Predicting Affective States from Acoustic Cues in Speech in the Wild

Timo K. Koch[1] & Ramona Schoedel[1]

[1] Ludwig-Maximilan-Universiät München

Author Note

Timo K. Koch, Department of Psychology, Psychological Methodss and Assessment, Ludwig-Maximilians-Universität München, Leopoldstr. 13, 80802 Munich

Ramona Schoedel, Department of Psychology, Psychological Methodss and Assessment, Ludwig-Maximilians-Universität München, Leopoldstr. 13, 80802 Munich

The authors made the following contributions. Timo K. Koch: Conceptualization, Writing - Original Draft Preparation, Writing - Review & Editing; Ramona Schoedel: Conceptualization, Writing - Review & Editing.

Correspondence concerning this article should be addressed to Timo K. Koch, Ludwig-Maximilians-Universität München, Department of Psycholgy, Leopoldstr. 13, 80802 Munich. E-mail: timo.koch@psy.lmu.de

Abstract

One or two sentences providing a **basic introduction** to the field, comprehensible to a scientist in any discipline.

Two to three sentences of **more detailed background**, comprehensible to scientists in related disciplines.

One sentence clearly stating the **general problem** being addressed by this particular study.

One sentence summarizing the main result (with the words "**here we show**" or their equivalent).

Two or three sentences explaining what the **main result** reveals in direct comparison to what was thought to be the case previously, or how the main result adds to previous knowledge.

One or two sentences to put the results into a more **general context**.

Two or three sentences to provide a **broader perspective**, readily comprehensible to a scientist in any discipline.

Keywords: Affect, Emotion, Speech, Acoustics, Machine Learning

Word count: 5000

Predicting Affective States from Acoustic Cues in Speech in the Wild

Introduction

Affect recognition is human skill. Growing effort to teach machines to do it across disciplines. Application in psychotherapy and huge privacy issues with all the passively collected speech data (Alexa, Siri etc.)

Affect prediction from acoustic properties of speech

Prior work has generated insights on affect prediction

Common approach in the prior literature is to map affective states based on the Circumplex Model of Affect, which suggests that emotions can be mapped onto a space with the two dimensions of valence (i.e., pleasure) and arousal (i.e., physical and psychological activation) (Russell, Weiss, & Mendelsohn, 1989).

Most results are based on either enacted or rated speech and only little on speech in the wild with self-reported affect.

In order to collect the required affective language, researchers usually rely either on enacted speech or use raters to assign emotion labels to utterances due to the aforementioned issues with collecting longitudinal self-report data on fluctuating states (Burkhardt, Paeschke, Rolfes, Sendlmeier, & Weiss, 2005; Gong et al., 2015; Schröder et al., 2008; Schuller, 2011). Both approaches come with multiple downsides.

Second, since language in most data sets has been manually labelled by raters instead of using self-reported scores, there is an ambiguity of ground truth due to the subjective nature of labelling (Schuller, 2011). In this manner, previous studies mostly assessed perceived personality or affect from external raters instead of self-reported personality or affective states (Mohammadi et al., 2010; Polzehl, 2015). Thereby, expressed affect or personality rather than the experienced affect or inherent personality were being assessed.

However, there is high subjectivity and uncertainty in the target labels because raters tend to disagree to some extent as to what the state or trait should be expressed in the language of others (Schuller, 2018).

Previous studies mostly analyzed language data, which was created in experimental settings rather than naturally occurring language, due to the challenges associated with collecting natural language in-the-wild (Schuller, 2011). Therefore, most datasets cited in the literature contain language data created in lab situations, for example from actors, who enact different affective states (Burkhardt et al., 2005; Polzehl, 2015). However, this comes at a disadvantage because the desired state may not be authentically acted out and it may be driven by how the actor believes the respective personality or emotion must be expressed (Schuller, 2018).

Further, the few available data sets containing naturally occurring speech data have less than 100 participants (Ivanov et al., 2011; Mairesse et al., 2007; Polzehl, 2015). Yet larger sample sizes are needed to discover robust effects.

However, new research tools help to collect speech data in the wild and in-situ self-reports of affect. In the form of the Experience Sampling Method (ESM) allow to assess self-reported affect in an ecologically valid way over a period of time (Servia-Rodríguez et al., 2017). ESM is a method growing in popularity among researchers to collect participants' self-reports on their activities, emotions and other situational variables (van Berkel, Ferreira, & Kostakos, 2017). Recent research by Sun and colleagues (2019) has applied ESM to collect longitudinal data on participants' affective states (i.e., happy mood) as well as their language use. They extracted acoustic features from audio snippets and linguistic features from real-life speech samples collected by a sound recorder in order to predict fluctuations in participants' happiness. Experience Sampling is particulary helpul when run on a smartphone. Again, new data collections methods, such as research apps for smartphones, allow to collect language in large quantities in-the-wild rather than from the lab

(Servia-Rodríguez et al., 2017).

In conclusion, previous findings on the prediction of affect from speech and recent methodological advances in the area of smartphone-based experience sampling motivate us to address the gap in the affect prediction literature based on speech data collected in the wild with self-reported affect annotations. Therefore, we collect speech data and self-reports on affect from participants using an experience-sampling module of the PhoneStudy App. We train cross-validated machine learning models on the extracted audio features to predict participants' self-reported affect, and investigate, which variables were most predictive in the models using interpretable machine learning methods.Thereby, we want to advance theories on affect in speech, elevate applications in automatic affect-detection from speech signals, and inform the discussion on the protection of privacy rights.

Method

Data collection

Data collection for this work is part of the PhoneStudy project at Ludwig-Maximilian-Universität München. Representative sample through panel.

The study comprises two two-week experience sampling phases (27.07.2020 to 09.08.2020; 21.09.2020 to 04.10.2020) during which participants receive two to four short questionnaires per day. The questionnaire is available for 15 minutes and participants are given another 15 minutes to complete it once they have started answering it.

The last experience sampling questionnaire of the day includes an audio logging task at the end. Here, participants are instructed to read out a series of given sentences while making an audio recording of their voice. We use the open source software Open Smile by Audeering (https://audeering.com/technology/opensmile/#features) to extract two feature sets (ComParE2016 and eGeMAPS) of voice parameters directly on the participant's device. We do not store the raw audio logging files. 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 emotional content: positive, negative, and neutral. These three emotional categories are presented consecutively in each audio logging task. The order of the categories is randomized per experience sampling questionnaire. For each emotional content category three sentences are randomly drawn from respective sets of sentences in the database created by Defren and colleagues. The audio recording is started by the participants via a button on the screen. Participants can stop the recording manually after a minimum of four seconds. Alternatively, the recording is stopped automatically after twelve seconds.

Further, we extract features directly on the participant's device. No audio files have to be transferred, only information on extracted features.

Predictive Modelling

We trained two separate models for valence and arousal. We used interpretability measures.

Model analysis

We also address two issues with relevance of the method in practice.

Number of features. How many features do we need? Does the larger feature set improve predictions? In Sun et al. the larger set did not improve predictions.

Content effects on acoustics. What effect does the content participants talk about have on the prediction performance?

Software & Open Science

We used R (Version 4.0.2; **???**) and the R-package *papaja* (Version 0.1.0.9997; **???**) for all our analyses.

Results

Prediction of valence and arousal

Interpretation

Model anlysis

Figure 3 shows the residuals for different values for valence and arousal.

Number of features.

Content effects on acoustics.

Discussion

Our results crated new insights.

References

Defren, S., de Brito Castilho Wesseling, P., Allen, S., Shakuf, V., Ben-David, B., & Lachmann, T. (2018). Emotional Speech Perception: A set of semantically validated German neutral and emotionally affective sentences. In *9th International Conference on Speech Prosody 2018* (pp. 714–718). ISCA. https://doi.org/10.21437/SpeechProsody.2018-145