Towards a Model of Arousal Change after Affective Word Pronunciation based on Electrodermal Activity and Speech Analysis

Claudia Marzi [1], Alberto Greco [2,3], Enzo Pasquale Scilingo [2,3], Nicola Vanello [2,3]

[1] *Institute for Computational Linguistics, National Research Council of Italy, Via G. Moruzzi 1 - 56124 Pisa, Italy.*

[2,3] *Dipartimento di Ingegneria dell'Informazione and Research Center "E. Piaggio", University of Pisa, Via G. Caruso 16 - 56122 Pisa, Italy.*

** Corresponding author e-mail: nicola.vanello@unipi.it*

Abstract

In this paper, we explore the possibility of building a model of subject arousal by exploiting the acquisition and the analysis of speech and electrodermal activity (EDA). Several issues have to be addressed to reach this goal as the estimation of the relationship between arousal and behavioural measures and the reliability of EDA signal during speech production. To accomplish this task, we will investigate the relation among EDA, speech activity and subject arousal, during isolated affective word pronunciation. Our results show that significant information on subject arousal can be obtained by analyzing EDA during the processing of out-of-context words with an emotional content in a reading aloud task. Based on a sample of eighteen Italian participants, we observed a significant relation between EDA features and self-reported arousal scores. Quantitative models relating EDA- and speech-derived features are proposed and discussed. We found that increasing values of tonic and phasic components of EDA signals correspond to increasing self-assessed arousal scores; Mel-frequency cepstral analysis of speech was also shown to carry relevant information about subject arousal, with a significant inverse relation to self-assessed scores. Our results suggest how the analysis of concurrent acquisition of EDA and speech features may offer a valid approach for the prediction of subject arousal during speech production, as well as a method for validating self-assessment ratings themselves.

Keywords: Speech, Electrodermal activity, Statistical models, Arousal, Word pronunciation

1. INTRODUCTION

Emotion recognition from speech and physiological signals is a complex task that has gained more and more attention. Affective Computing and emotion recognition methodologies are trying to boost the realization of effective human-machine interfaces [1], as well as improving decision support systems [2]. To achieve this goal many issues still remain to be solved. Concerning speech analysis, one source of complexity is the interplay between internal push factors and external pull factors [3]. While the former are related to the effect of the emotional state of a speaker, the latter are related to environmental factors or social rules that might influence speech production and hide the real emotional status of the speaker. Other approaches for emotion recognition try to overcome these limitations by exploiting signals related to both central and autonomic nervous system (ANS) functions, such as heart rate variability, respiration, and electrodermal activity (henceforth EDA) [4]. EDA refers to alterations in the conductance of the skin, due to changes in the sweat gland activity that are induced at a psychological level [5]. Since the sweat glands are directly controlled by the sympathetic branch of the ANS, the EDA is considered as an effective correlate of the sympathetic nervous system activity [6].

Despite several attempts have been made to create a robust emotion recognition system using ANS correlates, this task suffers from specificity issues and it presents its limits when applied to real world scenarios. For instance, EDA might be influenced by both respiration and speech, limiting its use and interpretation when speech activity is present [6]. Particularly, speech activity induces physiological irregular respiration that activates the sympathetic reflex and consequently affects the sweat gland dynamics. Accordingly, during experimental recordings, especially in the case of low intensity stimulation, subjects are usually asked to avoid body movements, irregular respiration and speech activity [6, 7].

In this paper, we explore the possibility of using EDA and speech features for the description of changes in the subject arousal level after emotional word pronunciation. More specifically, based on results of our pilot study, we will evaluate several models with the aim of predicting the arousal changes related to a reading aloud task using specific affective words. These models will include both EDA-related features, that are known to carry relevant information about sympathetic and parasympathetic nervous system activities, as well as speech related features. Our goal is to improve current knowledge about the relation between speech and EDA, thus resulting in a better comprehension of the influence of speech production on EDA, as well as in the development of emotion recognition systems merging both information channels. We thus propose a multimodal approach, which combines the prediction of perceived arousal by subjects when reading aloud out-of-context words

with an emotional content.

2. METHODS

In this study, we processed and modeled EDA and speech signals from healthy volunteers in order to predict their arousal state while reading aloud affective words. Though culture- and language-dependent, there are many words with an affective connotation that have an influence on cognitive processing. Arousal can be defined as the intensity or energy level that varies in degree of excitement or activation felt by a subject (from calm to highly exciting) towards a given word stimulus [8]. Our methodological approach consists in a processing pipeline that combines physiological signal processing, feature extraction and selection, binary classification, and statistical modeling. Particularly, we first processed both EDA and speech signals to extract features for a comprehensive characterization of both signals. Then, we applied a classification, which incorporates an embedded feature selection (FS) strategy to identify the subset of features that maximized the automatic discrimination of the two arousal classes. Finally, the features included in the selected subset were used as independent variables of different statistical models to highlight possible general relationships among the perceived arousal levels (self-assessment scores) and both speech- and EDA-related features.

2.1. Experimental protocol

We selected 30 Italian words from a database of a total of 1121 words [9], which contains a translation from English into Italian of the ANEW database (Affective Norms for English Words, [10]). Each word in the database is characterized by a distribution of arousal and valence ratings that were self-assessed by 1084 subjects participating in the study, on a Likert scale ranging from 1 to 9 by using a self-assessed manikin procedure.

For the present study, we selected thirty Italian words by controlling for both word length and elicited arousal level in the original large populations. Specifically, only single nouns were selected with a homogeneous word length (6–7 characters, mean 6.57, sd 0.50, mean syllables 2.83, sd 0.38). In addition, to obtain two homogeneous groups characterized by low and high arousal level elicited in the speakers, we selected fifteen words that were ranked in the study [9] as low arousal (mean <4.0, standard deviation (SD) <2.9) and fifteen words ranked as high arousal (mean >6.3, SD <2.2) only. Thirty numbers, matched by word length, were added to the dataset to be considered as neutral stimuli (e.g. *ventuno, settanta, novanta*, respectively "twenty-one", "seventy", "ninety").

Eighteen healthy volunteers, all of them Italian native speakers (12 females, 6 males), were enrolled to take part in our study. Each participant was instructed to read aloud the words that appeared on the screen of a PC. Speech signal was recorded in a quiet room with low reverberation by means of a high-quality microphone, sampled at 48 kHz with a resolution of 32 bits (AKG Perception P220 Condenser Microphone, M-Audio Fast-Track). After an initial resting session of 3 min, each word was shown for 2s and interspaced by 12s before showing the next word stimulus. Words belonging to the same arousal group (i.e., numbers, low arousal nouns, and high arousal nouns) were shown in succession, whereas the group order was randomized among subjects.

After the recording session, each subject was asked to score the arousal level that was elicited by the pronunciation of each word stimulus, on a Likert scale ranging from 1 (very low perceived arousal) up to 5 (very high arousal), with a corresponding Manikin scale shown on the top of the evaluation form. The Self-Assessment Manikin (SAM) measures arousal ranges from calm to excited, depicted as a sleepy figure and a figure with overtly open eyes, respectively.

The obtained arousal scores were adopted for each word in the successive analysis, instead of the mean arousal level that was indicated in the Italian database [9]. A correlation analysis was preliminary performed between the mean arousal levels indicated in the database, and the scores assessed by our participants, revealing a statistical significant correlation ($r=0.67$, $p<0.001$). It is noteworthy to highlight that the arousal levels in the Italian database are scored on a 9-point scale, which may differently account for inter-subject variability.

2.2. EDA processing and feature extraction

EDA signals reflect changes in the skin conductance due to eccrine sweat gland activity induced by physiological stimuli. EDA is commonly recorded by two Ag/AgCl electrodes place on the palm or fingers of the non-dominant hand or on the sole of the feet where the eccrine sweat glands are highly concentrated and respond to emotional or stress stimuli. The sympathetic nervous system (SNS) controls the activity of those sweat glands through the sudomotor nerve and therefore the EDA is considered as an effective way to monitor the sympathetic activity [6].

EDA is comprised of two main components: the phasic and the tonic components, which have a different time-scale and contain different information. The phasic component represents the response of the sudomotor nerve to an external stimulus. It incorporates a succession of relatively rapid evoked changes in the EDA signal known as skin conductance responses (SCRs). On the other hand, the tonic signal is the slow varying baseline of the EDA signal that reflects the general psycho-physiological state of a subject. When two consecutive SCRs are separated by a too short-time-interval, they overlap and it is not possible to perform a correct estimation of the single phasic response and to correctly estimate on the underlying sudomotor nerve activity (SMNA). Consequently, for a robust and effective characterization of the EDA signal, it is crucial to separate the two components reducing at the same time the effect of overlapping SCRs.

In this study, EDA was analyzed using the cvxEDA algorithm [11], which is able not only to disentangle the tonic and phasic signals but also to directly estimate the latent SMNA, thus overcoming the problem of overlapping SCRs (see more details in [11, 12]). Contrary to other decomposition methods, cvxEDA does not need any preprocessing (e.g., filtering) or postprocessing step, and is considered a rigorous and robust

Table 1: Features extracted from EDA phasic and tonic components

Feature	Description
Freq	Number of SMNA bursts wtw
Peak	Maximum peak value of the SMNA signal wtw
MeanAmp	Mean value of the SMNA peaks wtw
MeanPhasic	Mean value of the phasic component wtw
STDPhasic	Standard deviation of the phasic component wtw
MeanTonic	Mean value of the tonic component wtw
STDTonic	Standard deviation of the tonic component wtw
EDAsymp	EDA power spectrum over [0.045–0.25] Hz [14]

wtw = within time-window

model grounding on Bayesian statistics and mathematical regularized convex optimization.

Once the decomposition process was performed, to quantify the SNS activity, we extracted a set of features from the phasic and the SMNA signals over the 5-s time windows after the presentation of each word [13, 6]. Instead, concerning the slow tonic component, we used longer (20 s) time-windows, as we did to perform the EDA frequency analysis [14].

Particularly, we calculated the mean and the standard deviation of both the tonic (MeanTonic, STDTonic) and the phasic component (MeanPhasic, STDPhasic) within each analysis time-windows. The number of the SMNA peaks (Freq) as well as the maximum (Peak) and mean amplitude value (MeanAmp) of the SMNA signal. In addition, we performed a frequency analysis in order to calculate the EDAsymp index, defined as the spectral power of EDA signal on the frequency band between 0.045 and 0.25Hz (i.e., integrating part of the tonic and phasic bands). This frequency band has been proved to be a reliable index of the sympathetic nervous system activity [14]. For this reason, EDAsymp will be consistently investigated in the models taken into account, by considering it as a reliable predictor due to the proven correlation with the sympathetic neural activity, which is crucial in the arousal level perception.

In Table 1, the feature set is summarized along with the corresponding description.

2.3. Speech processing and feature extraction

For each spoken word, two sets of speech features were extracted. The first set included prosodic information derived by the speech fundamental frequency F_0, while the second set was related to speech spectral content and vocal tract shape, as obtained from the analysis of Mel-frequency cepstral coefficients (MFCCs). These coefficients allow to obtain an efficient time-frequency representation of speech signal power spectrum and are mainly related to vocal tract characteristics. In this work, we estimated statistical descriptors of the distribution of MFCCs in each word, as it will be described later.

Before feature extraction, audio signal has undergone a preprocessing stage. Specifically, a Wiener filter was applied to reduce low level ambient noise due to the PC. The Wiener filter was estimated using audio segments not containing speech. After noise reduction, an automatic segmentation of the speech signal corresponding to each single word was performed. We implemented a two-step segmentation algorithm, that was optimized for this specific scenario consisting in isolated words spoken in a low noise conditions: firstly, a rough segmentation was performed using the word presentation time markers; secondly, a voice activity detection (VAD) algorithm analyzed the speech segments extracted in the first step to detect each word time limits. The first step allowed to obtain a 3s long time window starting just after the stimulus presentation. To detect the time limit of each word, that was expected to be present in each pre-segmented audio signal, we looked at the time frames whose intensity was different from residual noise level and at the same time contained voiced sounds. To achieve this goal the VAD algorithm combines a voiced sound detection approach and a signal intensity thresholding algorithm based on Gaussian mixture model of signal and noise. The voiced sound detection step exploits the well-known SWIPE' algorithm [15] to estimate the speech F_0 along with its strength. Voiced segments were detected according to the joint analysis of F_0 strength and signal intensity. As a result, words were identified as time-contiguous audio signals, according to the signal/noise classification performed by the mixture model step, that contained voiced segments.

The SWIPE' algorithm was also employed to compute F_0-derived features.These comprise measure of F_0 values within each word, as median and median absolute deviation (MAD), as well as geometric features describing F_0 profile. These include features inspired by Taylor's Tilt intonational model [16], but estimated on every voiced segment, irrespectively of intonational event selection [17], as in equation 1, 2, and 3:

$$Amplitude^* = \frac{|A_{rise}| - |A_{fall}|}{|A_{rise}| + |A_{fall}|} \tag{1}$$

$$Duration^* = \frac{D_{rise} - D_{fall}}{D_{rise} + D_{fall}} \tag{2}$$

$$Tilt^* = \frac{Amplitude^* + Duration^*}{2} \tag{3}$$

where A_{rise} and A_{fall} are the F_0 changes during the rising and falling intervals of F_0 contour within a voiced segment, respectively, and D_{rise} and D_{fall} are the duration of the rising and falling sections. Of note, the F_0 values belonging to each subject were normalized with respect to the values estimated with the number speech task, to account for subject specific mean value differences.

MFCCs were estimated using 25 ms long time windows and a time shift of 10 ms. The analyzed frequency bandwidth ranged from 300 to 3700 Hz, and 20 filter banks were adopted. The median values of each coefficient ranging from the 1st to the 12th (M1–M12), estimated for each word, were retained for successive analysis. In addition, we calculated two features describing the overall distribution of cepstral coefficients in each word. Specifically, for each time frame, the index of the largest coefficient was estimated, resulting in a time vector of coefficient indexes. The median and MAD of this vector were then

Table 2: Features extracted from speech signal

Feature	Description
F0	median fundamental frequency F_0 of each word
MAD-F0	MAD fundamental frequency F_0 of each word
Tilt	Geometric description of the F_0 profile of each word
MedianMFCC	The median of the maximal MFCC coefficient across each word
MADMFCC	The MAD of the maximal MFCC coefficient across each word
M1, M3, ..., M12	The median values across each word of the 1st to the 12th MFCCs

estimated for each word (MedianMFCC and MADMFCC, respectively). The first feature is related to the shape of speech spectral envelope, while the second describes its variability and thus it is also related to the articulatory variability within each word. In Table 2, the speech feature set is summarized along with the corresponding description.

2.4. Feature selection and classification analysis

The feature set, comprised of the combination of EDA and speech features, was used as input of a pattern recognition system trying to discriminate the two arousal levels as defined in the Italian word database. However, the aim of a preliminary classification stage was not limited to a classification analysis but mainly to perform a data-driven exploratory feature ranking and selection.

Accordingly, we implemented an SVM-based feature selection by means of an L1-SVM algorithm. Particularly, the L1-regularization promotes sparse feature weights (i.e., with zero coefficients) thus reducing the number of features actually included in the model. The SVM-L1 is considered an embedded feature selection method, because the search of the optimal subset is built into the classifier construction and, therefore, is part of the learning process. Embedded methods have been proven to provide the best performance compared to other strategies because they are less computationally expensive and less prone to overfitting [18, 19, 20]. Of note, considering this processing stage as an exploratory and preliminary step, we labelled the words according to the a-priori arousal levels as defined in the ANEW dataset (scored by 1084 subjects) instead of considering the self-assessed arousal scores of the recruited subjects. In this way, we obtained a more generalized feature ranking, and we mitigated the impossibility for the L1-SVM to consider repeated measures and the within-subject variability (as for most of the learning models).

2.4.1. L1-SVM

SVM are characterized by a decision rule that only relies on a subset of the training data points (support vectors), but it is in general based on all available features in the input space. L1-SVM imposes an L1-norm (instead of canonical L2-norm) to the hyperplane normal β, thus constructing a maximum-margin hyperplane based on a limited subset of features in input space at the price though of making the optimization problem more difficult. Given x the feature vector of p dimension and y the class label, the L1-SVM can be described by the following minimization problem:

$$(\widehat{\beta}, \widehat{\beta_0}) \in \underset{\beta \in \mathbb{R}^P, \beta_0 \in \mathbb{R}}{\operatorname{argmin}} \|\beta\|_1 + (1/\lambda) \sum_{n=1}^{N} H_{\beta,\beta_0}(\mathbf{x}_n, y_n), \qquad (4)$$

where N is the number of observations, $H_{\beta,\beta_0}(\mathbf{x}, y) = \max(0, 1 - y(\beta^\top \mathbf{x} + \beta_0))$ is the hinge loss function, β and β_0 the model coefficients and intercept, and λ the regularization parameter.

It was shown that the reformulation of the penalization term with a square hinge loss function H^2_{β,β_0} permits to reduce the difficulty to solve the optimization problem:

$$(\widehat{\beta}, \widehat{\beta_0}) \in \underset{\beta \in \mathbb{R}^P, \beta_0 \in \mathbb{R}}{\operatorname{argmin}} \|\beta\|_1 + (1/\lambda) \sum_{n=1}^{N} H^2_{\beta,\beta_0}(\mathbf{x}_n, y_n) \qquad (5)$$

2.5. Statistical modeling

Following our goal of relating the interaction between EDA- and speech-derived features to the subject perceived arousal in a task of reading aloud affective nouns, we applied different statistical models to highlight possible general relations among the self-assessed arousal scores and both speech and EDA-related features.

Firstly, a linear regression approach was adopted to explore possible general relations between the acquired features and the arousal scores. Linear mixed-effects (LME) have been modeled by considering the arousal levels as dependent variable, with EDA features or speech related features as independent ones, with both participants and word stimuli as random effects. Finally, a non-linear approach has been added, with Generalized Additive Models (GAMs). These models provide greater flexibility than linear ones, since they are not forced to strict linearity assumptions. GAMs are a class of models that may replace the linear relationship between the response/target variable and predictors with smooth functions to take into account non-linearities in the data (see *mgcv* package in R, [21]).

We then evaluated a hierarchical linear model (HLM) [22] to better explain each word arousal level, using EDA and speech related features as predictors. Specifically, a two-level model was estimated, with the first level describing within subject variability across different words, and with the second level representing the relationship between predictors and predicted values at group level. This model thus allows to account for repeated measures and to describe, the inter- and intra- subject variabilities, under Gaussian assumptions. This is obtained by means of a parametrization of the covariance matrices at both levels through the definition of hyperparameters.

An Expectation Maximization procedure was used to jointly estimate the model parameters and the covariances at both levels [23]. Expectation Maximization is an iterative algorithm that allows to maximize the likelihood of the model. In the formulation of the algorithm for HLMs, the initial guess pertains

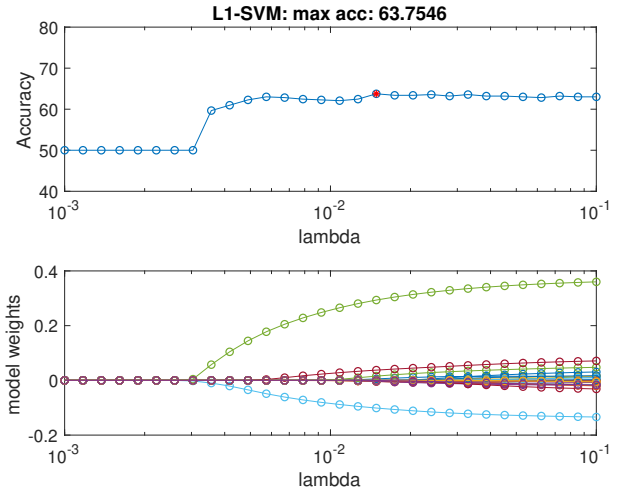

Figure 1: Upper sub-figure shows the accuracy trend as a function of the L1-penalty coefficient. Bottom sub-figure shows the feature weights as a function of the L1-penalty coefficient

the hyperparameters. We decided upon successful algorithm convergence when the sum of the squares of the gradient vectors, involved in the update of the hyperparameters, was lower than 10^{-7}. The statistical significance of a regression parameter at group level was assessed using a t-test statistics related to the conditional mean of the parameter given the observations. Regarding the goodness of fit of the model, while for HLM the classical R^2 cannot be defined, it is possible to use specific measures, focusing on first or second level covariances [24]. The performance of the HLM was here estimated using the proportion of explained variability at first level that is predicted by the model, i.e. R_1^2. The features entered in the HLM were chosen according to the feature selection process embedded in the L1-SVM procedure.

3. RESULTS

3.1. Classification and feature selection results

The binary classification of the a priori arousal classes using the linear L1-model showed poor accuracy. However, the trend of both classification accuracy and feature weights as a function of the regularization coefficient (i.e., λ) revealed interesting information as shown in Figure 1. In fact, increasing the lambda value and therefore reducing the L1-penalty (see eq.(5)), after a first step, the accuracy did not show significant variations as the sparsity level of feature weights decreased. In other words, including more features in the model, the accuracy showed only meaningless fluctuations. The features ranked according to the L1-SVM weights, correspondent to the maximum accuracy, are shown in Table 3. It is worthwhile noting that the major contribution was given by the first two features, i.e., EDA MeanTonic and speech MedianMFCC as also shown in the bottom plot of Figure 1).

3.2. Linear models with fixed and random effects

Table 3: Features ranking based on the L1-SVM output

Feature	|Weight|
MedianMFCC	**2.3E-01**
MeanTonic	**1.0E-01**
M8	**3.8E-02**
M6	**1.6E-02**
Peak	**5.5E-03**
F0	**4.3E-03**
MeanAMP	**3.6E-03**
Tilt	**3.6E-03**
MadF0	**9.4E-04**
M12	**4.5E-04**
MearPhasic	**4.4E-04**
M1	**3.2E-04**
STDTonic	**1.0E-04**
Freq	**7.1E-06**
STDPhasic	0
EDAsymp	0
MADMFCC	0
M2	0
M3	0
M4	0
M5	0
M7	0
M9	0
M10	0
M11	0

We firstly considered the linear fitting of self-assessed arousal scores by our participants with MedianMFCC (speech feature) and MeanTonic (EDA component) as predictors (the first two features in our L1-SVM-based ranking), with participants and words as random effects, to take into account for both inter-subject variability and words variability. The model presents an explained variance of 38% (conditional R^2), with a good statistical significance for meanTonic ($p<0.001$) as a positive predictor for self-assessed arousal scores. However, MedianMFCC as a negative predictor, results to be not significant ($p = 0.2$).

Statistical significance and explained variance increase when we separately model self-assessed arousal with either EDA- or speech-related features: meanTonic and EDAsymp, with participants and words as random effects, positively predict self-assessed arousal scores (with a conditional $R^2 = 0.40$). Whereas, related to speech, MedianMFCC is a negative significant predictor for self-assessed arousal when its values interact with participants as categorical independent variable, with words only as random effect (with a conditional $R^2 = 0.40$).

Despite our classification analysis, we considered EDAsymp as a reliable predictor since it has shown to be a suitable index of sympathetic function with stressors [14]. In fact, as shown in Figure 2, increasing values of EDAsymp correspond to the assessment of increasing ratings of arousal by our participants.

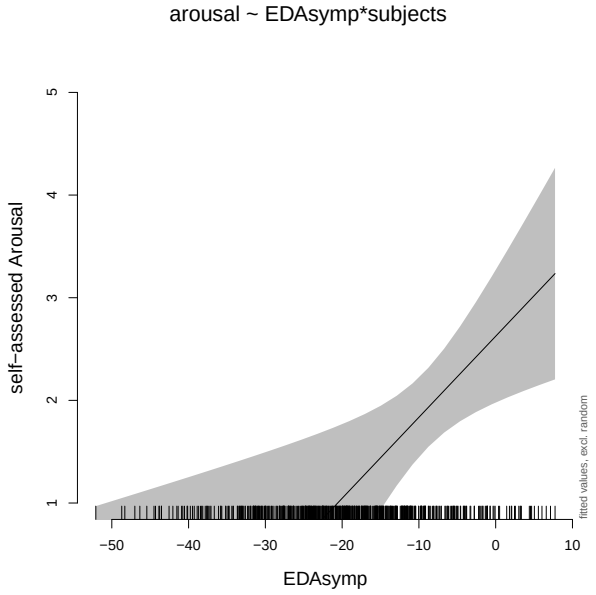

Figure 2: Summed effects for the linear regression model fitting the interaction of EDAsymp and participants predicting self-assessed arousal levels.

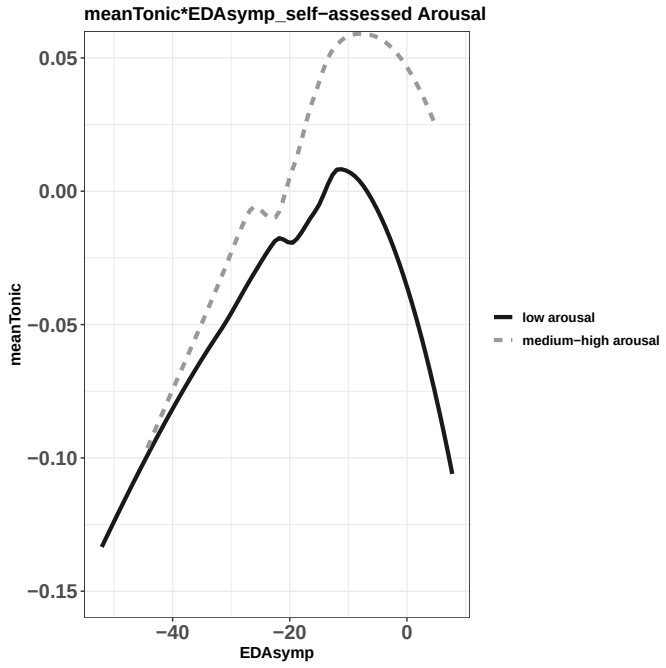

Figure 3: Regression plot of interaction between EDAsymp and MeanTonic fitting self-assessed arousal scores by considering two groups, i.e. low (solid line) and medium-high (dashed-line).

3.3. Generalized additive model

As a further analysis, we modeled self-assessed arousal scores with EDA- and speech-related features by using GAMs. The strength of GAMs is their ability to deal with highly non-linear and non-monotonic relations between the dependent variable and the set of explanatory variables.

Concerning EDA features, the relation between EDAsymp and MeanTonic is highly non-linear, as shown in a model fitting the interaction between the two features for self-assessed arousal scores grouped into two classes, with low (i.e. for arousal scores 1 and 2) versus medium-high (i.e. for scores ≥3) arousal ratings (see Figure 3). Splitting criteria between the two groups are defined on the basis of distributional data of arousal scores as assessed by our participants (i.e. mean 1.97, sd 1.14, median 2).

Accordingly, we modeled self-assessed arousal scores by fitting the non-linear interaction between EDAsymp and MeanTonic as independent variables, with participants as categorical variable, and words as random effects. The positive effect of EDAsymp on arousal is modulated by the positive effect of MeanTonic (the model fitting is expressed with an explained deviance of 50.8%).

Explained deviance significantly increases up to 80.4% when we consider the subdivision of self-assessed scores into the two arousal groups, i.e. low and medium-high arousal (see Figure 4 for the summed effects of EDAsymp and the MeanTonic component). This may suggest that the two contrasting groups, namely words with low versus medium-high arousal, emphasize the difference between arousal score 2 and 3 (on a 5-point scale), which, in our experiment, mark the difference between low and high arousal, as assessed by our participants.

It should be appreciated that for increasing values of EDAsymp and MeanTonic, in interaction, increasing values of self-assessed arousal are predicted, as confirmed by the GAM model parametric coefficients (see Table 4).

	est.	*sd error*	*t value*	*p-value*
Intercept(low)	1.45	0.09	16.02	p<0.001
Intercept(mh)	2.11	0.12	17.70	p<0.001
meanTonic(mh)	3.28	1.23	2.67	p<0.005
EDAsymp:meanTonic(mh)	0.15	0.06	2.70	p<0.005
s(participants)				p<0.001
s(words)				p<0.01
R^2 =0.79				
deviance explained = 80.4%				

Table 4: Parametric coefficients of a GAM fitted to arousal using EDAsymp and meanTonic as interacting predictors for the two arousal classes (low versus medium-high) as categorical variables, with participants and words as random effects.

Similarly, self-assessed arousal scores can be predicted by speech-related features. The most significant predictor resulting from our GAMs is MedianMFCC: when interacting with participants, it is a negative, significant predictor (p<0.01) in a GAM model with words as random effect, with a conditional R^2 = 0.44.

3.4. Hierarchical linear model

A hierarchical approach represents a principled solution to account for intercepts and slopes, that may vary independently by a grouping variable, such as participants in our study. The first five features selected by the L1-SVM approach were fit in the HLM, namely MedianMFCC, MeanTonic, M8, M6, Peak.

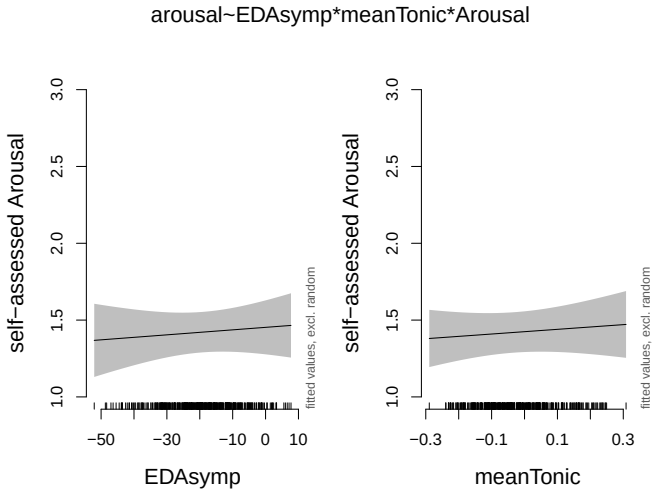

Figure 4: Summed effects for the GAM model fitting the interaction between EDAsymp and MeanTonic for self-assessed arousal scores.

These features allowed to obtain convergence of model estimation and reproducible estimate of the HLM parameters and hyperparameters, i.e. covariances. The dependent variable was self assessed arousal score. The HLM converged to a solution, obtaining a R_1^2=0.49. Only two regression coefficients resulted to be significantly different from 0. Specifically, a statistically significant (p=0.04) positive linear coefficient was found relating MeanTonic and arousal score. A negative linear relation (p=0.003) was also found between M8 and self assessed arousal. The analysis was repeated by adding EDAsymp as a regressor obtaining a R_1^2=0.50. In this case, a statistically significant (p=0.015) positive relationship between EDAsymp and the arousal score was found as well. In this latter model, the MeanTonic coefficient significantly decreased resulting outside hypothesis rejection region (p=0.1). Adding more regressors, i.e. features, belonging to the a priori selected list, resulted in frequent lack of convergence by changing initial conditions. This scenario is likely to be related to the convergence to local maxima, that we decided not to further explore.

4. DISCUSSION

This work focuses on the combination of EDA- and speech-derived features during an aloud reading task of emotional words to study the arousal perception.

Our results give evidence that the complementary information of speech- and EDA-related features can successfully model subject arousal during a reading-aloud task. Although EDA is often excluded by experimental paradigms in which the subjects are asked to speak, due to the influence of speech-induced irregular respiration, we have demonstrated the good reliability of EDA-related features measured during emotional reading. Noticeably, this work highlights the complexity of the estimation of arousal level from the physiological and behavioral features taken into account. Specifically, this is related both to the high inter- and intra- subject variability, as well as the nonlinear relationships among the observed measures.

It is worthwhile noting that for a comprehensive methodological approach, we merged the data-driven information extracted by an embedded feature selection algorithm with the physiological prior knowledge coming from previous studies. Particularly, we included in our models both EDA and speech features selected through an L1-SVM algorithm and the EDAsymp measure selected due to the proven correlation with the sympathetic neural activity.

Concerning the reliability of EDA-related features during speech activity, the analysis with both linear and non-linear regressions reveals a significant relation between arousal level and EDAsymp. However, the analysis performed with these models highlighted a high subject by subject variability, as confirmed by considering participants as categorical variable. This was found, for instance, when studying the relation between MeanTonic and EDAsymp predicting arousal scores. In addition, generalized additive models showed their highly non-linear interaction. Our results might be influenced by factors such as the individual variability in the changes of speech-related parameters and emotions, or anxiety levels [17]. A further source of variability can be due to the specific relevance of each word for each subject. In addition, each word was shown in isolation, namely out of context, thus allowing for possible different interpretations by each participant.

The proposed HLM allows to model subject-specific slopes and intercepts thus accounting for group level variability. At the same time, it allows to model intra-subject variabilities across repeated measures. Such a flexible approach might account for different sources of variability as those related to the difficulty in self reporting the elicited arousal level, or some other specific cognitive, behavioral factors affecting the task execution. In addition, it overcomes the limitations related to learning models, such as the SVM, unable to take into account repeated measures and marked within-subject variability. As a matter of fact, the HLM could highlight a positive correlation between word self-assessed arousal with both tonic and phasic components of EDA, as captured by EDAsymp. If, on the one hand, this appears to be quite trivial, due to the link between EDA and arousal, on the other, it reinforces our hypothesis that speech activity does not totally hide EDA emotional-related information. Besides, a sign of a complementary information achievable by analyzing speech- and EDA-related features was highlighted by the significance of one of the Mel-Frequency Cepstral Coefficients, the eighth in our analysis; similarly the relevance of MFCC-based features emerged from both the linear and additive models. However, only the HLM could significantly take into account both kinds of features in a unique statistical model.

Although the number of recruited subjects is limited, these results seem to indicate the benefits of modeling the subject-specific variability offered by the hierarchical linear modeling approach.

It is noteworthy to underline the difference between GAM and HLM statistical modeling, with the first model being able to capture the non linearity of the interaction between predictors, whereas the second one is able to efficiently model the

intra- and inter-subject variability in a small sample of participants. Our experimental analysis shows that both modeling approaches reveal interesting information about the phenomenon under study. Besides, a jointly analysis of both approaches emphasizes both their strengths and limitations. While HLM was able to highlight the relevance of a joint analysis of speech and EDA-related features, GAM revealed the highly non-linearity of some feature interactions in capturing the self-assessed arousal scores. However, no single approach could model the whole complexity of our target task.

One limitation of this work is the discrete ordinal nature of the arousal scoring. This is particularly relevant for linear models. In fact, not only the discrete nature of the predicted variable is not optimal when using regression models under Gaussian assumptions, but the relationship between changes in the arousal score could be non-linearly related to physiological measures. The generalized additive model showed some relevant non-linear relationships among the obtained features. Not only this non linear model could explain the largest variance observed across self-assessed scores, but highlighted also a relevant improvement of explained variance when arousal subgroups are introduced as categorical variable. This result suggests that the self-assessed values may suffer from a bias when low- or high-arousal words are pronounced, thus revealing that the actual arousal dynamics is not completely described by the self-assessed score. It should be considered that, in some cases, the self-assessed ratings may not be totally reflecting the actual arousal as perceived by participants.

In this perspective, our approach might offer a possible method for validating self-assessment ratings by participants. As a result, in our experimental study the difference between arousal score 2 and 3 marks the sharp contrast between low versus high arousal. Accordingly, a possible change of the experimental paradigm that might be considered in a follow-up of our study consists in the use of a wider arousal scale (e.g. ranging from 1 to 9) to a more reliable perceived arousal modeling, and to better contrast low versus high arousal effects.

Another issue affecting the changes of our observed features, is the choice of presenting words in blocks with homogeneous a-priori arousal levels. Our hypothesis has been that this experimental design should elicit a larger arousal change with respect to mixing words with different arousal levels. We are aware that both a possible cumulative effect of arousal level, and a correlation of the elicited responses to ensuing words might be significant. However, in our statistical modeling, words - together with participants - have been considered as random effects; and this has partially mitigated the effect of primacy and recency of word orders. In addition, future studies may focus on an experimental design where words with different arousal levels are mixed, thus improving current knowledge about ANS response to a reading aloud task.

Although these limitations, we have to stress that both the linear and non-linear models can fruitfully be used to explore the relationship among physiological measures and arousal level in speech task, exploiting the possibility of accounting for repeated measures across subjects. Finally, the HLM approach should also be extended to model noise covariance according to experimental hypotheses, as for instance dependency among successive measures.

5. CONCLUSION

This pilot study suggests that the informative content of EDA signal about subject arousal is not hidden by concurrent speech activity. In addition, both EDA activity, as described by the mean tonic component and by power at higher frequencies, and speech frequency content, as described by cepstral components, were found to change along with changing arousal level, during word pronunciation.

This work offers evidence for the possibility of building a quantitative model of arousal by jointly exploiting speech and EDA derived features. Specifically, it may represent a step forward with respect to a previous study [7] where arousal classification results were improved by jointly considering speech and EDA features, suggesting a starting approach to clarify and disentangle the relative contribution of speech to EDA signal change. By achieving this, a measure of subject psychological status could be achieved, overcoming some limitations of self administered tests. Nonetheless, some relevant issues are raised that could serve as bases for future developments.

Acknowledgment

The research leading to these results has received partial funding from the Italian Ministry of Education and Research (MIUR) in the framework of the CrossLab project (Departments of Excellence).

References

[1] R. Cowie, E. Douglas-Cowie, N. Tsapatsoulis, G. Votsis, S. Kollias, W. Fellenz, J. G. Taylor, Emotion recognition in human-computer interaction, IEEE Signal processing magazine 18 (1) (2001) 32–80.

[2] J. J. Gross, R. F. Muñoz, Emotion regulation and mental health, Clinical psychology: Science and practice 2 (2) (1995) 151–164.

[3] K. R. Scherer, Vocal affect signalling: A comparative approach, Advances in the study of behavior 15 (1985) 189–244.

[4] S. Jerritta, M. Murugappan, R. Nagarajan, K. Wan, Physiological signals based human emotion recognition: a review, in: 2011 IEEE 7th International Colloquium on Signal Processing and its Applications, IEEE, 2011, pp. 410–415.

[5] A. Greco, A. Lanata, L. Citi, N. Vanello, G. Valenza, E. P. Scilingo, Skin admittance measurement for emotion recognition: A study over frequency sweep, Electronics 5 (3) (2016) 46.

[6] W. Boucsein, Electrodermal activity, Springer Science & Business Media, 2012.

[7] A. Greco, C. Marzi, A. Lanata, E. P. Scilingo, N. Vanello, Combining electrodermal activity and speech analysis towards a more accurate emotion recognition system, in: 2019 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), IEEE, 2019, pp. 229–232.

[8] P. J. Lang, M. M. Bradley, B. N. Cuthbert, et al., Motivated attention: Affect, activation, and action, Attention and orienting: Sensory and motivational processes 97 (1997) 135.

[9] M. Montefinese, E. Ambrosini, B. Fairfield, N. Mammarella, The adaptation of the affective norms for English words (ANEW) for Italian, Behavior research methods 46 (3) (2014) 887–903.

[10] M. M. Bradley, P. J. Lang, Affective norms for English words (ANEW): Instruction manual and affective ratings, Tech. rep., Technical report C-1, the center for research in psychophysiology, University of Florida (1999).

[11] A. Greco, G. Valenza, A. Lanata, E. P. Scilingo, L. Citi, cvxeda: A convex optimization approach to electrodermal activity processing, IEEE Transactions on Biomedical Engineering 63 (4) (2015) 797–804.
[12] A. Greco, G. Valenza, E. P. Scilingo, Advances in Electrodermal activity processing with applications for mental health, Springer, 2016.
[13] M. Benedek, C. Kaernbach, A continuous measure of phasic electrodermal activity, Journal of neuroscience methods 190 (1) (2010) 80–91.
[14] H. F. Posada-Quintero, J. P. Florian, A. D. Orjuela-Cañón, T. Aljama-Corrales, S. Charleston-Villalobos, K. H. Chon, Power spectral density analysis of electrodermal activity for sympathetic function assessment, Annals of biomedical engineering 44 (10) (2016) 3124–3135.
[15] A. Camacho, J. G. Harris, A sawtooth waveform inspired pitch estimator for speech and music, The Journal of the Acoustical Society of America 124 (3) (2008) 1638–1652.
[16] P. Taylor, Analysis and synthesis of intonation using the tilt model, The Journal of the acoustical society of America 107 (3) (2000) 1697–1714.
[17] A. Guidi, N. Vanello, G. Bertschy, C. Gentili, L. Landini, E. P. Scilingo, Automatic analysis of speech F0 contour for the characterization of mood changes in bipolar patients, Biomedical Signal Processing and Control 17 (2015) 29–37.
[18] Y. Saeys, I. Inza, P. Larrañaga, A review of feature selection techniques in bioinformatics, bioinformatics 23 (19) (2007) 2507–2517.
[19] K. Yan, D. Zhang, Feature selection and analysis on correlated gas sensor data with recursive feature elimination, Sensors and Actuators B: Chemical 212 (2015) 353–363.
[20] T. N. Lal, O. Chapelle, J. Weston, A. Elisseeff, Embedded methods, in: Feature extraction, Springer, 2006, pp. 137–165.
[21] S. N. Wood, Fast stable restricted maximum likelihood and marginal likelihood estimation of semiparametric generalized linear models, Journal of the Royal Statistical Society (B) 73 (1) (2011) 3–36.
[22] R. Kass, D. Steffey, Approximate bayesian inference in conditionally independent hierarchical models (parametric empirical bayes models), Journal of the American Statistical Association 84 (407) (1989) 717–726.
[23] A. Dempster, D. Rubin, R. Tsutakawa, Estimation in covariance components models, Journal of the American Statistical Association 76 (374) (1981) 341–353.
[24] T. Snijders, R. Bosker, Modeled variance in two-level models, Sociological Methods & Research 22 (3) (1994) 342–363.