Pre-registration Protocol: Smartphone Sensing Panel Study - Predicting Affective States from Acoustic Voice Cues Collected with Smartphones

This pre-registration protocol deals with specific research questions and is completed before the data is accessed. Throughout this registration, we will refer to the corresponding basic registration protocol of the panel study. The basic protocol contains information on study procedures and further background information and can be found in the general pre-registration template here: http://dx.doi.org/10.23668/psycharchives.2901.

Working Title
Predicting Affective States from Acoustic Voice Cues Collected with Smartphones

Author(s) of the preregistration protocol
Timo Koch, Ramona Schoedel

Date
January 5, 2021

Background

Background Information (Optional; Short description of the theoretical background/introduction to research question)
The expression and recognition of emotions (i.e., short-lived and directed representations of affective states) through the acoustic properties of speech is a unique feature of human communication (Weninger et al., 2013). Researchers have identified acoustic features, which are predictable of affective states, and emotion detecting algorithms have been developed (Schuller, 2018). However, most studies used speech data produced by actors, who had

instructions to act out a given emotion, or speech samples labelled by raters, who were instructed to add affective labels to recorded utterances (e.g., from TV shows). Both, enacted and labelled speech, come with multiple downsides since these approaches assess expressed affect rather than the experience of actual affective states through voice. Further, there is only little research on the prediction of affective states from acoustic properties of speech collected in a natural everyday setting.

Research question(s)

In this work, we want to investigate if we can predict in-situ self-reported affective states from objective voice parameters collected with smartphones in everyday life. Further, we want to explore which acoustic features are most predictive for the prediction of the experience of affective states. Finally, we want to analyze how the affective quality of instructed spoken language (e.g., a sentence with negative affective valence) translates into objective markers in the acoustic signal, which then in turn could alter the predictions in our models.

Hypotheses
Please provide hypothesis for predicted results. If multiple hypotheses, uniquely number them (e.g. H1, H2a, H2b,) and refer to them the same way at other points in the registration document and in the manuscript.

Our study is exploratory in nature. Thus, we do not provide any confirmatory hypotheses. We pre-register our procedure as a transparent account of our work.

Variables

Which variables will be used? (see Variables in the basic protocol for an extensive overview of all available variables)
This section shall be used to unambiguously clarify which variables are used to operationalize the specified hypotheses. Please (a) list all variables that will be used in this study and (b) explicitly state the functional role of each variable (i.e., independent variable, dependent variable, covariate, mediator, moderator). It is important to (c) specify for each hypothesis how it is operationalized, i.e., which variables will be used to test the respective hypothesis

and how the hypothesis will be operationally defined in terms of these variables. This section is closely related to the statistical models used to test the hypotheses.

Data collection for this work was part of a six-month panel study based on the PhoneStudy research app at Ludwig-Maximilian-Universität München (LMU) from May until November 2020 (for more details see http://dx.doi.org/10.23668/psycharchives.2901). All data collection procedures were approved by the ethics board at LMU.

The study also comprised two two-week experience sampling phases (27.07.2020 to 09.08.2020; 21.09.2020 to 04.10.2020) during which participants received two to four short questionnaires per day. Here, self-reported experience of affective states was assessed. We assessed affective states based on the Circumplex Model of Affect, which suggests that affective states can be mapped onto a space with the two dimensions of valence (i.e., pleasure) and arousal (i.e., physical and psychological activation) (Russell et al., 1989). Valence and arousal were assessed in two separate items on six-point Likert scales among other psychological properties.

Further, the last experience sampling questionnaire of each day included an additional instruction, where participants were asked to read out a series of predefined affective sentences while making an audio recording of their voice. The sentences presented to the participants are based on a set of validated German neutral and emotionally affective sentences (Defren et al., 2018) and differ in their affective content: positive, negative, and neutral. These three affective categories are presented consecutively in each audio logging task. The order of the categories is randomized for each experience sampling questionnaire. For each affective category three sentences are randomly sampled (with replacement) from respective sets of sentences in the database created by Defren and colleagues. The audio recording is initiated by the participants via a button on the screen. Participants could stop the recording manually after a minimum of four seconds. Alternatively, the recording was stopped automatically after twelve seconds.

Next, we used the open-source software OpenSMILE to extract two feature sets, the extended Geneva Minimalistic Acoustic Parameter Set (eGeMAPS, Eyben et al., 2016) and the 2016 Interspeech Computational Paralinguistic Challenge feature set (ComParE2016, Schuller et al., 2016), of voice parameters directly on the participant's device. After feature extraction the voice records were automatically deleted.

At the moment of pre-registration, the data has already been collected and R scripts for data pre-processing are being prepared. However, we have not accessed the data yet. We expect to have up to 3000 to 4000 experience sampling events of participants' affect experience (valence and arousal) with corresponding acoustic features based on the first information from the panel compensation system.

Analysis Plan

Preprocessing

Inclusion criteria (e.g., criteria for including (1) participants (e.g., Do you only use a subsample?, (2) study days (e.g., only weekdays, certain number of study days), (3) any other criteria concerning data quality (e.g., only days with at least x% of logging data) etc. If you cannot specify these aspects now, please state why.
We will exclude experience samples, where participants did not provide information on valence and arousal. Further, we will exclude experience samples, if there is reason to believe that the questionnaire was not filled out thoroughly (e.g., extreme response styles, no variance in responses). Further, we exclude data from experience sampling events, if the respective acoustic features (e.g., voiced segments per second) indicate that no human voice was recorded.

Definition of variables based on smartphone sensing. Please specify your degrees of freedom in variable extraction procedures, e.g., ● *time information (e.g., what does night, daily, weekend exactly mean?)* ● *Aggregation measures (e.g., measures of central tendency/dispersion).* *If you cannot specify these aspects now, please state why.*
We are not using sensing based variables.

Further preprocessing steps (e.g., transformation of data, handling of missing data/outliers etc.)
The properties of participants' voices were transformed into acoustic features using the OpenSMILE algorithm (eGeMAPS and ComParE2016 feature sets) directly on participants' smartphones (the raw audio files will not be available). At the point of pre-registration, we do not know the amount of missing data throughout the experience sampling events (e.g., whether all instructed sentences were recorded). Depending on the extent of missing data, we will either exclude incomplete experience sampling instances or impute missing values (e.g., by using k-nearest neighbors).

Data Analysis

Statistical models
Please specify the statistical model (e.g. t-test, ANOVA, LMM) or algorithms that will be used to test each of your hypotheses. Give all necessary information about model specification (e.g., variables, interactions, planned contrasts) and follow-up analyses. Include model selection criteria (e.g., fit indices), corrections for multiple testing, and tests for statistical violations, if applicable. Please also indicate Inference Criteria (e.g., p-values, effect sizes, performance measures etc.).

We will train various machine learning regression models for the prediction of the outcome variables self-reported valence and arousal. We will use voice parameters (based on the eGeMAPS feature set) extracted from (i) positive affective sentences only, (ii) negative affective sentences only, (iii) neutral affective sentences only and (iv) a combination of positive, negative, and neutral affective sentences as predictor sets to train seperate machine learning models and compare their predictive performance within one single benchmark experiment for each outcome variable. Further, in a similar fashion to Weidman et al. (2020), we plan to run the benchmark analysis using all affective sentences also based on the much larger ComParE2016 feature set in order to compare the predictive performance.
We plan to compare the predictive performance of multiple algorithms, for example, Elastic Net regularized regression models (Zou & Hastie, 2005), non-linear tree-based Random Forest models (Breiman, 2001), boosted trees, Support Vector Machines, and a baseline model, which would predict the mean value from the training set for all cases in the test set. We will impute missing values and tune model hyperparameters in a nested cross-validation scheme and evaluate the predictive performance of our models. To prevent overlaps between training and test data, we will block participants in the resampling procedure ensuring that for one train/test set pair the given participant is either in the training set or in the test set. We might use dimensionality reduction by applying, for example, principal component analysis (PCA) to the features (particularly for the larger ComParE2016 feature set).
Our prediction models will be evaluated based on how accurate new (unseen) samples can be predicted. Model fit will be evaluated based on multiple statistical parameters, for example, Pearson correlation (r), Spearman correlation (ρ), root mean squared error (*RMSE*), mean absolute error (*MAE*), and the coefficient of determination (R^2). Further, we plan to run variance-corrected significance tests to determine if we can predict valence and arousal significantly above baseline levels (for example, from Nadeau & Bengio, 2003).
Further, we will use interpretable machine learning methods. Here, we aim to compute feature importance measures for single features and feature groups based on parameter groups in the OpenSMILE feature sets in order to investigate which acoustic features are predictive of the experience of affective states and accumulated local effects (ALE) plots and/ or partial dependence plots (PDP) in order to get insights into the direction of feature effects. Finally, in

order to analyze the effect of the affective substance of the sentences on the prediction of affect experience, we will analyze residuals in different value areas of valence and arousal.

Planned exploratory analysis (Optional)
-