Investigation and ordinal modelling of vocal features for stress detection in speech

George-Marios Kalatzantonakis-Jullien
School of Science & Technology
Hellenic Open University
Patras, Greece
std507567@ac.eap.gr

Nikolaos Stefanakis
Department of Music Technology and Acoustics
Hellenic Mediterranean University
Rethymno, Greece, GR-74100
nstefana@hmu.gr
Institute of Computer Science, Foundation for Research and Technology Hellas (FORTH), Heraklion, Greece

Giorgos Giannakakis
Institute of Computer Science
Foundation for Research and Technology Hellas (FORTH)
Heraklion, Greece
ggian@ics.forth.gr
School of Science and Technology, Hellenic Open University, Patras, Greece
Institute of AgriFood and Life Sciences
University Research Centre, Hellenic Mediterranean University, Heraklion, Greece

***Abstract*—This paper investigates a robust and effective automatic stress detection model based on human vocal features. Our study experimental dataset contains the voices of 58 Greek-speaking participants (24 male, 34 female, 26.9±4.8 years old), both in neutral and stressed conditions. We extracted a total of 76 speech-derived features after extensive study of the relevant literature. We investigated and selected the most robust features using automatic feature selection methods, comparing multiple feature ranking methods (such as RFE, mRMR, stepwise fit) to assess their pattern across gender & experimental phase factors. Then, classification was performed both for the entire dataset, and then for each experimental task, for both genders combined and separately. The performance was evaluated using 10-fold cross-validation on the speakers. Our analysis achieved a best classification accuracy of 84.8% using linear SVM for the social exposure phase and 74.5% for the mental tasks phase using the gaussian SVM classifier. The ordinal modelling improved significantly our results, yielding a best on-subject basis 10-fold cross-validation classification accuracy of 95.0% for social exposure using gaussian SVM and 85.9% for mental tasks using the gaussian SVM. From our analysis, specific vocal features were identified as being robust and relevant to stress along with parameters to construct the stress model. However, it is was observed the susceptibility of speech to bias and masking and thus the need for universal speech markers for stress detection.**

***Index Terms*—stress, voice, speech, pairwise transformation, Mel cepstral coefficients, emotion recognition, affective computing, biosignals, feature selection, mRMR, hyperparameter optimization**

I. Introduction

Speech is the main medium of human verbal communication, able to express linguistic (words) and paralinguistic information such as thoughts, ideas, moods and emotions in everyday life. The vocal behaviour parameters reflect the affective state as emotional changes activate physiological processes in the central and peripheral nervous system which, in turn, modulates the voice production process.

The mechanism of human speech production is described in [1] [2] and is presented in Figure 1. The vocal characteristics can be categorized into three components, namely *speech excitation* (source), *vocal tract* (filter-system) and *speech signal* (output). The emotional physiological response affects the first component by increasing tension in the vocal fold muscles, the second component by changing of the vocal tract articulators' position and, consequently, the third component due to its linkage with the other two components.

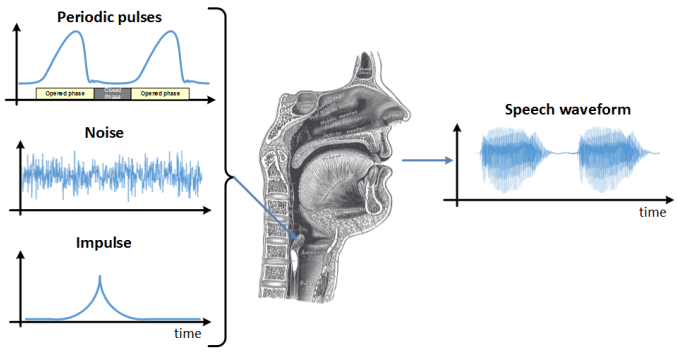

Fig. 1. Representation of the human speech production mechanism. It can be considered as a system with 3 possible inputs (periodic pulses or noise or impulse). These are modulated in the vocal tract by the articulators (system) producing the output speech waveform. Figure taken from [2].

Subsequently, stress conditions may cause variations to speech characteristics in relation to speech in neutral condi-

tions [3]. A scientific area called Voice Stress Analysis (VSA) has been established to differentiate normal from stressed voice signals by analyzing nonverbal aspects of speech such as intonation, voice quality, prosody, rhythm, and timing (pausing) [4].

The aim of this study is the identification of the most effective voice-derived features on stress detection and the construction of a reliable computational model for the anticipation of stress-related states.

The research topics investigated in this study can be summarized in the following aspects

1) identify the most robust and relevant voice features that efficiently represent stress state and discriminate control and stress states.
2) investigate the involvement of each voice feature in the stress model.
3) investigate the features' importance differences between genders.
4) investigate the construction of a stress model,able to anticipate stress states with efficient discriminatory ability.

II. Related work

In [5], an extensive review of theoretical and empirical models on human speech was provided, while in [6] the effects of stress on human voice were investigated.

The vocal fundamental frequency f_0 is a commonly used stress indicator. A great number of studies have been used it as a feature for stress detection, many of them concluding that stress conditions lead to an increase of the f_0 [1]. In life-threatening situations, the pitch increases were found to be the highest, often redoubled [7]. Also, there is a significant pitch increase during Lombard speech [8]. There are cases though where an f_0 decrease has been observed, for instance, [9] noted an increase in f_0 in 60.8% and a decrease in 13.6% of the speakers under physical stress, while 25.5% shown no change. Applying Linear Discriminant Analysis (LDA) in conversations from emergency call center databases [10] observed a systematic increase in pitch (more than one octave), as well as a shift in the f_0 contour. Their classification accuracy was 80% (84% for stressed males) and concluded that mean f_0, min f_0, and f_0 variation are, among others, significant stress indicators.

A very commonly used set of features for human speech analysis and recognition, Mel-Frequency Cepstral Coefficients (MFCC), were first introduced by [11]. MFCCs are a popular way to represent sound in a parametric way, and they are used as features in speech recognition as described in [12]. Essentially, they comprise a representation defined as the real cepstrum of a windowed short-time signal derived from the FFT of that signal [13]. MFCCs also seem to correlate with fatigue and sleep deprivation [7]. In [14], it was showed that their 3rd order polynomial SVM classifier performed better than humans in classification accuracy by 44.42% with a total accuracy of 86% in a cross-language (Chinese & English native speakers) and cross-gender study. They used 560 acoustic features MFCC & TEO (pitch considered to be less important). The differences across languages were very small, whereas the accuracy across genders decreased by 28.18%. In [15] found f_0 not to be a good indicator of stress when used on its own, though it improved a little the classification accuracy when combined with MFCC features with SVM (92.6 ± 1.6% over 92.4 ± 0.6%).

In the audio spectrum, the amplitudes decrease as the frequencies increase, so its shape tilts [16]. This spectral slope is approximated by linear regression [17] and has been used by some studies as a discriminatory feature for stress detection [18]. [19] consider the glottal structure to be subjective, less observable, but still measurable. [20] extracted the slope of the glottal source signal (more specifically, the three most probable slopes of speech spectrum, along the corresponding glottal slopes), as well the pitch standard deviation, and achieved an accuracy score of 92.06% using Random Forest classifier. Slope features were found to contribute more than f_0 and their main advantage is that they are not as easy to manipulate, and thus the subjects cannot easily conceal their stress [20].

Vowel production is affected by psychological disorders such as depression, PTSD, and suicidality by affecting speech production mechanisms (for example, increasing tension in vocal cords and vocal tract) [21]. Studies show that the f_1/f_2 area, defining the vowel space, decreases as the stress decreases [22] [23], while some other studies find less significant vowel change during stress [24] [25]. In a study on Daxi Hakk Chinese speakers, it was found that on stressed conditions there was a significant definite expansion of vowel space by increasing the f_2 for front vowels (especially /i/) and decreasing the f_2 for back vowels (especially /o/) [26]. The results in [27] show that mean f_1 (for /a/ /i/ and /u/) was significantly increased for highly aroused emotions, while at the same time mean f_2 is significantly reduced (for /a/). In the same study, they also pointed out that positive emotions (amusement, relief, pride, interest) resulted in higher mean f_2 than negative emotions (anxiety, fear, sadness, despair).

In [28], both speech and gestures features were utilized, in order to recognize stress states by the modulation of either speech and gestures (e.g. intonation for speech, speed and rhythm for gestures). In [29], pitch, intensity, formants, long term averaged spectral features and spectral features were selected for the automatic stress detection in emergency (telephone) calls leading to minimizing error rate to 4.2%.

III. Experimental Dataset and acquisition protocols

A. Acquisition protocol for stress recognition

An experimental protocol was designed and conducted to investigate vocal characteristics in stress conditions. Each of the participants was seated in front of a computer monitor which presented the stressful stimuli. Voices were recorded with a microphone attached to the subject's chest.

The experiment included neutral tasks (used as reference) and stressful tasks in which stress conditions were simulated and induced employing different types of stressors. These stressors were categorized into 4 different phases: *social*

exposure, *emotional recall*, *mental workload tasks*, *stressful videos presentation*. The experimental tasks along with their duration and affective state are presented in Table I.

TABLE I
EXPERIMENTAL TASKS EMPLOYED IN THIS STUDY

#	Experimental task	Duration (min)	Affective State
Social Exposure			
1	1.1 Neutral (reference)	1	N
2	1.2 Baseline Description	2	N
3	1.3 Interview	2	S
Emotional recall			
4	2.1 Neutral (reference)	2	N
5	2.2 Recall stressful event	2	S
Mental Tasks			
6	3.1 Reading letters/numbers (reference)	2	N
7	3.2 Stroop Colour-Word Test (SCWT)	2	S
8	3.3 PASAT Task	2	S
Stressful videos			
9	4.1 Calming video	2	R
10	4.2 Adventure video	2	S
11	4.3 Psychological pressure video	2	S

Note: Intended affective state N:neutral, S:stress, R:relaxed)

The phases, as shown in the table, were "Social exposure" (2 neutral and 1 stress task), "Emotional recall" (1 neutral and 1 stress task), "Mental tasks" (1 neutral and 2 stress tasks), and "Stressful stimuli" (1 relaxed and 2 stress tasks). In total there were 4 neutral states, 1 relaxed state, and 6 stressed states. Each phase started with a neutral/relaxed condition, acting as a baseline for the stressful tasks that followed.

The first phase (social exposure), included an interview where the participant was asked to describe himself.

In the second phase (emotional recall) participants were asked to recall in their memory and relive a traumatic stressful situation of their past life and act as it was happening this time.

The third phase (mental tasks), assessed the cognitive load of the participants with tasks that require mental effort. The first task was the Stroop color-word test (SCWT) [30], in which the participant is asked to read a color name inked with dissimilar color (ex. the word "BLACK" inked with red color). The second task was Paced Auditory Serial Addition Test (PASAT) [31], an arithmetic operation test for attentional processing assessment.

The fourth phase (stressful stimuli), included the presentation of 2-min video clips, with calming content for the induction of a relaxed state and intense content (adventure and action scenes involving heights, scenes with burglary and car accidents) for the induction of a stressed state.

B. Study dataset

The experimental dataset used in this study is described in [32]. The experiment took place at the premises of the FORTH research institute. It contains the voices of 58 Greek-speaking volunteer participants (34 females, 24 males) whose ages at the time of the recording were 26.9±4.8 years old. Each participant performed 11 total tasks (4 neutral, 6 stressed and 1 relaxed states) as shown in Table I. The study was approved by the FORTH Ethics Committee (FEC). All participants provided informed consent.

A neutral condition was presented at the beginning of each phase of the experiment which was used as a baseline for the subsequent stressful tasks. For the voice experiments of the current study, only the first (social exposure) and the third (mental tasks) phases were used as the other phases don't contain participants voices. This means, that for the social exposure, the tasks 2 (neutral) and 3 (stress condition) and for the mental tasks, the tasks 6 (neutral) and 7,8 (stress conditions) were used.

IV. METHODS

A. Speech signal preprocessing

Speech segments in the audio recordings were manually detected and isolated from silent segments using the Praat software [33] (see Fig. 2). All feature values calculated along a speech segment were averaged to produce a single feature vector for each segment.

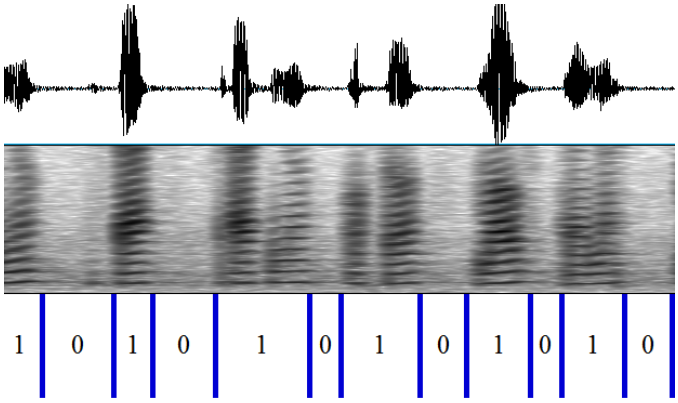

Fig. 2. Speech segments in each audio recording were manually isolated from silent segments using Praat.

B. Vocal features extraction

After the preprocessing and the voiced section segmentation, vocal features were extracted included prosodic, glottal source, and spectral envelope features. Prosodic features are the fundamental frequency (f_0), formants (f_1-f_5), Voice activity detection (VAD) [34], and the voiced/unvoiced boundaries (VUV). Voice quality features include Normalized amplitude quotient (NAQ), Quasi-open quotient (QOQ), the ratio of the first two harmonics of the glottal source spectrum (H1/H2), the parabolic spectral parameter (PSP), the estimation of maxima dispersion quotient (MDQ), the spectral slope (peakslope) and the shape parameter of the Liljencrants-Fant model of the glottal pulse dynamics (R_d). In total, 80 vocal features were extracted, and they are presented in Table II.

C. Feature selection

After constructing the feature matrix, we perform a set of feature selection algorithms to determine the most relevant

TABLE II
SPEECH-DERIVED FEATURES EXTRACTED AND USED IN THIS STUDY

Feature	Description
f_0	Fundamental frequency
f_1–f_5	Formants f_1–f_5
VUV	Voiced / unvoiced (voicing boundaries)
VAD	Voice Analysis Detection
NAQ	Normalized amplitude quotient
QOQ	Quasi-open quotient
H1/H2	The first 2 harmonics ratio
PSP	Parabolic spectral parameter
MDQ	Maxima dispersion quotient
Peak Slope	wavelet response spectral tilt/slope
R_d	Estimation of the shape parameter of the Liljencrants-Fant (LF) glottal model
creak	Creaky voice detection algorithm
MCEP (0-24)	Mel cepstral coefficients (MCEP0,MCEP1,..MCEP24) (25 features)
HMPDM (0-24)	Harmonic model and phase distortion mean
HMPDD (0-12)	Harmonic model and phase distortion deviations

features for the detection task. The feature selection is carried out by minimizing the classification error of an SVM classifier using 10-fold cross-validation. Many feature ranking methods were utilized and compared in this study, including *mRMR*, *step-wise fit*, and *Wilcoxon rank-sum test*. We used the *minimum Redundancy Maximum Relevance (mRMR)* [35] optimizing in terms of the *Mutual Information Quotient (MIQ)* criterion [36].

Assuming that q features $Z_q = z_{1:N,1:q}\ i = 1,..,q$ are selected from the total M of the G features matrix and these features form the subset Z_q for selecting the next best feature, it can be calculated by the equation

$$\max_{z_j \in x - Z_M} \left[I\left(z_j; c\right) - \frac{1}{M-1} \sum_{z_j \in Z_M} I\left(z_j; z_i\right) \right]$$

where the class labels are C and $I(x;y)$ is the mutual information (MI) function. The algorithm selects and ranks the most relevant to the class label and the least redundancy with the previously selected features.

The stepwise regression method [37] begins with an initial model, and then, modifies the model in successive steps by adding or removing features. In each step, the p-value of an F-statistic is computed to test whether a potential feature has a zero coefficient if added or removed from the model.

The Wilcoxon rank-sum test (also known as Mann-Whitney test) [38] is a non-parametric statistical method testing the hypothesis that for two random variables X and Y, the probability of X $>$ Y equals the probability of Y $>$ X

D. Ordinal transformation and analysis

As ordinal regression is performed, it is appropriate to take into account each participant's personalized values in the neutral state. This period corresponds to each subject's baseline, and using the mapping transformation to rankings [39] generates a common reference to each feature across subjects, providing data normalization.

In this case, the problem of stress detection can be viewed as a ranking problem. In order to transform into a 2-class classification problem (classes: no stress vs stress), we used the pairwise transformation introduced in [39] [40]. The pairwise transformation which maps the features matrix X (described in IV-B) and the class labels Y is described by the equation

$$T : \left\{ \begin{array}{l} X' = X\left(t_i\right) - X\left(t_j\right) \\ Y' = \operatorname{sign}\left\{\mathrm{Y}\left(t_i\right) - Y\left(t_j\right)\right\} \end{array} \right\}, \forall \text{ corresponding } i, j$$

where i, j refer to the indices of neutral and stress states respectively with all possible pairs of a specific subject of the feature matrix. The overall transformation procedure is described in Algorithm 1.

Algorithm 1: Pairwise transformation used in this study

Input:
X – feature matrix [cases x features]
Y – classes [1: non-stress, 2: stress]
Output:
X' – pairwise transformed feature matrix
Y' – classes [-1,1]
for each extracted data **do**
X_1, X_2 feature vectors of class Y_1, Y_2 respectively
Find indices i, j of all permutations without repetition of X_1, X_2
for each pair i, j **do**
$X' = X_1(i) - X_2(j)$
if $Y_i > Y_j$ then $Y' = 1$
if $Y_i < Y_j$ then $Y' = -1$
end
end

This transformation creates preference pairs of feature vectors $X(i) - X(j) = [f_1(i) - f_1(j), \ldots, f_M(i) - f_M(j)]$ and their labels $\operatorname{sign}\{Y(i) - Y(j)\}$. If $Y(i) > Y(j)$ then $X(i) \succ X(j)$ and this preference pair is a positive instance, otherwise, it is a negative instance $X(i) \prec X(j)$. The preference pairs and their corresponding labels after transformation can be considered as instances and labels in a new classification problem, which then can be performed with traditional classification schemes. This step is significant for the subsequent analysis as it addresses the inter-subject variability, taking into account the baseline of each subject of the neutral tasks.

E. Stress detection model and machine learning classification

Stress detection can be seen as a binary classification problem, i.e., classify a voiced segment into two classes based on whether or not stressful conditions occur during this time interval. Each voiced segment was assigned to a no stress or stress class based on the experimental task. The features produced by the feature extraction phase (Section IV-B were fed into classification schemes in order to provide automatic stress detection. The objective is to design a stress detector

for mapping $\mathcal{X}_i$ to y_i which can be formulated as a binary classification problem. In this study, we employ a plethora of classification schemes (22 in total) which are listed in Table III.

TABLE III
CLASSIFIERS USED IN THIS STUDY

Classifier	Parameters
SVM1	kernel:linear, scale:auto, constrain:1
SVM2	kernel:polynomial(2nd), scale:auto, constrain:1
SVM3	kernel:polynomial(3rd), scale:auto, constrain:1
SVM4	kernel:Gaussian, scale:2.1, constrain:1
SVM5	kernel:Gaussian, scale:8.5, constrain:1
SVM6	kernel:Gaussian, scale:34, constrain:1
LDA	DiscrimType: diaglinear
QDA	DiscrimType: diagquadratic
KNN1	neighbor:1 (Euclidean)
KNN2	neighbor:10 (Euclidean)
KNN3	neighbor:100 (Euclidean)
KNN4	neighbor:10 (cosine)
KNN5	neighbor:10 (Minkowski p=3)
KNN6	neighbor:10 (Eucl. dist. squared inverse distance weight)
TREE1	binary decision tree (max split 100)
TREE2	binary decision tree (max split 20)
TREE3	binary decision tree (max split 4)
ENS1	Ensemble Boosted Trees (AdaBoost max split 20)
ENS2	Ensemble Bagged Trees (random forest max split 16681)
ENS3	Ensemble random discriminant subspace (min 1 – max 36)
ENS4	Ensemble random KNN subspace (min 1 – max 36)
ENS5	Random undersampling Boosted Trees (max split 20)

F. Hyperparameter optimization

After an initial assessment of the out-of-the-box performances achieved by the different classifiers used, it can be deduced that the SVM classifiers, and particularly the ones with the Gaussian (SVM5) & polynomial of 3rd order (SVM3) kernels, consistently outperform the others in all scenarios. To try to further improve their performance, we tweak their hyperparameters (i.e., the classifier configuration variables) using grid search, initially on a larger scale search (10^{-2} to 10^{2}) followed by a smaller scale search centered on the best results of the first search. The hyperparameters optimized for SVM are *kernel scale* and *box constraints*. We also perform optimization for an ensemble classifier (AdaBoost), using the parameters *learning cycles* and *max split*.

RESULTS

The proposed methodology, as described in section IV, was applied to the study's voice dataset.

G. Evaluation of features ranking

Various ranking methods were compared in terms of their classification performance and the number of features used. Depending on the used method, speaker gender, and the selected task, the ranking results differ. So, the different methods used don't converge to the same selected features. Thus, to assess the importance of each feature based on all the ranking methods, we calculated a score that is the weighted sum of all occurrences of that feature in a specific ranking position multiplied by the weight of the position. The scoring formula for each feature g_j from the feature set G with j features is the following:

$$\text{score}\,(g_j) = \left(\frac{1}{\text{score}_{\max}}\right) \sum_{i=1}^{N} c_{ji} \times w_i$$

where i is the index of the ranking position, N is the number of ranking positions (which equals the number of features), c_{ji} is the number of times the feature j was ranked in position i and $w_i = [N, N-1, ..., 1]$ is the weight of that position. The sum is normalized by the maximum possible score.

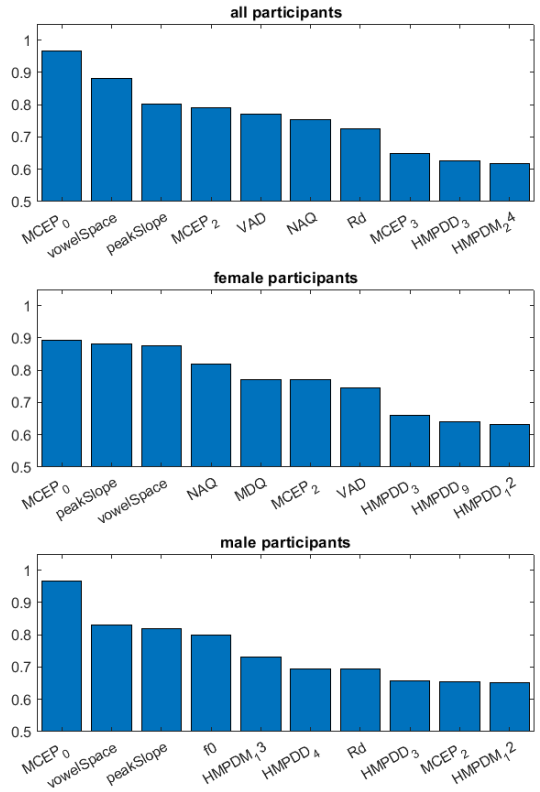

Fig. 3. Top-10 ranking features for all participants (upper figure), female speakers (middle figure) and male speakers (lower figure).

The maximum possible score is defined as

$$\text{score}_{\max} = C \times w_1 = C \times N$$

where C is the number of ranking methods tested. Essentially, scoremax is the score a feature would get if it ranked in the first position for all ranking methods.

Using the scoring method described above, we can assess the importance of the features when used for both genders, or male and female separately. Despite the output differences of each ranking method, some features seem to consistently score higher than others regardless of the method used indicating their importance.

It can be observed, as we can see in Figure 3, that MCEP_0, vowel space, VAD, and peak slope appear to be consistently among the most significant features across the

different conditions (gender, tasks). The features NAQ, MDQ, $HMPDD_9$ appear to be significant mainly for female voices (ranking 0.82, 0.77, and 0.64 respectively) while the same features rank a lot lower for male speakers (0.43, 0.37, and 0.17 respectively). On the other hand, the $HMPDM_{13}$ seems to be important for male speakers (0.73) but its score is low for female speakers (0.39). Formants f_1-f_5 seem to have very low importance in all cases, while the fundamental frequency f_0 appears to be significant for male participants (top 4^{th} position).

H. Feature and optimized model determination

Various feature selections methods (described in section IV-C) and classification schemes (listed in Table III) along with their hyperparameter optimization (described in section IV-F) were tested in terms of their classification accuracy for the whole dataset. This procedure adjusted the model parameters (features selection method, selected features, classifiers hyperparameters) for the analysis. The most effective combinations for all the experimental tasks are presented in Table IV.

TABLE IV
SELECTION CLASSIFICATION ACCURACIES (10-FOLD CROSS VALIDATION) WITH FEATURE SELECTION AND CONVENTIONAL ML TECHNIQUES

Combination	No of features selected	Classification Accuracy (%)
Stepwisefit, ENS1	45	94.36
Wilcoxon, SVM5	81	91.24
Svmrfe_ori, SVM3	55	91
Svmrfe_ker, SVM3	55	90.4
Stepwise fit, SVM3	45	90.0
Svmrfe_ori, KNN6	40	89.9
Stepwise fit, SVM4	20	89.2
Svmrfe_ori, KNN2	40	88.8
Entropy, TREE1	55	81
Svmrfe_ori, KNN5	40	88.1

It can be observed, that the 2 most effective combinations of feature selection and classification schemes are for all cases and participants are the (stepwise fit, AdaBoost (max split 100, learning rate 0.5, learning epochs 600), 45 selected features) with a classification accuracy of 94.36%, the (Wilcoxon, gaussian SVM (kernel scale 5, constraint 10) with an accuracy of 91.88%. For the case of Adaboost, the high performance will most likely come at the cost of high variance (overfitting), meaning that new data will not perform equally well.

I. Classification

The selected features subset was evaluated in terms of its ability to discriminate between non-stress and stress. A 10-fold cross-validation technique on a subject basis (which is closer to what a stress detector system is summoned to perform) was used with the classifiers listed in table III and the parameters determined in IV-H. The classification results for the social exposure and mental tasks phase are summarized in Table V.

It can be observed that the best achieved accuracies are 84.8% using the SVM1 classifier for social exposure and 74.5% for mental tasks using the gaussian SVM classifier.

J. Evaluation of ordinal modelling

Then, the pairwise transformation was applied to the data as described in Section IV-D. This transformation takes into consideration the baseline values (neutral state) of each participant, thus addresses the issue of inter-subject variability which is a common issue in affective studies like the present study. The effect of the transformation on the data distribution of 4 participants is presented in Figure 4.

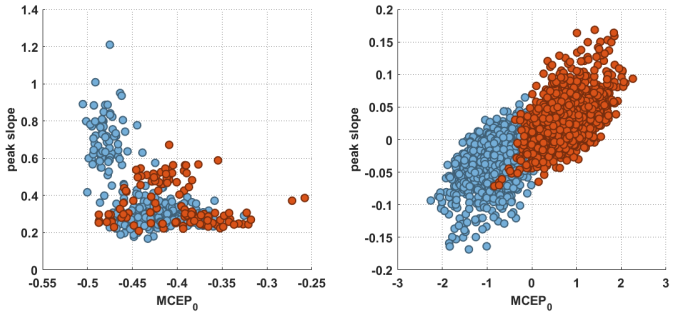

Fig. 4. Visualization of the effect of the pairwise transformation on the 2 top-ranked features $MCEP_0$ and peak slope of 4 participants for neutral (blue) and stress (red) state.

It is clear, that the transformation increases the separability of data taking into account the differential nature of stress response in relation to the neutral states of each participant.

Following the same pipeline, the classification accuracies for the social exposure and mental tasks are presented in Table VI.

It can be observed that using the pairwise transformation, the classification accuracies are significantly improved leading to a best achieved accuracy of 95.0% for social exposure using SVM6 (Gaussian SVM) and 85.9% for mental tasks using the SVM5 (Gaussian SVM).

V. DISCUSSION

In this study, we investigate the most robust vocal features involved in stress conditions and propose a ranking methodology for the construction of the stress model. We identified the most relevant features which are $MCEP_0$, peak slope, vowel space, VAD for all corpus. However, it seems that some features are gender-specific as NAQ, MDQ and $HMPDD_9$ appear to be significant only for female voices, while $HMPDM_{13}$ seems to be important only for male speakers.

The proposed methodology follows a pipeline of identification and selection of the most relevant to stress features, classification and hyperparameter optimization for the selection of the final stress model. Using this methodology, the proposed system yielded a best subject basis 10-fold cross-validation classification accuracy of 84.8% using SVM1 classifier for social exposure and 74.5% for mental tasks using the Gaussian SVM classifier.

TABLE V
SUBJECT BASIS 10-FOLD CROSS-VALIDATION CLASSIFICATION PERFORMANCES FOR SOCIAL EXPOSURE (TASKS 2,3) AND MENTAL TASKS (TASKS 6,7,8)

Classifiers	Social exposure			Mental tasks		
	Accuracy (%)	Sensitivity (%)	Specificity (%)	Accuracy (%)	Sensitivity (%)	Specificity (%)
SVM1	**84.8**	71.6	90.7	70.7	80.6	50.8
SVM2	83.3	76.1	86.4	72.9	77.7	63.9
SVM3	81.2	76.3	83.4	72.4	75.9	64.6
SVM4	69.6	0.0	100.0	66.2	100	0.1
SVM5	84.3	71.8	89.8	**74.5**	81.1	61.3
SVM6	84.4	62.0	94.0	71.8	83.5	48.2
LDA	71.6	61.9	76.0	61.4	62.8	58.2
QDA	69.5	0.0	100.0	66	100	0
KNN1	72.3	68.8	73.7	63.1	63.5	62.1
KNN2	77.9	67.4	82.6	66.3	64.5	69.8
KNN3	79.9	60.6	88.3	70.6	75	61.7
KNN4	78.6	65.4	84.5	68.4	74.2	57.1
TREE1	75.4	63.9	80.5	69.4	75.9	56.6
TREE2	79.6	61.1	87.7	70.2	78.7	53.7
TREE3	76.6	51.5	87.4	71.3	78	58.7
ENS1	79.0	64.9	85.0	72.1	80.7	55.5
ENS2	79.3	64.3	85.9	71.9	81.7	52.1
ENS3	83.6	60.2	93.8	71.2	88	38.4
ENS4	72.6	22.9	94.3	64.9	85.8	24.3
ENS5	76.9	71.1	79.4	71.5	74.1	66.6

Then, the pairwise transformation was employed in order to address the issue of inter-subject variability and to provide appropriate normalization for the analysis used. Using this technique, the discriminatory ability of the proposed system and the classification results were improved significantly. Specifically, it yielded a best subject basis 10-fold cross-validation classification accuracy of 95.0% for social exposure using SVM6 (Gaussian SVM) and 85.9% for mental tasks using the SVM5 (Gaussian SVM). A point of concern about the method is that, for each new subject (speaker) introduced, its corresponding baseline (neutral) speech segments have to be available in order for the transformation to take place.

It can be deduced that ordinal modelling can be an effective tool in stress-related studies like this taking into account a personalized baseline for each participant, addressing the crucial issue of inter-subject variability.

REFERENCES

[1] K. W. Godin and J. H. Hansen, "Physical task stress and speaker variability in voice quality," *EURASIP Journal on Audio, Speech, and Music Processing*, vol. 2015, no. 1, pp. 1–13, 2015.

[2] G. Giannakakis, D. Grigoriadis, K. Giannakaki, O. Simantiraki, A. Roniotis, and M. Tsiknakis, "Review on psychological stress detection using biosignals," *IEEE Transactions on Affective Computing*, 2019.

[3] B. D. Womack and J. H. Hansen, "N-channel hidden markov models for combined stressed speech classification and recognition," *IEEE Transactions on Speech and Audio Processing*, vol. 7, no. 6, pp. 668–677, 1999.

[4] L. J. Rothkrantz, P. Wiggers, J.-W. A. van Wees, and R. J. van Vark, "Voice stress analysis," in *International conference on text, speech and dialogue*. Springer, 2004, pp. 449–456.

[5] K. R. Scherer, T. Johnstone, and G. Klasmeyer, *Vocal expression of emotion.* Oxford University Press, 2003.

[6] K. R. Scherer, "Voice, stress, and emotion," in *Dynamics of stress*. Springer, 1986, pp. 157–179.

[7] M. Van Puyvelde, X. Neyt, F. McGlone, and N. Pattyn, "Voice stress analysis: a new framework for voice and effort in human performance," *Frontiers in Psychology*, vol. 9, p. 1994, 2018.

[8] J.-C. Junqua, "The influence of acoustics on speech production: A noise-induced stress phenomenon known as the lombard reflex," *Speech communication*, vol. 20, no. 1-2, pp. 13–22, 1996.

[9] K. W. Godin and J. H. Hansen, "Analysis and perception of speech under physical task stress," in *Ninth Annual Conference of the International Speech Communication Association*, 2008.

[10] G. Demenko and M. Jastrzębska, "Analysis of voice stress in call centers conversations," in *Speech Prosody 2012*, 2012.

[11] S. Davis and P. Mermelstein, "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences," *IEEE transactions on acoustics, speech, and signal processing*, vol. 28, no. 4, pp. 357–366, 1980.

[12] M. R. Hasan, M. Jamil, M. Rahman *et al.*, "Speaker identification using mel frequency cepstral coefficients," *variations*, vol. 1, no. 4, 2004.

[13] X. Huang, A. Acero, H.-W. Hon, and R. Reddy, *Spoken language processing: A guide to theory, algorithm, and system development*. Prentice hall PTR, 2001.

[14] X. Zuo and P. N. Fung, "A cross gender and cross lingual study on acoustic features for stress recognition in speech," in *Proceedings 17th International Congress of Phonetic Sciences (ICPhS XVII), Hong Kong*, 2011.

[15] H. Kurniawan, A. V. Maslov, and M. Pechenizkiy, "Stress detection from speech and galvanic skin response signals," in *Proceedings of the 26th IEEE International Symposium on Computer-Based Medical Systems*. IEEE, 2013, pp. 209–214.

[16] G. Peeters, "A large set of audio features for sound description (similarity and classification) in the cuidado project," *CUIDADO Ist Project Report*, vol. 54, no. 0, pp. 1–25, 2004.

[17] F. Morchen, A. Ultsch, M. Thies, and I. Lohken, "Modeling timbre distance with temporal statistics from polyphonic music," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 14, no. 1, pp. 81–90, 2005.

[18] J. H. L. Hansen, *Analysis and compensation of stressed and noisy speech with application to robust automatic recognition*. Georgia Institute of Technology, 1988.

[19] J. H. Hansen and S. Patil, "Speech under stress: Analysis, modeling and recognition," in *Speaker classification I*. Springer, 2007, pp. 108–137.

[20] O. Simantiraki, G. Giannakakis, A. Pampouchidou, and M. Tsiknakis, "Stress detection from speech using spectral slope measurements," in

TABLE VI
SUBJECT BASIS 10-FOLD CROSS-VALIDATION CLASSIFICATION PERFORMANCES FOR SOCIAL EXPOSURE (TASKS 2,3) AND MENTAL TASKS (TASKS 6,7,8) USING PAIRWISE TRANSFORMATION

	Social exposure			Mental tasks		
Classifiers	**Accuracy** (%)	**Sensitivity** (%)	**Specificity** (%)	**Accuracy** (%)	**Sensitivity** (%)	**Specificity** (%)
SVM1	94.8	94.8	94.8	79.4	90.41	73.33
SVM2	93.6	93.7	93.4	65.39	72.16	61.94
SVM3	93.8	93.8	93.7	83.02	82.52	83.59
SVM4	82.3	87.7	81.9	70.81	76.69	67.26
SVM5	94.8	94.8	94.8	**85.97**	85.92	86.03
SVM6	**95.0**	95.0	95.0	60.04	95.79	55.79
LDA	89.5	89.5	89.5	76.85	76.85	76.85
QDA	89.3	89.3	89.3	76.71	76.71	76.71
KNN1	88.6	88.6	88.6	72.01	72.01	72.01
KNN2	91.3	92.2	90.4	77.17	79.36	75.3
KNN3	91.7	91.9	91.6	79.18	79.47	78.89
KNN4	90.3	91.0	89.7	75.61	77.19	74.21
TREE1	86.1	86.2	86.1	75.42	75.54	75.43
TREE2	88.6	88.5	88.8	74.79	75.21	74.49
TREE3	88.6	90.1	87.6	68.11	70.6	66.91
ENS1	92.1	92.0	92.2	81.62	81.54	81.72
ENS2	90.0	90.8	89.2	78.35	80.19	77.09
ENS3	93.9	93.9	93.9	82.82	82.82	82.82
ENS4	81.1	82.3	79.9	64.38	65.5	63.42
ENS5	85.5	85.6	85.5	74.47	74.75	74.26

Pervasive Computing Paradigms for Mental Health. Springer, 2016, pp. 41–50.

[21] S. Scherer, L.-P. Morency, J. Gratch, and J. Pestian, "Reduced vowel space is a robust indicator of psychological distress: A cross-corpus analysis," in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2015, pp. 4789–4793.

[22] D. B. Fry, "Duration and intensity as physical correlates of linguistic stress," *The Journal of the Acoustical Society of America*, vol. 27, no. 4, pp. 765–768, 1955.

[23] W. R. Tiffany, "Nonrandom sources of variation in vowel quality," *Journal of Speech and Hearing Research*, vol. 2, no. 4, pp. 305–317, 1959.

[24] M. Fourakis, "Tempo, stress, and vowel reduction in american english," *The Journal of the Acoustical society of America*, vol. 90, no. 4, pp. 1816–1827, 1991.

[25] T. Gay, "Effect of speaking rate on vowel formant movements," *The journal of the Acoustical society of America*, vol. 63, no. 1, pp. 223–230, 1978.

[26] C. Qiu and J. Liang, "The effect of stress on vowel space in daxi hakka chinese," in *Sixteenth Annual Conference of the International Speech Communication Association*, 2015.

[27] M. Goudbeek, J. P. Goldman, and K. R. Scherer, "Emotion dimensions and formant position," in *Tenth Annual Conference of the International Speech Communication Association*, 2009.

[28] I. Lefter, G. J. Burghouts, and L. J. Rothkrantz, "Recognizing stress using semantics and modulation of speech and gestures," *IEEE Transactions on Affective Computing*, vol. 7, no. 2, pp. 162–175, 2015.

[29] I. Lefter, L. J. Rothkrantz, D. A. Van Leeuwen, and P. Wiggers, "Automatic stress detection in emergency (telephone) calls," *International Journal of Intelligent Defence Support Systems*, vol. 4, no. 2, pp. 148–168, 2011.

[30] J. R. Stroop, "Studies of interference in serial verbal reactions." *Journal of experimental psychology*, vol. 18, no. 6, p. 643, 1935.

[31] D. Gronwall, "Paced auditory serial-addition task: a measure of recovery from concussion," *Perceptual and motor skills*, vol. 44, no. 2, pp. 367–373, 1977.

[32] A. I. Korda, G. Giannakakis, E. Ventouras, P. A. Asvestas, N. Smyrnis, K. Marias, and G. K. Matsopoulos, "Recognition of blinks activity patterns during stress conditions using cnn and markovian analysis," *Signals*, vol. 2, no. 1, pp. 55–71, 2021.

[33] P. Boersma, "Praat: doing phonetics by computer [computer program]," *http://www. praat. org/*, 2011.

[34] T. Drugman, Y. Stylianou, Y. Kida, and M. Akamine, "Voice activity detection: Merging source and filter-based information," *IEEE Signal Processing Letters*, vol. 23, no. 2, pp. 252–256, 2015.

[35] C. Ding and H. Peng, "Minimum redundancy feature selection from microarray gene expression data," *Journal of bioinformatics and computational biology*, vol. 3, no. 02, pp. 185–205, 2005.

[36] G. Gulgezen, Z. Cataltepe, and L. Yu, "Stable and accurate feature selection," in *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*. Springer, 2009, pp. 455–468.

[37] N. R. Draper and H. Smith, *Applied regression analysis*. John Wiley & Sons, 1998, vol. 326.

[38] H. B. Mann and D. R. Whitney, "On a test of whether one of two random variables is stochastically larger than the other," *The annals of mathematical statistics*, pp. 50–60, 1947.

[39] R. Herbrich, T. Graepel, and K. Obermayer, "Support vector learning for ordinal regression," *IET Conference Proceedings*, 1999.

[40] J. Fürnkranz and E. Hüllermeier, "Pairwise preference learning and ranking," in *European conference on machine learning*. Springer, 2003, pp. 145–156.