Automatic multilingual arousal detection from voice applied to real product testing applications," in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2017, pp. 5155–5159.

[25] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg *et al.*, "Scikit-learn: Machine learning in python," *the Journal of machine Learning research*, vol. 12, pp. 2825–2830, 2011.

[26] J. Montiel, M. Halford, S. M. Mastelini, G. Bolmier, R. Sourty, R. Vaysse, A. Zouitine, H. M. Gomes, J. Read, T. Abdessalem *et al.*, "River: machine learning for streaming data in python," 2021.